Computer Physics Communications 137 (2001) 361–365

Computer Physics Communications

www.elsevier.com/locate/cpc

An accurate method for the generalized Fermi–Dirac integral

A. Natarajan *, N. Mohankumar

Health and Safety Division, Indira Gandhi Centre for Atomic Research, Kalpakkam, India 603102

Received 6 October 2000; received in revised form 1 January 2001; accepted 1 January 2001

One of the authors (NMK), dedicates this paper to his sister Santha, who is no more

Abstract

We present an accurate quadrature method for the evaluation of the generalized Fermi–Dirac integral, which is based on splitting the range of integration, Gauss–Legendre integration in the split intervals and correction for the poles of the integration in the interval containing the real part of the poles of the integrand. The present method will work for any values of the parameters k, θ and η of the integrand. © 2001 Elsevier Science B.V. All rights reserved.

PACS: 02.70.+d; 52.25.Fi; 97.10.Cv

Keywords: Degenerate electron gases; Stellar structure; Gauss quadrature with pole correction

1. Introduction

The generalized Fermi–Dirac integral (GFDI), defined by

$$F_k(\eta,\theta) = \int_0^\infty \psi(t,k,\eta,\theta)\,\mathrm{d}t,$$

$$\psi(t,k,\eta,\theta) = t^k \frac{\sqrt{1+(\theta t)/2}}{\mathrm{e}^{t-\eta}+1}$$

is needed extensively in the description of electronic properties of matter in astrophysics and condensed matter physics. The parameter η takes small negative to large positive real values. Values of the parameter k that are of frequent practical interest are the half-integers $-1/2$, $1/2$, $3/2$ and $5/2$. The calculation of transport coefficients in the partially relativistic region requires the above integral to be evaluated for positive θ values, of the order unity or less. Note that for half-integer values of k, the origin is a branch point for the integrand in the complex plane. Again, the integrand has a countable infinity of simple poles at z_j, where

$$z_j = \eta + \mathrm{i}(2j+1)\pi, \quad j = 0, \pm1, \pm2, \ldots.$$

There exists a very extensive literature that deals with different methods of evaluating the above integral and how they fare in terms of efficiency and accuracy. The methods differ with respect to the nature of approximation scheme used such as analytic series, polynomial fitting, rational approximation, etc. Direct numerical integration is possibly the best in terms of accuracy for all values of the parameters k, η and θ. We discuss here briefly some of the quadrature based methods and refer the reader to Pichon [1], Mohankumar and Natarajan [2] and Aparicio [3] for more details. Pichon [1] uses a variant of the generalized Gauss–Laguerre scheme for the evaluation of

* Corresponding author.
E-mail address: anat@igcar.ernet.in (A. Natarajan).

PII: S0010-4655(01)00153-9

the GFDI but his method needs a parameter adjustment for accuracy. The method of Sagar [4] incorporates the singularities of the integrand partially into the weight function. This involves the generation of Gaussian weights and nodes, for the modified weight function, and is computationally intensive. Further, the nodes and weights change with every set of parameter values, which is a disadvantage. The method of Gautschi [5] is similar in spirit to the method of Sagar and has the same limitations. Here, the Gauss–Laguerre weight function is modified by a multiplier function, which is essentially a product of the finite number of the dominant singularities from the infinite set $\{z_k\}$. This scheme is relatively more efficient.

All the quadrature methods just discussed are based on Gauss type quadrature. There exist non-Gaussian methods that are robust and accurate. The IMT quadrature method, an ingenious scheme introduced by Iri et al. [6], involves trapezoidal integration preceded by a change of variable of integration. The transformation of variables employed leads to a clever exploitation of the Euler–Maclaurin formula. By suitably adapting this method, Natarajan and Mohankumar [7] evaluated GFDI, with guaranteed double precision accuracy for any value of k, η and θ. Note that this method is capable of arbitrary precision and can serve as a reference method. However it requires 320 quadrature points which may be viewed as uneconomical.

Mohankumar and Natarajan [2] have later reported another scheme for evaluating GFDI. It uses a parabolic substitution $t = x^2$ to remove the branch point at the origin. The resulting integrand is analytic in the entire complex plane but for certain poles $\{zz'_j\}$, where $zz'_j = \pm z_j^{1/2}$. It also has a pair of branch points at $\pm \mathrm{i}(2/\theta)^{1/2}$ but these are of little consequence, as they are sufficiently away from the line of integration for the range of θ values needed in practice. The integrand tends to zero essentially like $\exp(-x^2)$ as x tends to $\pm\infty$. For integrands of this kind, it is known that trapezoidal integration, with correction for the effect of poles can be very accurate and efficient. The method of Mohankumar and Natarajan [2] exploits this property. The number of residue correction terms roughly increases as $(\eta)^{1/2}$. Though the method is well suited for even large values of η, the number of quadrature points increases like $(\eta + m)^{1/2}$, where m is of the order of 70. This is perhaps the only limitation of the trapezoidal method with correction for poles.

