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

I would like to thank my supervisor Prof Simon N. Chandler-Wilde for his endless support and patience. His deep insight, inspiring enthusiasm and constructive criticism have greatly helped me to accomplish this thesis.

Especially and with all my heart I want to thank my wife and my sons, for their unconditional love and support. To them I owe all.

I would also like to thank Prof Masatake Mori for his private communication regarding formula (2.8) in his paper [45]. Sadly, Prof Masatake Mori passed away on February 24, 2017, the date of my PhD viva...

Abstract

This thesis is concerned with the efficient (accurate and fast) computation, via modified trapezium rules, of some special functions which can be written as integrals of the form

$$\int_{-\infty}^{\infty} f(t)\,dt,$$

where

$$f(t) = e^{-\rho t^2} F(t), \quad \rho > 0,$$

and F is an even meromorphic function with simple poles in a strip surrounding the real line. Specifically, this thesis considers the approximation of the Fresnel integrals, the complementary error function of complex argument and the Faddeeva function, and the 2D impedance half-space Green's function for the Helmholtz equation.

The trapezium rule is exponentially convergent when F is analytic in a strip surrounding the real axis. In the case of meromorphic functions with simple poles, the trapezium rule can be modified to take into account the presence of these poles. The effect of truncating this modified trapezium rule is considered and specific approximations with explicit choices for step-size and number of quadrature points are given. Rigourous bounds for the errors are proven using complex analysis methods, and numerical calculations that demonstrate the accuracy of these approximations compared with the best known methods are also provided.

Table of contents

1 Introduction **1**

1.1 Special functions 1

1.2 The trapezium rule approximation 2

1.3 Numerical Examples 8

1.3.1 Example 1 8

1.3.2 Example 2 9

1.4 The contributions of this Thesis 10

2 Fresnel integrals **15**

2.1 Introduction 15

2.2 Summary of the main Results 19

2.3 The proposed approximation and its error bounds 20

2.3.1 Extensions of the error bounds 27

2.4 The approximations of $C(x)$ and $S(x)$ 29

2.5 Numerical results 32

3 The Faddeeva function **37**

3.1 Introduction 37

3.2 Summary of the main results 41

3.3 The proposed approximation and its error bounds 43

3.4 Numerical results 61

4 The 2D impedance half-space Green's function for the Helmholtz equation **67**

4.1 Introduction 67

4.2 Summary of the main results 74

4.3 The proposed approximation and its error bounds 77

4.3.1 Bounding the discretisation error 80

4.3.2 Bounding the truncation error 83

4.3.3 Choices of the step-size h 85
4.3.4 Bounding the total error 88
4.4 Numerical results 90

5 Concluding remarks and further work **97**
5.1 Concluding Remarks 97
5.2 Further work 98

References **101**

Appendix A Matlab codes **105**
A.1 Matlab codes to compute Fresnel integrals 105
A.2 Matlab code to compute Faddeeva function 107
A.3 Matlab code to compute P_β 109

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]. There exist also many mathematical libraries which contain routines and codes for computing special functions efficiently, and in particular we flag the very popular Press *et al.*[52] and the repository *GAMS* (*http://gams.nist.gov/*) which gives links to a range of software packages.

This thesis considers the computation of three special functions which can be written as integrals of the form

$$I := \int_{-\infty}^{\infty} f(t)\, dt, \tag{1.1}$$

where

$$f(t) = e^{-\rho t^2} F(t), \quad \rho > 0, \tag{1.2}$$

and F is an even meromorphic function with simple poles in a strip surrounding the real line. Namely, we approximate Fresnel integrals in Chapter 2, the complex error function and Faddeeva function in Chapter 3, and finally the 2D impedance half-space Green's function for the Helmholtz equation (see e.g. [19, 47]) in Chapter 4.

The approximations we propose are based on the trapezium rule approximation (1.6) below, and especially modifications to this rule to take into account the presence of simple poles of the integrand. To the best of our knowledge the earliest paper which applies the trapezium rule to computation of a special function is Turing [61] where he proposed an approximation of the zeta function. Other special functions can be represented in the form

(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 $w_1, ..., w_N$ and $x_1, ..., x_N$ 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 a full precision in double-precision arithmetic; while approximating the integral

$$\frac{a\, e^{-a^2}}{\pi} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{t^2 + a^2}\, dt = \operatorname{erfc}(a), \tag{1.5}$$

for $a = 0.1$ using J_N with $N = 1000$ gives only 3-digit accuracy, while the trapezium rule in (1.6), with $\alpha = 0$, gives 4-digit accuracy with 40 quadrature points (see Example 2 below).

1.2 The trapezium rule approximation

Numerical quadrature (numerical integration) is a well-established and fundamental topic in numerical analysis which deals with approximating integrals that cannot be computed exactly. Different quadrature methods are available to evaluate different types of integrals and we refer to Davis [18], Krommer [37] and Kress [36] for the classical theory of numerical integration. The trapezium rule (1.6) is perhaps the simplest and probably the oldest quadrature rule; it is reported recently by Ossendrijver [48] that the Babylonian astronomers used the trapezium rule in 350 BCE to calculate Jupiter's position by computing the area under a time-speed graph.

For any $h > 0$ and $\alpha \in [0,1)$, we define through this thesis the *trapezium rule approximation* to the integral I given by (1.1) by

$$I(h,\alpha) := h \sum_{k\in\mathrm{Z}} f((k-\alpha)h). \tag{1.6}$$

Remark 1.2.1. *Many authors use* **trapezoidal** *in place of* **trapezium**, *and reserve the terminology trapezium rule for* (1.6) *in the case* $\alpha = 0$. *For simplicity we will call the approximation rule* (1.6) *the trapezium rule, for all* $\alpha \in [0,1)$, *noting that*

$$h \sum_{k\in\mathrm{Z}} f((k-\alpha)h) = h \sum_{k\in\mathrm{Z}} \tilde{f}(kh),$$

where $\tilde{f}(t) = f(t-\alpha h)$, *i.e.* $I(h,\alpha)$ *is certainly the trapezium rule in the standard sense for a shifted function* $\tilde{f}$. *Many authors also refer to the formula* (1.6) *for* $\alpha = 1/2$ *as the* **midpoint rule**.

Approximating the integral (1.1) using the trapezium rule (1.6) gives an exponential rate of convergence when the integrand is analytic in a strip surrounding the real axis with sufficient decay at $\pm\infty$. The derivation of this result, using contour integration and Cauchy's residue theorem, dates back at least to Turing [61] and Goodwin [24]. The application of the trapezium rule for more general cases was developed later by Fettis [20], McNamee [44], Schwartz [55] and Stenger [57]. For a full history of the trapezium rule and its different applications we refer to the recent paper by Trefethen and Weideman [60].

To obtain accurate approximations it is desirable to modify the trapezium rule when the integrand is a meromorphic function with simple poles near the real axis, to take into account the presence of these poles. The required modification is to add a correction factor, which is related to the residues of the integrand at these poles, to the original trapezium rule. This correction factor can greatly improve both accuracy and rate of convergence. This kind of modification was first suggested in the context of developing methods for evaluating the complementary error function of complex argument by Chiarella and Reichel [15], Matta and Reichel [43], and Hunter and Regan [31]. The proposed modification follows naturally from the contour integration argument used by Turing [61] and Goodwin [24] to prove that the trapezium rule is exponentially convergent. This modification was developed later for more general cases in Bialecki [5], Mori [45] and La Porte [38].

We will present below in Propositions 1.2.1–1.2.4 the well-known results that will be the starting point for our analysis; and which show the exponential rate of convergence of the trapezium rule (1.6) with the suggested modification. For completeness we include the short proofs of these key results, and we put them in the context of the literature immediately

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 $S_H = \{z \in \mathrm{C} : |\mathrm{Im}(z)| < H\}$, we have that*

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

(ii) F is continuous on $\overline{S}_H \backslash \{z_1, z_2, z_3, ..., z_m\}$;

(iii) $F(z) = \mathrm{O}(1)$ 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

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

which is a meromorphic function with simple poles at $z = (k - \alpha)h$, $k \in \mathrm{Z}$, 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 $A_k = (k-\alpha+\frac{1}{2})h$ 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

$$\begin{aligned}\int_{C_H} f(z)g(z)\,dz &= 2\pi i \left(\sum_{k=1}^{n+j+1} \mathrm{Res}(fg,(k-\alpha)h) + \sum_{k=1}^{m} \mathrm{Res}(fg,z_k) \right) \\ &= 2\pi i \left(\frac{ih}{\pi} \sum_{k=-j}^{n} f((k-\alpha)h) + \sum_{k=1}^{m} g(z_k)R_k \right).\end{aligned}$$

Also, we have that

$$\lim_{n,j\to\infty} \int_{C_H} f(z)g(z)\,dz = \int_{\Gamma_H'} f(z)g(z)\,dz - \int_{\Gamma_H} f(z)g(z)\,dz,$$

since, using (iii) in Assumption 1.2.1,

$$\int_{-H}^{H} f(A_n+iy)g(A_n+iy)\,dy \to 0 \quad \text{and} \quad \int_{-H}^{H} f(-A_j+iy)g(-A_j+iy)\,dy \to 0,$$

as $n, j \to \infty$. □

Proposition 1.2.2. *If Assumption 1.2.1 holds, then the integral I as defined in* (1.1) *exists and*

$$I = \frac{1}{2}\left(\int_{\Gamma_H} f(z)\,dz + \int_{\Gamma_H'} f(z)\,dz \right) + \pi i \sum_{k=1}^{m} \mathrm{sign}(y_k)\,R_k, \tag{1.11}$$

where $y_k = \mathrm{Im}(z_k)$ and $R_k = \mathrm{Res}(f,z_k)$.

Proof. Let $z_1, z_2, ..., z_n$ be the simple poles of f in S_H with positive imaginary parts and $z_{n+1}, ..., z_m$ the simple poles with negative imaginary parts. Using Cauchy's residue theorem for the positively oriented rectangular contour C_+ with vertices [1] at $\pm\infty$ and $\pm\infty + iH$, we have that

$$\int_{C_+} f(z)\,dz = I - \int_{\Gamma_H} f(z)\,dz = 2\pi i \sum_{k=1}^{n} \mathrm{Res}(f,z_k). \tag{1.12}$$

Similarly, and using Cauchy's residue theorem for the negatively oriented rectangular contour C_- with vertices at $\pm\infty$ and $\pm\infty - iH$, we have that

$$\int_{C_-} f(z)\,dz = I - \int_{\Gamma_H'} f(z)\,dz = -2\pi i \sum_{k=n+1}^{m} \mathrm{Res}(f,z_k). \tag{1.13}$$

The result follows now by combining (1.12) and (1.13). □

[1] In more detail one first takes finite rectangular contour with vertices at $-A_j$, A_n, $-A_j + iH$ and $A_n + iH$ and then takes the limit as $j, n \to \infty$ as in Proposition 1.2.1

The following proposition is well-known from many papers. It is in Goodwin [24] for the case when $\alpha = 0$ and the integrand is analytic in S_H, in Chiarella and Reichel [15] for $\alpha = 0$ and in Hunter and Regan [31] for $\alpha = 0$ and $\alpha = 1/2$, where in the last two papers $F(t) = 1/(t^2 + a^2)$, for some $a \in \mathbb{C}$. For the more general case we can find similar results in Bialecki [5] and La Porte [38].

Proposition 1.2.3. *For $h > 0$ and $\alpha \in [0,1)$ let $E(h,\alpha) := I - I(h,\alpha)$. Then*

$$E(h,\alpha) = J_H + C(h,\alpha), \tag{1.14}$$

where

$$J_H := \frac{1}{2}\left(\int_{\Gamma_H} f(z)(1-g(z))\,dz + \int_{\Gamma_H'} f(z)(1+g(z))\,dz\right), \tag{1.15}$$

and

$$C(h,\alpha) := \pi i \sum_{k=1}^{m} (\operatorname{sign}(y_k) - g(z_k))\, R_k, \tag{1.16}$$

Proof. The result follows directly by combining the results of Propositions (1.2.1) and (1.2.2). □

Remark 1.2.2. *We can rewrite the expression for $C(h,\alpha)$ in the previous proposition as*

$$C(h,\alpha) := \pi i \sum_{k=1}^{m} \Phi(z_k,\alpha)\,,$$

where

$$\Phi(z_k,\alpha) := \frac{R_k}{1-e^{-2i\pi(\alpha+z_k/h)}} \times \begin{cases} 2\,e^{-2i\pi(\alpha+z_k/h)}, & \operatorname{Im}(z_k) < 0\,, \\ 1+e^{-2i\pi(\alpha+z_k/h)}, & \operatorname{Im}(z_k) = 0\,, \\ 2, & \operatorname{Im}(z_k) > 0, \end{cases} \tag{1.17}$$

with $R_k = \operatorname{Res}(f, z_k)$.

