Department of Mathematics and Statistics

Preprint MPCS-2020-03

14 October 2020

Computation of the Complex Error Function using Modified Trapezoidal Rules

by

Mohammad Al Azah and **Simon N Chandler-**

for $x > 0$,

$$(6) \qquad \mathrm{erfc}(z) = \frac{hze^{\ z^2}}{\ } \sum_{k=\ 1}^{1} \frac{e^{\ k^2h^2}}{z^2 + k^2h^2} + \frac{2H(\ =h\ \ x)}{1\ \ e^{2\ z=h}} + E(h):$$

Here the rst term is the trapezoidal rule approximation to (

> "is very accurate, provided for given z and N the optimal step size h is selected. It is not easy, however, to determine this optimal h a priori."

Our recommendations address this issue, detailing which approximation formula and what step size h should be used for each N and z.

The bounds we obtain in carrying out ii) prove that the absolute error in our approximation for $w(z)$ tends to zero exponentially with N, uniformly in the complex plane. This is a substantial improvement on the existing bound (7) which blows up when $x = \pi/h$, and does not capture the additional truncation errors due to replacing infinite by finite sums in the approximations (6) and (9).

Concretely, our proposed approximation to $w(z)$, for $z = x + iy$, with $x, y \geq 0$, is

$$w_N(z) := \begin{cases} w_N^{M}(z); & \text{if } y \geq \max(x, \pi/h); \\ w_N^{MT}(z); & \text{if } y < x \text{ and } 1/4 \leq \varphi(x/h) \leq 3/4; \\ w_N^{MM}(z); & \text{otherwise}; \end{cases} \tag{10}$$

where φ is defined by (8), $N \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$,

$$h := \sqrt{\frac{\pi}{(N+1)}}; \tag{11}$$

w

Indeed, we have previously used, in the restricted case $\arg(z) = \pi/4$, an approximation resembling $w_N^{MM}(z)$ when approximating Fresnel integrals [6], proving results in the spirit of Theorem 1.1.

Let us summarise the rest of the paper. In the largest §2 we derive the above formulae and error bounds. In §3 we review the existing, alternative approximate methods for computing $\mathrm{erfc}(z)$ and $w(z)$ for complex z, for none of which has an error bound been proved, similar to Theorem 1.1. In §4 we carry out numerical experiments that confirm the accuracy of $w_N(z)$, showing that its absolute error is $< 2 \times 10^{-15}$ throughout the complex plane with $N = 11$, and that the same bound holds for the relative error in the upper half-plane. We also show that our new approximation is competitive in accuracy and computing times with the methods that we survey in §3, specifically those of [24, 27, 26, 2].

We note that this paper is based, in significant part, on Chapter 3 of the first author's thesis [4].

2. The proposed approximation and its error bounds.

In this section we derive the approximation given by (10) based on modified trapezoidal rules. We also derive the error bounds of Theorem 1.1 that demonstrate that the absolute and relative errors in $w_N(z)$ both decrease exponentially as N increases.

2.1. The contour integral argument and its history.

Given any $f \in C(\mathbb{R})$ that decays sufficiently rapidly at infinity, let

$$I[f] := \int_{-\infty}^{\infty} f(t)\,dt;$$

and, for $h > 0$ and $\alpha \in [0, 1)$, define the generalised trapezoidal rule approximation to $I[f]$ by

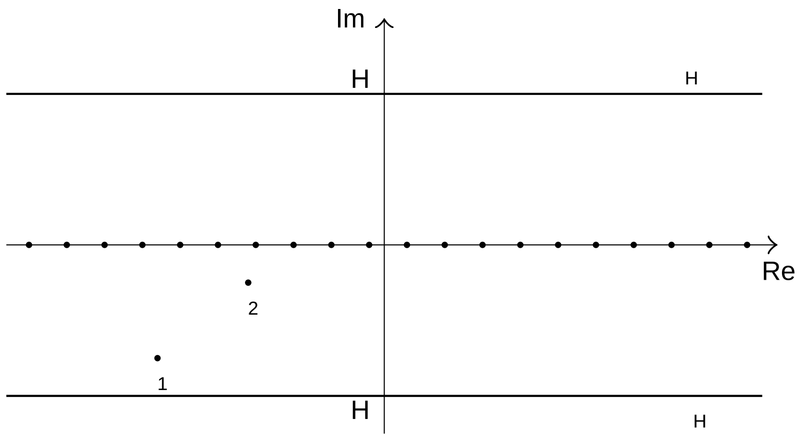
$$I_{h;\alpha}[f] := h \sum_{k \in \mathbb{Z}} f((k - \alpha)h). \tag{17}$$