Barring the IMT method as adapted by Natarajan and Mohankumar [7], all the above quadrature methods involve a global approximation of the integrand over the entire range of integration. The recent Gaussian quadrature method of Aparicio [3] uses a piecewise approximation with the range of integration split into four intervals. Gauss–Legendre quadrature and Gauss–Laguerre quadrature are employed for the first three and the fourth intervals, respectively. Also, the parabolic transformation is applied to the integrand of the first interval. This approach has led to very good computational economy. For small θ values, the method of Aparicio needs in all about 80 summation points. As η and θ values increase, the total number of quadrature points will increase. But there are few shortcomings in this method. The break points which determine end points of the four intervals, need to be worked out for an arbitrary choice of k, η and θ. This is done by stipulating the criterion that the errors on the four intervals be of similar order. However, this approach does not directly address the errors.

In the method proposed below, the error of the Gauss–Legendre quadrature is treated analytically in a direct and rigorous fashion. Compared to the method of Aparicio, the proposed method requires significantly less number of quadrature points due to the correction for the effect of singularities of the integrand. Further, it involves no effort to determine the break points, and needs only two intervals for $\eta > 100$. Finally, it uses Gauss–Legendre scheme in all the intervals.

2. The proposed method

We first truncate the range from $(0, \infty)$ to $(0, \eta + m)$, where m is an integer to be chosen appropriately. The integrand of GFDI reaches a peak value in the vicinity of η and afterwards falls rapidly. A choice of m as 70 is adequate for double precision computation, as shown in Appendix.

For smaller η values, the trapezoidal method with pole correction as reported by Mohankumar and Natarajan [2] is more efficient as will be discussed later in this paper. We need to consider only the case of η equal to 100 or more.

We split the truncated integration range $(0, \eta + m)$ into two intervals, $(0, \eta - m)$ and $(\eta - m, \eta + m)$. In the first interval $(0, \eta - m)$, we make a change of variables $t = x^2$, and then do a simple Gauss–Legendre integration. In the second interval $(\eta - m, \eta + m)$, we do a Gauss–Legendre integration, with correction for the poles as outlined below. The details of this method can be found in Chawla and Jain [8].

We first change the variable from t to x as given below.

$$t = \eta + mx; \quad t \in (\eta - m, \eta + m); \quad x \in (-1, 1),$$

$$T = t + \mathrm{i}\tau; \quad z = x + \mathrm{i}y; \quad T = \eta + mz,$$

$$\int_{\eta-m}^{\eta+m} \psi(t, k, \eta, \theta)\,\mathrm{d}t = \int_{-1}^{1} f(x, k, \eta, \theta)\,\mathrm{d}x,$$

$$f(x, k, \eta, \theta) = m\psi(t, k, \eta, \theta).$$

Also note that T and z, which are complex, are related by the same linear relation which links t and x. The branch point at $T = 0$ of the integrand is not contained in either of the above two intervals involving t and x. Also the other branch point of the integrand $T = -(2/\theta)$, lies outside these two intervals. The only singularities of the function $f(z, k, \eta, \theta)$ in the complex z-plane are the simple poles $\{zz_j\}$ where $zz_j = \pm\mathrm{i}(2j + 1)\pi/m$. Let $P_n(z)$ and $Q_n(z)$ denote the Legendre functions of the first and second kind, respectively and let $\{x_i\}$ denote the n zeros of $P_n(x)$ in $[-1, 1]$. Let L be a closed contour, enclosing the interval $[-1, 1]$. L encloses a finite number of poles of $f(z, k, \eta, \theta)$. For brevity we denote $f(z, k, \eta, \theta)$ by $f(z)$. We apply Residue theorem to the integral I below, which is evaluated anti clockwise to get

$$I = \frac{1}{2\pi\mathrm{i}} \oint_L \frac{f(z)\,\mathrm{d}z}{(z - s)P_n(z)},$$

$$I = \frac{f(s)}{P_n(s)} + \sum_{i=1}^{n} \frac{f(x_i)}{(x_i - s)P_n'(x_i)} + \sum_j \frac{\phi(zz_j)}{(zz_j - s)P_n(zz_j)}, \tag{1}$$

$$\phi(zz_j) = \mathrm{Res}\big[f(z)\big]|_{z=zz_j}.$$

Here $\phi(zz_j)$ denotes the residue of $f(z)$ at the poles zz_j of $f(z)$. The residue summation above must include the contribution from all the poles of $f(z)$ inside L. Let w_i denote the Gauss–Legendre weight corresponding to the node x_i. We note the following relations

$$w_i = \frac{1}{P_n'(x_i)} \int_{-1}^{1} \frac{P_n(x)\,\mathrm{d}x}{(x - x_i)}, \tag{2}$$

$$Q_n(z) = \frac{1}{2} \int_{-1}^{1} \frac{P_n(x)\,\mathrm{d}x}{(z - x)}. \tag{3}$$

