Multi-precision computation of the complex error function

Pascal Molin

Multi-precision computation of the complex error function

Pascal Molin*

March 2011

Abstract

We give a very simple algorithm to compute the error and complementary error functions of complex argument to any given accuracy.

Contents

1 Introduction **1**
1.1 Motivation, and a smooth integral formula 2

2 Computation **3**
2.1 Trapezoidal formula 4
2.2 Quadrature error 4
2.3 Truncation error 6
2.4 Proof of Theorem 2.3 6

3 Practical algorithm **7**
3.1 Small integer trick 7
3.2 Improvement for real argument 8
3.3 Extension for small real part 8

4 Timings **9**

1 Introduction

With standard normalization, the error function is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t$$

and the complementary error function as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} \, \mathrm{d}t.$$

*INRIA Nancy, project team CARAMEL.

Thanks to the relation $\operatorname{erf}(x) = 1 - \operatorname{erfc}(x)$ and the symmetry $\operatorname{erfc}(-x) = 2 - \operatorname{erfc}(x)$, it is sufficient to be able to compute erfc on the right half complex plane.

Many formula exist for this calculation, obtained either via integration or with asymptotic expansions near zero or infinity. In this note, we study an integration-based formula which is simple, efficient and easy to apply rigorously.

Although this formula is not new (we found in particular [MR71]), it seems that it has not been introduced in numerical systems so far. J. Weideman [Wei94] justifies this by the difficulties to find a correct stepsize, and some numerical instabilities. The rigorous treatment we give here aims at removing these prejudices. We also introduce a trick which accelerates the multiprecision computations.

1.1 Motivation, and a smooth integral formula

In [Mol10], we showed that the trapezoidal scheme yields a very fast method for the computation of integrals of the form

$$\int_{\mathbb{R}} F(u) e^{-Q(u^2)} \, \mathrm{d}u \tag{1}$$

where $F(x) \in \mathbb{R}(x)$ is a rational fraction with no pole on $\mathbb{R}$ and $Q(x) \in \mathbb{R}_+[x]$ is a polynomial.

If we focus on the simplest case

$$f(\alpha) = \int_{\mathbb{R}} \frac{e^{-u^2}}{u - \alpha} \, \mathrm{d}u, \; \alpha \notin \mathbb{R}, \tag{2}$$

we have the following formula, which enables us to compute the complementary error function to any accuracy.

Proposition 1.1

For all x such that $\operatorname{Re}(x) > 0$,

$$\operatorname{erfc}(x) = \frac{-ie^{-x^2}}{\pi} f(ix). \tag{3}$$

Proof: f is analytic outside $\mathbb{R}$, derivating there under the integral and integrating by parts yields the differential equation

$$f'(\alpha) + 2\alpha f(\alpha) + 2\sqrt{\pi} = 0.$$

Writing $f(\alpha) = \lambda(\alpha) e^{-\alpha^2}$, λ satisfies $\lambda'(\alpha) = -2\sqrt{\pi} e^{\alpha^2}$, so that for $\operatorname{Im}(\alpha) > 0$ we have $\lambda(\alpha) = K - 2\sqrt{\pi} \int_0^\alpha e^{u^2} \, \mathrm{d}u$ with $K \in \mathbb{C}$. The limit $\lim_{x\to\infty} f(ix) e^{-x^2} = 0$ determines K to be $2\sqrt{\pi} \int_0^{i\infty} e^{u^2} \, \mathrm{d}u$. As a consequence, $\lambda(ix) = 2\sqrt{\pi} \int_{ix}^{i\infty} e^{-u^2} \, \mathrm{d}u = i\pi \operatorname{erfc} x$, hence the result. □

More precisely, the general theory ensures that $f(\alpha)$ can be evaluated to a precision of p binary digits with a number of points $n \sim \frac{p \log 2}{\pi}$.

The results of [Mol10] also assert that this formula is optimal amongst integration-based formulas, which means that any other integration formula will need a larger number of points (for theoretic reasons due to the uncertainty principle).