Definition 1.2.1. *The* **modified trapezium rule***, denoted by $I^*(h,\alpha)$, is defined as*

$$I^*(h,\alpha) := I(h,\alpha) + C(h,\alpha), \tag{1.18}$$

where $I(h,\alpha)$ and $C(h,\alpha)$ are given by (1.6) *and* (1.16)*, respectively.*

The following proposition demonstrates the exponential convergence rate of the modified trapezium rule $I^*(h,\alpha)$. This result has appeared in many papers for different cases of the

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

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

$$|E^*(h,\alpha)| \leq \frac{2\sqrt{\pi}M_H(F)\,e^{\rho H^2 - 2\pi H/h}}{\sqrt{\rho}(1-e^{-2\pi H/h})}, \tag{1.19}$$

where

$$M_H(F) := \sup_{x\in\mathbb{R},|y|=H} |F(x+iy)|.$$

Proof. Using (1.14) and (1.18) we have that

$$\begin{aligned} E^*(h,\alpha) &= J_H \\ &= \frac{1}{2}\left(\int_{\Gamma_H} f(z)(1-g(z))\,dz + \int_{\Gamma_H'} f(z)(1+g(z))\,dz\right). \end{aligned}$$

Using (1.15), (1.8) and (1.9) we have that

$$\begin{aligned} |J_H| &\leq \frac{e^{-2\pi H/h}}{1-e^{-2\pi H/h}}\left(\int_{-\infty}^{\infty}|f(x+iH)|\,dx + \int_{-\infty}^{\infty}|f(x-iH)|\,dx\right) \\ &\leq \frac{2M_H(F)\,e^{\rho H^2-2\pi H/h}}{1-e^{-2\pi H/h}}\int_{-\infty}^{\infty} e^{-\rho x^2}\,dx \\ &= \frac{2\sqrt{\pi}M_H(F)\,e^{\rho H^2-2\pi H/h}}{\sqrt{\rho}(1-e^{-2\pi H/h})}. \end{aligned} \tag{1.20}$$

□

Remark 1.2.3. *Note that the value of H that minimises the expression $\rho H^2 - 2\pi H/h$ in* (1.19) *is*

$$H = \frac{\pi}{\rho h}. \tag{1.21}$$

The integrands in the integral representations of the special functions considered in this thesis are even, and hence we will choose $\alpha = 0$ or $1/2$ in the modified trapezium rule (1.18) to take advantage of the symmetry of the trapezium rule for these special values of α which reduces the cost of computing by half. In applications, the approximation formula (1.18) is truncated after N terms and the recommended choice of N will be discussed later for the three special functions. We define the truncated modified trapezium rule in the following.

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*

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

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

$$I_N^*(h,\alpha) := I_N(h,\alpha) + C(h,\alpha). \tag{1.23}$$

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

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

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

$$E_N^*(h,\alpha) = E^*(h,\alpha) + T_N(h,\alpha). \tag{1.25}$$

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]):

$$I = \int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi} = 1.77245385090551602 73\ldots. \tag{1.26}$$

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 $I_N(h,0)$ for different values of N and h. We choose $h=\sqrt{\pi/(N+1)}$. This value of h is chosen in [38] to equalize the discretization and truncation errors of the modified trapezium rule with $\alpha=0$. We can see from the table that the trapezium rule approximation gives 3-digit accuracy with only one quadrature point for $h\approx 1.253$ and 16-digit accuracy with 11 quadrature points for $h\approx 0.512$.

N	$h=\sqrt{\pi/(N+1)}$	$I_N^*(h,0)$
1	1.253	**1.77**43
3	0.886	**1.7724**60
7	0.627	**1.772453850**33
11	0.512	**1.772453850905516**

Table 1.1 Approximation of the integral (1.26)

1.3.2 Example 2

Let $a>0$ and

$$I=\frac{a\,e^{-a^2}}{\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2}}{t^2+a^2}\,dt=\operatorname{erfc}(a), \tag{1.27}$$

where erfc is the complementary error function. For $a=0.1$, and using the built in function in *Mathematica 11*, we find that

$$I=0.88753708398171510 77...$$

The integrand here is a meromorphic function with two simple poles at $t=\pm ia$ so, for $H=\pi/h$ as per Remark 1.2.3,

$$\begin{aligned} I_N^*(h,0) &= I_N(h,0)+C(h,0)\\ &= \frac{a\,h\,e^{-a^2}}{\pi}\left(\frac{1}{a^2}+2\sum_{k=1}^{N}\frac{e^{-k^2h^2}}{k^2h^2+a^2}\right)+\frac{2}{1-e^{2\pi a/h}}, \quad \text{if} \quad a<H. \end{aligned}$$

Tables 1.2 and 1.3 below show computed values of $I_N(h,0)$ and $I_N^*(h,0)$ for different values of N and h. We choose $h=0.7\,(N+1)^{-2/3}$ in Table 1.2 as recommended by [60], and $h=\sqrt{\pi/(N+1)}$ in Table 1.3 as recommended by [38]. We can see from the first table that the trapezium rule needs 20 quadrature points to reach 2-digit accuracy for $h=0.092$ and 80 quadrature points to reach 7-digit accuracy for $h=0.037$, while in Table 1.3 the modified trapezium rule achieves with only 1 quadrature point 3-digit accuracy and 15-digit accuracy with 9 quadrature points.

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

Table 1.2 Approximating the integral (1.27) using the trapezium rule for $a = 0.1$.

N	$h = \sqrt{\pi/(N+1)}$	$I_N^*(h,0)$
1	1.253	**0.887**486
3	0.886	**0.8875370**406
6	0.670	**0.887537083979**8
9	0.492	**0.887537083981715**

Table 1.3 Approximating the integral (1.27) using the modified trapezium rule for $a = 0.1$.

1.4 The contributions of this Thesis

The above examples illustrate the large potential of the modified trapezium rule. However, there are a number of difficulties and issues that arise when applying these methods. The main contribution of this thesis is to address and solve the following issues, in particular as they arise when computing three particular special functions:

1. To apply the modified trapezium rule to a particular case there are a number of implementation choices to be made, including: the value of α in (1.6) since the wrong choice leads to numerical instability if the poles of the integrand lie on or near the real line; the optimal choice of the parameter H in the modified trapezium rule (1.18); the optimal truncation of the infinite sums in (1.6) to the finite sums in (1.22), in other words the optimal choice of N in (1.22) and, related to this, the choice of how the step-size h should depend on N and other parameters. That there is some challenge in computing the optimal parameters for the modified trapezium rule was recognised by Weideman [62] who, in the context of developing approximations for the complementary error function of complex argument, observed that the modified trapezium rule is

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

   The choices of the parameters α, H, h and N are interdependent and influence the numerical error in complex ways. So the second important and complementary issue to address and solve is

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 $w(z) = e^{-z^2}\mathrm{erfc}(-iz)$ (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 $\mathrm{erfc}(z)$ with $z = x + iy$, starting from the integral representation

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

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.

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

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.

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

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

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

where $\phi(t)$ denotes the fractional part of t, i.e. $\phi(t) = t - [t]$. 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

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

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 $|x - \pi/h| \le \varepsilon$, for some $\varepsilon > 0$, by taking $H = \pi/h \pm \varepsilon$ 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

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

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

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

with

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

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

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

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 $w(z)$ based on the truncated trapezium rule in (1.23). Our starting point is the method for computation of the complementary error function of complex argument due to Matta and Reichel [43] and Hunter and Regan [31]. We show through theoretical and numerical calculations in Sections §3.3 and §3.4 that the proposed approximation is exponentially convergent as a function of N, the number of quadrature points, uniformly in the first quadrant. We compare our approximation with the best known methods, and we show that our approximation attains an accuracy of 10^{-16} with $N = 12$ quadrature points.

In Chapter 4, we propose an approximation to the 2D impedance half-space Green's function for the Helmholtz equation (4.1). This problem is important in outdoor sound propagation applications (see e.g.[19, 47]). Building on a method of computing this function in La Porte [38] using the modified trapezium rule (1.18) with $\alpha = 0$, we propose an improved approximation based on the modified trapezium rule (1.18), and we show through theoretical and numerical calculations in Sections §4.3–§4.4 that this approximation is more stable than the approximation of La Porte [38] and more accurate than the approximation of Chandler-Wilde and Hothersall [14].

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 $C(x)$, $S(x)$, and $F(x)$ be the Fresnel integrals defined by

$$C(x) := \int_0^x \cos\left(\tfrac{1}{2}\pi t^2\right)\,dt, \quad S(x) := \int_0^x \sin\left(\tfrac{1}{2}\pi t^2\right)\,dt, \tag{2.1}$$

and

$$F(x) := \frac{\mathrm{e}^{-\mathrm{i}\pi/4}}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{\mathrm{i}t^2}\,dt. \tag{2.2}$$

Our definitions in (2.1) are those of [2] and [46, §7.2(iii)], and F, C and S are related through

$$\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}F(x) = \tfrac{1}{2} - C\left(\sqrt{2/\pi}\,x\right) + \mathrm{i}\left(\tfrac{1}{2} - S\left(\sqrt{2/\pi}\,x\right)\right). \tag{2.3}$$

The derivation of our approximation makes use of the relationship between the Fresnel integral and the error function, that

$$F(x) = \tfrac{1}{2}\mathrm{erfc}(\mathrm{e}^{-\mathrm{i}\pi/4}x) = \tfrac{1}{2}\,\mathrm{e}^{\mathrm{i}x^2}\,w\left(\mathrm{e}^{\mathrm{i}\pi/4}x\right) \tag{2.4}$$

where erfc is the complementary error function, defined by

$$\mathrm{erfc}(z) := \frac{2}{\sqrt{\pi}} \int_z^\infty \mathrm{e}^{-t^2}\,dt,$$

and $w(z)$ is the Faddeeva function, defined by

$$w(z) := \mathrm{e}^{-z^2}\mathrm{erfc}(-\mathrm{i}z).$$

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

$$w(z) = \frac{\mathrm{i}}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-t^2}}{z-t}\,dt = \frac{\mathrm{i}z}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-t^2}}{z^2-t^2}\,dt, \quad \mathrm{Im}(z)>0. \tag{2.5}$$

Combining (2.4) and (2.5) gives an integral representation for $F(x)$, that

$$F(x) = \frac{x}{2\pi}\,\mathrm{e}^{\mathrm{i}(x^2+\pi/4)}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-t^2}}{x^2+\mathrm{i}t^2}\,dt, \quad x>0. \tag{2.6}$$

Fresnel integrals arise in applications throughout science and engineering, especially in problems of wave diffraction and scattering [6, §8.2], [10], so that methods for the efficient and accurate computation of these functions are of wide application. The purpose of this chapter is to present new approximations for the Fresnel integrals, based on applications of the truncated modified trapezium rule approximation (1.23) to the integral representation (2.6) for $F(x)$.

