Fast Evaluation of Special Functions by the Modified Trapezium Rule

MOHAMMAD AL AZAH

Department of Mathematics and Statistics
University of Reading

This dissertation is submitted for the degree of
Doctor of Philosophy

School of Mathematical, Physical and
Computational Sciences

April 2017

I would like to dedicate this thesis to my loving parents, especially to my mother who sadly passed away on 2015 . . .

Declaration

I confirm that this is my own work and that the use of all material from other sources has been properly and fully acknowledged. Chapter 2 is based on the paper [3], joint work with Chandler-Wilde and La Porte, for which I was the principal contributor.

MOHAMMAD AL AZAH
April 2017

Acknowledgements

Table of contents

Chapter 1

Introduction

1.1 Special functions

Special functions arise in the mathematical sciences as non-elementary solutions of differential equations, and these solutions can be represented in different ways. Computing these special functions efficiently is of major interest for scientific applications and we can find formulas for approximating many of them in Abramowitz and Stegun [2] and Luke [42].

(1.1) and evaluated effectively using the trapezium rule (1.6): this method of approximation has been proposed for the incomplete gamma function in [4]; for Bessel functions in [20, 29, 56], for the Airy function in [23], for the gamma function in [53]; and for the error function in [15, 43, 31, 45].

It is well-known [18] that integrals of the form (1.1) with f is given by (1.2) can be approximated by the Hermite-Gaussian quadrature rule, denoted by J_N, which is given by

$$J_N := \frac{1}{\sqrt{\rho}} \sum_{i=1}^{N} w_i F(x_i/\sqrt{\rho}); \tag{1.3}$$

where $w_1, \dots, w_N$ and $x_1, \dots, x_N$ are the weights and abscissae, respectively. The Hermite-Gaussian quadrature rule is very accurate, and sometimes outperforms the trapezium rule, when the function F is smooth; but the accuracy deteriorate when F is meromorphic with simple poles near the real axis. For example, approximating the integral

$$\int_{-\infty}^{\infty} e^{-t^2} \cos(t)\, dt \tag{1.4}$$

using J_N with $N = 12$ (see http://www.chebfun.org/examples/quad/HermiteQuad.html) gives

before Propositions 1.2.3 and 1.2.4. We assume in the following results that the function F in (1.2) satisfies the following assumption.

Assumption 1.2.1. For $H > 0$ and $S_H = \{z \in \mathbb{C} : |\mathrm{Im}(z)| < H\}$, we have that

(i) F is meromorphic with simple poles at $z_j \in S_H$, $\mathrm{Im}(z_j) \neq 0$ and $j = 1, \ldots, m$;

(ii) F is continuous on $\bar{S}_H \setminus \{z_1, z_2, z_3, \ldots, z_m\}$;

(iii) $F(z) = O(1)$ as $|\mathrm{Re}(z)| \to \infty$ uniformly for $|\mathrm{Im}(z)| \leq H$.

Given $h > 0$ and $\alpha \in [0, 1)$, define the function $g(z)$ by

$$g(z) := i \cot\left(\pi\left(\frac{z}{h} + \alpha\right)\right); \tag{1.7}$$

which is a meromorphic function with simple poles at $z = (k - \alpha)h$, $k \in \mathbb{Z}$, which has the properties that, for $z = x + iy$ with $y > 0$,

$$|1 - g(z)| \leq \frac{2e^{-2\pi y/h}}{1 - e^{-2\pi y/h}}; \tag{1.8}$$

and for $z = x + iy$ with $y < 0$,

$$|1 + g(z)| \leq \frac{2e^{2\pi y/h}}{1 - e^{2\pi y/h}}. \tag{1.9}$$

We will make use in the following results of the signum function, $\mathrm{sign}(t)$, which is defined by $\mathrm{sign}(t) = 1$ for $t > 0$, $\mathrm{sign}(0) = 0$ and $\mathrm{sign}(t) = -1$ for $t < 0$. We will make use also of the paths Γ_H and Γ'_H in the complex plane which are defined as the lines $\mathrm{Im}(z) = H$ and $\mathrm{Im}(z) = -H$, respectively, traversed in the direction of increasing $\mathrm{Re}(z)$.

Proposition 1.2.1. If Assumption 1.2.1 holds, then $I(h;\alpha)$ as defined in (1.6) exists as the limit

$$\lim_{n,j\to\infty} h \sum_{k=-j}^{n} f((k-\alpha)h);$$

and has the value

$$I(h;\alpha) = \frac{1}{2}\left(\int_{\Gamma_H} f(z)g(z)\,dz - \int_{\Gamma'_H} f(z)g(z)\,dz\right) + \pi i \sum_{k=1}^{m} g(z_k) R_k. \tag{1.10}$$

where $R_k = \mathrm{Res}(f; z_k)$.

Proof. Let $A_k = (k - a + \frac{1}{2})h$ for $k \in \mathbb{N}$ and define C_H as the positively oriented rectangular contour with vertices at $-A_j \pm iH$ and $A_n \pm iH$. Using Cauchy's residue theorem for C_H (which encloses $j+n+1$ simple poles of the integrand) we find that

$$\int_{C_H} f(z)g(z)\,dz = 2\pi i \sum_{k=1}^{n+j+1} \mathrm{Res}(fg;(k-a)h) + \sum_{k=1}^{m} \mathrm{Res}(fg;z_k$$

The following proposition is well-known from many papers. It is in Goodwin [24] for the case when $a = 0$ and the integrand is analytic in S_H, in Chiarella and Reichel [15,1.956(Reichef)1(1.02

integrand. For example, in Hunter [29, 30] we find this result for the case where the integrand is even and analytic in S_H and $a = 0$; in Hunter and Regan [31] for $a = 0$ and $a = 1/2$ with $F(t) = 1/(t^2 + a^2)$, for some $a \in \mathbb{C}$; in Theorem 2.2 of Bialecki [5] for $a = 0$ when the integrand is meromorphic with poles of arbitrary order, in Theorem 2.3.2 of La Porte [38] for $a = 0$, and recently in Theorem 5.1 of [60] for the case where $a = 0$ and the integrand is analytic in S_H.

Proposition 1.2.4. For $h > 0$ and $a \in [0, 1)$ let $E(h; a) := I - I(h; a)$. If Assumption (1.2.1) holds, then

$$|E(h;a)| \quad 2^{\text{pan [}}$$

Definition 1.2.2. For $h > 0$ and $\alpha = 0$ or $1/2$, we denote by $I_N(h;\alpha)$ the **truncated trapezium rule** defined by

$$I_N(h;0) := h\,f(0) + 2h\sum_{k=1}^{N} f(kh) \quad \text{and} \quad I_N(h;1/2) := 2h\sum_{k=0}^{N} f((k+1/2)h). \qquad (1.22)$$

We denote also by $I_N(h;\alpha)$ the **truncated modified trapezium rule** defined by

$$I_N(h;\alpha) := I_N(h;\alpha) + C(h;\alpha). \qquad (1.23)$$

Note that the truncation of $I(h;\alpha)$ induces the additional error

$$T_N(h;\alpha) := 2h\sum_{k=N+1}^{\infty} f((k+\alpha)h), \qquad (1.24)$$

which will be considered in the coming chapters. The total error in approximating the integral I (1.1) by $I_N(h;\alpha)$ will be denoted by $E_N(h;\alpha)$ where

$$E_N(h;\alpha) = E(h;\alpha) + T_N(h;\alpha). \qquad (1.25)$$

1.3 Numerical Examples

To give a flavour and preview of the extraordinary efficiency of the modified trapezium rule we present here two examples that demonstrate the convergence rate of the rule (1.18). In the first example the integrand is an entire function; and in the second example the integrand is a meromorphic function. In both examples, we approximate the integral by $I_N(h;\alpha)$ with $\alpha = 0$.

1.3.1 Example 1

The following integral is a famous example (see Goodwin [24]):

$$I = \int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi} = 1.7724538509055160273\ldots \qquad (1.26)$$

The integrand here is an entire function and hence we have that $C(h;0) = 0$ so that

$$I(h;0) = I(h;0) = h\sum_{k=-\infty}^{\infty} e^{-k^2h^2} \quad \text{and} \quad I_N(h;0) := I_N(h;0) = 1 + 2h\sum_{k=1}^{N} e^{-k^2h^2}.$$

Table 1.1 shows the computed values of $I_N(h;0)$

N	$h = 0{:}7(N+1)^{2=3}$	$I_N(h;0)$
10	0:142	0:910749
20	0:092	**0**:**88**9598
40	0:059	**0.8875**7706
80	0:037	**0.88753706**862

2. To derive completely rigorous and explicit bounds on both the absolute and relative errors when approximating particular special functions by the truncated modified trapezium rule. The bounds we obtain justify theoretically the choices that we recommend for the parameters α, H, h and N, and prove exponential (or near exponential) convergence as $N \to \infty$. These theoretical predictions are supported by systematic and comprehensive numerical experiments.

The largest part of this thesis is concerned with the application of the truncated modified trapezium rule (1.23) (with $\alpha = 0$ or $\alpha = 1/2$) to the computation of the complex error function $w(z) = e^{-z^2}\operatorname{erfc}(-iz)$ (Chapter 3), and with the related problem of computing Fresnel integrals (Chapter 2). The application of the modified trapezium rule (1.18) with $\alpha = 0$ to compute the complementary error function, denoted by $\operatorname{erfc}(z)$ with $z = x + iy$, starting from the integral representation

$$\operatorname{erfc}(z) = \frac{z e^{-z^2}}{\pi} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{z^2 + t^2}\, dt, \quad x > 0, \tag{1.28}$$

was proposed by Chiarella and Reichel [15] and Matta and Reichel [43] who proposed to use $I(h;0)$ given by (1.18) with $H = \pi/h$, i.e.

$$\operatorname{erfc}(z) \approx \frac{h e^{-z^2}}{\pi z} + \frac{2hz e^{-z^2}}{\pi} \sum_{k=1}^{\infty} \frac{e^{-k^2h^2}}{z^2 + k^2h^2} + \frac{2\mathbf{H}(H-x)}{1 - e^{2\pi z/h}}, \tag{1.29}$$

where $\mathbf{H}$ is the Heaviside step function. This proposal was refined later by Hunter and Regan [31]. In particular, Hunter and Regan [31] noted that (1.29) blows up if the simple poles of the integrand at $t = \pm iz$ coincide with any quadrature point at kh. They proposed to use the approximation $I(h;1/2)$ with $H = \pi/h$, i.e.

$$\operatorname{erfc}(z) \approx \frac{2hz e^{-z^2}}{\pi} \sum_{k=1}^{\infty} \frac{e^{-(k-1/2)^2h^2}}{z^2 + (k-1/2)^2h^2} + \frac{2\mathbf{H}(H-x)}{1 + e^{2\pi z/h}}, \tag{1.30}$$

when (1.29) fails or suffers from numerical instability. They proposed precisely the approximation

$$\operatorname{erfc}(z) \approx \begin{cases} I(h;0), & \text{if } 1/4 \le \{y/h\} \le 3/4 \\ I(h;1/2), & \text{otherwise}, \end{cases} \tag{1.31}$$

where $f(t)$ denotes the fractional part of t, i.e. $f(t) = t - [t]$. They also proved, essentially applying Proposition 1.2.4 with $H = \pi/h$, and noting for

$$F(t) = \frac{ze^{-z^2}}{\pi(z^2+t^2)}$$

it holds that

$$M_H(F) \le \frac{|ze^{-z^2}|}{\pi|x^2-\pi^2/h^2|},$$

that the error in this approximation is

$$\frac{|ze^{-z^2}|e^{-\pi^2/h^2}}{\sqrt{\pi}\,|x^2-\pi^2/h^2|(1-e^{-2\pi^2/h^2})}. \tag{1.32}$$

Clearly this error bound blows up when $x = \pi/h$, and so is inadequate as a bound for $x \approx \pi/h$. This can be fixed by finding an improved version for $|x - \pi/h| \ge \epsilon$, for some $\epsilon > 0$, by taking $H = \pi/h - \epsilon$ in Proposition 1.2.4, but the bounds obtained with this modification are still unsatisfactory as they don't imply small absolute and relative errors as $h \to 0$ uniformly in $z = x + iy$.

Mori [45] studied the approximation $I(h;0)$ in (1.29) specifically for $z = x > 0$. He bounded the error in this approximation by (1.32) and by another bound obtained from Proposition 1.2.4 with $H = \pi/h + 1/\sqrt{2}$, namely that the error is

$$\frac{xe^{-x^2}e^{1/2}e^{-\pi^2/h^2}}{\sqrt{\pi}\,|x^2-(\pi/h+1/\sqrt{2})^2|(1-e^{-2\pi/h(\pi/h+1/\sqrt{2})})}. \tag{1.33}$$

Mori [45] used the minimum of the bounds (1.32) and (1.33), i.e. he used (1.32) for $x > b$, (1.33) for $0 < x \le b$, where b is the value (given by (2.8) and (2.9) in [45], but here we correct a calculation error in [45])

$$b := \left[\frac{1}{1+\lambda}\left(\left(\frac{\pi}{h}+\frac{1}{\sqrt{2}}\right)^2 + \lambda\left(\frac{\pi}{h}\right)^2\right)\right]^{1/2}, \tag{1.34}$$

with

$$\lambda := \frac{1-e^{-2\pi^2/h^2}e^{1/2}}{1-e^{-2\pi/h(\pi/h+1/\sqrt{2})}}, \tag{1.35}$$

for this value of β the two bounds (1.32) and (1.33) coincide. Mori [45] also bounded the relative error, using that

$$\mathrm{erfc}(x) \geq \frac{2e^{-x^2}}{\sqrt{\pi}(x+\sqrt{x^2+2})}; \quad x \geq 0. \tag{1.36}$$

Mori [45] showed further that the relative error in (1.29) is

$$\leq \frac{\beta(\beta+\sqrt{\beta^2+2})}{(\beta^2-\pi^2/h^2)(1-e^{-2\pi^2/h^2})} e^{-\pi^2/h^2}, \tag{1.37}$$

for all $z = x \geq 0$.

The work in this thesis extends and improves significantly, by more sophisticated and delicate analysis, the previous works. In Chapter 2 we propose methods for computing Fresnel integrals based on the truncated modified trapezium rule in (1.23) where $\alpha = 1/2$. We construct approximations in Sections §2.3 and §2.4 which we prove are exponentially convergent as a function of N, the number of quadrature points, obtaining completely explicit error bounds in Theorems 2.3.3 and 2.3.5 which show that accuracies of 10^{-15} uniformly on the real line are achieved with $N = 12$, this confirmed by computations in Section §2.5. The approximations we obtain are attractive in that they maintain small relative errors for small and large argument, are analytic on the real axis (echoing the analyticity of the Fresnel integrals), and are straightforward to implement.

In Chapter 3 we propose a method for computing the complex error function $w(z)$

Matlab codes are provided (see Listings A.1, A.2, A.3 and A.4) for computing all these functions, and these codes are easily adaptable to other programming languages.

Chapter 2

Fresnel integrals

2.1 Introduction

Let $C(x)$,

It also depends on the integral representation [2, (7.1.4)] that

$$w(z) = \frac{i}{\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{z-t}\,dt = \frac{iz}{\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{z^2-t^2}\,dt; \quad \mathrm{Im}(z) > 0. \qquad ()$$

[redacted]

where the size of N controls the accuracy of the approximation, $L = 2^{1/4}N^{1/2}$ and the coefficients are computed as

$$a_n := \quad 1$$

2.2 Summary of the main Results

Based on the truncated modified trapezium rule (1.23) with $a = 1/2$ and $H = A_N$ (given by (2.13)), the approximation to $F(x)$ we propose is

$$F_N(x) := \frac{1}{2} + \frac{i}{2}\tan\left(A_N x e^{i\pi/4}\right) + \frac{x}{A_N} e^{i(x^2+\pi/4)} \sum_{k=1}^{N} \frac{e^{-t_k^2}}{x^2 + i t_k^2} \tag{2.11}$$

$$= \frac{1}{\exp\left(2A_N x e^{i\pi/4}\right)}$$

in modified form into this strip. This implies exponentially convergent error estimates, presented in §2.3.1 and §2.4, for the difference between the coefficients in the Maclaurin series of F, C, and S and those in the corresponding series for F_N, C_N and S_N. In turn (see §2.4), this implies that the approximations all retain small relative error for $|x|$ small, and the computations in §2.5 demonstrate this.