2 Computation

We adopt the following notation for the numerical approximation to p binary digits of absolute precision:

Definition 2.1

For $a, b \in \mathbb{C}$,

$$a =_p b \Leftrightarrow |a - b| \leqslant 2^{-p}. \tag{4}$$

Moreover, the consideration of a residue term leads to the following:

Definition 2.2

Let x be a complex number with $\mathrm{Re}(x) \geqslant 0$, and $h > 0$, and define $\delta(x, h)$ to be

$$\delta(x,h) = 1 \text{ if } \mathrm{Re}(x) < \frac{\pi}{h}; \tag{5}$$

$$= 0 \text{ otherwise.} \tag{6}$$

The numerical formula is then

Theorem 2.3

Let x be a complex number such that $\mathrm{Re}(x) \geqslant 1$, and suppose $p \geqslant 2$. Then for all $h > 0$ and $n \in \mathbb{N}$ such that $h \leqslant \frac{\pi}{\sqrt{\mathrm{arcsinh}(2^p\sqrt{\pi})+2}}$ and $nh \geqslant \sqrt{p \log 2}$ we have

$$\mathrm{erfc}(x) =_p \frac{e^{-x^2}}{\pi} \left(\frac{h}{x} + 2hx \sum_{k=1}^{n} \frac{e^{-(kh)^2}}{x^2 + (kh)^2} \right) - \frac{2\delta(x+1,h)}{e^{\frac{2\pi x}{h}} - 1}. \tag{7}$$

We give a complete proof in the next paragraphs.

2.1 Trapezoidal formula

For $\alpha \in \mathbb{C}$ such that $\mathrm{Im}(\alpha) > 0$ we define $g(x) = \frac{e^{-x^2}}{x-\alpha}$. The trapezoidal formula for $f = \int_{\mathbb{R}} g$ reads

Lemma 2.4

Let $\alpha \in \mathbb{C}$ such that $\mathrm{Im}(\alpha) > 0$. For all $h > 0$ and $n \geqslant 1$, we have

$$f(\alpha) = h \sum_{k=-n}^{n} \frac{e^{-(kh)^2}}{kh - \alpha} + E_t(n,h) - E_q(h) \tag{8}$$

where $E_q(h) = \sum_{k\neq 0} \hat{g}(\frac{k}{h})$ and $E_t(N,h) = \sum_{|k|>N} g(kh)$.

Proof: This is the Poisson formula written for g on the lattice $h\mathbb{Z}$, isolating the first Fourier term $\hat{g}(0) = f(\alpha)$. □

2.2 Quadrature error

Lemma 2.5

Let $\alpha \in \mathbb{C}$ such that $\mathrm{Im}(\alpha) > 0$ and $\tau > 0$ such that $\tau \neq \mathrm{Im}(\alpha)$, then we have for all $X > 0$

$$\left|\hat{g}(-X) + \delta 2i\pi e^{-\alpha^2 - 2i\pi X\alpha}\right| \leqslant \frac{\sqrt{\pi}}{\left|\tau - \mathrm{Im}(\alpha)\right|} e^{\tau^2 - 2\pi X\tau} \tag{9}$$

$$\left|\hat{g}(X)\right| \leqslant \frac{\sqrt{\pi}}{\left|\tau + \mathrm{Im}(\alpha)\right|} e^{\tau^2 - 2\pi X\tau} \tag{10}$$

where $\delta = 0$ if $\tau < \mathrm{Im}(\alpha)$ and $\delta = 1$ if $\tau > \mathrm{Im}(\alpha)$.