In the context of developing methods for evaluating the complementary error function of complex argument (by (2.4), evaluating $F(x)$ for x real is just a special case of this larger problem), Chiarella and Reichel [15], Matta and Reichel [43], and Hunter and Regan [31] proposed modifications of the trapezium rule that follow naturally from the contour integration argument used to prove that the trapezium rule is exponentially convergent. The most appropriate form of this modification is that in [31] where the modified trapezium rule approximation

$$F(x) \approx \frac{xh}{\pi}\,\mathrm{e}^{\mathrm{i}(x^2+\pi/4)}\sum_{k=1}^{\infty}\frac{\mathrm{e}^{-\tau_k^2}}{x^2+\mathrm{i}\tau_k^2} + R(h,x), \ \text{ for } x>0, \tag{2.7}$$

is proposed. Here the correction term $R(h,x)$ is defined by

$$R(h,x) := \begin{cases} 1/(\exp(2\pi\mathrm{e}^{-\mathrm{i}\pi/4}x/h)+1), & \text{if } 0<x<\sqrt{2}\,\pi/h, \\ 0.5/(\exp(2\pi\mathrm{e}^{-\mathrm{i}\pi/4}x/h)+1), & \text{if } x=\sqrt{2}\,\pi/h, \\ 0, & \text{if } x>\sqrt{2}\,\pi/h. \end{cases}$$

The approximation (2.7) clearly coincides with $F_N(x)$, given by (2.11), for $0<x<\sqrt{2}\,\pi/h$, if the range of summation in (2.7) is truncated to $1,\ldots,N$ and the choice (2.41) for h is made. Hunter and Regan [31] proved that the magnitude of the error in (2.7) is

$$\leq \frac{x\mathrm{e}^{-\pi^2/h^2}}{\sqrt{\pi}\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)\,|x^2/2-\pi^2/h^2|}, \tag{2.8}$$

for $x > 0$, provided $x \neq \sqrt{2}\pi/h$. Similar estimates, it appears arrived at independently, are derived by Mori [45], in which paper the emphasis is on computing $\mathrm{erfc}(x)$ for real x.

The approximation (2.7) is the starting point for the method we propose in this chapter. Our main contributions (see §2.2 for detail) are: (i) to point out that the approximation proposed in (2.7) for $0 < x < \sqrt{2}\,\pi/h$ in fact provides an accurate (and real-analytic) approximation to the entire function F on the whole real line; (ii) to provide an optimal formula for the choice of the step-size h as a function of N, the number of terms retained in the sum in (2.7); (iii) to prove that, with this choice of h, the resulting approximations are exponentially convergent as a function of N, uniformly on the real line (this in contrast to (2.8) which blows up at $x = \sqrt{2}\,\pi/h$).

Naturally, there exist already a number of effective schemes for computation of Fresnel integrals, and we briefly summarise now the best of these. An effective computational method for smaller values of $|x|$ is to make use of the power series for $C(x)$ and $S(x)$ (see (2.65) below). These converge for all x, and very rapidly for smaller x, and so are widely used for computation. For example, the algorithm in the standard reference [52] uses these power series for $|x| \leq 1.5$. For this range, after the first two terms, these series are alternating series of monotonically decreasing terms, and the error in truncation has magnitude smaller than the first neglected term. Thus, for $|x| \leq 1.5$, the errors in computing $C(x)$ and $S(x)$ by these power series truncated to N terms are $\leq 2 \times 10^{-16}$ and $\leq 2.3 \times 10^{-17}$, respectively, for $N = 14$.

For $|x| > 1.5$, [52] recommends computation using the representations in terms of erfc which follow from (2.3) and (2.4), and the continued fraction representation for $e^{z^2}\mathrm{erfc}(z) = w(\mathrm{i}z)$ given as [46, (7.9.2)]. Methods for evaluation of $w(z)$ based on continued fraction representations for larger complex z (which can be applied to evaluate $F(x)$ and hence $C(x)$ and $S(x)$) are also discussed in Gautschi [22] and are finely tuned, to form TOMS "Algorithm 680", in Poppe and Wijers [50], which achieves relative errors of 10^{-14} over "nearly all" the complex plane by using 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] presents an alternative method of computation (the derivation starts from the integral representation (2.5)) which approximates $w(z)$ by

$$W_N(z) := \frac{1}{\sqrt{\pi}(L-\mathrm{i}z)} + \frac{2}{(L-\mathrm{i}z)^2}\sum_{n=0}^{N-1} a_{n+1}\left(\frac{L+\mathrm{i}z}{L-\mathrm{i}z}\right)^n, \tag{2.9}$$

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

$$a_n := \frac{1}{2M} \sum_{j=-M+1}^{M-1} (L^2 + t_j^2) e^{-t_j^2} e^{-in\theta_j}, \quad n = 1, .., N, \tag{2.10}$$

with $M = 2N$, $t_j = L\tan(\theta_j/2)$ and $\theta_j = \pi j/M$ for $j = -M+1, ..., M-1$. Using (2.9) with $N = 36$ to compute $F(x) = \mathrm{e}^{\mathrm{i}x^2} w(\mathrm{e}^{\mathrm{i}\pi/4}x)/2$ gives a relative error $\leq 10^{-15}$ uniformly on the real line. Weideman [62] argues carefully and persuasively that, in terms of operation counts, the work required to compute $w(z)$ with the 10^{-14} relative accuracy of Algorithm 680 [50] is much smaller using the approximation (2.9) for intermediate values of $|z|$ (values in approximately the range $1.5 \leq |z| \leq 5$ for the case $\arg(z) = \pi/4$ which we require).

All these approximations described above are polynomial or rational approximations (or piecewise polynomial/rational approximations, proposing different approximations on different regions). Many other authors describe approximations of these types for computing the Fresnel integrals specifically with real arguments. The best of these in terms of accuracy is Cody [16], where numerical coefficient values are given for piecewise rational approximations to $C(x)$ and $S(x)$ for $0 \leq x \leq 1.6$, and for piecewise rational approximations to $f(x)$ and $g(x)$ in (2.60) and (2.61), for $x \geq 1.6$. These approximations, in their respective regions of validity, achieve relative errors $\leq 10^{-15.58} \approx 2.7 \times 10^{-16}$, this using rational approximations which are ratios of polynomials of degree ≤ 6; in total five different approximations are used on different subintervals of the real axis. Single rational approximations, based on a "polar" version of (2.60) and (2.61), are computed in [27], but these are of limited accuracy (absolute errors $\leq 4 \times 10^{-8}$).

We end this introduction by outlining the remainder of this chapter. Section 2.2 gives a summary of the main results and contributions in this chapter. In §2.3 we derive the approximation (2.11) to $F(x)$ and prove rigorous bounds on $E_N(x)$, including (2.44) and (2.49). In §2.3 we deduce from this the approximations (2.14) and (2.15) and bounds on the errors $C(x) - C_N(x)$ and $S(x) - S_N(x)$, especially bounds for x small, and survey other methods for computing Fresnel integrals. In §2.5 we show numerical results, comparing our new approximations with the error bounds derived in the earlier sections and with certain rival methods for computing Fresnel integrals.

2.2 Summary of the main Results

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

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

$$= \frac{1}{\exp\left(2A_N x \mathrm{e}^{-\mathrm{i}\pi/4}\right)+1} + \frac{x}{A_N}\mathrm{e}^{\mathrm{i}(x^2+\pi/4)}\sum_{k=1}^{N}\frac{\mathrm{e}^{-t_k^2}}{x^2+\mathrm{i}t_k^2}, \tag{2.12}$$

where

$$t_k := \frac{(k-1/2)\,\pi}{\sqrt{(N+1/2)\,\pi}}, \quad A_N := t_{N+1} = \sqrt{(N+1/2)\pi}. \tag{2.13}$$

The corresponding approximations to $C(x)$ and $S(x)$ (obtained by substituting this approximation in (2.3) and separating real and imaginary parts) are

$$C_N(x) := \frac{1}{2}\frac{\sinh(\sqrt{\pi}A_N x)+\sin(\sqrt{\pi}A_N x)}{\cos(\sqrt{\pi}A_N x)+\cosh(\sqrt{\pi}A_N x)} + \frac{\sqrt{\pi}x}{A_N}\left(a_N\left(\frac{\pi}{2}x^2\right)\sin\left(\frac{\pi}{2}x^2\right) - b_N\left(\frac{\pi}{2}x^2\right)\cos\left(\frac{\pi}{2}x^2\right)\right) \tag{2.14}$$

and

$$S_N(x) := \frac{1}{2}\frac{\sinh(\sqrt{\pi}A_N x)-\sin(\sqrt{\pi}A_N x)}{\cos(\sqrt{\pi}A_N x)+\cosh(\sqrt{\pi}A_N x)} - \frac{\sqrt{\pi}x}{A_N}\left(a_N\left(\frac{\pi}{2}x^2\right)\cos\left(\frac{\pi}{2}x^2\right) + b_N\left(\frac{\pi}{2}x^2\right)\sin\left(\frac{\pi}{2}x^2\right)\right), \tag{2.15}$$

where

$$a_N(s) := s\sum_{k=1}^{N}\frac{\mathrm{e}^{-t_k^2}}{s^2+t_k^4}, \quad b_N(s) := \sum_{k=1}^{N}\frac{t_k^2\,\mathrm{e}^{-t_k^2}}{s^2+t_k^4}. \tag{2.16}$$

These approximations, designed for computation of $F(x)$, $C(x)$ and $S(x)$ for all $x \in \mathbb{R}$, are attractive in several respects.

- The approximation F_N is proven in Theorems 2.3.3 and 2.3.5 to converge to F approximately in proportion to $\exp(-\pi N)$, uniformly on the real line with respect to both absolute and relative error, and this predicted rate of exponential convergence is observed in numerical experiments in §2.5.

- The approximations $F_N(z)$, $C_N(z)$ and $S_N(z)$ to the entire functions F, C, and S, are analytic in the strip $|\mathrm{Im}(z)| < \sqrt{(N+1/2)\pi/2}$ and the error bounds we prove extend

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 normalisation of $F(x)$ is such that

$$F(-x) = 1 - F(x), \tag{2.17}$$

so that, in particular, $F(0) = 1/2$. It is clear from (2.11) that the same holds for $F_N(x)$, *i.e.*,

$$F_N(-x) = 1 - F_N(x). \tag{2.18}$$

Similarly, where an overline denotes a complex conjugate,

$$\overline{F(z)} = F(\mathrm{i}\bar{z}) \text{ and } \overline{F_N(z)} = F_N(\mathrm{i}\bar{z}). \tag{2.19}$$

Both these symmetries can be deduced as a consequence of the structure of C and S and their approximations: by inspection of (2.14) and (2.15) we see that

$$C_N(x) = x f_C(x^4), \quad S_N(x) = x^3 f_S(x^4), \tag{2.20}$$

where f_C and f_S are analytic in a neighbourhood of the real line and are real-valued for real arguments. This is the same structure as C and S (see (2.65) below). In particular, (2.20) implies that C_N and S_N, like C and S, are odd functions.

- The final attractive feature is that these approximations are straightforward to code. Listing A.1 shows the simple Matlab code used to evaluate $F_N(x)$ for all the computations in this paper. Of course this code is easily converted to other programming languages.

2.3 The proposed approximation and its error bounds

In this section we derive the approximation $F_N(x)$ to $F(x)$ and derive error bounds for this approximation demonstrating that both the absolute and relative errors converge exponentially to zero as N increases, uniformly on the real line, and that $N = 12$ is enough to achieve errors

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

$$F(x) := \int_{-\infty}^{\infty} f(t)\,dt, \text{ where } f(t) := \mathrm{e}^{\mathrm{i}(x^2+\pi/4)}\,\frac{x}{2\pi}\,\frac{\mathrm{e}^{-t^2}}{x^2+\mathrm{i}t^2}, \tag{2.21}$$

and we have suppressed in our notation the dependence of $f(t)$ on x. The integrand in (2.21) is even and meromorphic with simple poles at $t = \pm z_0$, with $z_0 := \mathrm{e}^{\mathrm{i}\pi/4}x$. The residues at these poles are

$$R_1 = \mathrm{Res}\,(f, z_0) = \frac{1}{4i\pi} \quad \text{and} \quad R_2 = \mathrm{Res}\,(f, -z_0) = -R_1. \tag{2.22}$$

Using (1.16) and Remark 1.2.2, we find that

$$C(h, 1/2) = (1 + \mathrm{i}\tan(\pi z_0/h))/2. \tag{2.23}$$

Applying the trapezium rule (1.6) with $\alpha = 1/2$ and step-size $h > 0$ to (2.6) leads to

$$I(h, 1/2) = \frac{xh}{\pi}\,\mathrm{e}^{\mathrm{i}(x^2+\pi/4)} \sum_{k=1}^{\infty} \frac{\mathrm{e}^{-\tau_k^2}}{x^2+\mathrm{i}\tau_k^2}, \quad x > 0, \tag{2.24}$$

where

$$\tau_k := (k - 1/2)h. \tag{2.25}$$

When $x > 0$ is large this approximation is very accurate. However, this approximation becomes increasingly poor as $x \to 0^+$.