These approximations inherit symmetries of the Fresnel integrals. In particular, our

$< 10^{15}$. From (2.6) we have that, for $x > 0$,

$$F(x) := \int_{\infty}^{\infty} f(t)\,dt; \quad \text{where}\ f(t) := e^{i(x^2+\pi/4)} \quad x$$

Here

$$d_1(x) := \frac{x e^{-\pi^2/h^2}}{\sqrt{\pi}\,|\pi^2/h^2 - x^2/2|\,\left(1 - e^{-2\pi^2/h^2}\right)}; \tag{2.29}$$

$$d_2(x) := \frac{4hx e^{-\pi^2/h^2}}{\sqrt{\pi}\,|\pi/h + x/\sqrt{2}|\,\left(1 - e^{-2\pi^2/h^2}\right)}\left(1 + 2\sqrt{\pi}\, e^{-b\pi^2/h^2}\right); \tag{2.30}$$

with $b = \dfrac{15 - 10\sqrt{2}}{16} \approx 0.0536$, and

$$d_3(x) := d_1(x) + \frac{e^{-\sqrt{2}\pi x/h}}{1 - e^{-\sqrt{2}\pi x/h}}. \tag{2.31}$$

Proof. Applying Proposition 1.2.4, for $0 < x < \sqrt{2}\pi/h$, with $H = \pi/h$, and noting for

$$F(t) = \frac{x e^{-i(x^2+\pi/4)}}{2\pi(x^2 + it^2)} + \ldots = d_1$$

j664945 10.037 Td 978 0 Td [(e)]TJ/F92 Ts(9ds

Thus, and applying (1.8), similarly to (2.29) we deduce that

$$\left|\int_{\Gamma'} f(z)(1+g(z))\,dz\right| \le \frac{\xi e^{-\pi^2/h^2}}{\sqrt{\pi}\,\varepsilon|\pi/h+\xi/\sqrt{2}|\,(1-e^{-2\pi^2/h^2})}. \tag{2.35}$$

To bound the integral over γ we note that, for $z = X + iY = z_0 + \varepsilon e^{i\theta} \in \gamma$, (2.34) is true and $Y \ge H$. Further, $|e^{-z^2}| = e^P$, where

$$P = Y^2 - X^2 = 2\xi\varepsilon\sin(\theta - \pi/4) - \varepsilon^2\cos(2\theta) < 2\xi\varepsilon + \varepsilon^2 \le 2\sqrt{2}H\varepsilon + (2\sqrt{2}+1)\varepsilon^2,$$

since $\xi/\sqrt{2} - H < \varepsilon$. From these bounds and (1.8), defining $\alpha = \varepsilon/H \in (0,1)$, we deduce that

$$\left|\int_{\gamma} f(z)(1+g(z))\,dz\right| \le \frac{2\xi\exp((2\sqrt{2}\alpha + (2\sqrt{2}+1)\alpha^2 - 2)\pi^2/h^2)}{\varepsilon|\pi/h+\xi/\sqrt{2}|\,(1-e^{-2\pi^2/h^2})}. \tag{2.36}$$

For $\xi/\sqrt{2} - H < \varepsilon$ we can bound e_h using (2.33), (2.35), (2.36), and the triangle inequality, to get that

$$|e_h| \le d_2(\xi) := \frac{4h\xi e^{-\pi^2/h^2}}{\sqrt{\pi}\,\pi|\pi/h+\xi/\sqrt{2}|\,(1-e^{-2\pi^2/h^2})}\left(1 + 2\sqrt{\pi}\,e^{-\beta\pi^2/h^2}\right), \tag{2.37}$$

where

$$\beta = 1 - 2\sqrt{2}\alpha - (2\sqrt{2}+1)\alpha^2. \tag{2.38}$$

Proposition 2.3.1. For $x > 0$,

$$|T_N(h;1/2)| \le \frac{(2ht_{N+1}+1)x}{2\pi t_{N+1}\sqrt{x^4+t_{N+1}^4}}\, e^{-t_{N+1}^2}.$$

Proof.

$$
\begin{aligned}
|T_N(h;1/2)| &\le \frac{hx}{\pi}\sum_{m=N+1}^{\infty}\frac{e^{-t_m^2}}{\sqrt{x^4+t_m^4}}\\
&\le \frac{x}{2\pi\sqrt{x^4+t_{N+1}^4}}\left(2he^{-t_{N+1}^2}+2h\sum_{m=N+2}^{\infty}e^{-t_m^2}\right)\\
&\le \frac{x}{2\pi\sqrt{x^4+t_{N+1}^4}}\left(2he^{-t_{N+1}^2}+2\int_{t_{N+1}}^{\infty}e^{-t^2}dt\right)\\
&\le \frac{x}{2\pi\sqrt{x^4+t_{N+1}^4}}\left(2he^{-t_{N+1}^2}+\frac{e^{-t_{N+1}^2}}{t_{N+1}}\right) = \frac{(2ht_{N+1}+1)x}{2\pi t_{N+1}\sqrt{x^4+t_{N+1}^4}}\, e^{-t_{N+1}^2}.
\end{aligned}
$$

To arrive at the last line we have used that, for $x > 0$,

$$2\int_x^{\infty} e^{-t^2}dt = \frac{e^{-x^2}}{x} - \int_x^{\infty}\frac{e^{-t^2}}{t^2}dt < \frac{e^{-x^2}}{x}. \tag{2.40}$$

□

At this point we make a choice of h to approximately equalise $D_h(x)$ in Theorem 2.3.1 and the bound on $T_N(h;1/2)$ in Proposition 2.3.1, choosing h so that $\pi/h = t_{N+1} = (N+1/2)h$, giving that

$$h = \sqrt{\pi/(N+1/2)}; \tag{2.41}$$

in which case $t_{N+1} = A_N = \sqrt{(N+1/2)\pi}$, and $t_k = t_k$, where t_k is defined by (2.13). Making this choice of h we see that

$E_N(x) = F(x) - F_N(x) = e_h$

Td [978 0 Td [(2)]2503h-h [(h)]TJ/F1021.95544u64 For x > 0,
jTN(h
2h +
h 2
9.9552 Tf 4.638 0 Td 973

Theorem 2.3.2. For $h = \sqrt{\pi/(N+1/2)}$ so that $H = \pi/h = A_N$ we have that

$$|E_N(x)| \le h_N(x) := D_h(|x|) + \frac{(2\pi+1)|x|}{2\pi A_N\sqrt{x^4+A_N^4}}\, e^{-A_N^2}, \tag{2.42}$$

where

$$D_h(x) = \begin{cases} \dfrac{x e^{-A_N^2}}{\sqrt{\pi}\,(A_N^2 - x^2/2)\left(1-e^{-2A_N^2}\right)}, & 0 \le \dfrac{x}{\sqrt{2}} \le \frac{3}{4}A_N; \\[2ex] \dfrac{4x e^{-A_N^2}\left(1+2\sqrt{\pi}\, e^{-bA_N^2}\right)}{\sqrt{\pi}\,A_N(A_N + x/\sqrt{2})\left(1-e^{-2A_N^2}\right)}, & \frac{3}{4}A_N < \dfrac{x}{\sqrt{2}} < \frac{5}{4}A_N; \\[2ex] \dfrac{x e^{-A_N^2}}{\sqrt{\pi}\,(x^2/2 - A_N^2)\left(1-e^{-2A_N^2}\right)} + \dfrac{e^{-\sqrt{2}A_N x}}{1-e^{-\sqrt{2}A_N x}}, & \dfrac{x}{\sqrt{2}} \ge \frac{5}{4}A_N. \end{cases} \tag{2.43}$$

Theorem 2.3.3. For $x > 0$,

$$|F(x) - F_N(x)| = |E_N(x)| \le h_N(x) \le c_N \frac{e^{-\pi N}}{\sqrt{N+1/2}}, \quad \text{for } x \in \mathbb{R}, \tag{2.44}$$

where

$$c_N = \frac{20\sqrt{2}\, e^{-\pi/2}}{9\pi\left(1-e^{-2A_N^2}\right)}\left(1+2\sqrt{\pi}\, e^{-bA_N^2}\right) + \frac{(2\pi+1)e^{-\pi/2}}{2\sqrt{2}\,\pi^{3/2}A_N},$$

which decreases as N increases, with

$$c_1 \le 0.825 \quad \text{and} \quad \lim_{N\to\infty} c_N = \frac{20\sqrt{2}\, e^{-\pi/2}}{9\pi} \approx 0.208. \tag{2.45}$$

Proof. It is easy to see that $D_h(x)$ is increasing on $[0, \frac{5}{4}\sqrt{2}A_N)$ and decreasing on $[\frac{5}{4}\sqrt{2}A_N, \infty)$. Further, where $D_h(\frac{5}{4}\sqrt{2}A_N^-)$ denotes the limiting value of $D_h(x)$ as $x \to \frac{5}{4}\sqrt{2}A_N$ from below, since $2A_N^{-1} > e^{-A_N^2}$,

$$\begin{aligned} D_h\left(\tfrac{5}{4}\sqrt{2}A_N^-\right) &= \frac{20\sqrt{2}\, e^{-A_N^2}}{9\sqrt{\pi}\,A_N\left(1-e^{-2A_N^2}\right)}\left(1+2\sqrt{\pi}\, e^{-bA_N^2}\right) \\ &> \frac{20\sqrt{2}\, e^{-A_N^2}}{9\sqrt{\pi}\,A_N\left(1-e^{-2A_N^2}\right)} + \frac{e^{-5A_N^2/2}}{1-e^{-5A_N^2/2}} = D_h\left(\tfrac{5}{4}\sqrt{2}A_N\right). \end{aligned}$$

Similarly, $xD_h(x)$ is increasing on $[0, \frac{5}{4}\sqrt{2}A_N)$ and decreasing on $[\frac{5}{4}\sqrt{2}A_N, \infty)$. Thus, for $x > 0$,

$$D_h(x) \le D_h\left(\tfrac{5}{4}\sqrt{2}A_N\right) \quad \text{and} \quad xD_h(x) \le \tfrac{5}{4}\sqrt{2}A_N D_h\left(\tfrac{5}{4}\sqrt{2}A_N\right). \tag{2.46}$$

Moreover,

$$\frac{x}{\sqrt{x^4 + A_N^4}} \le \frac{1}{\sqrt{2}A_N} \quad \text{and} \quad \frac{x^2}{\sqrt{x^4 + A_N^4}} < 1, \quad \text{for } x > 0. \tag{2.47}$$

Combining (2.42), (2.46) and (2.47) we reach the result. □

Remark 2.3.1. We have shown the bounds (2.42) and (2.44) for $x > 0$, but the symmetries (2.17) and (2.18) imply that $E_N(-x) = -E_N(x)$, so that (2.42) and (2.44) hold also for $x < 0$, and, by continuity, also for $x = 0$ (and in fact $E_N(0) = h_N(0) = 0$).

The following result from [3, Theorem 4] will be used to bound the relative error of $F_N(x)$.

Lemma 2.3.4. For the Fresnel integral $F(x)$ we have that

$$|F(x)| \ge \begin{cases} \dfrac{1}{2 + 2\sqrt{\pi}x}, & \text{for } x \ge 0 \\[2ex] \dfrac{1}{2}, & \text{for } x \le 0. \end{cases} \tag{2.48}$$

Theorem 2.3.5. For the Fresnel integral $F(x)$ and its approximation $F_N(x)$ we have that

$$\frac{|F(x) - F_N(x)|}{|F(x)|} \le \frac{h_N(x)}{|F(x)|} \le \begin{cases} c_N e^{-\pi N}, & \text{for } x \ge 0; \\[1ex] 2c_N \dfrac{e^{-\pi N}}{\sqrt{N + 1/2}}, & \text{for } x \le 0; \end{cases} \tag{2.49}$$

where

$$c_N = \frac{10\sqrt{2}\left(4 + 5\sqrt{2\pi}A_N\right)\left(1 + 2\sqrt{\pi}e^{-\beta A_N^2}\right)}{9\sqrt{\pi}e^{\pi/2}A_N\left(1 - e^{-2A_N^2}\right)} + \frac{(2\pi + 1)}{\pi e^{\pi/2}A_N}\left(\frac{1}{\sqrt{2}A_N} + \sqrt{\pi}\right),$$

which decreases as N increases, with $c_1 \le 10.4$ and $\lim_{N\to\infty} c_N = 100e^{-\pi/2}/9 \approx 2.3$.

Proof. Combining (2.42), (2.46), (2.47) and (2.48) we see, for $x > 0$, that

$$\frac{h_N(x)}{|F(x)|} \le \left(2 + \frac{5}{2}\sqrt{\pi} \right.$$

This implies (2.49) for $x > 0$. The bound for $x \quad 0$ follows immediately from (2.48), (2.44) and Remark 2.3.1. □

The above estimates use (2.42) and (2.43)

0 arg(z) p=2; moreover, it is clear from (2.12) that the same holds for $F_N(z)$ and hence for $E_N(z)$. Thus (2.52) implies that (2.44) holds for 0 arg(z) p=2, and (2.17) and (2.18) then imply that (2.44) holds also for p arg(z) 3p=4.

It is clear from the derivations above that, if h is given by (2.41), then I (h; 1=2) also satisfies the bound (2.44), i.e.,

$$jF(z) \quad I\ (h; 1=2)j \quad c_N\, p\frac{e^{\ \ pN}}{N+1=2}; \tag{2.53}$$

this holding in the first instance for real z, then for imaginary z, and finally for all z in the first and third quadrants. The bound (2.44) cannot hold in the second or fourth quadrant because $E_N(z) = F(z) \quad F_N(z)$ has poles there. This issue does not hold for $F(z) \quad I\ (h; 1=2)$, which is an entire function, but (2.53) cannot hold in the whole compleW 0 0 12Tf 7.531 0 Td [(tcomple).8n7

Thus, for $z = x + iy$ in the second and fourth quadrants with $|y| \le A_N/(2\sqrt{2})$,

$$|F(z) - F_N(z)| \le \hat{c}_N e^{-xy} \frac{e^{-\pi N}}{\sqrt{N+1/2}}; \tag{2.55}$$

where

$$\hat{c}_N := c_N + \frac{\sqrt{2}(2\pi+1)}{\pi^{3/2}\exp(\pi/2)\sqrt{N+1/2}}. \tag{2.56}$$

The sequence $\hat{c}_N$ is decreasing with $\hat{c}_1 \approx 1.14$ and $\lim_{N\to\infty} \hat{c}_N = \lim_{N\to\infty} c_N \approx 0.208$.