Proof: We have $\hat{g}(\pm X) = \int_{\mathbb{R}} \frac{e^{-u^2 \mp 2\pi uX}}{u-\alpha}\,\mathrm{d}u$, and we shift the contour to $\mathbb{R} \pm i\tau$ thanks to the residue theorem. The first line corresponds to $\mathbb{R} \to \mathbb{R} + i\tau$, which goes though the pole at α if $\tau > \mathrm{Im}(\alpha)$, and the second line to the shift $\mathbb{R} \to \mathbb{R} + i\tau$. We then bound with the L^1 norm, with $\int_{\mathbb{R}} \left|e^{-(u\pm i\tau)^2 \mp 2i\pi(u\pm i\tau)X}\right| \mathrm{d}u = e^{\tau^2 - 2\pi\tau X} \int_{\mathbb{R}} e^{-u^2}\,\mathrm{d}u = \sqrt{\pi} e^{\tau^2 - 2\pi\tau X}$. □

Lemma 2.6

Assume $X_0 \geqslant \frac{1}{\pi}\sqrt{\mathrm{arcsinh}(2^p\sqrt{\pi})}$ such that $\pi X_0 \notin]\mathrm{Im}(\alpha) - 1, \mathrm{Im}(\alpha) + 1[$, then

$$\forall X \geqslant X_0, \left|\sum_{k\neq 0} \hat{g}(kX) + \delta(x, 1/X_0) \frac{2i\pi e^{-\alpha^2}}{e^{-2i\pi\alpha X} - 1}\right| \leqslant 2^{-p}. \tag{11}$$

For all $X \geqslant \frac{1}{\pi}\sqrt{\mathrm{arcsinh}(2^p\sqrt{\pi})} + \frac{2}{\pi}$,

$$\left|\sum_{k\neq 0} \hat{g}(kX) + \delta(\mathrm{Im}(\alpha) + 1, 1/X) \frac{2i\pi e^{-\alpha^2}}{e^{-2i\pi\alpha X} - 1}\right| \leqslant 2^{-p}.$$

Proof: Fix $\tau = \pi X_0$ in the preceding lemma; the hypothesis allows to remove the terms $\left|\tau \pm \mathrm{Im}(\alpha)\right|$. Summing on kX, $k \neq 0$ we obtain

$$\left| \sum_{k\neq 0} \hat{g}(kX) + \delta \frac{2i\pi e^{-\alpha^2}}{e^{-2i\pi\alpha X} - 1} \right| \leqslant 2\sqrt{\pi} \sum_{k>0} e^{-\pi^2 (kX_0)^2} \leqslant \frac{\sqrt{\pi}}{\sinh(\pi^2 X_0^2)} \tag{12}$$

hence the first assertion of the lemma. Now take X as in the second part, and set $X_2 = X - \frac{2}{\pi}$, $X_1 = X - \frac{1}{\pi}$. If $\mathrm{Im}(\alpha) \leqslant \pi X_1$, then we use the first part with $X_0 = X$ and $\delta(\mathrm{Im}(\alpha), 1/X) = \delta(\mathrm{Im}(\alpha), 1/X_1) = 1$. If $\mathrm{Im}(\alpha) > \pi X_1$ we use the first part with $X \geqslant X_0 = X_2$ and $\delta(\mathrm{Im}(\alpha), 1/X_0) = \delta(\mathrm{Im}(\alpha), 1/X_1) = 0$. Finally, we use $\delta(x, 1/X_1) = \delta(x+1, 1/X)$. □

2.3 Truncation error

Lemma 2.7

Assume $\mathrm{Im}(\alpha) > 1$, *and let* $p \geqslant 2$. *Then for all* n *such that* $nh \geqslant \sqrt{p \log 2}$, *we have*

$$\left| h \sum_{|k|>n} \frac{e^{-(kh)^2}}{kh + \alpha} \right| \leqslant 2^{-p} \tag{13}$$

Proof: Since $\mathrm{Im}(\alpha) \geqslant 1$ we have $\left|(kh)^2 + \alpha\right| \geqslant 1$, and the left-hand side is bounded above by $2h \int_n^\infty e^{-(th)^2}\, \mathrm{d}t = 2\,\mathrm{erfc}(nh)$. The upper bound $\mathrm{erfc}(x) \leqslant \frac{2e^{-x^2}}{\sqrt{\pi}(x + \sqrt{x^2 + \frac{4}{\pi^2}})}$ gives the result (we have $\sqrt{\pi}(x + \sqrt{x^2 + \frac{4}{\pi^2}}) > 4$ for $x \geqslant \sqrt{2\log 2}$). □