Let

$$I^*(h, 1/2) := I(h, 1/2) + C(h, 1/2), \tag{2.26}$$

where $C(h, 1/2)$ and $I(h, 1/2)$ are given by (2.23) and (2.24), respectively, then we have the following result.

Theorem 2.3.1. *Let $e_h^* := F(x) - I^*(h, 1/2)$. Then, for $x > 0$,*

$$|e_h^*| \le \Delta_h(x), \tag{2.27}$$

where

$$\Delta_h(x) := \begin{cases} \delta_1(x), & 0 \le \frac{x}{\sqrt{2}} \le \frac{3}{4}\frac{\pi}{h}, \\ \delta_2(x), & \frac{3}{4}\frac{\pi}{h} < \frac{x}{\sqrt{2}} < \frac{5}{4}\frac{\pi}{h}, \\ \delta_3(x), & \frac{x}{\sqrt{2}} \ge \frac{5}{4}\frac{\pi}{h}. \end{cases} \tag{2.28}$$

Here

$$\delta_1(x) := \frac{x\,\mathrm{e}^{-\pi^2/h^2}}{\sqrt{\pi}\,|\pi^2/h^2 - x^2/2|\,\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}, \tag{2.29}$$

$$\delta_2(x) := \frac{4hx\,\mathrm{e}^{-\pi^2/h^2}}{\sqrt{\pi}\,\pi|\pi/h + x/\sqrt{2}|\,\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi^2/h^2}\right), \tag{2.30}$$

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

$$\delta_3(x) := \delta_1(x) + \frac{\mathrm{e}^{-\sqrt{2}\,\pi x/h}}{1-\mathrm{e}^{-\sqrt{2}\,\pi x/h}}. \tag{2.31}$$

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

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

it holds that

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

gives

$$|e_h^*| \le \frac{x\,\mathrm{e}^{-\pi^2/h^2}}{\sqrt{\pi}\,|\pi^2/h^2 - x^2/2|\,\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}. \tag{2.32}$$

Since, applying (1.8),

$$|C(h,1/2)| \le \frac{\mathrm{e}^{-\sqrt{2}\,\pi x/h}}{1-\mathrm{e}^{-\sqrt{2}\,\pi x/h}}, \quad x > 0,$$

the bound (2.29) also implies that $|e_h^*| \le \delta_3(x)$ for $x > \sqrt{2}\,\pi/h$.

Setting $H = \pi/h$, select ε in the range $(0,H)$ and consider the case that $\left|x/\sqrt{2} - H\right| < \varepsilon$. In this case we observe that the derivation of results of Proposition 1.2.4 can be modified to show that

$$e_h^* = \int_{\Gamma_H^*} f(z)(1+g(z))\,dz, \tag{2.33}$$

where the contour Γ_H^* passes above the pole in f at z_0; precisely, Γ_H^* is the union of Γ' and γ, where $\Gamma' = \{t + iH : t \in \mathbb{R} \text{ and } |(t+iH) - z_0| > \varepsilon\}$ and γ is the circular arc $\gamma = \{z_0 + \varepsilon \mathrm{e}^{\mathrm{i}\theta} : \theta_0 \le \theta \le \pi - \theta_0\}$, where $\theta_0 = \sin^{-1}((H - x/\sqrt{2})/\varepsilon) \in (-\pi/2, \pi/2)$. For $z \in \Gamma'$ it holds that

$$|x^2 + \mathrm{i}z^2| = |z_0 - z|\,|z_0 + z| \ge \varepsilon\,|x/\sqrt{2} + H|. \tag{2.34}$$

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

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

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

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

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

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

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

$$|e_h^*| \le \delta_2(x) := \frac{4hx\,\mathrm{e}^{-\pi^2/h^2}}{\sqrt{\pi}\,\pi|\pi/h+x/\sqrt{2}|\,\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi^2/h^2}\right), \tag{2.37}$$

where

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

Noting that $\beta > 0$ if and only if $0 < \alpha < \alpha_0$ where $\alpha_0 = (1+2\sqrt{2})^{-1} \approx 0.2612$, we choose $\alpha < \alpha_0$ to be $\alpha = 1/4$. With this choice it follows from (2.37) that $|e_h^*| \le \delta_2(x)$ for $\frac{3\pi}{4h} < \frac{x}{\sqrt{2}} < \frac{5\pi}{4h}$, and the proof is complete. □

When the infinite sum in $I^*(h,1/2)$ given by (2.26) is truncated after N terms, this induces the truncation error

$$T_N(h,1/2) := 2h\sum_{m=N+1}^{\infty} f(\tau_m), \quad x > 0, \tag{2.39}$$

where $\tau_m := (m-1/2)h$ and

$$f(t) = \frac{x\,\mathrm{e}^{\mathrm{i}(x^2+\pi/4)}\,\mathrm{e}^{-t^2}}{2\pi(x^2+\mathrm{i}t^2)}.$$

The following proposition bounds $T_N(h,1/2)$.

Proposition 2.3.1. *For $x > 0$,*

$$|T_N(h,1/2)| \leq \frac{(2h\tau_{N+1}+1)x}{2\pi\tau_{N+1}\sqrt{x^4+\tau_{N+1}^4}}\,\mathrm{e}^{-\tau_{N+1}^2}.$$

Proof.

$$\begin{aligned}
|T_N(h,1/2)| &\leq \frac{hx}{\pi}\sum_{m=N+1}^{\infty}\frac{\mathrm{e}^{-\tau_m^2}}{\sqrt{x^4+\tau_m^4}} \\
&\leq \frac{x}{2\pi\sqrt{x^4+\tau_{N+1}^4}}\left(2h\mathrm{e}^{-\tau_{N+1}^2}+2h\sum_{m=N+2}^{\infty}\mathrm{e}^{-\tau_m^2}\right) \\
&\leq \frac{x}{2\pi\sqrt{x^4+\tau_{N+1}^4}}\left(2h\mathrm{e}^{-\tau_{N+1}^2}+2\int_{\tau_{N+1}}^{\infty}\mathrm{e}^{-t^2}dt\right) \\
&\leq \frac{x}{2\pi\sqrt{x^4+\tau_{N+1}^4}}\left(2h\mathrm{e}^{-\tau_{N+1}^2}+\frac{\mathrm{e}^{-\tau_{N+1}^2}}{\tau_{N+1}}\right) = \frac{(2h\tau_{N+1}+1)x}{2\pi\tau_{N+1}\sqrt{x^4+\tau_{N+1}^4}}\,\mathrm{e}^{-\tau_{N+1}^2}.
\end{aligned}$$

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

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

□

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

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

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

$$E_N(x) = F(x) - F_N(x) = e_h^* + T_N(h,1/2)$$

and that

$$|T_N(h,1/2)| \leq \frac{(2\pi+1)x}{2\pi A_N\sqrt{x^4+A_N^4}}\,\mathrm{e}^{-A_N^2}.$$

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

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

where

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

Theorem 2.3.3. *For $x>0$,*

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

where

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

which decreases as N increases, with

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

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

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

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

$$\Delta_h(x) \leq \Delta_h\left(\tfrac{5}{4}\sqrt{2}A_N^-\right) \text{ and } x\Delta_h(x) \leq \tfrac{5}{4}\sqrt{2}A_N\Delta_h\left(\tfrac{5}{4}\sqrt{2}A_N^-\right). \tag{2.46}$$

Moreover,

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

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 $E_N(0) = \eta_N(0) = 0$).*

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*

$$|F(x)| \geq \begin{cases} \dfrac{1}{2+2\sqrt{\pi}x}, & \text{for } x \geq 0 \\ \dfrac{1}{2}, & \text{for } x \leq 0. \end{cases} \tag{2.48}$$

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

$$\frac{|F(x)-F_N(x)|}{|F(x)|} \leq \frac{\eta_N(x)}{|F(x)|} \leq \begin{cases} c_N^* \mathrm{e}^{-\pi N}, & \text{for } x \geq 0, \\ 2c_N \dfrac{\mathrm{e}^{-\pi N}}{\sqrt{N+1/2}}, & \text{for } x \leq 0, \end{cases} \tag{2.49}$$

where

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

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

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

$$\frac{\eta_N(x)}{|F(x)|} \leq \left(2+\tfrac{5}{2}\sqrt{2\pi}A_N\right)\Delta_h\left(\tfrac{5}{4}\sqrt{2}A_N^-\right) + \frac{(2\pi+1)}{2\pi}\frac{\mathrm{e}^{-A_N^2}}{A_N}\left(\frac{1}{\sqrt{2}A_N}+\sqrt{\pi}\right).$$

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

The above estimates use (2.42) and (2.43) to bound the maximum absolute and relative errors in the approximation $F_N(x)$. These inequalities, additionally, imply that $F_N(x)$ is particularly accurate for $|x|$ small. For $|x| \leq A_N/\sqrt{2} = \sqrt{(N+1/2)\pi/2}$, it follows from (2.42) and (2.43) that

$$|F(x) - F_N(x)| \leq \eta(x) \leq \tilde{c}_N |x| \frac{\mathrm{e}^{-\pi N}}{2N+1} \tag{2.50}$$

where

$$\tilde{c}_N = \frac{8}{3\pi^{3/2}\mathrm{e}^{\pi/2}\left(1-\mathrm{e}^{-2A_N^2}\right)} + \frac{(2\pi+1)}{\pi^2 \mathrm{e}^{\pi/2} A_N}, \tag{2.51}$$

which decreases as N increases, with $\tilde{c}_1 \approx 0.17$ and $\lim_{N\to\infty} \tilde{c}_N = 8/(3\pi^{3/2}\mathrm{e}^{\pi/2}) \approx 0.10$.

2.3.1 Extensions of the error bounds

One attractive feature of the modified trapezium rule approximation $I^*(h,1/2)$ given by (2.26) is that, in contrast to $I(h,1/2)$, it is an entire function of $z = x + \mathrm{i}y$. This is not immediately obvious: $I^*(h,1/2) = I(h,1/2) + C(h,1/2)$, and $C(h/1/2)$ has simple pole singularities at $z = \mathrm{e}^{-\mathrm{i}\pi/4}\tau_k$, $k \in \mathbb{Z}$. But $I(h/1/2)$ also has simple poles at the same points and it is an easy calculation to see that the residues add to zero, so that the singularities cancel out. Since $F_N(z) = I^*(h,1/2) - T_N(h,1/2)$, with h given by (2.41), it follows that the singularities of $F_N(z)$ are those of $T_N(h,1/2)$, *i.e.*, simple poles at $\pm\mathrm{e}^{-\mathrm{i}\pi/4}t_k$, for $k = N+1, N+2, \ldots$. Thus $F_N(z)$ is a meromorphic function and, in particular, is analytic in the strip $|\mathrm{Im}(z)| < A_N/\sqrt{2}$ and in the first and third quadrants of the complex plane.

We will note two important consequences of this analyticity and the bounds that we have already proved. In these arguments we will use an extension of the maximum principle for analytic functions to unbounded domains, that if $f(z)$ is analytic in an open quadrant in the complex plane, let us say $Q = \{z \in C : 0 < \arg(z) < \pi/2\}$, and is continuous and bounded in its closure, then

$$\sup_{z\in Q} |f(z)| \leq \sup_{z\in \partial Q} |f(z)|, \tag{2.52}$$

where ∂Q denotes the boundary of the quadrant. (This sort of extension of the maximum principle to unbounded domains is due to Phragmen and Lindelöf; see, *e.g.*, [17].)

The first consequence is that, from (2.3), (2.44), and (2.19), it follows that the bound (2.44) holds on both the real and imaginary axes. Further, from (2.4) and the asymptotics of $\mathrm{erfc}(z)$ in the complex plane [2, (7.1.23)], it follows that $F(z) \to 0$, uniformly in $\arg(z)$, for

$0 \le \arg(z) \le \pi/2$; 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 $0 \le \arg(z) \le \pi/2$, and (2.17) and (2.18) then imply that (2.44) holds also for $\pi \le \arg(z) \le 3\pi/4$.