We observe above that the bound (2.44) on $E_N(z) = F(z) - F_N(z)$ holds for all complex z in the first and third quadrants of the complex plane, and on the boundaries of those quadrants, the real and imaginary axes, while the bound (2.55) holds in the second and fourth quadrants for $|\mathrm{Im}(z)| \le A_N/(2\sqrt{2})$. A significant implication of these bounds is that they imply that the coefficients in the Maclaurin series of $F_N(z)$ are close to those of $F(z)$. Precisely, at least for $|z| < A_N/\sqrt{2}$,

$$F(z) = \sum_{n=0}^{\infty} a_n z^n \quad \text{and} \quad F_N(z) = \sum_{n=0}^{\infty} b_n z^n;$$

with a_n

and are given explicitly in (2.14) and (2.15). We note the similarity between (2.14) and (2.15) and the formulae [46, (7.5.3)-(7.5.4)]

$$C(x) = \tfrac{1}{2} + f(x)\sin\left(\tfrac{1}{2}\pi x^2\right) - g(x)\cos\left(\tfrac{1}{2}\pi x^2\right); \tag{2.60}$$

$$S(x) = \tfrac{1}{2} - f(x)\cos\left(\tfrac{1}{2}\pi x^2\right) - g(x)\sin\left(\tfrac{1}{2}\pi x^2\right); \tag{2.61}$$

which express $C(x)$ and $S(x)$ in terms of the auxiliary functions, $f(x)$ and $g(x)$, for the Fresnel integrals [46, §7.2(iv)]. Indeed, it follows from [46, (7.7.10)-(7.7.11)] that, for $x > 0$, $f(x)$ and $g(x)$ have the integral representations

$$f(x) = \frac{\sqrt{\pi}\,x^3}{2}\int_0^\infty \frac{e^{-t^2}}{\left(\frac{\pi}{2}x^2\right)^2 + t^4}\,dt \quad \text{and} \quad g(x) = \frac{x}{\sqrt{\pi}}\int_0^\infty \frac{t^2 e^{-t^2}}{\left(\frac{\pi}{2}x^2\right)^2 + t^4}\,dt;$$

and, recalling that A_N is linked to the quadrature step-size through (2.41), it is clear that, for $x > 0$, $\sqrt{\pi}\,x a_N\left(\frac{\pi}{2}x^2\right)/A_N$ and $\sqrt{\pi}\,x b_N\left(\frac{\pi}{2}x^2\right)/A_N$ can be viewed as quadrature approximations to these integrals.

The approximations (2.14) and (2.15) inherit the accuracy of $F_N(x)$ on the real line: from (2.58) and (2.59) we see, for $x \in \mathbb{R}$, that

$$|C(x) - C_N(x)| \le \sqrt{2}\,|E_N(\sqrt{\pi/2}\,x)| \quad \text{and} \quad |S(x) - S_N(x)| \le \sqrt{2}\,|E_N(\sqrt{\pi/2}\,x)|. \tag{2.62}$$

where $E_N(x) = F(x) - F_N(x)$. Thus the error bounds of the previous section can be applied. In particular, from (2.44) and (2.50) it follows that both $|C(x) - C_N(x)|$ and $|S(x) - S_N(x)|$ are

$$2c_N \frac{e^{-\pi N}}{\sqrt{2N+1}}, \quad \text{for } x \in \mathbb{R}; \tag{2.63}$$

and

$$\sqrt{\pi}\,\tilde{c}_N|x|\frac{e^{-\pi N}}{2N+1}, \quad \text{for } |x| \le \sqrt{N+1/2}. \tag{2.64}$$

Here $c_N < 0.83$ and $\tilde{c}_N < 0.18$ are the decreasing sequences of positive numbers defined by (2.14) and (2.51), respectively.

These bounds show that $C_N(x)$ and $S_N(x)$ are exponentially convergent as $N \to \infty$, uni-
.955iv

the power series [46, §7.6(i)]

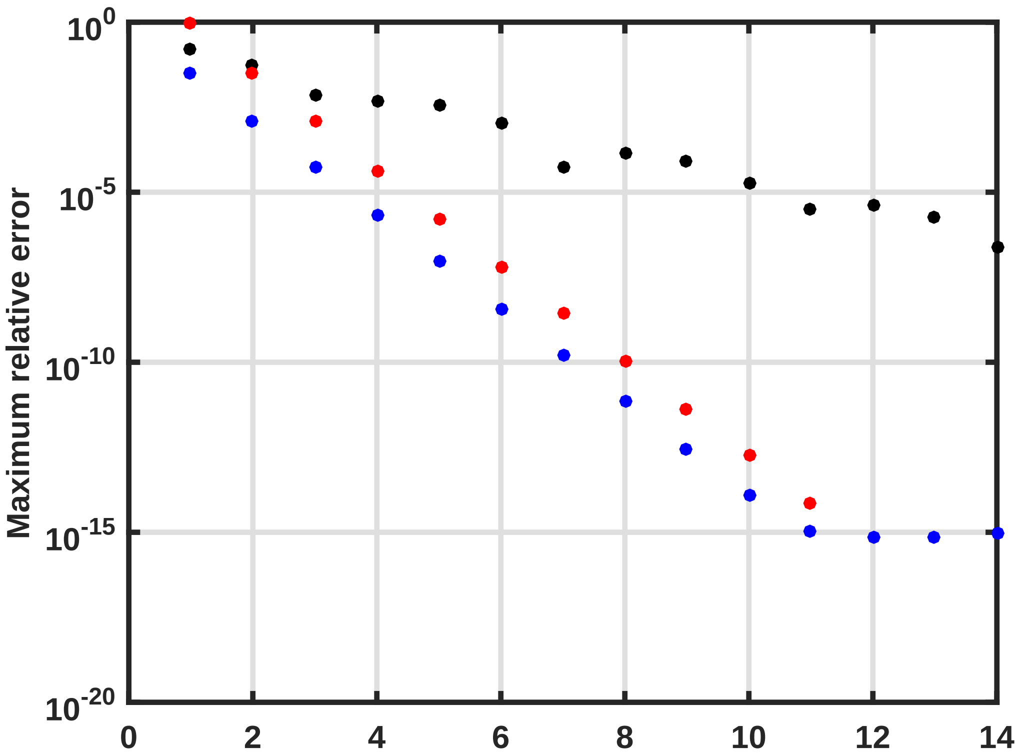
$$C(x) = \sum_{n=0}^{\infty} \frac{(-1)^n \left(\frac{1}{2}\pi\right)^{2n} x^{4n+1}}{(2n)!(4n+1)}; \quad S(x) = \sum_{n=0}^{\infty} \frac{(-1)^n \left(\frac{1}{2}\pi\right)^{2n+1} x^{4n+3}}{(2n+1)!(4n+3)}: \tag{2.65}$$

It follows from the analyticity of $F_N(x)$ inF6603J/F6r[(å)]TJ/F69 8.9664 t 0.9discussed TJ/F65

equally spaced numbers between 0 and 1,000. The average elapsed times were 11.1 and 15.6 seconds, respectively, so that F(x,12) is almost 50% faster.

In Figure 2.2 we see that the theoretical error bounds are upper bounds as claimed, and that these bounds appear to capture the x-dependence of the errors fairly well, for example that $E_N(x) = O(x)$ as $x \to 0$, $= O(x^{-1})$ as $x \to \infty$, and that $E_N(x)$ reaches a maximum at about $x = \sqrt{2}A_N = \sqrt{\pi(2N+1)}$ (≈ 7.7 when $N = 9$).

Turning to $C(x)$ and $S(x)$, in Figure 2.3 we have plotted the maximum values of the absolute and relative errors in $S_N(x)$ and $C_N(x)$, computed using fresnelCS in Table A.2. As accurate values for $C(x)$ and $S(x)$ we use $C_{20}(x)$ and $S_{20}(x)$ for $x > 1.5$ while, for $0 < x < 1.5$ (following [52]) we approximate by the series (2.65) truncated after 15 terms, evaluated by the Horner algorithm. Exponential convergence is seen in Figure 2.3: the absolute errors are $\leq 4.5 \times 10^{-16}$ for $N \geq 11$, the maximum relative error in $C_N(x)$ is $\approx 3.6 \times 10^{-15}$ for $N = 11$ but that in $S_N(x)$ as large as 2.7×10^{-13}. These errors may be entirely acceptable, but the truncated power series (2.65) must achieve smaller errors for small x and is cheaper to evaluate. (Evaluating at 10^7 equally spaced points between 0 and 1.5 takes 2.9 times longer in Matlab with fresnelCS than evaluating 15 terms of both the series (2.65) via Horner's algorithm.)


10^0
10^-5
10^-10
10^-15
10^-20
Maximum relative error
0
2
4
6
8
10
12
14
N

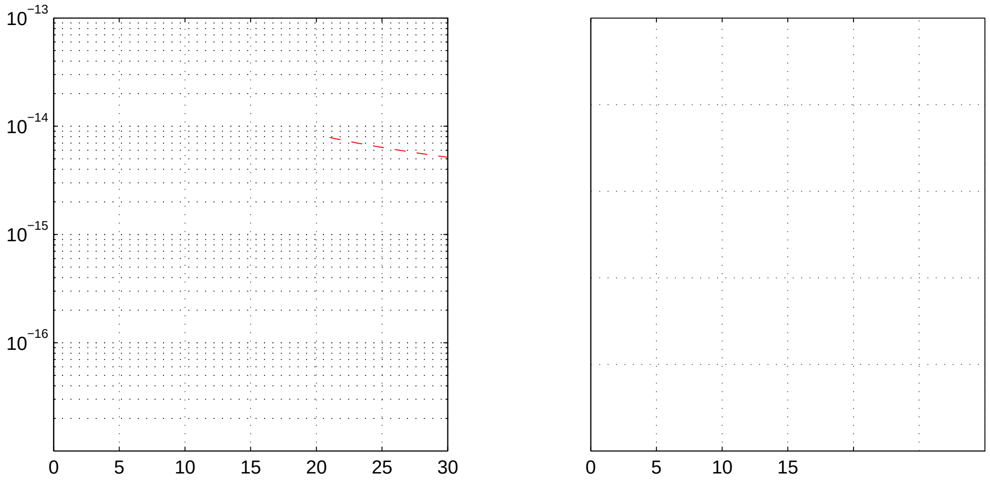


Fig. 2.2 Left hand side: Absolute error, $|F(x) - F_N(x)|$ ( ), and its upper bound $h_N(x)$ given by (2.42) ( ), plotted against x. Right hand side: Relative error, $|F(x) - F_N(x)|/|F(x)|$ ( ), and its upper bound $2(1+\sqrt{p\,x})h_N(x)$ ( ), plotted against x. In both figures $N = 9$ and the exact value for $F(x)$ is approximated by F_{20}

Chapter 3

The Faddeeva function

3.1 Introduction

The complex error function is defined by [46, (7.2.1)]

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt;$$

quadrants can be obtained using the symmetries [50, (3.1) and (3.2)]

$$w(-z) = e^{-z^2} - w(z) \quad \text{and} \quad w(\bar{z}) = \overline{w(-z)}. \tag{3.5}$$

Chiarella and Reichel [15] and Matta and Reichel [43] first proposed to compute erfc(z) for complex z by $I(h;0)$ given by (1.18) with $H = \pi/h$ starting from the integral representation, which follows from (3.4), that

$$\operatorname{erfc}(z) = \frac{z e^{-z^2}}{\pi}\int_{-\infty}^{\infty} \frac{e^{-t^2}}{z^2+t^2}\,dt; \quad \operatorname{Re}(z) > 0. \tag{3.6}$$

Hunter and Regan [31] discussed the stability of these approximations when z is near one of the quadrature points, and proposed to use the formula $I(h;0)$, if $|\phi(y/h) - 0.5| \leq 0.25$, otherwise to use formula $I(h;1/2)$ given by (1.18) with $H = \pi/h$, where $y = \operatorname{Im}(z)$ and

$$\phi(t) = t - [t] \in [0;1) \tag{3.7}$$

is the function that gives the fractional part of t. This criterion and proposal is our main starting point for the methods developed in this chapter to approximate $w(z)$.

There are a number of other effective schemes for computation of $w(z)$, and we briefly summarise here the best of these. Gautschi [22] proposed an approximation for complex z based on continued fractions and this approximation is the basis of ACM TOM Algorithm 680 in Poppe and Wijers [50] which achieves a relative error of 10^{-14} over nearly all the complex plane by Taylor expansions of degree up to 20 in an ellipse around the origin, convergents of up to order 20 of continued fractions outside a larger ellipse, and a more expensive mix of Taylor expansion and continued fraction calculations in between.

Weideman [62] proposed a rational approximation (the derivation starts from the integral representation (3.4)) to compute $w(z)$, for $\operatorname{Im}(z) > 0$. The approximation proposed is

$$w(z) \approx \frac{1}{\sqrt{\pi}(L-iz)} + \frac{2}{(L-iz)^2}\sum_{n=0}^{N-1} a_{n+1}\left(\frac{L+iz}{L-iz}\right)^n; \tag{3.8}$$

where the size of N controls the accuracy of the approximation, $L = 2^{-1/4}N^{1/2}$ and the coefficients are computed as

$$a_n := \frac{1}{2M}\sum_{j=-M+1}^{M-1}(L^2+t_j^2)e^{-t_j^2}e^{-in\theta_j}; \quad n = 1;\ldots;N; \tag{3.9}$$

with $M = 2N$, $t_j = L\tan(\theta_j/2)$ and $\theta_j = \pi j/M$ for $j = -M+1, \ldots, M-1$. Weideman [62] argued that, for intermediate values of $|z|$, and as measured by operation counts, the work required to compute $w(z)$ to 10^{-14} relative accuracy is much smaller for the approximation (3.8) than for ACM TOMS Algorithm 680 in [50].

Remark 3.1.1. Weideman [62] also compared his method to the modified trapezium rule approximation developed in [43, 31] and commented that the trapezium rule "**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.**" As we will see shortly, we address this comment

z1 86.208 T.9552 Tf 9.325 305.227 1 8 275

$\mathrm{erfcx}(y) = e^{y^2}\mathrm{erf}(y)$ and

$$S_1 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2}\, e^{(a^2k^2 + x^2)};$$
$$S_2 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2}\, e^{(ak + x)^2};$$
$$S_3 := \sum_{k=1}^{\infty} \frac{1}{a^2k^2 + y^2}\, e^{(ak\ x)^2}; \qquad (3.14)$$
$$S_4 := \sum_{k=1}^{\infty} \frac{ak}{a^2k^2 + y^2}\, e^{(ak + x)^2};$$
$$S_5 := \sum_{k=1}^{\infty} \frac{ak}{a^2k^2 + y^2}\, e^{(ak\ x)^2}.$$

2)12 1 a)12

TJ/TJ/F66 8277(2.965.977 3.455103)]TJ/F66 11.9538384f1287f939f 11.4p5 -116.8620.755(v)552 T056 2= Tf 11.

13

The authors have supplied us with the Matlab implementation of this method [64] in the form of a Matlab function Faddeyeva_v2(z,M), where the parameter M is the number of accurate significant figures required, and the code enforces a choice of M in the range $4 \le M \le 13$. In this Matlab implementation the choice $a = 1/2$ is made and the sums in (3.14) are truncated, the number of terms retained depending in a complicated way on M. Zagloul and Ali [63] argued, using numerical calculations, that the approximation (3.11), with appropriate choices for a and truncation of (3.14)

$$A_m := \frac{\sqrt{\pi}(2m-1)}{2^{2M}h} \sum_{n=-N}^{N} e^{\alpha^2/4 - n^2h^2} \sin\left(\frac{\pi(2m-1)(nh+\alpha/2)}{2^M h}\right); \tag{3.19}$$

and

$$B_m := \frac{i}{\sqrt{\pi}\,2^{M-1}} \sum_{n=-N}^{N} e^{\alpha^2/4 - n^2h^2} \cos\left(\frac{\pi(2m-1)(nh+\alpha/2)}{2^M h}\right). \tag{3.20}$$

Abrarov and Quine [1] argued, based on numerical calculations, that the approximation (3.16) is more accurate and faster (using the same number of summation terms in (3.16) as in (3.8)) than the approximation (3.8). We will be investigating these claims in Section §3.4 and we will be comparing the efficiency (accuracy and speed) of $w_N(z)$ given in (3.21) with the approximations (3.8), (3.11) and (3.16).

We end this introduction by outlining the remainder of this chapter. Section 3.2 gives summary of the main results; §3.3 is concerned with the proposed approximation and its error bounds and §3.4 explores, using the theoretical and numerical calculations, the accuracy of our approximation in comparison with the approximations (3.8), (3.11) and (3.17).

3.2 Summary of the main results

The main contributions of this chapter are: (i) to propose a family of approximations to $w(z)$, based on the truncated modified trapezium rules defined in (1.22) adopting (at least for $0 \leq \arg(z) < \pi/4$) the proposals of Hunter and Regan [31], but making explicit the choice of the step-size h as a function of N, the number of quadrature points addressing the criticism in Remark 3.1.1 by Weideman [62]; (ii) to prove completely explicit and rigorous bounds on both the absolute and relative errors as a function of N, uniform in $z = x + iy$, with $x, y \geq 0$; and (iii) to demonstrate through the bounds and numerical experiments the high accuracy and efficiency of our approximation in comparison with the approximations (3.8), (3.12), (3.13) and (3.17).