We note that $I_{h;\alpha}[f] = I_{h;0}[f_\alpha]$, where $f_\alpha(t) := f(t - \alpha h)$ for $t \in \mathbb{R}$, and that $I_{h;0}[f]$ is the trapezoidal rule approximation to $I[f]$ and $I_{h;1/2}[f]$ its composite midpoint rule approximation.

The approximation (17) for $I[f]$ converges exponentially when the integrand is analytic in a strip surrounding the real axis and has sufficient decay at $\pm\infty$. The derivation of this result, using contour integration and Cauchy's residue theorem, dates back, for a particular case, at least to Turing [23], and has been analysed in more general cases by Goodwin [12], McNamee [16], Schwartz [20] and Stenger [21]. For a detailed history and discussion see Trefethen and Weideman [22].

The rate of exponential convergence depends on the width of the strip of analyticity around the real axis, and the accuracy of $I_{h;\alpha}[f]$ deteriorates when f has singularities close to the real line. But, in the case when these singularities are poles, the contour integral method for establishing the exponential convergence of $I_{h;\alpha}[f]$, that we will recall in Proposition 2.1 below, leads naturally to corrections for modifying the trapezoidal rule and recovering rapid convergence, these corrections expressed in terms of residues of f at these poles. This appears to have been observed explicitly first by Chiarella and Reichel [8], in the context of evaluating (5) (and see Matta and Reichel [15], Hunter and Regan [13], and Mori [17]), and has been developed into a general theory by Bialecki [7] (and see La Porte [14]).

It is convenient to recall in a proposition the standard arguments ([8, 15, 13] and cf. [22, pp. 402–403]) that are made to prove exponential convergence, since we will



Im
H
H
Re
2
1
H
H

The representations (6) and (9

bounds hold also for $y = 0$ and $y = \pi/h$ since the left hand sides of the bounds depend continuously on y on $[0, \pi/h]$ (recall that $I_{h,\tau}[f_z]$ is an entire function of z and that $w(z)$ is bounded below on $y \geq 0$ by (32)).

Now suppose that $y > \pi/h$ and take $H = \pi/h - \epsilon$ for some $\epsilon \in (0, \pi/h)$. Then $I_{h,\tau,H}[f_z] = I_{h,\tau}[f_z]$, and since $y/(y^2 - H^2)$ and $(1+\sqrt{2}y)y/(y^2-H^2)$ are decreasing as functions of y on $(H, \infty]$, it follows from (38) and (39) with $H = \pi/h - \epsilon$ that

$$|w(z) - I_{h,\tau}[f_z]| \leq \frac{2\sqrt{2}\, e^{-\pi^2/h^2+\epsilon^2}}{(1 - e^{-2\pi^2/h^2+2\pi\epsilon/h})\,\epsilon(2\pi - \epsilon h)} \tag{42}$$

and

$$\frac{|w(z) - I_{h,\tau}[f_z]|}{|w(z)|} \leq \frac{2\sqrt{2}\,(h+\sqrt{2}\pi)\, e^{-\pi^2/h^2+\epsilon^2}}{h\,(1 - e^{-2\pi^2/h^2+2\pi\epsilon/h})\,\epsilon(2\pi - \epsilon h)}. \tag{43}$$

If $\pi/h > 1/\sqrt{2}$ we can again choose $\epsilon = 1/\sqrt{2}$, obtaining the bounds (35) and (36) for $y > \pi/h$; these bounds hold also for $y = \pi/h$ since the left hand sides of the bounds depend continuously on y on $[\pi/h, \infty)$. ∎