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.*,

$$|F(z) - I^*(h,1/2)| \le c_N \frac{\mathrm{e}^{-\pi N}}{\sqrt{N+1/2}}, \tag{2.53}$$

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 $E_N(z) = F(z) - F_N(z)$ has poles there. This issue does not hold for $F(z) - I^*(h,1/2)$, which is an entire function, but (2.53) cannot hold in the whole complex plane because this, by Liouville's theorem [17], would imply that $F(z) - I^*(h,1/2)$ is a constant. What does hold is that $\mathrm{e}^{-\mathrm{i}z^2}(F(z) - I^*(h,1/2))$ is bounded in the second and fourth quadrants, this a consequence of the definition of $I^*(h1/2)$ and the asymptotics of $\mathrm{e}^{z^2}\mathrm{erfc}(z)$ at infinity. Thus it follows from (2.52), and since $|\mathrm{e}^{-\mathrm{i}z^2}| = 1$ if z is real or pure imaginary, that

$$|F(z) - I^*(h,1/2)| \le c_N \mathrm{e}^{-xy} \frac{\mathrm{e}^{-\pi N}}{\sqrt{N+1/2}}, \tag{2.54}$$

for $z = x + \mathrm{i}y$ in the second and fourth quadrants.

We can use the bound (2.54) to obtain a bound on $E_N(x)$ in the second and fourth quadrants. Clearly, where $T_N(h,1/2)$ is defined by (2.39), with h given by (2.41), for $z = x + \mathrm{i}y$ in the second and fourth quadrants,

$$|F(z) - F_N(z)| \le c_N \mathrm{e}^{-xy} \frac{\mathrm{e}^{-\pi N}}{\sqrt{N+1/2}} + |T_N(h,1/2)|.$$

Further, arguing as below (2.39), if $|y| \le A_N/(2\sqrt{2})$ so that

$$|z^2 + \mathrm{i}t_k^2| \ge \left(\frac{A_N}{\sqrt{2}} - |y|\right)\left(\left(\frac{A_N}{\sqrt{2}} - |y|\right)^2 + \left(\frac{A_N}{\sqrt{2}} + |x|\right)^2\right) \ge \frac{A_N}{2\sqrt{2}}\left(A_N^2/8 + |x|^2\right),$$

which implies that $|z^2 + \mathrm{i}t_k^2| \ge |z|A_N/(2\sqrt{2})$, then

$$|T_N(h,1/2)| \le \mathrm{e}^{-xy}\frac{(2\pi+1)\sqrt{2}}{\pi A_N^2}\mathrm{e}^{-A_N^2} = \mathrm{e}^{-xy}\frac{\sqrt{2}(2\pi+1)}{\pi^{3/2}\exp(\pi/2)(N+1/2)}\mathrm{e}^{-\pi N}.$$

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

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

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 $a_n = F^{(n)}(0)/n!$, $b_n = F_N^{(n)}(0)/n!$. Thus, where $M_N = \sup_{|z|<\sqrt{\pi/2}} |E_N(z)|$, it follows from Cauchy's estimate [17, Theorem 2.14] and the bounds (2.44) and (2.55) that, for $N \ge 4$ so that $A_N/(2\sqrt{2}) \ge \sqrt{\pi/2}$,

$$|a_n - b_n| = \frac{|E_N^{(n)}(0)|}{n!} \le M_N \left(\frac{2}{\pi}\right)^{n/2} \le \hat{c}_N \left(\frac{2}{\pi}\right)^{n/2} \frac{\mathrm{e}^{-\pi(N-1/2)}}{\sqrt{N+1/2}}. \tag{2.57}$$

2.4 The approximations of $C(x)$ and $S(x)$

From (2.3) we see that, for x real,

$$C(x) = \mathrm{Re}\left(\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}(\tfrac{1}{2} - F(\sqrt{\pi/2}\,x))\right),\ S(x) = \mathrm{Im}\left(\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}(\tfrac{1}{2} - F(\sqrt{\pi/2}\,x))\right). \tag{2.58}$$

Clearly, given the approximation $F_N(x)$ to $F(x)$, these relationships can be used to generate approximations for the Fresnels integrals $C(x)$ and $S(x)$. These approximations are defined, for $x \in \mathbb{R}$, by

$$\begin{aligned} C_N(x) &= \mathrm{Re}\left(\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}(\tfrac{1}{2} - F_N(\sqrt{\pi/2}\,x))\right), \\ S_N(x) &= \mathrm{Im}\left(\sqrt{2}\,\mathrm{e}^{\mathrm{i}\pi/4}(\tfrac{1}{2} - F_N(\sqrt{\pi/2}\,x))\right), \end{aligned} \tag{2.59}$$

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

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

and, recalling that A_N is linked to the quadrature step-size through (2.41), it is clear that, for $x > 0$, $\sqrt{\pi}xa_N\left(\frac{\pi}{2}x^2\right)/A_N$ and $\sqrt{\pi}xb_N\left(\frac{\pi}{2}x^2\right)/A_N$ 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

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

and

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

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 $N \to \infty$, uniformly on the real line, so that very accurate approximations can be obtained with very small values of N ((2.63) shows that both $|C_N(x) - C(x)|$ and $|S_N(x) - S(x)|$ are $\le 1.4 \times 10^{-16}$ on the real line for $N \ge 11$). In §2.5 we will confirm the effectiveness of these approximations by numerical experiments, checking the accuracy of (2.14) and (2.15) by comparison with

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 $F_N(x)$ in the complex plane, discussed in §2.3.1, that $F_N(x)$ has a power series convergent in $|x| < A_N/\sqrt{2}$, and from (2.59) that $C_N(x)$ and $S_N(x)$ have convergent power series representations in $|x| < A_N/\sqrt{\pi}$. From the observations below (2.20) it is clear that, echoing (2.65), these take the form

$$C_N(x) = \sum_{n=0}^{\infty} \mathfrak{c}_n x^{4n+1}, \quad S_N(x) = \sum_{n=0}^{\infty} \mathfrak{s}_n x^{4n+3}. \tag{2.66}$$

Further, it follows from (2.59) and (2.57) that the coefficients $\mathfrak{c}_n$ and $\mathfrak{s}_n$ are close to the corresponding coefficients of $C(x)$ and $S(x)$, with the difference having absolute value

$$\leq \sqrt{2}\,\hat{c}_N \frac{\mathrm{e}^{-\pi(N-1/2)}}{\sqrt{N+1/2}}, \tag{2.67}$$

for $N \geq 4$, where $\hat{c}_N \leq \hat{c}_4 < 0.77$ is the decreasing sequence of positive numbers given by (2.56). This implies that, near zero, where $C(x)$ has a simple zero and $S(x)$ a zero of order three, the approximations $C_N(x)$ and $S_N(x)$ retain small relative error. For $C_N(x)$ this follows already from (2.64) but to see this for $S_N(x)$ we need the stronger bound implied by (2.67) that, for $|x| < 1$,

$$|S(x) - S_N(x)| \leq \sqrt{2}\,\hat{c}_N \frac{\mathrm{e}^{-\pi(N-1/2)}}{\sqrt{N+1/2}} \sum_{n=0}^{\infty} |x|^{4n+3} = \frac{|x|^3}{1-|x|^4} \frac{\sqrt{2}\,\hat{c}_N\,\mathrm{e}^{-\pi(N-1/2)}}{\sqrt{N+1/2}}. \tag{2.68}$$

Listing A.2 shows the Matlab implementing (2.14) and (2.15) that we use in the next section. To evaluate $(\sinh t \pm \sin t)/(\cosh t + \cos t)$, with $t = \sqrt{\pi} A_N x$, in (2.14) and (2.15), we note that, for $|t| \geq 39$, $\cosh(t) + \cos(t)$ and $\exp(t)/2$ have the same value in double precision arithmetic, as do $\sinh t \pm \sin t$ and $\mathrm{sign}(t)\exp(t)/2$. Thus this expression evaluates as $\mathrm{sign}(t)$ in double precision arithmetic for $39 \leq |t| \lessapprox 710$. To avoid underflow and reduce computation time, we evaluate it as $\mathrm{sign}(t)$ for $|t| \geq 39$. For small t there is an additional issue of loss of precision in evaluating $\sinh t - \sin t$ for $|t|$ small. This is avoided in Table A.2 by using $\sinh t - \sin t = 2t^3/3! + 2t^7/7! + \ldots$ for $|t| < 1$, truncating after four terms as the 5th term is negligible in double precision.

2.5 Numerical results

In this section we show the results of numerical computations that confirm and illustrate the theoretical error bounds in §2.3 and §2.4, and that explore the accuracy and efficiency of our new methods, through qualitative and quantitative comparisons with certain of the other computational methods described in §2.3.

In Figure 2.1, we plot the maximum of absolute and relative errors of $F_N(x)$ given by (2.11) and its error bounds (2.44) and (2.49), as a function of N, in comparison with the approximation $F_N^*(x) := \mathrm{e}^{\mathrm{i}x^2} W_N(\mathrm{e}^{\mathrm{i}\pi/4}x)/2$, with $W_N(z)$ given by (2.9). The maximums are taken over 40,000 equally spaced points between 0 and 1,000. It can be seen from Figure 2.1 that:

(i) the exponential convergence predicted by the bounds (2.44) and (2.49) is achieved, indeed these bounds overestimate their respective maximum errors by at most a factor of 10;

(ii) it appears that, with N as small as 12, we achieve maximum absolute and relative errors in $F_N(x)$ which are $< 2.9 \times 10^{-16}$ and $< 9.3 \times 10^{-16}$, respectively;

(iii) the convergence rate of the approximation $F_N^*(x) := \mathrm{e}^{\mathrm{i}x^2} W_N(\mathrm{e}^{\mathrm{i}\pi/4}x)/2$, with $W_N(z)$ given by (2.9), is slower than that of $F_N(x)$;

(iv) the approximation $F_N(x)$ given by (2.11), with $N \leq 14$, is significantly more accurate and more efficient than the approximation $F_N^*(x)$.

Figure 2.1 explores the accuracy of the approximation $F_N(x)$. Let us comment on efficiency. Most straightforward is a comparison of the Matlab function `F(x,N)` in Listing A.1 with computation of $F(x)$ as $F_N^*(x) := \mathrm{e}^{\mathrm{i}x^2} W_N(\mathrm{e}^{\mathrm{i}\pi/4}x)/2$ using `cef.m` from [62] implementing (2.9). Both `F(x,N)` and `cef(x,N)` are optimised for efficiency when `x` is a large vector. The main cost in computation of $F(x)$ via `cef` when x is a large vector is a complex vector exponential (for $\mathrm{e}^{\mathrm{i}x^2}$), and the N complex vector multiplications and N additions required to evaluate the polynomial (2.9) using Horner's algorithm. In comparison, evaluation of $F(x)$ using `F(x,N)` in Table A.1 requires 2 complex vector exponentials, and slightly more than N real vector multiplications/divisions, real vector additions, complex vector multiplications, and complex vector additions. Thus computing $F(x)$ via `F(x,N)` requires a substantially lower operation count than computing via `cef`.

To test whether `F(x,N)` is faster we have compared computation times in Matlab (version 7.8.0.347 (R2009a) on a laptop with dual 2.4GHz P8600 Intel processors) between `exp(i*x.^2).*cef(exp(i*pi/4)*x,N)/2` and `F(x,12)` when `x` is a length 10^7 vector of

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.)