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

$$w_N(z) := \begin{cases} I_N(h;1/2), & y \geq \max(x, \pi/h); \\ I_N(h;0), & y < x \text{ and } |f(x/h) - 1/2| \leq 1/4; \end{cases}$$

x 0 Td [(;)]T90 G [-2

where φ is defined by (3.7),

$$I_N(h;1/2) := \frac{2ihz}{\pi}\sum_{k=0}^{N}\frac{e^{-t_k^2}}{z^2-t_k^2}; \tag{3.22}$$

$$I_N(h;1/2) := \frac{2e^{-z^2}}{1+e^{-2i\pi z/h}} + I_N(h;1/2); \tag{3.23}$$

$$I_N(h;0) := \frac{2e^{-z^2}}{1-e^{-2i\pi z/h}} + \frac{ih}{\pi z} + \frac{2ihz}{\pi}\sum_{k=1}^{N}\frac{e^{-t_k^2}}{z^2-t_k^2}; \tag{3.24}$$

$$h = \sqrt{\frac{\pi}{N+1}}; \quad t_k := (k+1/2)h \quad \text{and} \quad t_k := kh. \tag{3.25}$$

The main error estimate that we prove is

Theorem 3.2.1. Suppose $w_N(z)$ is given by (3.21). Then, for $z = x+iy$ with

The approximation w_N is proven in Theorem 3.2.1 (where we give completely explicit error bounds) to converge exponentially, uniformly in the first quadrant with respect to both absolute and relative errors, and this predicted rate of exponential convergence is observed in numerical experiments in Section §3.4 below (we know of no other rigorous error bounds for approximations for $w(z)$ in the whole quadrant $\mathrm{Re}(z), \mathrm{Im}(z) \geq 0$).

This approximation is straightforward to code. Listing A.3 shows the Matlab code used to evaluate w_N for all the computations in this paper.

The approximation w_N is very competitive in accuracy and operation counts with other methods, as discussed in Section §3.4.

3.3 The proposed approximation and its error bounds

In this section we derive the approximation $w_N(z)$ given by (3.21) and its error bounds which demonstrate that the absolute and relative errors are both converging exponentially as N (the number of quadrature points) increases.

We can rewrite (3.4) as

$$w(z) = \int_{-\infty}^{\infty} f(t)\, dt, \tag{3.32}$$

where

$$f(t) = e^{-t^2} F(t) \quad \text{and} \quad F(t) = \frac{iz}{\pi(z^2 - t^2)}. \tag{3.33}$$

Note that the function $e^{-t^2}F(t)$ is even and meromorphic with simple poles at $t = \pm z$. The residues at these two simple poles are

$$R_1 = \mathrm{Res}(f; z) = \frac{ie^{-z^2}}{2\pi} \quad \text{and} \quad R_2 = \mathrm{Res}(f; -z) = -R_1. \tag{3.34}$$

Using (1.16) and Remark 1.2.2, we have

$$C(h;a) = \frac{2e^{-z^2}}{1 - e^{-2i\pi(a+z/h)}} \quad \text{so that} \quad |C(h;a)| \leq \frac{2e^{-2\pi y/h}}{1 - e^{-2\pi y/h}} e^{y^2 - x^2}. \tag{3.35}$$

Applying the trapezium rule (1.6) to the integral in (3.32) leads to

$$I(h;a) = h \sum_{k \in \mathbb{Z}} \frac{ize^{-(k-a)^2h^2}}{\pi(z^2 - (k-a)^2h^2)}. \tag{3.36}$$

Let

$$I\ (h;a) := I(h;a) + C(h;a); \quad \text{for} \quad a = 0;1/2; \tag{3.37}$$

where $C(h;a)$ and $I(h;a)$ are given by (3.35) and (3.36) we have

$$|w(\zeta) - I(h;a)| \leq \frac{2\sqrt{\pi}\, M_H(F)\, e^{H^2 - 2\pi H/h}}{1 - e^{-2\pi H/h}}; \tag{3.44}$$

where F is given by (3.33) and

$$M_H(F) := \sup_{t \in \mathbb{R}} |F(t + iH)|. \tag{3.45}$$

For $H > 0$ and $z = t + iH$, we have

$$|z^2 - \zeta^2| = |z - \zeta||z + \zeta| \geq |y - H||y + H| = H^2 - y^2;$$

and hence we have, for $H = \pi/h$, that

$$|w(\zeta) - I(h;a)| \leq d_1(y) := \frac{2\sqrt{2}\, y\, e^{-\pi^2/h^2}}{\sqrt{\pi}(\pi^2/h^2 - y^2)\left(1 - e^{-2\pi^2/h^2}\right)}. \tag{3.46}$$

Similarly and using the bound in (3.35) for $C(h;a)$, we have for $y \geq \frac{5}{4}H$, that

$$|w(\zeta) - I(h;a)| \leq d_1(y) + |C(h;a)| \leq d_3(y). \tag{3.47}$$

Select ϵ in the range $(0, H)$ and consider the case that $|y - H| < \epsilon$. We can easily show that

$$w(\zeta) - I(h;a) = \int_{C_H} f(z)(1 - g(z))\, dz; \tag{3.48}$$

where f is given by (3.33), $g(z) = -i\cot(\pi z/h + a\pi)$ and the contour C_H, passing above the pole of f at $z = \zeta$, is the union of C_H and γ, where $C_H = \{t + iH : t \in \mathbb{R} \text{ and } |(t + iH) - \zeta| > \epsilon\}$ and $\gamma = \{\zeta + \epsilon e^{i\theta} : \theta_0 \leq \theta \leq \pi - \theta_0\}$, where $\theta_0 = \sin^{-1}((H - y)/\epsilon) \in (-\pi/2, \pi/2)$.

For $z \in C_H$, it holds that

$$|z^2 - \zeta^2| = |z - \zeta||z + \zeta| \geq \epsilon|y + H|. \tag{3.49}$$

Thus, using (1.8), similarly to (3.46) we deduce that

$$\left|\int_{C_H} f(z)(1 - g(z))\, dz\right| \leq \frac{2\sqrt{2}\, y\, e^{-\pi^2/h^2}}{\sqrt{\pi}\, \epsilon(\pi/h + y)\left(1 - e^{-2\pi^2/h^2}\right)}. \tag{3.50}$$

To bound the integral over γ we note, for $z = X + iY \in \gamma$, that (3.49) is true and $Y \geq H$. Further,

$$|e^{-z^2}| = e^{P};$$

where

$$
\begin{aligned}
P &= Y^2 \quad X^2 \\
&= y^2 \quad x^2 \quad e^2 \cos(2q) + 2e \sqrt{y^2 + x^2} \sin(q \quad \tan^{\,1}(y=
\end{aligned}
$$

where

$$D_h\left(\frac{\pi}{h}\right) = \frac{4\sqrt{2}h\left(1+2\sqrt{\pi}e^{-b\pi^2/h^2}\right)}{\pi^{3/2}\left(1-e^{-2\pi^2/h^2}\right)}; \tag{3.56}$$

and b is given by (3.43).

Proof. It is easy to show, using (3.39) , that $D_h(y)$ and $yD_h(y)$ are increasing functions of y for $0 \le y < \pi/h$, in particular

$$D_h\left(\frac{3\pi}{4h}\right) = \frac{3\sqrt{2}h}{14\pi(1-e^{-2\pi^2/h^2})} < D_h\left(\frac{\pi}{h}\right) = \frac{4\sqrt{2}h\left(1+2\sqrt{\pi}e^{-b\pi^2/h^2}\right)}{\pi^{3/2}\left(1-e^{-2\pi^2/h^2}\right)}. \tag{3.57}$$

Also we have, using (3.53), that

$$\begin{aligned}\frac{|w(z)-I(h;a)|}{|w(z)|} &\le (1+\sqrt{\pi}|z|)|w(z)-I(h;a)| \\ &\le (1+\sqrt{2\pi}y)|w(z)-I(h;a)|;\end{aligned} \tag{3.58}$$

and the two results follow. □

In the following proposition we bound $|w(z)-I(h;a)|$ and $|w(z)-I(h;a)|/|w(z)|$.

Proposition 3.3.3. Suppose that $I(h;a)$ is given by (3.36). Then, for $h>0$ and $z=x+iy$ with $x\ge 0$ and $y \ge \max(x,\pi/h)$, we have

$$|w(z)-I(h;a)| \le \left(D_h\left(\frac{5\pi}{4h}\right)+\frac{2e^{1/4}}{1-e^{-2\pi^2/h^2}}\right) e$$

where

$$M_H(F) := \sup_{t\in\mathbb{R}} |F(t+iH)| \le \frac{\sqrt{2}y}{\pi(y^2-H^2)}. \tag{3.63}$$

Since $\frac{y}{y^2-H^2}$ and $\frac{y^2}{y^2-H^2}$ are both decreasing functions of y on (H,∞), we have

$$\frac{y}{y^2-H^2} \le \frac{H+\epsilon}{\epsilon^2+2\epsilon H} \le \frac{H+\epsilon}{2\epsilon H} \le \frac{5}{8\epsilon} \quad \text{and} \quad \frac{y^2}{y^2-H^2} \le \frac{25}{32\epsilon}H. \tag{3.64}$$

Thus, we have

$$|w(z) - I(h;a)| \le \frac{5\sqrt{2}}{4\sqrt{\pi}}$$

Similarly, using (3.53) and since $yD_h(y)$ and $\frac{2y e^{y^2 - 2\pi y/h}}{1 - e^{-2\pi y/h}}$ are both increasing functions of y for $H \le y < \pi e$, we have that

$$\frac{|w(z) - I(h;a)|}{|w(z)|} \le (1+\sqrt{2\pi}y)\left(|w(z) - I(h;a)| + |C(h;a)|\right)$$

$$\le 1 + \frac{5\sqrt{2}\pi^{3/2}}{}$$

where $D_h\left(\frac{5\pi}{4h}\right)$ is given by (3.61).

Proof. Define

$$E_h(z) = w(z) - I(h;a) \quad \text{and} \quad e_h(z) = E_h(z)/w(z);$$

on $G := \{z \in \mathbb{C} : 0 < \arg(z) < \pi/4\}$. Since $w(z)$ and $I(h;a)$ are both entire functions of z and, using (3.53), $w(z) \neq 0$ for all $z \in G$, $E_h(z)$ and $e_h(z)$ are analytic on G and continuous on its closure. From the asymptotic expansion of $w(z)$ in the complex plane (see [22, (2.6)]) it follows that $w(z) \to 0$ as $|z| \to \infty$, uniformly for $0 < \arg(z) < \pi/4$. Moreover it follows from (3.37) and (3.35) that the same holds for $I(h;a)$ and hence for $E_h(z)$. Thus we have, using Lemma 3.3.1, that

$$\sup_{z \in G} |E_h(z)| = \sup_{z \in \partial G} |E_h(z)|.$$

Let $z = re^{i\pi/4}$ with $r \geq 0$. Then, using Proposition 3.3.1, we have that

Now, for $z \in G$, using (3.53) and (3.71),

$$|e_h(z)| \le (1+\sqrt{\pi}|z|)|E_h(z)| \le P e^{|z|};$$

where $P := M D_h\left(\frac{5\pi}{4h}\right) - e^{-\pi^2/h^2}$ and $M := \max(1+\sqrt{\pi}|z|)e^{-|z|}$, for $z \in G$. Thus we have, using Lemma 3.3.1, that

$$\sup_{z\in G}|e_h(z)| = \sup_{z\in\partial G}|e_h(z)|. \tag{3.75}$$

Let $z = re^{i\pi/4}$ with $r \ge 0$. Then, we have, using Proposition 3.3.1, that $yD_h(y)$ is increasing on $\left[0, \frac{5}{4}\frac{\pi}{h}\right]$ and decreasing on $\left[\frac{5}{4}\frac{\pi}{h}, \infty\right)$ with $D_h\left(\frac{5}{4}\frac{\pi}{h}\right) > D_h\left(\frac{5}{4}\frac{\pi}{h}\right)$; thus we have

$$|e_h(z)| \le \left(1 + \frac{5\sqrt{2}\pi^{3/2}}{4h}\right)\left(D_h\left(\frac{5}{4}\frac{\pi}{h}\right) - e^{-\pi^2/h^2}\right). \tag{3.76}$$

Let $z = x + i\varepsilon$ with $0 < \varepsilon < \pi/h$. Then we have, using (3.53) and Proposition 1.2.4, that

$$\begin{aligned}|e_h(z)| &\le (1+\sqrt{\pi}|z|)|E_h(z)| \\ &\le \frac{2|z|(1+\sqrt{\pi}|z|)\,e^{-\pi^2/h^2}}{\pi(1-e^{-2\pi^2/h^2})}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{|z^2-(t+i\pi/h)^2|}\,dt.\end{aligned}$$

Taking the limit $\varepsilon \to 0^+$, since both sides in the above bound are continuous for $0 < \varepsilon < \pi/h$, we obtain

$$|e_h(x)| \le \frac{2x(1+\sqrt{\pi}x)\,e^{-\pi^2/h^2}}{\pi(1-e^{-2\pi^2/h^2})}\int_{-\infty}^{\infty} G(t)\,dt, \quad x \ge 0, \tag{3.77}$$

where

$$G(t) = \frac{e^{-t^2}}{|x^2-(t+i\pi/h)^2|}.$$

Note

$$\int_{-\infty}^{\infty} G(t)\,dt = \int_{-\infty}^{-x/2} G(t)\,dt + \int_{-x/2}^{x/2} G(t)\,dt + \int_{x/2}^{3x/2} G(t)\,dt + \int_{3x/2}^{\infty} G(t)\,dt. \tag{3.78}$$

Since, for $x \ge 0$ and $t \in \mathbb{R}$,

$$|x^2-(t+i\pi/h)^2| = |x-t-i(\pi/h)||x+t+i(\pi/h)| \ge \frac{\pi}{h}\sqrt{x^2+(\pi/h)^2},$$

we have

$$\int_{x/2}^{3x/2} G(t)\,dt \le \frac{h}{\pi\sqrt{x^2+(\pi/h)^2}}\int_{x/2}^{3x/2} e^{-t^2}\,dt \le \frac{h x e^{-x^2/4}}{\pi\sqrt{x^2+(\pi/h)^2}}, \tag{3.79}$$

$$\int_{3x/2}^{\infty} G(t)\,dt \le \frac{h}{\pi\sqrt{x^2+\pi^2/h^2}}\int_{3x/2}^{\infty} e^{-t^2}\,dt \le \frac{h e^{-9x^2/4}}{3\pi x\sqrt{x^2+\pi^2/h^2}}, \tag{3.80}$$

with f given by (3.33) and t_k and t_k are given by (3.25).