It follows immediately from the definition (29) that, for $x \in \mathbb{R}$, $y > 0$,

$$|C_{h,\tau}[f_z]| \leq \frac{2e^{-2\pi y/h}}{1 - e^{-2\pi y/h}}\, e^{y^2 - x^2}. \tag{44}$$

Since $|I_{h,\tau}[f_z]| \leq |I_{h,\tau}[f_z]| + |C_{h,\tau}[f_z]|$, the following corollary follows from the above proposition, (44), and (32).

Corollary 2.3. If $z = x + iy$ with $x = y \geq 0$ and $h < \pi\sqrt{2}$, then

$$|w(z) - I_{h,\tau}[f_z]| \leq c_a \frac{e^{-\pi^2/h^2}}{1 - e^{-2\pi^2/h^2+\sqrt{2}\pi/h}} \tag{45}$$

and

$$\frac{|w(z) - I_{h,\tau}[f_z]|}{|w(z)|} \leq \frac{c_r}{h}\, \frac{e^{-\pi^2/h^2}}{1 - e^{-2\pi^2/h^2+\sqrt{2}\pi/h}}; \tag{46}$$

where

$$c_a := \frac{2(2e+\sqrt{\pi})}{\sqrt{\pi e}} \approx 4.934 \quad \text{and} \quad c_r := \frac{2\sqrt{2}\,\pi(1+\sqrt{\pi})(2e+\sqrt{\pi})}{\sqrt{\pi e}} \approx 60.77. \tag{47}$$

Proof. For $0 \leq x = y \leq \pi/h$ these bounds follow immediately from the sharper bounds (33) and (34). Suppose now that15.349 5.358 Td [(;)]TJ/F43 9.9626 Tf -276.101 -23.887 Td [(wher)51(e)]TJ/F8 9.9

Proposition 2.2 tells us that I

for x 0, where

$$G_x(t) := e^{\quad t^2} jF_x(t + i \quad =h)j =$$

$h < \sqrt{2}$, it follows from Corollary 2.3 and Proposition 2.5, on noting that the bounds in Proposition 2.5 are smaller than those in Corollary 2.3, that E_h and e_h satisfy the bounds claimed in the proposition when $z \in \partial$, i.e. for z on $\{x + i x : x \geq 0\}$ and on the positive real axis. Thus the proposition follows by Lemma 2.4 if we can show that $E_h(z)$ and $e_h(z)$ do not grow too rapidly as $z \to \infty$ in .

But, if $h < \sqrt{2}$, it follows from (31) and (38) applied with $H = 3$ that, for some constant $C > 0$ independent of z, $|E_h(z)| \leq C|z|$ if $z \in$ with $y \leq 2$. Similarly, since $I_{h;\,;H}[f_z] = I_{h;}[f_z]$ if $y > H$ and $I_{h;}[f_z] = I_{h;}[f_z] + C_{h;}[f_z]$, it follows from (38) applied with $H = 1$ and (44) that, for some constant $\tilde{C} > 0$ independent of z, $|E_h(z)| \leq \tilde{C}|z|$ if $z \in$ with $y \geq 2$. Thus $|E_h(z)| \leq C|z|$ for $z \in$, where $C := \max(C; \tilde{C})$, so that also, applying (32), $|e_h(z)| \leq C|z|(1 + \sqrt{\ }|z|)$ for $z \in$.
Thus the proposition follows by applying Lemma 2.4. □

The following corollary summarises and simplifies, at the cost of a little sharpness, the results of Propositions 2.2 and 2.6 and of this subsection.

Corollary 2.7. Suppose that $z = x + iy$ with $x \geq 0$, $y \geq 0$, and $h < \sqrt{2}$. Then the bounds (45) and (46) hold with c_a and c_r given by (47) if $y \geq \max(x; =h)$. The same bounds hold as bounds on $|w(z) - I_{h;}[f_z]|$ and $|w(z) - I_{h;}[f_z]|/|w(z)|$, respectively, with the same values of c_a and c_r, if $y \leq \max(x; =h)$.