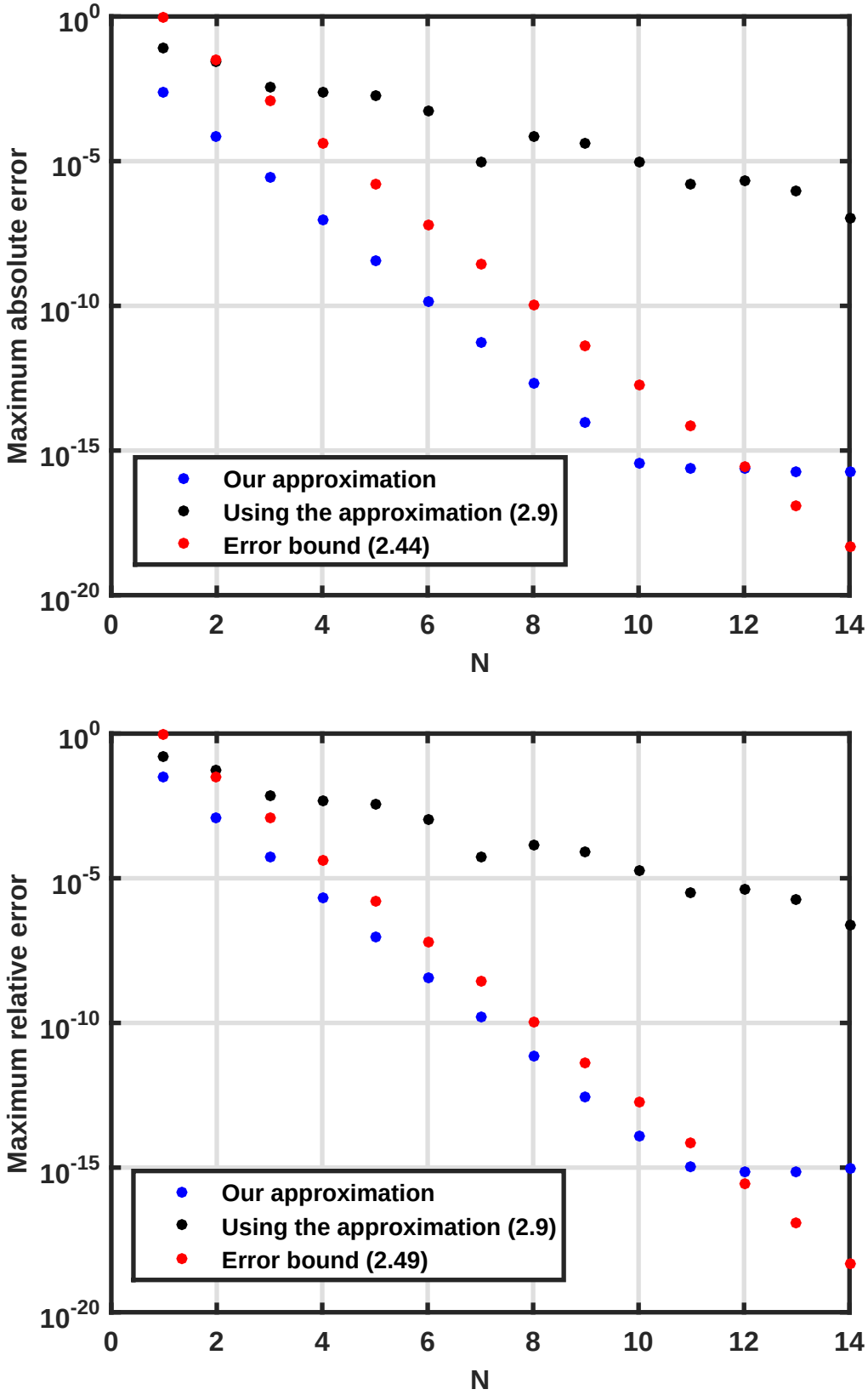

Fig. 2.1 Accuracy of our approximation (2.11) and its error bounds (2.44) and (2.49), as a function of N, in comparison with Weideman's approximation (2.9).

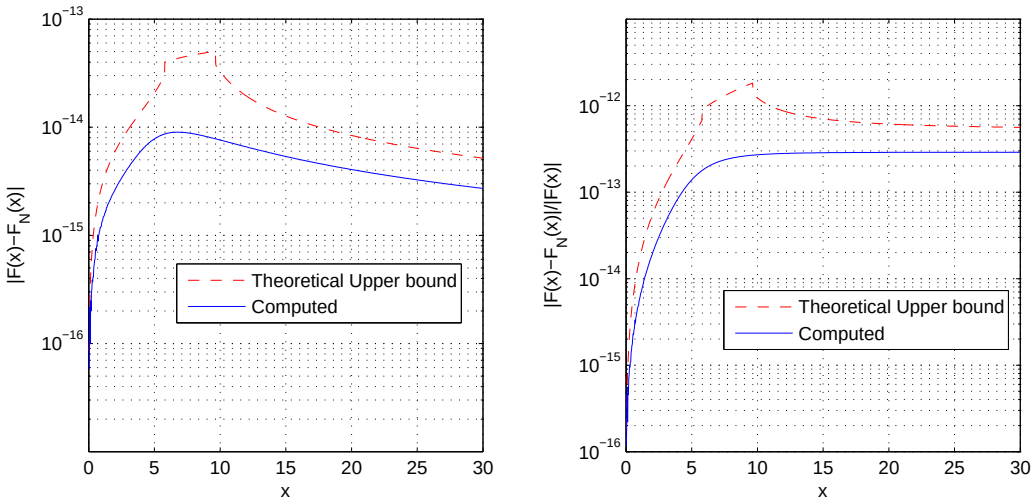

Fig. 2.2 Left hand side: Absolute error, $|F(x)-F_N(x)|$ (−), and its upper bound $\eta_N(x)$ given by (2.42) (−−), plotted against x. Right hand side: Relative error, $|F(x)-F_N(x)|/|F(x)|$ (−), and its upper bound $2(1+\sqrt{\pi}x)\eta_N(x)$ (−−), plotted against x. In both figures $N=9$ and the exact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

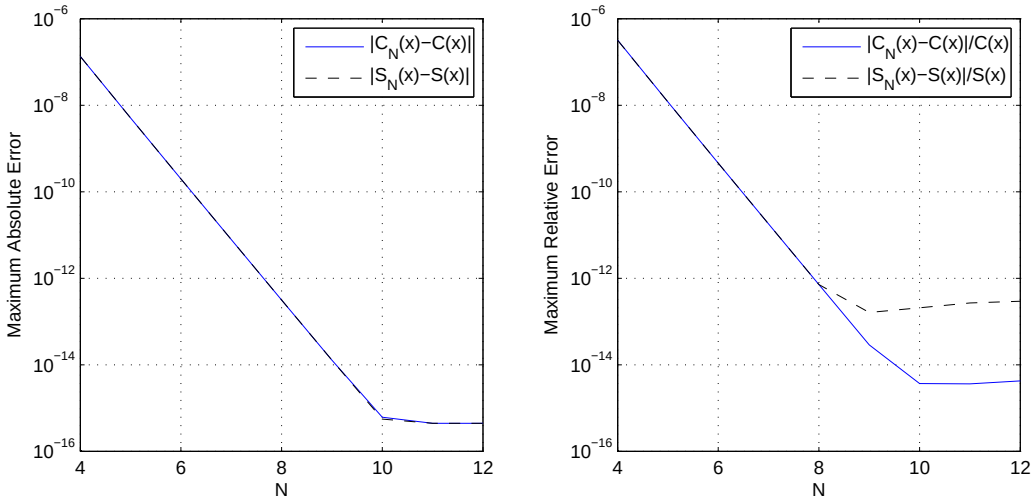

Fig. 2.3 Left hand side: Maximum values of $|C_N(x)-C(x)|$ and $|S_N(x)-S(x)|$ plotted against N on $0 \le x \le 20$. Right hand side: Maximum values of $|C_N(x)-C(x)|/C(x)$ and $|S_N(x)-S(x)|/S(x)$ plotted against N on $0 \le x \le 20$.

Chapter 3

The Faddeeva function

3.1 Introduction

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

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt, \tag{3.1}$$

and the complementary error function is defined by [46, (7.2.2)]

$$\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2}\, dt = 1 - \operatorname{erf}(z), \tag{3.2}$$

where $z = x + iy$. This chapter is concerned with approximating $\operatorname{erfc}(z)$ through approximating the Faddeeva function, denoted by $w(z)$, which is defined as [46, (7.2.3)]

$$w(z) := \mathrm{e}^{-z^2} \operatorname{erfc}(-iz). \tag{3.3}$$

The Faddeeva function, also known as the Plasma Dispersion function [62], is encountered directly in many applications (e.g. [33, 54]) making the development of efficient computational methods of great importance. It is well known [2, (7.1.4)] that $w(z)$ can be represented as

$$w(z) = \frac{\mathrm{i}z}{\pi} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-t^2}}{z^2 - t^2}\, dt, \quad \operatorname{Im}(z) > 0, \tag{3.4}$$

and this representation is the starting point of our approximation. It is sufficient to devise methods to compute $w(z)$ for z in the first quadrant since the values of $w(z)$ in the other

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

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

Chiarella and Reichel [15] and Matta and Reichel [43] first proposed to compute $\mathrm{erfc}(z)$ 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

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

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 $y = \mathrm{Im}(z)$ 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 $\mathrm{Im}(z) > 0$. 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

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

with $M = 2N$, $t_j = L\tan(\theta_j/2)$ and $\theta_j = \pi j/M$ for $j = -M+1,...,M-1$. 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 head-on in this chapter.*

Zagloul and Ali [63] proposed another method of approximating $w(z)$, which forms the basis of the recently published *ACM TOM Algorithm 916*, based on the following representation of $\mathrm{erf}(z)$, with $z = x+iy$, namely

$$\mathrm{erf}(z) = \mathrm{erf}(x) + \frac{2e^{-x^2}}{\sqrt{\pi}}\int_0^y e^{t^2}\sin(2xt)\,dt + \frac{2ie^{-x^2}}{\sqrt{\pi}}\int_0^y e^{t^2}\cos(2xt)\,dt. \tag{3.10}$$

Specifically, they proposed to approximate $w(z)$ for $z = x+iy$ with $x \geq 0$ and $y \geq 0$, by

$$w(z) \approx u(x,y) + iv(x,y), \tag{3.11}$$

where

$$u(x,y) := e^{-x^2}\mathrm{erfcx}(y)\cos(2xy) + \frac{2a\sin^2(xy)}{\pi y}e^{-x^2} + \frac{ay}{\pi}\left(-2\cos(2xy)S_1 + S_2 + S_3\right), \tag{3.12}$$

$$v(x,y) := -e^{-x^2}\mathrm{erfcx}(y)\sin(2xy) + \frac{a\sin(2xy)}{\pi y}e^{-x^2} + \frac{a}{\pi}\left(2y\sin(2xy)S_1 - S_4 + S_5\right), \tag{3.13}$$

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

$$\begin{aligned} S_1 &:= \sum_{k=1}^{\infty} \left(\frac{1}{a^2k^2+y^2} \right) e^{-(a^2k^2+x^2)}, \\ S_2 &:= \sum_{k=1}^{\infty} \left(\frac{1}{a^2k^2+y^2} \right) e^{-(ak+x)^2}, \\ S_3 &:= \sum_{k=1}^{\infty} \left(\frac{1}{a^2k^2+y^2} \right) e^{-(ak-x)^2}, \\ S_4 &:= \sum_{k=1}^{\infty} \left(\frac{ak}{a^2k^2+y^2} \right) e^{-(ak+x)^2}, \\ S_5 &:= \sum_{k=1}^{\infty} \left(\frac{ak}{a^2k^2+y^2} \right) e^{-(ak-x)^2}. \end{aligned} \tag{3.14}$$

The authors have supplied us with their *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 $4 \leq M \leq 13$. 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 the infinite sums (3.14), is more accurate and faster than *ACM TOMS Algorithm 680*. We will explore this further in §3.4.