2.4 Proof of Theorem 2.3

We combine the preceeding lemmas for $\alpha = ix$, writing

$$-ih \sum_{k=-n}^{n} \frac{e^{-(kh)^2}}{kh + ix} = \frac{h}{x} + 2xh \sum_{k=1}^{n} \frac{e^{-(kh)^2}}{x^2 + (kh)^2}$$

The value $\frac{1}{h} = X = \frac{\sqrt{\mathrm{arcsinh}(2^p\sqrt{\pi})}}{\pi} + \frac{2}{\pi}$ ensures the hypothesis of Lemma 2.6 is satisfied, so that

$$\left| E_q(h) + \frac{2i\pi\delta(x,h)}{e^{\frac{2\pi x}{h}} - 1} \right| \leqslant 2^{-p}.$$

The lemma 2.7 bounds $\left|E_t(n,h)\right|$, and this proves the computation of $-\frac{i}{\pi} f(ix)$ to absolute precision $\frac{2}{\pi} 2^{-p} < 2^{-p}$ under the hypothesis of Theorem 2.3.

The fact that we obtain the right relative precision on $\mathrm{erfc}(x)$ follows from the bounds

$$\frac{1}{2\,|x| + 1} \leqslant \left| \frac{\mathrm{erfc}(x)}{e^{-x^2}} \right| \leqslant \frac{1}{2\,|x| - 1}$$

valid for any $x \in \mathbb{C}$ with $|x| \geqslant 1$.

3 Practical algorithm

The formula of Theorem 2.3 is simple. We decribe here how to evaluate it efficiently for multiprecision values.

First, we write $\lambda = \frac{x}{h}$ to put formula (7) into the form

$$\operatorname{erfc}(x) =_p \frac{e^{-x^2}}{\pi}\left(\frac{1}{\lambda} + 2\lambda \sum_{k=1}^{n} \frac{(e^{-h^2})^{k^2}}{\lambda^2 + k^2}\right) - \delta\frac{2}{e^{2\pi\lambda} - 1} \tag{14}$$

(with $h \sim \frac{\pi}{\sqrt{p \log 2}}$ and $n \sim \frac{p \log 2}{\pi}$).

We then write $U_k = e^{-(kh)^2}$, $V_k = e^{-(2k+1)h^2}$ and $D_k = \lambda^2 + k^2$, so that the main sum becomes

$$\sum_{k=1}^{n} \frac{U_k}{D_k},$$

subject to the recursions

$$U_{k+1} = U_k V_k \tag{15}$$

$$V_{k+1} = e^{-2h^2} V_k \tag{16}$$

$$D_{k+1} = D_k + 2k + 1. \tag{17}$$

3.1 Small integer trick

Finally, thanks to the loose condition we have on h, we can improve the computation if we constraint the factor e^{-2h^2} to be exactly a small precision rational $u/2^v$, as soon as $h = \sqrt{-\log \sqrt{u/2^v}}$ satifies Theorem 2.3.

This way, the computation of each term of the sum is reduced to the following significant operations

- a multiprecision multiplication $U_k V_k$;
- a small multiplication $V_k u/2^v$;
- a multiprecision division U_k / D_k;

and we have

Theorem 3.1 (*complexity*)

The complex error function can be evaluated to p binary digits in complexity $\frac{p \log 2(1+\lambda)}{\pi} M(p) + o(pM(p))$, where $M(p)$ denotes the complexity of a multiplication of size p and λ is the number of multiplications needed to perform a division.

Remark : Of course this complexity can be lowered for negative real part, since there the result approaches 2 and the precision required to compute the difference can be decreased.