Proof. The first claim of the corollary follows from Proposition 2.6 and (33) and (34), and the second follows from (35) and (36). □

2.3. Truncating the infinite series. Propositions 2.2 and 2.6 together provide accurate trapezoidal-rule-baeep38 (z) on $(0;1\ )$ and noting (53), that

$$(57) \qquad 2h \sum_{k=M}^{1} e^{\ s_k^2} \quad 2$$

Proof. From (25) and (55), for $0 \le y \le x$,

$$|T^N_{h,\phi}[f_z]| \le \frac{2h|z|}{\pi} \sum_{k=N+1}^{\infty} \frac{e^{-s_k^2}}{|z^2 - s_k^2|} \le \frac{2\sqrt{2}hx}{\pi(x + s_{N+1})} \sum_{k=N+1}^{\infty} \frac{e^{-s_k^2}}{|z - s_k|}. \tag{63}$$

Thus, and noting (32), the bounds (61) and (62) hold if $x = 0$.

Choose α with $0 < \alpha < 1$. Given $x > 0$ let M be the smallest integer $\ge N + 1$ such that $s_M > \alpha x$, so that, if $M > N + 1$, $s_k \le \alpha x$ and $|z - s_k| \ge (1 - \alpha)x$ for $k < M$. If $M > N + 1$ it follows, using the bound (57), that

$$2hx \sum_{k=N+1}^{M-1} \frac{e^{-s_k^2}}{|z - s_k|} \le \frac{2h}{1-\alpha} \sum_{k=N+1}^{\infty} e^{-s_k^2} \le \frac{2hs_{N+1} + 1}{(1}$$

Further, if x 0 and

the nth convergent of the beautiful Laplace continued fraction representation for $w(z)$ (specifically suggesting $n = 9$). Gautschi notes that: i) by construction the nth convergent is asymptotically accurate, with error $O(|z|^{-2n-1})$ as $|z| \to \infty$, uniformly in the first quadrant; ii) the nth convergent converges to $w(z)$ as $n \to \infty$ if and only if $\mathrm{Im}(z) > 0$; iii) remarkably, for $\mathrm{Im}(z) > 0$, the nth convergent coincides with the approximation obtained by approximating (3) by an n-point Gauss-Hermite rule. For smaller z Gautschi [11] proposed (rational) approximations that are truncated Taylor expansions with the coefficients approximated by convergents of continued fractions.

This methodology, carefully tuned, is the basis of TOMS Algorithm 680 (Poppe and Wijers [18]) which achieves a relative error of 10^{-14} over nearly all the complex plane using, in the first quadrant: i) Maclaurin polynomials of degree ≤ 55 for the odd function $\mathrm{erf}(-iz)$ (substituted into (2)) in an ellipse around the origin; ii) the convergents (78) with $n \leq 18$ outside a larger ellipse; iii) the more expensive mix of Taylor expansion and continued fraction calculation proposed by Gautschi [11] in between. This algorithm has been used as a benchmark by several later authors.

Weideman [24, 25] proposed (the derivation starts from (3)) the single rational approximation

$$w(z) \approx \frac{1}{\sqrt{\pi}(L - iz)} + \frac{2}{(L - iz)^2} \sum_{n=0}^{N-1} a_{n+1} \left(\frac{L + iz}{L - iz} \right)^n, \quad \text{for } \mathrm{Im}(z) \geq 0, \tag{79}$$

where the size of N controls the accuracy of the approximation, $L := 2^{-1/4} N^{1/2}$, and the a_n are discrete Fourier coefficients that can be precomputed by the FFT. He argues, based on operation counts, that, for intermediate values of $|z|$, the work required to compute $w(z)$ to 10^{-14} relative accuracy is much smaller for (79) than for the Poppe and Wijers algorithm [18].