Abrarov and Quine [1] proposed recently a rational approximation to compute $w(z)$, with $z = x + iy$, based on the integral representation

$$w(z) := \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-t^2} e^{2(ixt-yt)}\, dt, \quad y > 0. \tag{3.15}$$

The obtained approximation is

$$w(z) \approx \psi(z + i\alpha/2), \tag{3.16}$$

where $\alpha \geq 2$,

$$\psi(z) := \sum_{m=1}^{2^{M-1}} \frac{A_m + zB_m}{C_m^2 - z^2}, \quad \mathrm{Im}(z) > 0, \tag{3.17}$$

with

$$C_m := \frac{\pi(2m-1)}{2^{M+1}h}, \tag{3.18}$$

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

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 $0 \le \arg(z) < \pi/4$) 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 $x, y \ge 0$; 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 $x, y \ge 0$, is

$$w_N(z) := \begin{cases} I_N(h,1/2), & y \ge \max(x, \pi/h), \\ I_N^*(h,0), & y < x \quad \text{and} \quad |\phi(x/h) - 1/2| \le 1/4, \\ I_N^*(h,1/2), & \text{otherwise}, \end{cases} \tag{3.21}$$

where ϕ is defined by (3.7),

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

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

$$I_N^*(h,0) := \frac{2\mathrm{e}^{-z^2}}{1-\mathrm{e}^{-2\mathrm{i}\pi z/h}} + \frac{\mathrm{i}h}{\pi z} + \frac{2\mathrm{i}hz}{\pi}\sum_{k=1}^{N}\frac{e^{-\tau_k^2}}{z^2-\tau_k^2}, \tag{3.24}$$

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

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 $x,y \geq 0$, we have*

$$|w(z)-w_N(z))| \leq c_N\,\mathrm{e}^{-\pi N}, \quad \textit{and} \tag{3.26}$$

$$\frac{|w(z)-w_N(z)|}{|w(z)|} \leq c_N^*\sqrt{N+1}\,\mathrm{e}^{-\pi N}, \tag{3.27}$$

where

$$c_N := \frac{25\sqrt{2}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{3\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)} + \frac{10\sqrt{2}(1+2\pi)}{\pi^2\mathrm{e}^{\pi}} \tag{3.28}$$

and

$$c_N^* := \left(\frac{1+\sqrt{2}\,\pi\sqrt{N+1}}{\sqrt{N+1}}\right)c_N, \tag{3.29}$$

with

$$\beta = \frac{11-5\sqrt{2}}{8} \approx 0.4911. \tag{3.30}$$

Further, c_N and c_N^* decrease as N increases with $c_1 \approx 0.58$, $c_1^* \approx 3.01$,

$$\lim_{N\to\infty} c_N = \frac{10\sqrt{2}(1+2\pi)}{\pi^2\,e^{\pi}} \approx 0.45 \quad \textit{and} \quad \lim_{N\to\infty} c_N^* = \frac{20(1+2\pi)}{\pi\,e^{\pi}} \approx 2.0. \tag{3.31}$$

The approximation (3.21) is attractive in at least three respects:

- 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

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

Using (1.16) and Remark 1.2.2, we have

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

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

$$I(h, \alpha) = h \sum_{k \in \mathbb{Z}} \frac{iz\, e^{-(k-\alpha)^2 h^2}}{\pi(z^2 - (k - \alpha)^2 h^2)}. \tag{3.36}$$

Let

$$I^*(h,\alpha) := I(h,\alpha) + C(h,\alpha), \quad \text{for} \quad \alpha = 0, 1/2, \tag{3.37}$$

where $C(h,\alpha)$ and $I(h,\alpha)$ are given by (3.35) and (3.36), respectively. Then we have the following remark.

Remark 3.3.1. *An attractive property of $I^*(h,\alpha)$ as defined by* (3.37) *is that, in contrast to $I(h,\alpha)$ given by* (3.36)*, it is an entire function of z; $I(h,\alpha)$ has simple poles at $z = (k-\alpha)h$, $k \in \mathbb{Z}$, but $C(h,\alpha)$ has simple poles at the same points and it is an easy calculation to show that the residues add to zero so that the singularities are removable.*

Proposition 3.3.1. *Suppose that $I^*(h,\alpha)$ is given by* (3.37)*. Then, for $h > 0$ and $z = x + iy$, with $y \geq x \geq 0$, we have*

$$|w(z) - I^*(h,\alpha)| \leq \Delta_h(y)\, \mathrm{e}^{-\pi^2/h^2}, \tag{3.38}$$

where

$$\Delta_h(y) := \begin{cases} \delta_1(y), & 0 < y \leq \frac{3\pi}{4h} \\ \delta_2(y), & \frac{3\pi}{4h} < y < \frac{5\pi}{4h} \\ \delta_3(y), & y \geq \frac{5\pi}{4h}, \end{cases} \tag{3.39}$$

with

$$\delta_1(y) := \frac{2\sqrt{2}y}{\sqrt{\pi}\left(1 - e^{-2\pi^2/h^2}\right)|\pi^2/h^2 - y^2|}, \tag{3.40}$$

$$\delta_2(y) := \frac{8\sqrt{2}hy\left(1 + 2\sqrt{\pi}\, e^{-\beta\pi^2/h^2}\right)}{\pi^{3/2}(\pi/h + y)\left(1 - e^{-2\pi^2/h^2}\right)}, \tag{3.41}$$

$$\delta_3(y) := \delta_1(y) + \frac{2\, e^{-2\pi y/h}}{1 - e^{-2\pi y/h}}\, e^{y^2 - x^2}, \tag{3.42}$$

and

$$\beta = \frac{11 - 5\sqrt{2}}{8} \approx 0.4911. \tag{3.43}$$

Proof. By Remark 1.2.3 it is easy to show this result for $0 < x \leq y \leq \frac{3}{4}H$ (which then implies the same result for $y = x = 0$ by Remark 3.3.1). For this range of y, using Proposition 1.2.4,

we have

$$|w(z) - I^*(h,\alpha)| \le \frac{2\sqrt{\pi}\, M_H(F)\, e^{H^2 - 2\pi H/h}}{1 - e^{-2\pi H/h}}, \tag{3.44}$$

where F is given by (3.33) and

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

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

$$|z^2 - \zeta^2| = |z-\zeta|\,|z+\zeta| \ge |y-H|\,|y+H| = H^2 - y^2,$$

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

$$|w(z) - I^*(h,\alpha)| \le \delta_1(y) := \frac{2\sqrt{2}y\, e^{-\pi^2/h^2}}{\sqrt{\pi}(\pi^2/h^2 - y^2)\left(1 - e^{-2\pi^2/h^2}\right)}. \tag{3.46}$$

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

$$|w(z) - I^*(h,\alpha)| \le \delta_1(y) + |C(h,\alpha)| \le \delta_3(y). \tag{3.47}$$

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

$$w(z) - I^*(h,\alpha) = \int_{C_H} f(\zeta)(1 - g(\zeta))\, d\zeta, \tag{3.48}$$

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

For $\zeta \in C_H^*$, it holds that

$$|z^2 - \zeta^2| = |z-\zeta|\,|z+\zeta| \ge \varepsilon\,|y+H|. \tag{3.49}$$

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

$$\left| \int_{C_H^*} f(\zeta)(1 - g(\zeta))\, d\zeta \right| \le \frac{2\sqrt{2}y\, e^{-\pi^2/h^2}}{\sqrt{\pi}\,\varepsilon(\pi/h + y)\left(1 - e^{-2\pi^2/h^2}\right)}. \tag{3.50}$$

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

$$|e^{-\zeta^2}| = \mathrm{e}^P,$$

where

$$\begin{aligned} P &= Y^2 - X^2 \\ &= y^2 - x^2 - \varepsilon^2\cos(2\theta) + 2\varepsilon\sqrt{y^2+x^2}\sin(\theta - \tan^{-1}(y/x)) \\ &< y^2 + \varepsilon^2 + 2\sqrt{2}\,\varepsilon\, y \\ &< (2\sqrt{2}+2)\varepsilon^2 + 2\varepsilon(1+\sqrt{2})H + H^2, \text{ since } |y-H| < \varepsilon. \end{aligned}$$

From these bounds we deduce, for $a = \varepsilon/H \in (0,1)$ and $H = \pi/h$, that

$$\left| \int_\gamma f(\zeta)(1-g(\zeta))\,d\zeta \right| \leq \frac{4|z|\,\exp[((2+2\sqrt{2})a^2 + (2+2\sqrt{2})a - 1)\pi^2/h^2]}{\varepsilon|\pi/h + y|\left(1 - e^{-2\pi^2/h^2}\right)}. \tag{3.51}$$

Combining (3.50) and (3.51), and using the triangle inequality, will give

$$|w(z) - I^*(h,\alpha)| \leq \frac{2h|z|\left(1 + 2\sqrt{\pi}\,e^{-\beta\pi^2/h^2}\right)e^{-\pi^2/h^2}}{a\sqrt{\pi}\,\pi|\pi/h + y|\left(1 - e^{-2\pi^2/h^2}\right)}, \tag{3.52}$$

where $\beta = 2 - (2+2\sqrt{2})a - (2+2\sqrt{2})a^2$. Note that $\beta > 0$ if and only if $0 < a < a_0$, where $a_0 = -\dfrac{(\sqrt{2}-1)(1+\sqrt{2}-\sqrt{7+6\sqrt{2}})}{2} \approx 0.31499$. The result follows by choosing $a = 1/4$. □

We show in [3, Theorem 6] a lower bound for the complementary error function $\operatorname{erfc}(z)$ which can be rewritten using (3.3) as

$$|w(z)| \geq \frac{1}{1 + \sqrt{\pi}|z|}, \quad \operatorname{Im}(z) \geq 0. \tag{3.53}$$

This is a sharp bound since $w(0) = 1$ and $w(z) \sim \dfrac{i}{\sqrt{\pi}z}$ as $z \to \infty$ (see [22, (2.6)]). We will use this bound in the following propositions.

Proposition 3.3.2. *Suppose that $I^*(h,\alpha)$ is given by* (3.37)*. Then, for $h > 0$ and $z = x + iy$ with $0 \leq x \leq y < \pi/h$, we have*

$$|w(z) - I^*(h,\alpha)| \leq \Delta_h\left(\frac{\pi}{h}\right)e^{-\pi^2/h^2} \text{ and} \tag{3.54}$$

$$\frac{|w(z) - I^*(h,\alpha)|}{|w(z)|} \leq \left(1 + \frac{\sqrt{2}\,\pi^{3/2}}{h}\right)|w(z) - I^*(h,\alpha)|, \tag{3.55}$$

where

$$\Delta_h\left(\frac{\pi}{h}\right)=\frac{4\sqrt{2}\,h\left(1+2\sqrt{\pi}\mathrm{e}^{-\beta\pi^2/h^2}\right)}{\pi^{3/2}\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}, \tag{3.56}$$

and β *is given by* (3.43).

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

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

Also we have, using (3.53), that

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

and the two results follow. □

In the following proposition we bound $|w(z)-I(h,\alpha)|$ and $|w(z)-I(h,\alpha)|/|w(z)|$.

Proposition 3.3.3. *Suppose that* $I(h,\alpha)$ *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*

$$|w(z)-I(h,\alpha)| \le \left(\Delta_h\left(\frac{5\pi^-}{4h}\right)+\frac{2\,e^{1/4}}{1-e^{-2\pi^2/h^2}}\right)e^{-\pi^2/h^2}, \tag{3.59}$$

$$\frac{|w(z)-I(h,\alpha)|}{|w(z)|} \le \left(1+\frac{5\sqrt{2}\pi^{3/2}}{4h}\right)|w(z)-I(h,\alpha)|, \tag{3.60}$$

where $\Delta_h\left(\frac{5\pi^-}{4h}\right)$ *is the limiting value of* $\Delta_h(y)$ *as* y *approaches* $\frac{5\pi}{4h}$ *from below, given by*

$$\Delta_h\left(\frac{5}{4}\frac{\pi^-}{h}\right)=\frac{40\sqrt{2}\,h\left(1+2\sqrt{\pi}\mathrm{e}^{-\beta\pi^2/h^2}\right)}{9\pi^{3/2}\left(1-\mathrm{e}^{-2\pi^2/h^2}\right)}. \tag{3.61}$$

Proof. Let $H=\pi/h$ and $0<\varepsilon<H/4$. Then, we need to consider the two cases when $y\ge\max(x,H+\varepsilon)$ or $\max(x,H)\le y<H+\varepsilon$.