3.2 Improvement for real argument

The trick above can be improved if we apply it to the denominator D_k and simplify the division instead of the multiplication. We write

$$\operatorname{erfc}(x) =_p \frac{e^{-x^2}}{\pi\lambda}\left(1 + 2\sum_{k=1}^{n}\frac{(e^{-h^2})^{k^2}}{1+k^2/\lambda^2}\right) - \frac{2\delta}{e^{2\pi\lambda}-1} \tag{18}$$

and choose h to have exactly $\lambda^{-2} = u2^{-v}$, $u, v \in \mathbb{N}^2$ with v much smaller than p, so that the computation becomes

$$\operatorname{erfc}(x) =_p \frac{e^{-x^2}}{\pi\lambda}\left(1 + 2^{v+1}\sum_{k=1}^{n}\frac{(e^{-h^2})^{k^2}}{2^v+k^2u}\right) - \frac{2\delta}{e^{2\pi\lambda}-1}. \tag{19}$$

In order to compute with this formula, the only condition we have is that $v > \log_2(x^2/h^2)$. For the multiprecision range[1], this condition is very mild.

Remark : This trick assumes x to be real, otherwise there is no reason why both its real and imaginary parts should be exactly representable using small integers.

3.3 Extension for small real part

The theorem 2.3 can be extended to $\operatorname{Re}(x) < 1$ if we shift the path of integration. Indeed, the function f defined for $\operatorname{Im}(\alpha) > 0$ in (2) extends analytically to $\operatorname{Im}(\alpha) > -d$ for any $d > 0$ with

$$f(\alpha) = \int_{\mathbb{R}-id}\frac{e^{-u^2}}{u-\alpha}\,\mathrm{d}u. \tag{20}$$

We then have

Theorem 3.2

Let x be a complex number such that $\operatorname{Re}(x) \geqslant 0$. Then for all $h > 0$ and $n \in \mathbb{N}$ such that $h < \frac{\pi}{\sqrt{\operatorname{arcsinh}(2^p\sqrt{\pi})+2}}$ and $nh \geqslant \sqrt{(p+1)\log 2}$, we have

$$\operatorname{erfc}(x) =_p \frac{e^{-x^2+1}}{\pi}\left(\frac{1}{\lambda} + 2\sum_{k=1}^{n}\frac{(\lambda\cos(2kh)+k\sin(2kh))e^{-(kh)^2}}{\lambda^2+k^2}\right) - \frac{2\delta(x+1,h)}{e^{2\pi\lambda}-1} \tag{21}$$

where $\lambda = \frac{x+1}{h}$.

[1] We remark that current multiprecision library do not allow x to exceed 10^{10}.

Proof: Using (20), the main Riemann sum for $f(\alpha)$ takes the form

$$\begin{aligned} S(\alpha) &= h\sum_{k=-n}^{n} \frac{e^{-(kh)^2}e^{2idkh}e^{d^2}}{kh-\tilde{\alpha}},\ \tilde{\alpha}=\alpha+id \\ &= -\frac{he^{d^2}}{\tilde{\alpha}} + he^{d^2}\sum_{k=1}^{n} e^{-(kh)^2}\left(\frac{2ikh\sin(2dkh)+2\tilde{\alpha}\cos(2dkh)}{(kh)^2-\tilde{\alpha}^2}\right). \end{aligned}$$

If $\mathrm{Re}(x+d) \geqslant 1$, the truncation error is bounded by $2\int_{nh}^{\infty} e^{-t^2+d^2} \leqslant e^{d^2}E_t(n,h)$, so that for $d=1$ it is enough to take $nh > \sqrt{(p+3)\log 2}$.

The quadrature error, once corrected by the residue

$$\frac{2\pi\delta(x+d,h)e^{x^2}}{e^{\frac{2\pi(x+d)}{h}}-1},$$

is bounded by $\frac{\sqrt{\pi}}{\tau\pm\mathrm{Re}(x+d)}e^{(\tau\pm d)^2-2\pi X\tau}$. For $X>0$ the denominator is greater than d and this can be bounded by $\frac{\sqrt{\pi}}{d}e^{-(\pi X-d)^2}$ so that the value X_1 of Lemma 2.6 is enough; for $X<0$ we apply the same bounds as in Lemma 2.6. This gives the result, fixing the value $d=1$. □

4 Timings

We did a PARI/gp implantation which proves to be quite efficient. In Table 1 we compared its running time on a few inputs to MPFR [FHL+07] and Maple 14. It turns out that the integration formula is not very interesting in the asymptotic range (near zero or infinity), where the asymptotic expansions of erf and erfc are very accurate. However, it sould be considered in the transition range, when the modulus of the argument comes around the square root of the precision (as in the line 10^4 digits, $x=200$). When the precision increase, this range gets wider : at 10^5 digits our formula starts to be better from $x=10$.