Zagloul and Ali proposed a method (see TOMS Algorithm 916 [27] and the refinements in [26], and cf. [19] and [3, (7.1.29)]) starting from

$$\mathrm{erf}(z) = \mathrm{erf}(x) + \frac{2e^{-x^2}}{\sqrt{\pi}} \int_0^y e^{t^2} \sin(2xt)\, dt + \frac{2i\, e^{-x^2}}{\sqrt{\pi}} \int_0^y e^{t^2} \cos(2xt)\, dt, \tag{80}$$

for $z = x + iy$. They approximate

$$w(z) \approx u(x, y) + i\, v(x, y), \quad x, y \geq 0, \tag{81}$$

where

$$u(x, y) := e^{-x^2} \mathrm{erfcx}(y) \cos(2xy) + \frac{2a \sin^2(xy)}{\pi y} e^{-x^2} + \frac{ay}{\pi} \left(-2\cos(2xy)\, S_1 + S_2 + S_3 \right),$$

$$v(x, y) := -e^{-x^2} \mathrm{erfcx}(y) \sin(2xy) + \frac{a \sin(2xy)}{\pi y} e^{-x^2} + \frac{a}{\pi} \left(2y \sin(2xy)\, S_1 - S_4 + S_5 \right),$$

$\mathrm{erfcx}(y) := e^{y^2} \mathrm{erf}(y)$, and S_j, $j = 1, \ldots, 5$, are the following summations reminiscent of the trapezoidal rule approximations (6):

$$\begin{aligned}
&S_1 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2} e^{-(a^2k^2 + x^2)}, \quad S_2 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2} e^{-(ak + x)^2},\\
&S_3 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2} e^{-(ak - x)^2}, \quad S_4 := \sum_{k=1}^{\infty} \frac{ak}{a^2k^2 + y^2} e^{-(ak + x)^2},\\
&S_5 := \sum_{k=1}^{\infty} \frac{ak}{a^2k^2 + y^2} e^{-(ak - x)^2}.
\end{aligned} \tag{82}$$

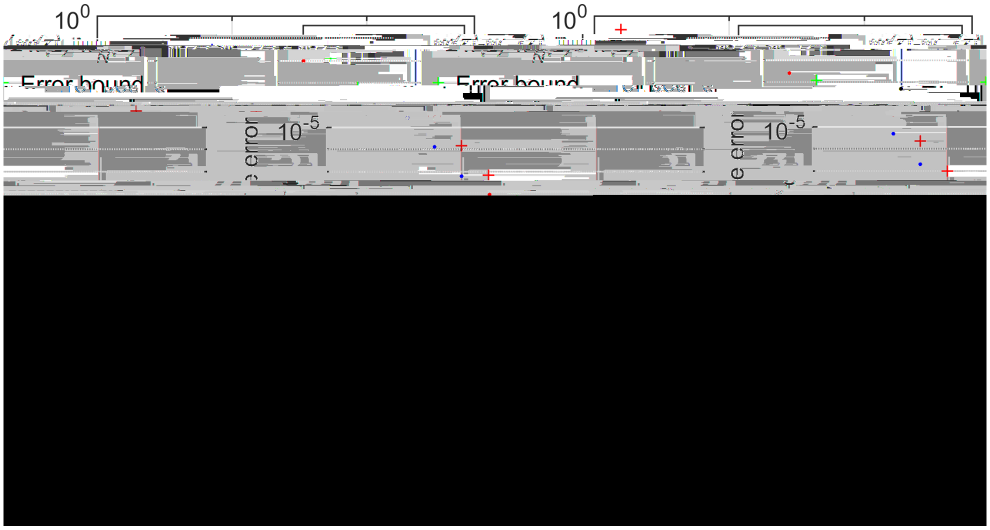


Fig. 2. *Maximum absolute and relative errors in the approximation* ([10](#)) *and the error bounds of Theorem [1.1](#), plotted against* N.

The authors have supplied us with their Matlab implementation Faddeyeva_v2(z,M) [26], where the parameter M is the number of accurate significant figures required, and the code enforces $4 \le M \le 13$. In this code the choice $a = 1/2$ is made and the sums in (82) are truncated, the number of terms retained depending on M. Zagloul and Ali [27] (and see [26]) present numerical evidence that the approximation (81), with $a = 1/2$ and appropriate truncation of the infinite sums (82), is more accurate and faster than TOMS Algorithm 680 [18].