We will call the error in approximating $I(h;a)$ by $I_N(h;a)$ the truncation error, given by

$$T_N(h;a) := 2h \sum_{k=N+1}^{\infty} f((k+a)h). \tag{3.88}$$

Proposition 3.3.5. Suppose t_k is given by (3.25) and $|z \quad t_k| \quad h/4$ for $k = N+1; N+2; \ldots$ and $z = x+iy$ with $0 \quad y < x$. Then, for $h > 0$,

$$|T_N(h;0)| \qquad 2\sqrt{}$$

For the second summation we have that

$$
\begin{aligned}
2h\sum_{k=M}^{\infty}\frac{e^{-t_k^2}}{|z-t_k|} &\le \frac{4}{h}\left(2h\sum_{k=M}^{\infty}e^{-t_k^2}\right)\\
&\le \frac{4}{h}\left(2he^{-t_M^2}+2h\sum_{k=M+1}^{\infty}e^{-t_k^2}\right)\\
&\le \frac{4}{h}\left(2he^{-t_M^2}+2\int_{t_M}^{\infty}e^{-t^2}\,dt\right)\\
&\le \frac{4}{h}\,\frac{1+2ht_M}{t_M}\,e^{-t_M^2}\\
&\le \frac{4}{h}\,\frac{1+2ht_M}{\theta x}\,e^{-t_M^2}.
\end{aligned}
$$

Note that $(1+2ht)e^{-t^2}$ is a decreasing function of t for $t \ge t_0$, where $t_0 := 2h/(1+\sqrt{1+8h^2})$ and $t_0 < h < t_{N+1}$. Thus we have that

$$
2h\sum_{k=M}^{\infty}\frac{e^{-t_k^2}}{|z-t_k|} \le \frac{4}{h}\,\frac{1+2ht_{N+1}}{\theta x}\,e^{-t_{N+1}^2}. \tag{3.93}
$$

We have, using $\frac{1}{\sqrt{x^2+t_{N+1}^2}} \le \frac{1}{t_{N+1}}$ and (3.91), (3.92) and (3.93), that

$$
|T_N(h;0)| \le \left(\frac{\sqrt{2}(1+2ht_{N+1})}{\pi t_{N+1}}\right)\left(\frac{1}{(1-\theta)t_{N+1}}+\frac{4}{h\theta}\right)e^{-t_{N+1}^2}.
$$

Choose θ such that

$$
\frac{1}{(1-\theta)t_{N+1}} = \frac{4}{h\theta},
$$

i.e.

$$
\theta = \frac{4t_{N+1}}{h+4t_{N+1}}.
$$

Then we have that

$$
|T_N(h;0)| \le \frac{2\sqrt{2}(1+2ht_{N+1})(h+4t_{N+1})}{\pi h t_{N+1}^2}\,e^{-t_{N+1}^2}. \tag{3.94}
$$

Similarly, we have, using $\frac{x}{\sqrt{x^2+t_{N+1}^2}} \le 1$ and (3.91), (3.92) and (3.93), that

$$
x|T_N(h;0)| \le \frac{2\sqrt{2}(1+2ht_{N+1})(h+4t_{N+1})}{\pi h t_{N+1}}\,e^{-t_{N+1}^2}. \tag{3.95}
$$

□

In a similar way we can prove the following result for $T(h;1/2)$.

Proposition 3.3.6. Suppose t_k is given by (3.25) and $|z - t_k| \ge h/4$ for $k = N+1; N+2; \dots$ and $z = x + iy$ with $0 \le y < x$. Then, for $h > 0$,

$$|T_N(h;1/2)| \le \frac{2\sqrt{2}(1+2ht_{N+1})(h+4t_{N+1})}{\pi h t_{N+1}^2} e^{-t_{N+1}^2}; \quad \text{and} \tag{3.96}$$

$$\frac{|T_N(h;1/2)|}{|w(z)|} \le (1+\sqrt{2\pi}\, t_{N+1})\, |T_N(h;1/2)|. \tag{3.97}$$

Proof. Suppose that $0 < \theta < 1$, then we have, using (3.88) with $a = 1/2$, that

$$|T_N(h;1/2)| \le \sqrt{\ }$$

Then we have that

$$|T_N(h; 1/2)| \leq \frac{2\sqrt{2}(1 + 2ht_{N+1})(h + 4t_{N+1})}{\sqrt{\pi}ht_{N+1}^2} e^{-t_{N+1}^2}. \qquad (3.101)$$

Similarly, we have, using1

Proposition 3.3.7. Suppose $a = 0$ or $a = 1/2$ and $z = x + iy$ with $y \geq x \geq 0$. Then, for $h > 0$,

$$|T_N(h;a)| \leq \frac{(1+2ht_{N+1})}{\pi t_{N+1}^2} e^{-t_{N+1}^2}; \quad \text{and} \tag{3.108}$$

$$\frac{|T_N(h;a)|}{|w(z)|} \leq \frac{(1+2ht_{N+1})(1+2\sqrt{\pi}\, t_{N+1})}{\pi t_{N+1}^2} e^{-t_{N+1}^2}. \tag{3.109}$$

Proof. Suppose t_k and t_{-k} be given by (3.25) and $F(t)$ is given by (3.33). Then, for $z = x + iy$ with $y \geq x \geq 0$,

$$|z^2 - t_k^2|^2 = y^4 + t_k^4 + x^4 + 2x^2y^2 + 2t_k^2(y^2 - x^2) \geq |-z^2 - t_k^2|^2.$$

Thus, we have

$$|T_N(h;a)| \leq 2h \sum_{k=N+1}^{\infty} e^{-t_k^2} |F(t_k)|;$$

and, using (3.53),

$$\begin{aligned}
\frac{|T_N(h;a)|}{|w(z)|} &\leq (1+\sqrt{\pi}|z|)\left(2h \sum_{k=N+1}^{\infty} e^{-t_k^2}|F(t_k)|\right) \\
&\leq (1+\sqrt{2\pi}\, y)\left(2h \sum_{k=N+1}^{\infty} e^{-t_k^2}|F(t_k)|\right); \quad y \geq 0.
\end{aligned}$$

Since

$$|z^2 - t_k^2|^2 = y^4 + t_k^4 + x^4 + 2x^2y^2 + 2t_k^2(y^2 - x^2) \geq y^4 + t_k^4;$$

$$\begin{aligned}
|T_N(h;a)| &\leq \frac{2\sqrt{2}hy}{\pi} \sum_{k=N+1}^{\infty} \frac{e^{-t_k^2}}{\sqrt{y^4 + t_k^4}} \\
&\leq \frac{\sqrt{2}y}{\pi\sqrt{y^4 + t_{N+1}^4}}\left(2he^{-t_{N+1}^2} + 2\int_{t_{N+1}}^{\infty} e^{-t^2}dt\right) \\
&\leq \frac{\sqrt{2}y(1+2ht_{N+1})}{\pi t_{N+1}\sqrt{y^4 + t_{N+1}^4}} e^{-t_{N+1}^2}.
\end{aligned}$$

Moreover

$$\frac{y}{\sqrt{y^4 + t_{N+1}^4}} \leq \frac{1}{\sqrt{2}\, t_{N+1}} \quad \text{and} \quad \frac{y^2}{\sqrt{y^4 + t_{N+1}^4}} \leq 1.$$

3.59

3.Proof.

Since

$$|w(z) - I_N(h;a)| \le |w(z) - I(h;a)| + |T_N(h;a)|;$$

and

$$\frac{|w(z) - I_N(h;a)|}{|w(z)|} \le \frac{|w(z) - I(h;a)|}{|w(z)|} + \frac{|T_N(h;a)|}{|w(z)|};$$

the first result follows by combining (3.125) and (3.127) and the second result follows by combining (3.126) and (3.128). □

Remark 3.3.3. Using Proposition 3.3.3 and Theorem 3.3.3, we can easily show, for $z = x + iy$ with $0 < x \ \ y < \pi/h$, that

$$|w(z) - I_N(h;a)| \le b_N e^{-\sqrt{N}} \tag{3.129}$$

and

$$\frac{|w(z) - I_N(h;a)|}{|w(z)|} \le \hat{b}_N \sqrt{N+1}\, e^{-\sqrt{N}}; \tag{3.130}$$

where $\tilde{b}_N$ and $\hat{b}_N$ are given by (3.122) and (3.123), respectively.

3.4 Numerical results

In this section we show numerical calculations that illustrate and confirm the theoretical results (Theorems 3.3.2)

(this section) will 0 g 062ombining (p100[(j)91 and the second r48TJ/ [([(j)91)25(w)- ining (jw Th02 11.9552 Tf 5.977 0 Td1[(ET[(j)91)25(w)-wf 10.005 3.391 Td []T

(iii) the approximation w_N, with $N \geq 14$, is significantly more accurate than the approximation (3.8) from Weideman [62];

Figure 3.2 below shows that $w_N(z)$ is very accurate as $|z| \to \infty$, and with N as small as 9 the computed relative error is 10^{-12}, which confirms the calculations in Figure 3.1.

We will comment now on the accuracy and the efficiency of computing $w(z)$ using the approximation $w_N(z)$ given by (3.21) and its code `w(z,N)` in Listing A.3 in comparison with the approximations (3.8), (3.11) and (3.16) and their codes. We do not have access to exact values for $w(z)$ and so we use four different accurate approximations to $w(z)$:

(i) Our own approximation $w_N(z)$ with $N = 20$ computed by the call `W(z,20)` to the code in Listing A.3;

(ii) Weideman's approximation (3.8) with $N = 40$ (this choice of N gives maximum accuracy for this approximation), implemented by the call `cef(z,40)` in Table 1 [62];

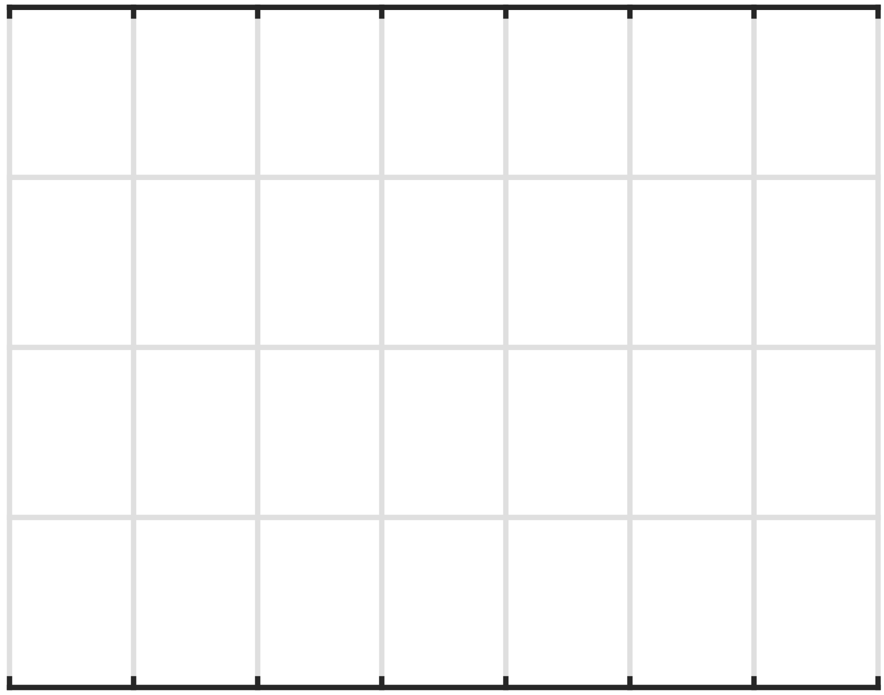
(iii) The approximation (3.11) of Zagloul and Ali [63], implemented in the Matlab code [64], supplied to us by the author, computed by the call `Faddeyeva_v2(z,M)` with $M = 13$ (the maximum value permitted by the code), where M is the number of accurate significant figures required, which must be in the range $4 \leq M \leq 13$;

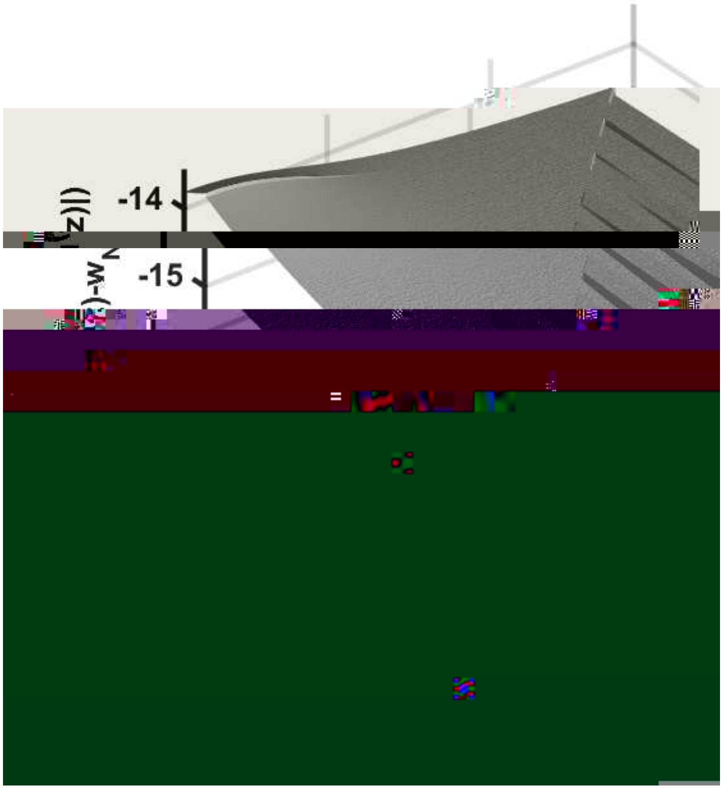
(iv) The approximation (3.16) of Abrarov and Quine [1], implemented as the the Matlab function `comperf(z)` of Abrarov and Quine [1, Appendix], which uses the method (3.16) with $a = 2.75$ and $M = 5$.

The maximum absolute errors and computation times are shown in Table 3.1 (using Matlab (R2015a) on a laptop with Intel core i7-4510U 2.00 GHz processor) for `w(z,N)` in Table A.3, `cef(z,40)` in Table 1 of Weideman [62], `comperf(z)` of Abrarov and Quine [1, Appendix] and the method of M. Zaghloul and A. Ali [63] as implemented in `Faddeyeva_v2(z,13)` of [64]. The calculations are implemented for $z = 10^p e^{iq}$, with $p = -6(0.0006)6$ and $q = 0(\pi/400)\pi/2$ giving in total 4020201 values. It can be seen from Table 3.1 that the approximation w_N given by (3.21), with N as small as 11, is as accurate as most accurate version of the approximation (3.11) in Zagloul and Ali [63] as implemented in [64] with $a = 1/2$ and $M = 13$, and significantly more accurate than the Matlab code of Abrarov and Quine [1] based on (3.16) with $a = 2.75$ and $M = 5$, and at least as accurate as Weideman's approximation (3.8) with $N = 40$

Algorithm	Maximum absolute error	Computation time in seconds
w(z,11)	$1{:}11 \quad 10^{15}$	0.64
cef(z,40)	$1{:}30 \quad 10^{15}$	1.46
comperf(z)	$5{:}53 \quad 10^{10}$	0.90
Faddeyeva_v2(z,13)	$3{:}92 \quad 10^{15}$	0.51

Table 3.1 Accuracy and computation times of the Matlab codes of the approximations (3.21), (3.8), (3.11) and (3.16).


N



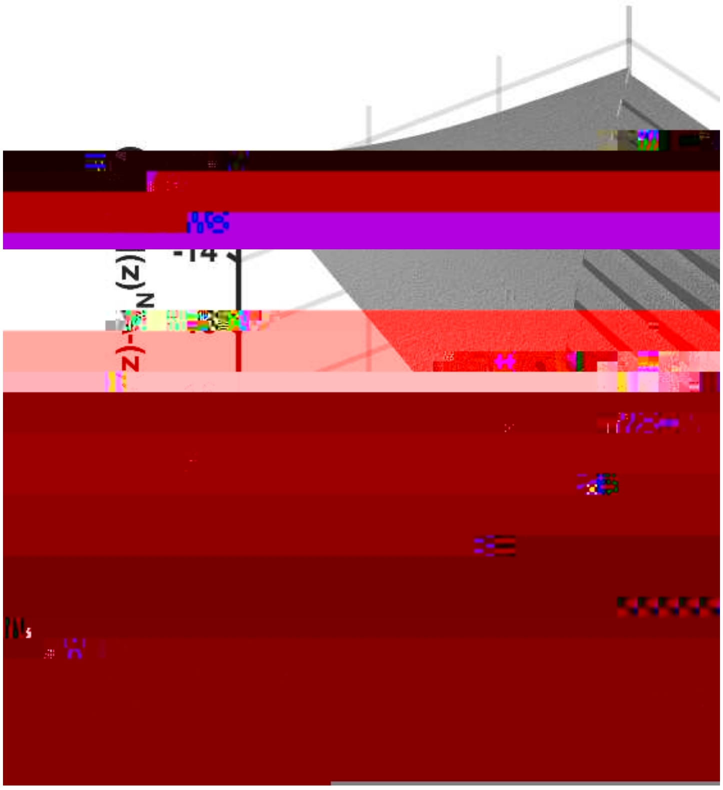


Fig. 3.2 The surfaces of the absolute (top) and relative (bottom) errors of the approximation $w_N(z)$ given by (3.21) with $N = 9$, where the exact value of $w(z)$ is computed by $w_{20}(z)$.