We also remark that the integration formula gives a running time which depends only on the required precision, and to a minor extent of the nature (real or complex) of the argument, contrary to Maple and MPFR.

References

[FHL+07] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier, and Paul Zimmermann. MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Trans. Math. Softw.*, 33, June 2007.

[Mol10] Pascal Molin. Intégration numérique par la méthode double-exponentielle. http://hal.archives-ouvertes.fr/hal-00491561/fr/, june 2010. 48 pages.

Algorithm 1: Computation of the complementary error function

Input: x such that $\mathrm{Re}(x) \geqslant 1$
Input: binary precision $p \geqslant 1$
Output: z such that $\left|(\mathrm{erfc}\, x - z)/z\right| < 2^{-p}$
begin choose parameters
  $h_0 = \pi/(2 + \sqrt{\mathrm{arcsinh}(2^p\sqrt{\pi})})$;
  $u_0 = e^{-2h_0^2}$;
  $u = \lceil 2^v u_0 \rceil$;
  $n = \lceil \sqrt{p \log 2}/h_0 \rceil$;
end
begin main multiprecision computation
  $h = \sqrt{-\log(u/2^v)/2}$;
  $\lambda = x/h$;
  $U \leftarrow 1$;
  $V \leftarrow \sqrt{u/2^v}$;
  $D \leftarrow \lambda^2$;
  $z \leftarrow U/D$;
  for k=*1 to* n **do**
    $U \leftarrow U \times V$; /* $U_k = e^{-(kh)^2}$ */
    $V \leftarrow V \times u/2^v$; /* $V_k = e^{-(2k+1)h^2}$ */
    $D \leftarrow D + 2k - 1$; /* $D_k = \lambda^2 + k^2$ */
    $z \leftarrow z + U/D$;
  end
  $z \leftarrow z \times 2\lambda$;
  $z \leftarrow z + 1/\lambda$; /* term for $k = 0$ */
  $z \leftarrow z \times e^{-x^2}/\pi$;
end
begin residue term
  if $\mathrm{Re}(x + 1) < \pi/h$ **then**
    $z \leftarrow z - 2/(e^{2\pi\lambda} - 1)$;
  end
end
return z

digits	value	Maple 14	MPFR 3.0.0	integration
100	3	1ms	0.3ms	0.1ms
100	200	0.6ms	0.04ms	0.1ms
100	10 000	0.5ms	0.04ms	0.1ms
100	$\pi + i$	5.6ms	$\star$	0.3ms
100	$\pi + 1\,000i$	2ms	$\star$	0.3ms
1 000	3	5ms	9.9ms	9.8ms
1 000	200	30ms	1.9ms	9.3ms
1 000	10 000	30ms	1.3ms	9.3ms
1 000	$\pi + i$	60ms	$\star$	23ms
1 000	$\pi + 1\,000i$	50ms	$\star$	23ms
10 000	3	0.08s	0.246s	2.280s
10 000	200	16.78s	47.840s	2.290s
10 000	10 000	3.48s	0.301s	2.280s
10 000	$\pi + i$	14.26s	$\star$	7.440s
10 000	$\pi + 1\,000i$	16.34s	$\star$	7.462s

Table 1: timings (on a Intel Core2 Quad, 2.40GHz)

[MR71] F. Matta and A. Reichel. Uniform computation of the error function and other related functions. *Mathematics of Computation*, 25(114):pp. 339–344, 1971.

[Wei94] J. A. C. Weideman. Computation of the complex error function. *SIAM Journal on Numerical Analysis*, 31(5):pp. 1497–1518, 1994.