Abrarov et al. [2] (and see [1]) proposed recently another method for computing $w(z)$ using modified rational approximations, namely

$$w(z) \approx \begin{cases} {}_1(z); & \text{if } z \in D_1; \\ {}_2(z); & \text{if } z \in D_2; \\ {}_3(z); & \text{if } z \in D_3; \end{cases} \tag{83}$$

where $D_1 := \{z = x + iy : |z| < 8 \text{ and } y > 0.05x\}$, $D_2 := \{z = x + iy : |z| < 8 \text{ and } y \le 0.05x\}$, $D_3 := \{z : |z| \ge 8\}$,

$$ {}_1(z) := \sum_{m=1}^{M} \frac{A_m + B_m(z + i\;\;/2)}{C_m^2 \;\; (z + i\;\;/2)^2}; \qquad {}_2(z) := e^{\;\;z^2} + z \sum_{m=1}^{M+2} \frac{\;_m \;\;\; _m z^2}{\;_m \;\;\; _m z^2 + z^4}; \tag{84}$$

the coefficients A_m, B_m, C_m, $_m$, $_m$, $_m$ and $_m$ are specified in [2], and ${}_3(z)$ is approximately (78) with $n = 10$ (see [2, Equation (9)]). Abrarov et al. [2] present numerical evidence to show that (83) achieves an accuracy of 10^{-13} using $= 2.75$ and $M = 23$ in (84).

4. Numerical results. In this section we show calculations that illustrate and support Theorem 1.1, and that compare the accuracy and efficiency of our approximation $w_N(z)$ given by (10 using the Matlab code wTrap(z,N) provided in Table 1 of the supplementary material to this paper [5]. The maximum values we plot are discrete maxima taken over the $16{,}020{,}801$ points $z = 10^p e^{i\theta}$, with $p = -6(0.0006)6$ and $\theta = 0(\pi/1600)\pi/2$, a superset of the $40{,}401$ test values in Weideman [24, 25]. To compute errors we use as the exact values of $w(z)$ the independent approximation (79) with $N = 45$, implemented through a call cef(z,45) to the Matlab code in [24, Table 1]. (The results in [24, 25, Figure 8], [6, Figure 2] suggest that $N = 45$ in (79) is ample for accuracies close to machine precision, and we obtain almost identical results in Figure 2 and Table 1 below if we use, instead, $w_{20}(z)$ given by (10) as the exact value.)

We observe in Figure 2 the rate of exponential convergence predicted by Theorem 1.1. The approximation $w_N(z)$ achieves, with $N = 11$ over this set of discrete points in the first quadrant, maximum absolute and relative errors which are $< 2 \times 10^{-15}$.

Algorithm	Maximum abs. error	Maximum rel. error	Computing time (seconds)
wTrap(z,11)	1.67×10^{-15}	1.89×10^{-15}	$4.29\ (\pm 0.08)$
cef(z,40)	2.11×10^{-15}	2.15×10^{-15}	$4.20\ (\pm 0.02)$
fadsamp(z)	3.86×10^{-14}	3.86×10^{-14}	$5.74\ (\pm 0.04)$
Faddeyeva_v2(z,13)	4.07×10^{-15}	1.71×10^{-13}	$11.00\ (\pm 0.11)$

Table 1

Maximum absolute and relative errors for the Matlab codes implementing the approximations (10), (79), (83), and (81). The computing times are mean and s.d. of 25 executions.

In Table 1 we compare the accuracy and efficiency of our approximation and Matlab code with Matlab implementations of the approximations (79), (81), and (83). Results are shown in Table 1 for:

1. Our approximation $w_N(z)$ with $N = 11$ implemented by the call wTrap(z,11) to the Matlab code provided in [5, Table 1];
2. Weideman's approximation (79) with $N = 40$ (this choice of N ensures high accuracacble