The function $Q_n(z)$ is single valued and analytic in the z-plane with the interval $[-1, 1]$ deleted. We multiply both sides of Eq. (1) by $P_n(s)\,\mathrm{d}s$ and integrate over $(-1, 1)$. By making use of Eqs. (2) and (3) we get,

$$\int_{-1}^{1} f(s)\,\mathrm{d}s - \left\{ \sum_{1}^{n} w_i f(x_i) + \sum_j \frac{(-2)\phi(zz_j)Q_n(zz_j)}{P_n(zz_j)} \right\} = \frac{1}{\pi\mathrm{i}} \oint_L \frac{f(z)Q_n(z)\,\mathrm{d}z}{P_n(z)} = E_d. \tag{4}$$

The first sum within the curly brackets of above equation is the Gauss–Legendre quadrature sum and the second sum is the correction to the quadrature sum for the effect of the poles of the integrand, which are enclosed within L. Correction terms for a pole and its conjugate need to be summed up together. For a given order of quadrature, the residue correction terms are added till the residue correction sum reaches saturation.

The discretization error for the quadrature E_d after correction for the poles, is given by the contour integral term on the rhs of above equation. One can show that E_d tends to zero very rapidly as a function of n. Precise estimate of E_d can be found in Chawla and Jain [8], who use the symbol E_{Gn} instead of E_d.

3. Results and discussions

Table 1 gives the number of quadrature and residue summation terms needed to achieve a relative error less than 10^{-14} for various values of k, η and θ. n_1 and

Table 1
This gives the number of quadrature and residue summation terms needed to achieve a relative error less than 10^{-14} for various values of k, η and θ. n_1 and n_2 are the orders of Gauss–Legendre quadrature in the first and second intervals, respectively and n_p is the number of residue summation terms used in the second interval

k	η	θ	n_1	n_2	n_p
0.5	100	0.01	14	25	7
0.5	100	0.10	14	25	7
0.5	100	1	17	25	7
0.5	1000	0.01	16	25	6
0.5	1000	0.1	22	25	6
0.5	1000	1	33	25	6
0.5	10000	0.01	22	25	6
0.5	10000	0.1	33	25	6
0.5	10000	1	42	25	6
1.5	100	0.01	16	25	8
1.5	100	0.10	17	25	8
1.5	100	1	17	25	7
1.5	1000	0.01	14	25	7
1.5	1000	0.1	18	25	7
1.5	1000	1	22	25	7
1.5	10000	0.01	21	25	6
1.5	10000	0.1	24	25	6
1.5	10000	1	30	25	6
2.5	100	0.01	13	25	8
2.5	100	0.10	13	25	8
2.5	100	1	24	25	8
2.5	1000	0.01	24	25	7
2.5	1000	0.1	30	25	7
2.5	1000	1	32	25	7
2.5	10000	0.01	24	25	6
2.5	10000	0.1	27	25	6
2.5	10000	1	30	25	6

n_2 are the orders of Gauss–Legendre quadrature in the first and second intervals, respectively and n_p is the number of residue summation terms used in the second interval. The Legendre functions of the first and second kind, $P_n(z)$ and $Q_n(z)$ for purely imaginary arguments, are evaluated using the routines, Double $pli(i,z)$ and Double $qli(l,z)$, found in Baker [9] after converting the original C routines to FORTRAN. For $k = 0.5$, n_1 increases from 14 to 42 as η increases from 100 to 10 000. In most cases, n_1 is of the order of 30. In the second interval which has a fixed length $2m$, 25 quadrature points and a maximum of 8 residue correction terms suffice. Thus the present method is more economical than that of Aparicio, since his method needs about 80 quadrature summations.