Chapter 4

The 2D impedance half-space Green's function for the Helmholtz equation

4.1 Introduction

This chapter is concerned with the problem of calculating sound propagation from a mono-frequency coherent line source above an impedance plane. The interest in this problem has been motivated by the development of boundary element methods (BEMs) for the calculation of outdoor sound propagation for many applications (e.g. [26], [11], [12] and [13]). These

$d' = |\mathbf{r} - \mathbf{r}_0'|$ be the distance from the image source to the receiver and $\rho' = kd'$, where k is the wave number that satisfies $k = 2\pi/\lambda$ where λ is the wavelength.

The problem is to calculate the acoustic pressure at $\mathbf{r}$, denoted by $G_\beta(\mathbf{r};\mathbf{r_0})$, due to the source at $\mathbf{r_0}$, where β is the normalised admittance of the impedance plane with $\mathrm{Re}(\beta) > 0$. $G_\beta(\mathbf{r};\mathbf{r_0})$ (the Green's function) satisfies the following conditions:

(i) the **Helmholtz equation**, that is

$$\nabla^2 G_\beta(\mathbf{r};\mathbf{r_0}) + k^2 G_\beta(\mathbf{r};\mathbf{r_0}) =$$

deform the path L to the steepest descent path (see [35], [8] and [14]), and obtain [14]

$$P_\beta(\mathbf{r};\mathbf{r_0}) = P_\beta^{(\Gamma)} + P_\beta^{(\sigma)}; \tag{4.11}$$

where

$$P_\beta^{(\Gamma)} = \frac{\beta e^{i\rho}}{\pi} \int_{-\infty}^{\infty} e^{-\rho t^2} F(t)\,dt; \tag{4.12}$$

with

$$F(t) := \frac{\beta + \gamma(1+it^2)}{\sqrt{t^2-2i}\,(t^2-\zeta_1^2)(t^2-\zeta_2^2)}; \qquad -\frac{\pi}{2} < \arg\sqrt{t^2-2i} < \frac{\pi}{2}; \tag{4.13}$$

$$\zeta_1 := \sqrt{i\alpha_+}; \qquad \frac{\pi}{4} < \arg\sqrt{i\alpha_+} < \frac{3\pi}{4}; \tag{4.14}$$

$$\zeta_2 := \sqrt{i\alpha_-}; \qquad 0 < \arg\sqrt{i\alpha_-} < \frac{\pi}{2}; \tag{4.15}$$

$$\alpha_\pm := 1+\beta\gamma \pm \sqrt{1-\beta^2}\sqrt{1-\gamma^2}; \qquad \operatorname{Re}\sqrt{1-\beta^2} \geq 0; \tag{4.16}$$

and

$$P_\beta^{(\sigma)} := \frac{\beta e^{i\rho}}{\pi}\,\frac{\pi e^{-i\rho\alpha_+}}{2\sqrt{1-\beta^2}}\,\delta_\sigma; \tag{4.17}$$

where

$$\delta_\sigma := \begin{cases} 2; & \operatorname{Im}\beta < 0; \operatorname{Re}\alpha_+ < 0 \\ 1; & \operatorname{Im}\beta < 0; \operatorname{Re}\alpha_+ = 0; \\ 0; & \text{otherwise} \end{cases} \tag{4.18}$$

The integral representation (4.12) from [14] will be the starting point for our proposed approximation of P_β.

Numerical computation of the solution of the problem (

has been widely cited and applied (e.g. [41], [25], [7], [47] and [39]) as a well-established method for solving this problem. In particular, it is used in many papers as an efficient method for the solution of outdoor sound propagation problems via the BEM (e.g. [32], [34], [49], [51]). The following representations for $P_b(r;r_0)$ is derived and used in [14]:

$$P_b(\mathbf{r};\mathbf{r_0}) = \frac{b\, e^{ir}}{p} \int_0^\infty t^{\,1/2} e^{\,rt} f(t)\, dt; \quad \mathrm{Im}(b) > 0 \quad \text{or} \quad \mathrm{Re}(a_+) > 0; \tag{4.19}$$

and

$$P_b(\mathbf{r};\mathbf{r_0}) = \frac{b\, e^{ir}}{p} \int^\infty$$

ct >TJ/F95 8.96642 Tf 15525308 4.938 Td []TJ/F102 11.955b

and $\mathbf{H}$ is the Heaviside step function defined by

$$\mathbf{H}(t) := \begin{cases} 1; & t > 0; \\ 1/2; & t = 0; \\ 0; & t < 0; \end{cases} \tag{4.32}$$

and

$$d_+^{(1)} := \begin{cases} 2e^{-2i\pi z_1/h}; & y_1 < 0\,; \\ 1 + e^{-2i\pi z_1/h}; & y_1 = 0\,; \\ 2; & y_1 > 0. \end{cases} \tag{4.33}$$

La Porte [38] proved a bound on $|P_b - P_b^{h,N,H}|$ derived largely from Proposition 1.2.4 and using, for F given by (4.13), that

$$M_H(F) := \sup_{x \in \mathbb{R}, |y| = H} |F(x+iy)| \le \frac{|b|+g}{\sqrt{1-H^2}} \hat{M}; \tag{4.34}$$

where

$$\hat{M} := \max\left\{3, \frac{2\max(x_1^2, x_2^2) + 2}{|H^2 - y_1^2||H^2 - y_2^2|}\right\}; \tag{4.35}$$

and $x_j = \mathrm{Re}(z_j)$ and $y_j = \mathrm{Im}(z_j)$ for $j = 1, 2$.

La Porte [38] showed, using numerical calculations, that the approximation in (4.28) achieves with $N = 11$ higher accuracy than the approximation (4.25), with $n = 40$ and $m = 22$, in Chandler-Wilde and Hothersall [14] for $0.5 \le r \le 8.54$, $0 \le g \le 1$ and $0.1 \le |b| \le 1$.

This chapter of the thesis builds on the work of La Porte [38] but extends this work significantly. The main issues with the approximation $P_b^{h,N,H}$ in (4.28) are that: (i) the approximation formula blows up if the simple pole at $z_1 = \sqrt{ia_+}$ coincides with a quadrature point at kh and is inaccurate in floating point arithmetic when z_1 is close to kh; (ii) the expression (4.31) blows up when $a = 2$ and is inaccurate in floating point arithmetic when a is close to 2; and (iii) the bound (4.34) blows up when $H = \mathrm{Im}(z_1)$ or $H = -\mathrm{Im}(z_1)$. In this chapter of the thesis we address all these issues: we propose an approximation which is stable for numerical calculations for $r > 0$, $0 \le g \le 1$ and b with $\mathrm{Re}(b) > 0$; we prove a rigorous and uniform error bound for this approximation; and finally we show through systematic numerical experiments that this approximation is at least as accurate as the approximation (4.28) in La Porte [38] and is more accurate and more efficient than the approximation of Chandler-Wilde and Hothersall [14].

Recently, O’Neil et al. [47] propose a method of computing $P_b(r;r_0)$, for 0 b 1, based on the following representation for $P_b(\mathbf{r};\mathbf{r_0})$:

$$P_b(\mathbf{r};\mathbf{r_0}) = I_1 + I_2;$$

where

$$I_1 := \frac{ikb}{2p} Z_0^1 H_0^{(1)}(kj\mathbf{r} \quad \mathbf{r_0}j)e^{ikbh}\, dh; \quad \mathbf{r_0} = (x_0; \quad (y_0 + h));$$

and

$$I_2 := \frac{ikb}{2p} Z_{¥}^{¥} \frac{e^{p\overline{l^2 \quad k^2}(y+y_0)}\, e^{p\overline{l^2 \quad k^2} \quad ikb}}{p\overline{l^2 \quad k^2}\; p\overline{l^2 \quad k^2} \quad ikb}\, e^{il(}$$

experiments to demonstrate the accuracy of the proposed approximation $P_{b,N}$ in comparison with the approximations of Chandler-Wilde and Hothersall [14] and La Porte [38].

Let $P_b(r, r_0)$ be given by equations (4.11)–(4.18) and $H := \min(0.9, \tilde{A}_N)$ with $\tilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}r)}$, and recall that

with

$$C_N := (|b|+1)\left[\frac{384\sqrt{10}(4|b|+7)(1+4\sqrt{\pi r})\,r^{3/2}}{\pi^{3/2}(N+1)^2\left(1-e^{-2\pi(N+1)/\sqrt{3}}\right)}+20\left(1+\frac{1}{\tilde{A}_N}\right)\right]; \tag{4.46}$$

$$\tilde{C}_N := (|b|+1)\left[\frac{781\sqrt{10\pi}(4|b|+7)(1+4\sqrt{\pi r})}{\sqrt{r}\left(1-e^{-0.9\pi/h_N}\right)(N+1)^{1/3}}+\tilde{K}_N\right] \quad \text{and}$$

$$\tilde{K}_N := 8K_N\left(1+\frac{r^{1/3}}{a\pi^{1/3}H^{1/3}(N+1)^{1/3}}\right); \tag{4.47}$$

where K_N is given by (4.120)

Remark 4.3.1. The advantage of choosing the branch cut for $\sqrt{i(\alpha_+ - 2)}$ as in (4.51) is that a cut from 2 to $+\infty$ on the positive real axis in the α_+-plane is implied. This is convenient, since $\alpha_+ \geq 2$ is impossible unless $\beta \geq 1$. Thus, $\sqrt{i(\alpha_+ - 2)}$, considered as a function of β, is analytic in the cut half-plane.

The formulas (4.51) and (4.52) for R_1 and R_2 are not numerically stable in floating point arithmetic when β and γ are close to zero, close to 1, or when $\beta = \gamma$. We simplify them in the following lemma to make them more stable in numerical calculations.

Lemma 4.3.1. For $0 \leq \gamma \leq 1$ and β in the cut half-plane, we have that

$$R_1 = \frac{i e^{i\rho\alpha_+}}{4\sqrt{1-\beta^2}}, \tag{4.54}$$

$$R_2 = \frac{i e^{i\rho\alpha_-}}{4\sqrt{1-\beta^2}}\, W, \tag{4.55}$$

where

$$W := \begin{cases} +1, & \text{if } \beta \end{cases}$$

where W is given by (4.56),

$$d_+ := \begin{cases} 2e^{-2i\pi(a+z_1/h)}, & y_1 < 0\,; \\ 1 + e^{-2i\pi(a+z_1/h)}, & y_1 = 0\,; \\ 2, & y_1 > 0; \end{cases} \tag{4.73}$$

and $\mathbf{H}$ is the Heaviside step function given by (4.32).

4.3.1 Bounding the discretisation error

This section is concerned with bounding, for $h > 0$ and $a = 0$ or $a = 1/2$,

$$E(h;a) := I - I(h;a);$$

where I and $I(h;a)$ are given by (4.50) and (4.70), respectively.

Since F given by (4.13) is meromorphic for $|\mathrm{Im}(t)| < 1$, we will be defining H throughout this chapter as

$$H := \min\left\{0.9, \frac{\pi}{\rho h}\right\}. \tag{4.74}$$

Then, we have the following result.

Proposition 4.3.1. Let $h > 0$ and $H := \min\left\{0.9, \frac{\pi}{\rho h}\right\}$. Then

$$|E(h;a)| \leq \frac{D(H)\, e^{\rho H^2/4 - \pi H/h}}{1 - e^{-\pi H/h}}; \tag{4.75}$$

where

$$D(H) := \frac{512\sqrt{10\pi}(|\beta|+1)(4|\beta|+7)(1+4\sqrt{\pi\rho})}{\sqrt{\rho}\, H^4} + \frac{2\pi}{|1-\beta^2|^{1/2}}. \tag{4.76}$$

Proof. Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ be given by (4.14) and (4.15), respectively. Select $\epsilon \in (0, H/4)$ and consider the case $|H - |y_1|| \geq \epsilon$ and $|H - |y_2|| \geq \epsilon$. Then, using Proposition 1.2.4, we have that

$$|E(h;a)| \leq \frac{2\sqrt{\pi}\, M_H(F)}{\sqrt{\rho}(1 - e^{-2\pi H/h})} e^{\rho H^2 - 2\pi H/h}, \tag{4.77}$$

and using equation (4.34) and noting $x_j^2 \le |z_j|^2 \le 2+2|b|$ with $j=1,2$, it holds that

$$
\begin{aligned}
M_H(F) &\le \frac{\sqrt{10}(|b|+1)}{\sqrt{1+H}} \max\left(3, \frac{2\max(x_1^2,x_2^2)+3}{\varepsilon^2(|y_1|+H)(y_2+H)}\right)\\
&\le \sqrt{10}(|b|+1)\max\left(3, \frac{7+4|b|}{\varepsilon^2(|y_1|+H-4\varepsilon)(y_2+H-4\varepsilon)}\right)\\
&\le \sqrt{10}(|b|+1)\max\left(3, \frac{7+4|b|}{\varepsilon^2(H-4\varepsilon)^2}\right). \qquad (4.78)
\end{aligned}
$$

We consider now the case $|H-|y_1||<\varepsilon$ or $|H-y_2|<\varepsilon$. Let D be the region in the complex plane defined by

$$
D := \{z : 0 < \mathrm{Im}(z) < H\} \setminus \bigcup_{j=1,2} B_\varepsilon(z_j), \qquad (4.79)
$$

where, for $j=1,2$,

$$
B_\varepsilon(z_j) := \begin{cases} \{z : |z-z_j| < \varepsilon\}, & \text{if } |\mathrm{Im}(z_j)-H| < 2\varepsilon;\\ \emptyset, & \text{otherwise} \end{cases} \qquad (4.80)
$$

and let, for $j=1,2$,

$$
\gamma_j = \{z \in \partial D : |z-z_j| = \varepsilon\} \quad \text{and} \quad \Gamma_H = \{z \in \partial D : z = t+iH,\ t \in \mathbb{R}\},
$$

where ∂D is the boundary of D. Then we can show, recalling that γ, F and $C(h;\alpha)$ are given by (1.7), (4.13) and (4.72), respectively, that

$$
|E(h;\alpha)| \le \int_{\Gamma_H} e^{-\rho z^2}F(z)(1-\gamma(z))\,dz + \sum_{j=1}^{2}\int_{\gamma_j} e^{-\rho z^2}F(z)(1-\gamma(z))\,dz + |C(h;\alpha)|. \qquad (4.81)
$$

If $H-\varepsilon < |y_1| \le H$ or $H-\varepsilon < y_2 \le H$, then, using (1.8), (1.9) and (4.72), it holds that

$$
\begin{aligned}
|C(h;\alpha)| &\le \frac{\pi}{2|1-\beta^2|^{1/2}}\left(\frac{2e^{-2\pi|y_1|/h}}{1-e^{-2\pi|y_1|/h}} + \frac{2e^{-2\pi y_2/h}}{1-e^{-2\pi y_2/h}}\right)\\
&\le \frac{\pi}{2|1-\beta^2|^{1/2}}\left(\frac{4e^{-2\pi(H-4\varepsilon)/h}}{1-e^{-2\pi(H-4\varepsilon)/h}}\right)
\end{aligned}
$$

and

$$\frac{2\sqrt{\pi}\,M_H(F)}{\sqrt{r}\,(1-e^{-2\pi H/h})}\,e^{-rH^2-2\pi H/h} + \frac{512\sqrt{10\pi}\,(|b|+1)(7|b|+4)}{\sqrt{r}\,H^4(1-e^{-\pi H/h})}\,e^{-rH^2/4-\pi H/h}. \quad (4.94)$$

Thus, the result follows by combining, with $\varepsilon = H/8$, (4.82), (4.92) and (4.94). □

4.3.2 Bounding the truncation error

This section will give bounds on the truncation error $T_N(h;a)$ as defined in (1.24) for $a = 0$ or $a = 1/2$. We will present two results on the truncation errors $T_N(h;0)$ and $T_N(h;1/2)$ and then we propose a scheme for choosing the step-size h. This scheme will be used to simplify further the bounds on $T_N(h;0)$ and $T_N(h;1/2)$.