Table 2
Economy of trapezoidal method with correction for the poles for $\eta \leqslant 200$. n is the number of trapezoidal summation terms and n_p is the number residue correction terms

k	η	θ	$n+n_p$
0.5, 1.5, 2.5	10	0.01, 0.1, 1.0	$17+5$
0.5, 1.5, 2.5	60	0.01, 0.1, 1.0	$22+9$
0.5, 1.5, 2.5	100	0.01, 0.1, 1.0	$25+12$
0.5, 1.5, 2.5	200	0.01, 0.1, 1.0	$32+16$

The pole correction scheme used is numerically stable whenever the real part of the pole lies in the middle of the interval. That is the reason for choosing the second interval evenly as $(\eta-m, \eta+m)$. However, for $\eta < m$, such a splitting of the ranges is not possible, and we need to split the truncated quadrature range $\eta+m$, into three intervals, namely, $(0, \eta-m_1)$, $(\eta-m_1, \eta+m_1)$ and $(\eta+m_1, \eta+m)$, where $0 < m_1 < \eta$. As usual, in the second interval we have to do Gauss–Legendre quadrature with pole correction. If η is negative, the poles of the integrand lie outside the range of integration $(0,\infty)$ and we do not need correction at all. Thus for $\eta < 100$ even though the present method can work, we note that the trapezoidal method is simpler to use as Table 2 indicates. For $\eta = 10$, we need just 17 summation terms and 5 correction terms. For $\eta = 200$, we need 32 summation terms and 16 correction terms.

4. Conclusions

We have indicated a Gauss–Legendre integration method for the evaluation of GFDI which is superior to all the existing methods for $\eta > 100$. The economy of the method is a consequence of the use of piecewise approximation and the Gauss–Legendre quadrature with appropriate correction for the singularities of the

integrand. More importantly, the method has rigorous bounds for the quadrature error. Finally, the choice of the break points which determine the subintervals is quite simple.

The FORTRAN listing of the present method as well as the listing of trapezoidal method with pole correction, can be obtained from the authors upon request.

Appendix

We need to show that a choice of m around 70 is sufficient to make the relative truncation error less than 10^{-14}. The integral which is neglected after making the transformation $t = x^2$ is given by,

$$E_t = 2 \int_{\sqrt{\eta+m}}^{\infty} \frac{x^{2k+1}\sqrt{1+\theta x^2/2}}{e^{x^2-\eta}+1}\,dx.$$

We will first assume that θ is of the order of unity. In this case we neglect the term unity in the square root term and also in the denominator of the above integrand, leading to

$$E_t \simeq \sqrt{2\theta}e^{\eta} \int_{x_0}^{\infty} x^{2k+2}e^{-x^2}\,dx; \quad x_0 = \sqrt{\eta+m}.$$

Consider the integral term of the above equation,

$$I_j = \int_{x_0}^{\infty} x^j e^{-x^2}\,dx.$$

Integrating by parts we get the recursion

$$I_j = \tfrac{1}{2}x^{j-1}e^{-x_0^2} + \tfrac{1}{2}(j-1)I_{j-2}.$$

As an example, with k set to $(5/2)$, we use the above recursion in the expression for E_t and get

$$E_t \simeq \sqrt{2\theta}e^{-m}S,$$

$$S = x_0^6/2 + x_0^4(3/2) + 3x_0^2 + 3.$$

The above expression shows that E_t is dominated by the term e^{-m}. Hence when E_t is divided by the actual value of the integral, we will get the relative error less than 10^{-14}. As an example, for a choice of $\eta = 10\,000$, $\theta = 1$ and $m = 70$, the above quantity E_t divided by the value of the integral is of the order of the order of 10^{-35} which is much less than 10^{-14}. Here we have assumed that θ is of the order of unity. If θ is less than unity, we replace the square root term by x^2. The estimate in this case is obtained similarly and here again we can show that due to the presence of the term e^{-m}, the relative error of the omitted integral is much less than 10^{-14}. The later case has been dealt with in the appendix of Mohankumar and Natarajan [2].

References

[1] B. Pichon, Comput. Phys. Commun. 55 (1989) 127.
[2] N. Mohankumar, A. Natarajan, Astrophys. J. 458 (1996) 233.
[3] J.M. Aparicio, Astrophys. J. Suppl. Ser. 117 (1998) 627.
[4] R.P. Sagar, Comput. Phys. Commun. 66 (1991) 271.
[5] W. Gautschi, Comput. Phys. Commun. 74 (1993) 233.
[6] M. Iri, S. Moriguti, Y. Takasawa, J. Comput. Appl. Math. 17 (1987) 3.
[7] A. Natarajan, N. Mohankumar, Comput. Phys. Commun. 76 (1993) 48.
[8] M.M. Chawla, M.K. Jain, Math. Comput. 23 (1968) 82.
[9] L. Baker, C Mathematical Function Library, McGraw-Hill, New York, 1991.