Recall that $z_1 = x$ +7 Td [(=)]TJ/F69 11.J/F66633.886 Td [(1)]TJ/F102 19 055211.J/F66633+7 0 T

Combining the above inequalities, we find that

$$|F(t_k)| \le \frac{8x_2(|b|+1)\sqrt{1+t_k^4}}{h\sqrt[4]{1+t_k^4}(t_k+|x_1|)(t_k+x_2)} \tag{4.101}$$

$$\le \frac{8(|b|+1)\sqrt[4]{1+t_k^4}}{h(t_k+|x_1|)} \tag{4.102}$$

$$\le \frac{8(|b|+1)(1+t_k)}{h(t_k+|x_1|)}, \tag{4.103}$$

where the last line comes from

$$\frac{1+t}{\sqrt{2}} \le (1+t^4)^{1/4} \le \left((1+t^2)^2\right)^{1/4} \le (1+t).$$

Also, note that

$$\frac{d}{dt}\left(\frac{1+t}{|x_1|+t}\right) = \frac{|x_1|-1}{(t+|x_1|)^2},$$

thus we have that

$$|F(t_k)| \le \frac{8}{h}(|b|+1)\begin{cases} \dfrac{1+t_{N+1}}{|x_1|+t_{N+1}}, & \text{if } |x_1| \le 1\,; \\ 1, & \text{otherwise}\,; \end{cases} \tag{4.104}$$

but

$$\frac{1+t_{N+1}}{|x_1|+t_{N+1}} \le \left(1+\frac{1}{t_{N+1}}\right),$$

and hence the result follows. □

Proposition 4.3.2. Let $h>0$, $N\in\mathbb{N}$, $F(t)$ be given by (4.13) and $t_k = kh$ with $|t_k - z_1| \ge h/4$ for $k = N+1, N+2, \ldots$. Then, for

$$T_N(h;0) := 2h\sum_{k=N+1}^{\infty} e^{-rt_k^2}F(t_k),$$

we have

$$|T_N(h;0)| \le \frac{8(|b|+1)(1+2hrt_{N+1})}{hrt_{N+1}}\left(1+\frac{1}{t_{N+1}}\right)e^{-rt_{N+1}^2}. \tag{4.105}$$

Proof. Using Lemma 4.3.2 we find that

$$
\begin{aligned}
|T_N(h;0)| &\le \frac{8M_N(|b|+1)}{h}\left(2h\sum_{k=N+1}^{\infty} e^{-rt_k^2}\right)\\
&= \frac{8M_N(|b|+1)}{h}\left(2he^{-rt_{N+1}^2}+2h\sum_{k=N+2}^{\infty} e^{-rt_k^2}\right)\\
&\le \frac{8M_N(|b|+1)}{h}\left(2he^{-rt_{N+1}^2}+2\int_{t_{N+1}}^{\infty} e^{-rt^2}\,dt\right)\\
&\le \frac{8M_N(|b|+1)}{h}\left(2he^{-rt_{N+1}^2}+\frac{e^{-rt_{N+1}^2}}{rt_{N+1}}\right)\\
&= \frac{8M_N(|b|+1)(1+2hrt_{N+1})}{hrt_{N+1}}\, e^{-rt_{N+1}^2}.
\end{aligned}
$$

To arrive at the last line we have used that, for $x > 0$ and $r > 0$,

$$2\int_x^{\infty} e^{-rt^2}\,dt = 2\left(\frac{e^{-rx^2}}{2rx}-\int_x^{\infty}\frac{e^{-rt^2}}{2rt^2}\,dt\right) < \frac{e^{-rx^2}}{rx}. \tag{4.106}$$

□

Remark 4.3.2. We can show in a similar way, for $t_k = (k+1/2)h$ with $|t_k - z_1| \le h/4$ and $k = N+1, N+2, \ldots$, that

$$|F(t_k)| \le \frac{8}{h}(|b|+1)\left(1+\frac{1}{t_{N+1}}\right). \tag{4.107}$$

Also, since $\tilde{t}_{N+1} = t_{N+1} + h/2$, it holds that

$$\left(1+\frac{1}{\tilde{t}_{N+1}}\right) \le \left(1+\frac{1}{t_{N+1}}\right);$$

and hence we have that

$$|T_N(h;1/2)| \le \frac{8(|b|+1)(1+2hrt_{N+1})}{hrt_{N+1}}\left(1+\frac{1}{t_{N+1}}\right)e^{-rt_{N+1}^2}. \tag{4.108}$$

4.3.3 Choices of the step-size h

This section is concerned with proposing explicit recommendations on how to choose the step-size h, following the recommendations in La Porte [38].

For $r > 0$, $H := \min(0.9, \pi/(rh))$ and $t_{N+1} = (N+1)h$ with $N \in \mathbb{N}$, we define two possible choices, h_N and $\tilde{h}_N$, for the step-size h. For both we choose the step-size to satisfy the right hand equations in (4.109) and (4.111) below, i.e. to equalise the exponents in our

Lemma 4.3.3. If $b > 0$ and a given by (4.114), then

$$\frac{1}{1+3b} \leq a \leq \frac{1}{1+b}. \tag{4.115}$$

Given $r > 0$ and $N \in \mathbb{N}$ we choose $h > 0$ as follows.

Remark 4.3.3. Let $H := \min(0.9; \tilde{A}_N)$ with $\tilde{A}_N := \sqrt{2p(N+1)/(\sqrt{3}r)}$, and set

$$h := \begin{cases} h_N := \sqrt{\dfrac{\sqrt{3}p}{2r(N+1)}}, & \tilde{A}_N \leq 0.9 \\[2ex] h_N = a\left(\dfrac{pH}{r(N+1)^2}\right)^{1/3}; & \text{otherwise}, \end{cases} \tag{4.116}$$

where $a \in [1/(1+3b); 1/(1+b)]$ is given by

$$a = \sqrt[3]{\frac{1}{2} + \sqrt{\frac{1}{4} + b^3}} + \sqrt[3]{\frac{1}{2} - \sqrt{\frac{1}{4} + b^3}}, \tag{4.117}$$

and

$$b = \frac{r^{2/3}H^{4/3}}{12p^{2/3}(N+1)^{2/3}}. \tag{4.118}$$

The following result bounds the expression

$$\frac{1+2hr\,t_{N+1}}{hr\,t_{N+1}}$$

for the choice of h given in Remark 4.3.3 which will be used to simplify further the bound (4.105) in Proposition 4.3.2.

Lemma 4.3.4. Let $r > 0$, $N \in \mathbb{N}$ and $\tilde{A}_N := \sqrt{2p(N+1)/(\sqrt{3}r)}$, and h be given as in Remark 4.3.3. Then, for $t_{N+1} = (N+1)h$,

$$\frac{1+2hr\,t_{N+1}}{r\,ht_{N+1}} \leq \begin{cases} \dfrac{5}{2}; & \text{if } \tilde{A}_N \leq 0.9; \\[2ex] (N+1)^{1/3} K_N; & \text{otherwise}, \end{cases} \tag{4.119}$$

where

$$K_N := \frac{2}{(N+1)^{1/3}} + \frac{2}{r^{1/3} p^{2/3} H^{2/3}} + \frac{r H^2}{8p^2 (N+1)^{4/3}}. \tag{4.120}$$

Proof. For $h = h_N = \sqrt{\frac{\sqrt{3}\, p}{2r(N+1)}}$, we have

$$\frac{1 + 2r\, h t_{N+1}}{r\, h t_{N+1}} = 2 + \frac{1}{r(N+1)(h_N)^2} = 2 + \frac{2}{\sqrt{3}\, p} \quad 2.5.$$

For $h = h_N = a \left(\frac{pH}{r(N+1)^2 t_{N+1}} \right)^{2/3} = (2 \quad N+1)^{1/p}$

$r^{1/} (pH)^{2/3} \; 1(N+1)^{1/p}$

and hence the first bound follows.

Now we consider the case $H = 0.9$ and $h = h_N = a\left(\frac{\sqrt{H}}{r\,(N+1)^2}\right)^{1/3}$. Using (

$P_{b;N}$ given by (4.36) in comparison with the approximations (4.25) and (4.28). Systematic numerical calculations are implemented for $q_0 = 0^o(10^o)90^o$, $|b| = 0.1(0.1)0.999$ and $\arg(b) = \; 89^o(8.9^o)89^o$, and the Faddeeva function in $P^{(2)}_{n;m}$ given by (4.24) is computed by Wiedeman's approximation (3.8), implemented by the call `cef(z,40)` in Table 1 [62].

For convenience, we denote in this section the approximation (4.28) in La Porte [38] by $P^{(1)}_N$ and our approximation $P_{b;N}$ given by (4.36) by $P^{(2)}_N$. We do not have access to exact values for P_b and so using different accurate approximations to P_b:

(i) Our approximation $P_{b;N}$ given by (4.36) with $N = 100$, computed by the Matlab code in Listing A.4;

(ii) Chandler-Wilde and Hothersall's approximation $P_{100;100}$ given by (4.25) computed by a Matlab

(ii) With $N = 11$, our approximation P_b

$r = kd^{0}$

$r = kd^0$	d^0	E_{approx}	$E_9^{(4)}$	$E_{11}^{(4)}$	$E_{21}^{(4)}$
0:5	0:0796	5:8 10^{4}	1:7 10^{5}	3:0 10^{6}	9:8 10^{9}
0:75	0:119	8:1 10^{5}	2:7 10^{6}	7:1 10^{7}	2:5 10^{9}
1:125	0:179	7:1 10^{6}	1:3 10^{6}	2:8 10^{7}	4:9 10^{10}
1:688	0:269	3:5 10^{7}	5:3 10^{7}	9:4 10^{8}	7:6 10^{11}
2:531	0:403	8:3 10^{9}	1:8 10^{7}	2:7 10^{8}	8:6 10^{12}
3:793	0:604	8:4 10^{11}	5:2 10^{8}	6:1 10^{9}	7:1 10^{13}
5:70	0:906	7:0 10^{13}	1:3 10^{8}	1:1 10^{9}	4:0 10^{14}
8:54	1:36	4:0 10^{13}	2:5 10^{9}	1:7 10^{10}	6:7 10^{15}
12:814	2:039	4:0 10^{13}	5:2 10^{10}	2:2 10^{11}	1:1 10^{14}
19:222	3:059	3:9 10^{13}	9:7 10^{11}	2:7 10^{12}	3:2 10^{15}
28:833	4:589	3:9 10^{13}	1:9 10^{11}	3:1 10^{13}	5:0 10^{15}
43:249	6:883	3:9 10^{13}	4:3 10^{12}	3:7 10^{14}	3:7 10^{15}
64:873	10:325	3:9 10^{13}	1:7 10^{11}	4:1 10^{14}	6:0 10^{15}
97:31	15:487	3:9 10^{13}	1:5 10^{11}	6:8 10^{14}	7:8 10^{15}
145:96	23:230	3:9 10^{13}	7:4 10^{12}	5:8 10^{14}	6:1 10^{15}
218:95	34:847	3:9 10^{13}	1:0 10^{11}	5:0 10^{14}	6:2 10^{15}
328:42	51:633	3:9 10^{13}	4:1 10^{12}	2:2 10^{14}	1:3 10^{14}
492:63	78:404	3:9 10^{13}	3:0 10^{12}	1:8 10^{14}	2:9 10^{15}
738:95	117:608	3:9 10^{13}	2:9 10^{12}	1:2 10^{14}	7:6 10^{15}
1108:4	176:407	3:9 10^{13}	2:0 10^{12}	9:9 10^{15}	4:9 10^{15}

Table 4.3 Maximum values of E_{approx} and $E_N^{(4)}$ given by (4.129) and (4.133), respectively, with $N = 9;11;21$, for $q_0 = 0^o(10^o)90^o$, $jbj = 0{:}1(0{:}1)0{:}999$ and $arg(b) = 89^o(8{:}9^o)89^o$.

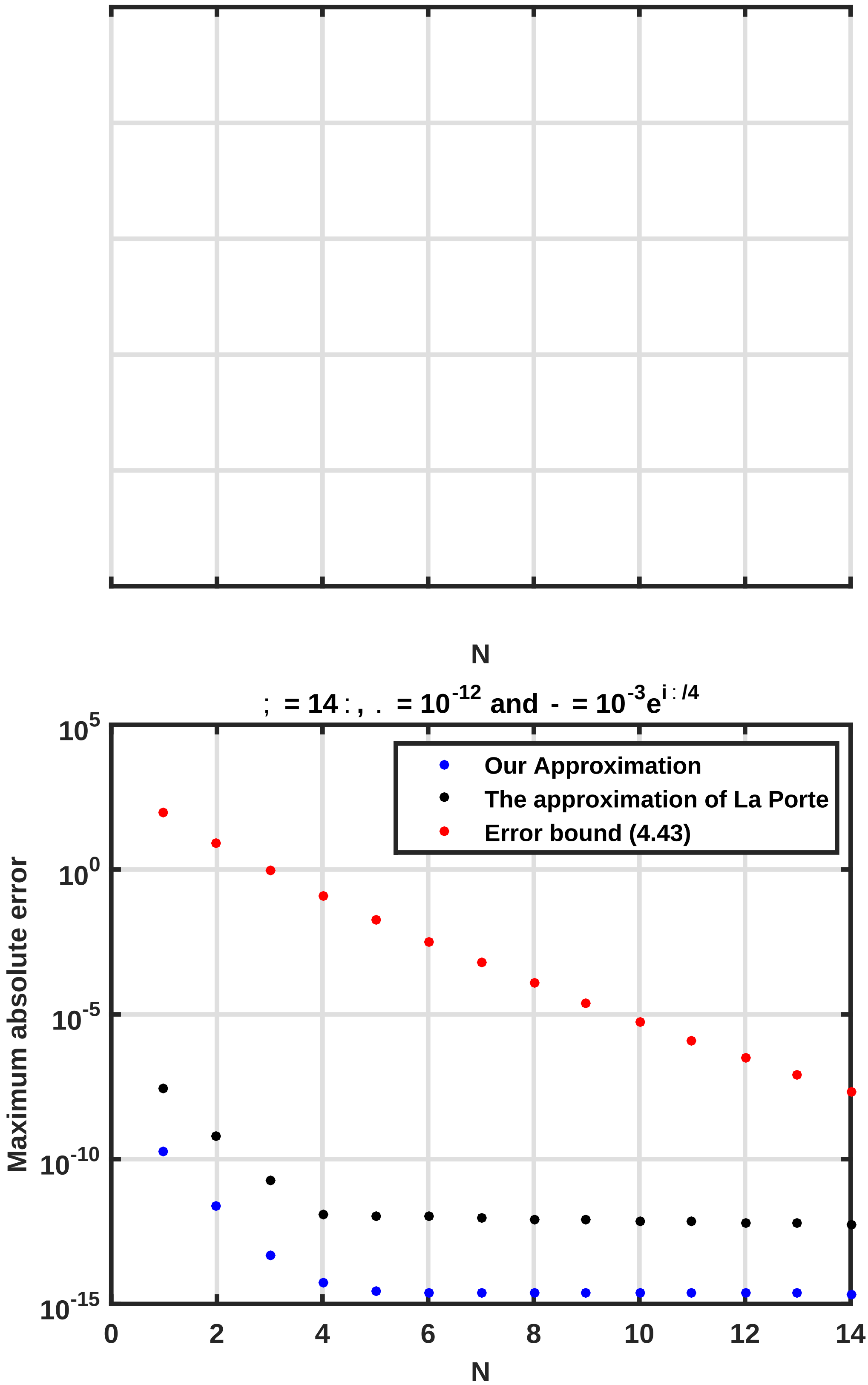


Fig. 4.2 Accuracy of our approximation (4.36) and its upper bound (4.43), as a function of N, in comparison with La Porte's approximation (4.28).

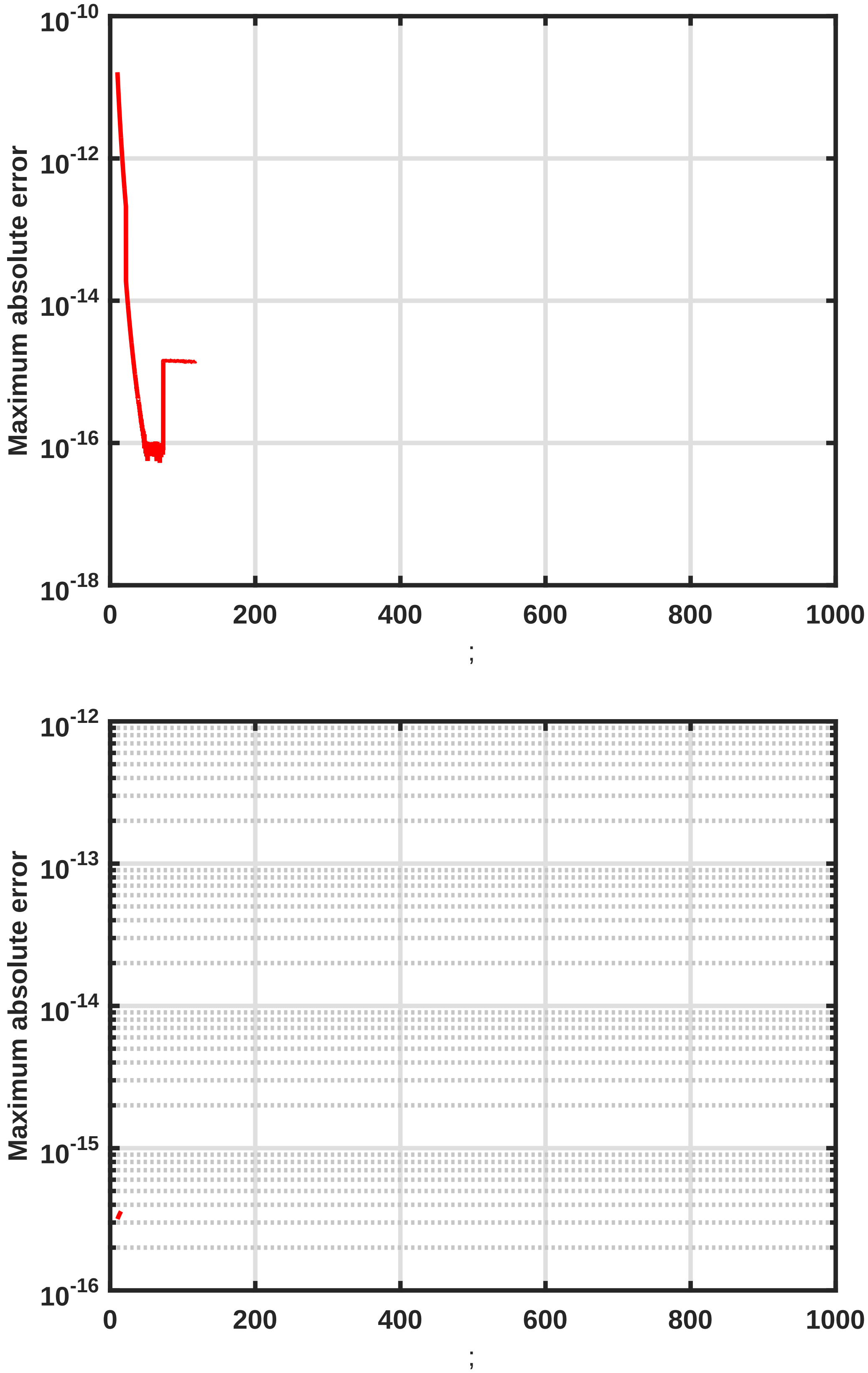


Fig. 4.3 Accuracy of our approximation (4.36), as a function of r , in comparison with La Porte's approximation (4.28).

Chapter 5

In Chapter 4, building on the works of Chandler-Wilde and Hothersall [14] and La Porte [38], we extended and improved the approximation of La Porte [38] by proposing a more stable (in floating point arithmetic) approximation of the 2D impedance half-space Green's function of the Helmholtz equation. We proved a uniform bound on the absolute error of this approximation and we showed, using systematic numerical calculations, that our approximation is more accurate and more efficient than the approximation of Chandler-Wilde and Hothersall [14].

We have achieved our objectives in this thesis and we hope that the presented approximations will be of great benefit for the wide range of applications of these three special functions.

5.2 Further work

It was shown in this thesis that the truncated modified trapezium rule given by (1.23) is an accurate and efficient method to approximate three special functions which can be written as integrals of the form

$$I := \int_{-\infty}^{\infty} e^{-\rho t^2} F(t)\, dt, \quad \text{for} \quad \rho > 0, \tag{5.1}$$

where F is an even meromorphic function with simple poles in a strip surrounding the real line. It is of interest to investigate further to what extent the methods of this thesis are applicable to other special functions. In particular, we summarize below suggested extensions to the work of this thesis, motivated by our theoretical and numerical results, as follows:

(i) The Voigt function, denoted by $V(x,y)$, is defined as $V(x,y) = \mathrm{Re}(w(z))$, and its derivatives satisfy that

$$\frac{\partial V}{\partial x} = -2\,\mathrm{Re}($$

(iii) Additionally, it is interesting to investigate to what extent the methods of Chapter 4 are applicable to the 3D impedance half-space Green's function for the Helmholtz equation [14], to the 2D case of an infinite periodic array of point sources above an impedance plane [28], and the related important 2D case of an infinite periodic array of point sources in free space [40]. In all three cases integral representations of the form (5.1) are relevant with F meromorphic.

References

[1] Abrarov, S. and Quine, B. M. (2015). Sampling by incomplete cosine expansion of the sinc function: Application to the Voigt/complex error function. Applied Mathematics and Computation, 258:425–435.

[2] Abromowitz, M. and Stegun, I. A. (1968). Handbook of Mathematical Functions. Dover.

[3] Alazah, M., Chandler-Wilde, S. N., and La Porte, S. (2014). Computing Fresnel integrals via modified trapezium rules. Numerische Mathematik, 128(4):635–661.

[4] Allasia, G. and Besenghi, R. (1986). Numerical calculation of incomplete gamma functions by the trapezoidal rule. Numerische Mathematik, 50(4):419–428.

[5] Bialecki, B. (1989). A modified sinc quadrature rule for functions with poles near the arc of integration. BIT Numerical Mathematics, 29(3):464–476.

[6] Bowman, J., Senior, T., and Uslenghi, P. (1969). Electromagnetic and acoustic scattering by simple shapes. Amsterdam: North Holland.

[7] Brambley, E. and Gabard, G. (2014). Reflection of an acoustic line source by an impedance surface with uniform flow. Journal of Sound and Vibration, 333(21):5548–5565.

[8] Chandler-Wilde, S. N. (1988). Ground Effects in Enviromental Sound Propagation. PhD thesis, University of Bradford.

[9] Chandler-Wilde, S. N. (2016). Private communication.

[10] Chandler-Wilde, S. N., Hewett, D., Langdon, S., and Twigger, A. (2015). A high frequency boundary element method for scattering by a class of nonconvex obstacles. Numerische Mathematik, 129(4):647–689.

[11] Chandler-Wilde, S. N. and Hothersall, D. (1985). Sound propagation above an inhomogeneous impedance plane. Journal of Sound and Vibration, 98(4):475–491.

[12] Chandler-Wilde, S. N. and Hothersall, D. (1988a). Integral equations in traffic noise simulation. In

[14] Chandler-Wilde, S. N. and Hothersall, D. (1995). Efficient calculation of the Green function for acoustic propagation above a homogeneous impedance plane. *Journal of Sound and Vibration*, 180(5):705–724.

[15] Chiarella, C. and Reichel, A. (1968). On the evaluation of integrals related to the error function. *Mathematics of Computation*, 22(101):137–143.

[16] Cody, W. (1968). Chebyshev approximations for the Fresnel integrals. *Mathematics of Computation*, 22(102):450–453.

[17] Conway, J. B. (1978). *Functions of one complex variable*. Springer.

[18] Davis, P. J. and Rabinowitz, P. (2007). *Methods of numerical integration*. Dover.

[19] Durán, M., Hein, R., and Nédélec, J.-C. (2007). Computing numerically the Green's function of the half-plane Helmholtz operator with impedance boundary conditions. *Numerische Mathematik*, 107(2):295–314.

[20] Fettis, H. E. (1955). Numerical calculation of certain definite integrals by Poisson's summation formula. *Mathematical Tables and Other Aids to Computation*, pages 85–92.

[21] Filippi, P. (1983). Extended sources radiation and Laplace type integral representation: Application to wave propagation above and within layered media. *Journal of Sound and Vibration*, 91(1):65–84.

[22] Gautschi, W. (1970). Efficient computation of the complex error function. *SIAM Journal on Numerical Analysis*, 7(1):187–198.

[23] Gil, A., Segura, J., and Temme, N. M. (2002). Computing complex Airy functions by numerical quadrature. *Numerical Algorithms*, 30(1):11–23.

[24] Goodwin, E. (1949). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$. *Mathematical Proceedings of the Cambridge Philosophical Society*, 45(02):241–245.

[25] Grubeša, S., Jambrošić, K., and Domitrović, H. (2012). Noise barriers with varying cross-section optimized by genetic algorithms. *Applied Acoustics*, 73(11):1129–1137.

[26] Habault, D. (1985). Sound propagation above an inhomogeneous plane: boundary integral equation methods. *Journal of Sound and Vibration*, 100(1):55–67.

[27] Heald, M. A. (1985). Rational approximations for the Fresnel integrals. *Mathematics of Computation*, 44(170):459–461.

[28] Horoshenkov, K. V. and Chandler-Wilde, S. N. (2002). Efficient calculation of two-dimensional periodic and waveguide acoustic Green's functions. *The Journal of the Acoustical Society of America*, 111(4):1610–1622.

[29] Hunter, D. (1964). The calculation of certain Bessel functions. *Mathematics of Computation*, 18(85):123–128.

[30] Hunter, D. (1968). The evaluation of a class of functions defined by an integral. *Mathematics of Computation*, 22(102):440–444.

[31] Hunter, D. and Regan, T. (1972). A note on the evaluation of the complementary error function. Mathematics of Computation, 26(118):539–541.

[32] Jean, P. and Gabillet, Y. (2000). Using a boundary element approach to study small screens close to rails. Journal of sound and vibration, 231(3):673–679.

[33] Jiménez-Mier, J. (2001). An approximation to the plasma dispersion function. Journal of Quantitative Spectroscopy and Radiative Transfer, 70(3):273–284.

[34] Kang, J. (2002). Numerical modelling of the sound fields in urban streets with diffusely reflecting boundaries. Journal of Sound and Vibration, 258(5):793–813.

[35] Kawai, T., Hidaka, T., and Nakajima, T. (1982). Sound propagation above an impedance boundary. Journal of Sound and Vibration, 83(1):125–138.

[36] Kress, R. (1998). Numerical analysis. Springer-Verlag, New York.

[37] Krommer, A. R. and Ueberhuber, C. W. (1998). Computational integration. SIAM.

[38] La Porte, S. (2007). Modified Trapezium Rule Methods for the Efficient Evaluation of Green's Functions in Acoustics. PhD thesis, Brunel University.

[39] Li, J., Sun, G., and Zhang, R. (2016). The numerical solution of scattering by infinite rough interfaces based on the integral equation method. Computers & Mathematics with Applications, 71(7):1491–1502.

[40] Linton, C. (1998). The Green's function for the two-dimensional Helmholtz equation in periodic domains. Journal of Engineering Mathematics, 33(4):377–401.

[41] Liu, S. and Li, K. M. (2012). Efficient computation of the sound fields above a layered porous ground. The Journal of the Acoustical Society of America, 131(6):4389–4398.

[42] Luke, Y. L. (1969). Special functions and their approximations, volume 2. Academic press.

[43] Matta, F. and Reichel, A. (1971). Uniform computation of the error function and other related functions. Mathematics of Computation, pages 339–344.

[44] McNamee, J. (1964). Error-bounds for the evaluation of integrals by the Euler-Maclaurin formula and by Gauss-type formulae. Mathematics of Computation, 18(87):368–381.

[45] Mori, M. (1983). A method for evaluation of the error function of real and complex variable with high relative accuracy. Publications of the Research Institute for Mathematical Sciences, 19(3):1081–1094.

[46] Olver, F. W. (2010). NIST Handbook of Mathematical Functions Hardback and CD-ROM. Cambridge University Press.

[47] O'Neil, M., Greengard, L., and Pataki, A. (2014). On the efficient representation of the half-space impedance Green's function for the Helmholtz equation. Wave Motion, 51(1):1–13.

Appendix A

Matlab codes

A.1 Matlab codes to compute Fresnel integrals

Listing A.1 Matlab code to evaluate $F_N(x)$ given by (2.12)

```
function f = fresnel(x,N)
select = x>=0;
f = zeros(size(x));
if any(select), f(select) = F(x(select),N); end
if any(~select), f(~select) = 1  F(  x(~select),N); end
function f = F(x,N)
h = sqrt(pi/(N+0.5));
t = h*((N:  1:1)  0.5);  AN = pi/h;
t2 = t.*t; t4 = t2.*t2; et2 = exp(  t2);
rooti = exp(i*pi/4);
z = rooti*x; x2 = x.*x; x4 = x2.*x2; z2 = i*x2;
S = (  et2(1)./(x4+t4(1))).*(z2+t2(1));
for n = 2:N
    S = S + (  et2(n)./(x4+t4(n))).*(z2+t2(n));
end
ez = exp((2*AN*i*rooti)*x);

```

Listing A.2 Matlab code to evaluate $C_N(x)$ and $S_N(x)$ given by (2.14) and (2.15)

```
function [C,S] = fresnelCS(x,N)
h = sqrt(pi/(N+0.5));
t = h*((N:  1:1)  0.5);  AN = pi/h; rootpi = sqrt(pi);
t2 = t.
```

A.2 Matlab code to compute Faddeeva function

```
end
function f = w3(z,N)
z2 = z.*z; az = (2i/AN)*z;
a = h*((N:  1:1)+0.5); a2 = a.^2; et2 = exp( a2);
S1 = et2(1)./(z2 a2(1));
for n = 2 : N
    S1 = S1 + et2(n)./(z2 a2(n));
end
h0 = 0.5*h;
S0 = exp( h0.^2)./(z2 h0.^2);
f = az.*(S0 + S1);
end
end
```

A.3 Matlab code to compute P_b

```
if V1 == V2
    V = 1;
else
    V =  1;
end
Cm =  2*V*exp( 1i*rho.*am).*heaviside(H imag(z2))./(1+exp
   ( 2*1i*pi*z2./h));
TC = pi*(Cp + Cm)./(2*sqrt(1 beta.^2));
t = h.*((N: 1:1)+ 0.5); t2 = t.^2; h0 = 0.5.*h; et2 =  exp
   ( t2.*rho);
s1 = beta + gamma.*(1+1i*t2); s2 = sqrt(t2 2*1i);
s3 = t2 1i*ap; s4 = t2 1i*am;
S1 = et2(1).*s1(1)./(s2(1).*s3(1).*s4(1));
for n = 2 : N
    S1 = S1 + et2(n).*s1(n)./(s2(n).*s3(n).*s4(n));
end
A =  (beta + gamma.*(1+1i*h0.^2)).*exp( rho.*h0.^2);
B = sqrt(h0.^2  2*1i).*(h0^2  z1.^2).*(h0^2  z2.^2);
I = (beta.*exp(1i*rho)/pi).*2*h.*(S1 + A./B);
f2 = I + (beta.*exp(1i*rho)/pi).*TC;
end
end
```