For $y\ge\max(x,H+\varepsilon)$, we have, using Proposition 1.2.4, that

$$|w(z)-I(h,\alpha)|\le\frac{2\sqrt{\pi}M_H(F)\,e^{-H^2}}{1-e^{-2H^2}}, \tag{3.62}$$

where

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

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

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

Thus, we have

$$|w(z)-I(h,\alpha)| \le \frac{5\sqrt{2}}{4\sqrt{\pi}\,\varepsilon(1-\mathrm{e}^{-2\pi^2/h^2})}\, e^{-\pi^2/h^2}, \tag{3.65}$$

and, using (3.53) and (3.64),

$$\begin{aligned} \frac{|w(z)-I(h,\alpha)|}{|w(z)|} &\le (1+\sqrt{2\pi}\,y)|w(z)-I(h,\alpha)| \\ &\le \left(\frac{5\sqrt{2}}{4\sqrt{\pi}\,\varepsilon}+\frac{25\pi}{8\varepsilon\,h}\right)\frac{e^{-\pi^2/h^2}}{1-\mathrm{e}^{-2\pi^2/h^2}}. \end{aligned} \tag{3.66}$$

We consider now the case when $\max(x,H) \le y < H+\varepsilon$. Using the bound in (3.35) and Proposition 3.3.1, we find that

$$\begin{aligned} |w(z)-I(h,\alpha)| &\le |w(z)-I^*(h,\alpha)| + |C(h,\alpha)| \\ &\le \Delta_h(y)\, e^{-\pi^2/h^2} + \frac{2\,e^{y^2-2\pi y/h}}{1-e^{-2\pi y/h}} \end{aligned}$$

Since $\Delta_h(y)$ given by (3.39) and $\dfrac{2\,e^{y^2-2\pi y/h}}{1-e^{-2\pi y/h}}$ are both increasing functions of y for $H \le y < H+\varepsilon$, we have that

$$\begin{aligned} |w(z)-I(h,\alpha)| &\le \Delta_h\left(\frac{5\pi}{4h}^-\right) e^{-\pi^2/h^2} + \frac{2\,e^{-\pi^2/h^2}\,e^{\varepsilon^2}}{1-e^{-2\pi^2/h^2-2\varepsilon\pi/h}} \\ &\le \Delta_h\left(\frac{5\pi}{4h}^-\right) e^{-\pi^2/h^2} + \frac{2\,e^{-\pi^2/h^2}\,e^{\varepsilon^2}}{1-e^{-2\pi^2/h^2}}. \end{aligned}$$

Choosing $\varepsilon = 1/2$, in which case $e^{\varepsilon^2} < 2$, gives that

$$|w(z)-I(h,\alpha)| \le \left(\Delta_h\left(\frac{5\pi}{4h}^-\right) + \frac{2\,e^{1/4}}{1-e^{-2\pi^2/h^2}}\right) e^{-\pi^2/h^2}. \tag{3.67}$$

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

$$\begin{aligned}\frac{|w(z)-I(h,\alpha)|}{|w(z)|} &\le (1+\sqrt{2\pi}y)\,(|w(z)-I^*(h,\alpha)|+|C(h,\alpha)|)\\ &\le \left(1+\frac{5\sqrt{2}\pi^{3/2}}{4h}\right)\left(\Delta_h\left(\frac{5\pi}{4h}^-\right)+\frac{2\,e^{1/4}}{1-e^{-2\pi^2/h^2}}\right)e^{-\pi^2/h^2}.\end{aligned} \tag{3.68}$$

Further, with $\varepsilon = 1/2$ and noting $5/\sqrt{2\pi} < 2e^{1/4}$, we can show that the bound (3.67) is greater than the bound (3.65) and the first result follows.

Similarly, with $\varepsilon = 1/2$ and noting

$$\frac{5}{\sqrt{2\pi}}+\frac{25\pi}{4h} < 2e^{1/4}\left(1+\frac{5\sqrt{2}\pi^{3/2}}{4h}\right), \quad \text{for} \quad h>0, \tag{3.69}$$

we see that the bound (3.68) is greater than the bound (3.66) and the second result follows. □

The following extension of the maximum modulus principle to unbounded domains [17, Corollary 4.2] will be used to prove bounds for $|w(z)-I^*(h,\alpha)|$ and $|w(z)-I^*(h,\alpha)|/|w(z)|$, with z in the lower half of the first quadrant.

Lemma 3.3.1. *Let $a \ge 1/2$ and put*

$$G := \left\{z \in \mathrm{C} : |\arg(z)| < \frac{\pi}{2a}\right\}.$$

Suppose that f is analytic on G and continuous in $\overline{G}$ and that there is a constant M such that $|f(z)| \le M$ for all $z \in \partial G$. If there are positive constants P and $b < a$ such that

$$|f(z)| \le P\,e^{|z|^b} \tag{3.70}$$

for all z with $|z|$ sufficiently large, then $|f(z)| \le M$ for all $z \in \overline{G}$.

Proposition 3.3.4. *Suppose that $I^*(h,\alpha)$ is given by* (3.37)*. Then, for $h>0$ and $z = x+\mathrm{i}y$ with $0 \le y < x$, we have*

$$|w(z)-I^*(h,\alpha)| \le \Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right)\mathrm{e}^{-\pi^2/h^2} \text{ and} \tag{3.71}$$

$$\frac{|w(z)-I^*(h,\alpha)|}{|w(z)|} \le 2\left(1+\frac{5\sqrt{2}\,\pi^{3/2}}{4h}\right)\Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right)e^{-\pi^2/h^2}, \tag{3.72}$$

where $\Delta_h\left(\frac{5\pi}{4h}^-\right)$ *is given by* (3.61).

Proof. Define

$$E_h(z) = w(z) - I^*(h,\alpha) \quad \text{and} \quad e_h(z) = E_h(z)/w(z),$$

on $G := \{z \in \mathbb{C} : 0 < \arg(z) < \pi/4\}$. Since $w(z)$ and $I^*(h,\alpha)$ 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 $I^*(h,\alpha)$ 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 $z = r\mathrm{e}^{\mathrm{i}\pi/4}$ with $r \geq 0$. Then, using Proposition 3.3.1, we have that $\Delta_h(y)$ given by (3.39) 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 $\Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right) > \Delta_h\left(\frac{5}{4}\frac{\pi}{h}\right)$; thus we have

$$|E_h(z)| \leq \Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right) \mathrm{e}^{-\pi^2/h^2}. \tag{3.73}$$

Let $z = x + i\varepsilon$ with $0 < \varepsilon < \pi/h$. Then, using Proposition 1.2.4,

$$|E_h(z)| \leq \frac{2|z|\, e^{-\pi^2/h^2}}{\pi(1 - e^{-2\pi^2/h^2})} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{|z^2 - (t + i\pi/h)^2|}\, dt.$$

Taking the limit $\varepsilon \to 0^+$, since $E_h(z)$ is continuous, we obtain

$$|E_h(x)| \leq \frac{2x\, e^{-\pi^2/h^2}}{\pi(1 - e^{-2\pi^2/h^2})} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{|x^2 - (t + i\pi/h)^2|}\, dt, \quad x \geq 0.$$

Note, for $x \geq 0$,

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

and

$$\frac{x}{\sqrt{x^2 + \pi^2/h^2}} \leq 1. \tag{3.74}$$

Thus, we have

$$|E_h(x)| \leq \frac{2h\, e^{-\pi^2/h^2}}{\pi^{3/2}(1 - e^{-2\pi^2/h^2})} \leq \Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right) \mathrm{e}^{-\pi^2/h^2}, \quad x \geq 0,$$

and the first bound (3.71) follows.

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

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

where $P := M\Delta_h\left(\frac{5\pi}{4h}^-\right)\mathrm{e}^{-\pi^2/h^2}$ and $M := \max(1+\sqrt{\pi}|z|)\mathrm{e}^{-|z|}$, 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 $z = r\mathrm{e}^{\mathrm{i}\pi/4}$ with $r \ge 0$. Then, we have, using Proposition 3.3.1, that $y\Delta_h(y)$ 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 $\Delta_h\left(\frac{5}{4}\frac{\pi}{h}^-\right) > \Delta_h\left(\frac{5}{4}\frac{\pi}{h}\right)$; thus we have

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

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

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

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

and

$$\int_{-\infty}^{-x/2} G(t)\,dt \le \frac{h}{\pi\sqrt{x^2+\pi^2/h^2}} \int_{x/2}^{\infty} e^{-t^2}\,dt \le \frac{h\,e^{-x^2/4}}{\pi x\sqrt{x^2+\pi^2/h^2}}, \tag{3.81}$$

To arrive at the last lines in (3.80) and (3.81) we have used that, for $a>0$,

$$2\int_a^{\infty} \mathrm{e}^{-t^2} dt = \frac{e^{-a^2}}{a} - \int_a^{\infty} \frac{e^{-t^2}}{t^2} dt < \frac{e^{-a^2}}{a}. \tag{3.82}$$

Additionally, we have, for $-x/2 \le t \le x/2$, that

$$|x^2-(t+i\pi/h)^2| = |x-t-i(\pi/h)||x+t+i(\pi/h)| \ge \frac{x^2+4(\pi/h)^2}{4}.$$

Thus we have

$$\int_{-x/2}^{x/2} G(t)\,dt \le \frac{4}{x^2+4(\pi/h)^2} \int_{-x/2}^{x/2} e^{-t^2}\,dt \le \frac{4\sqrt{\pi}}{x^2+4(\pi/h)^2}. \tag{3.83}$$

Moreover,

$$\frac{1}{\sqrt{x^2+(\pi/h)^2}} \le \frac{h}{\pi}, \quad \frac{x^2}{x^2+4(\pi/h)^2} \le 1, \quad \frac{x}{x^2+4(\pi/h)^2} \le \frac{h}{4\pi}. \tag{3.84}$$

Combining (3.77), (3.74), (3.79)–(3.84) we find for $x \ge 0$, that

$$|e_h(x)| \le \frac{2(h+4\sqrt{2}\pi^{3/2})(8.5h+3\pi^{3/2})}{3\pi^3(1-e^{-2\pi^2/h^2})}\, e^{-\pi^2/h^2}. \tag{3.85}$$

Further, we can show that

$$\frac{2(h+4\sqrt{2}\pi^{3/2})(8.5h+3\pi^{3/2})}{3\pi^3(1-e^{-2\pi^2/h^2})} \le 2\left(1+\frac{5\sqrt{2}\,\pi^{3/2}}{4h}\right)\Delta_h\left(\frac{5}{4}\frac{\pi}{h}^{-}\right),$$

and the second result (3.72) follows. □

In applications, the approximation $I^*(h,\alpha)$ in (3.37) is truncated after N terms and the resulting approximation formula, denoted by $I_N^*(h,\alpha)$, will be

$$I_N^*(h,\alpha) := I_N(h,\alpha) + C(h,\alpha), \tag{3.86}$$

where $C(h,\alpha)$ is given by (3.35) and

$$I_N(h,\alpha) := \begin{cases} hf(0) + 2h\displaystyle\sum_{k=1}^{N} f(\tau_k), & \alpha = 0 \\ 2h\displaystyle\sum_{k=0}^{N} f(t_k), & \alpha = 1/2, \end{cases} \tag{3.87}$$

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

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

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

Proposition 3.3.5. *Suppose τ_k is given by* (3.25) *and $|z-\tau_k| \geq h/4$ for $k = N+1, N+2, \ldots$ and $z = x+iy$ with $0 \leq y < x$. Then, for $h > 0$,*

$$|T_N(h,0)| \leq \frac{2\sqrt{2}\,(1+2h\tau_{N+1})\,(h+4\tau_{N+1})}{\pi h\,\tau_{N+1}^2}\, e^{-\tau_{N+1}^2}, \quad \textit{and} \tag{3.89}$$

$$\frac{|T_N(h,0)|}{|w(z)|} \leq (1+\sqrt{2\pi}\tau_{N+1})\,|T_N(h,0)|. \tag{3.90}$$

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

$$\begin{aligned} |T_N(h,0)| &\leq 2h \sum_{k=N+1}^{\infty} \frac{|z| e^{-\tau_k^2}}{\pi |z+\tau_k||z-\tau_k|} \\ &\leq \frac{\sqrt{2}x}{\pi\sqrt{x^2+\tau_{N+1}^2}} \left(2h \sum_{k=N+1}^{\infty} \frac{e^{-\tau_k^2}}{|z-\tau_k|} \right) \\ &= \frac{\sqrt{2}x}{\pi\sqrt{x^2+\tau_{N+1}^2}} \left(2h \sum_{k=N+1}^{M-1} \frac{e^{-\tau_k^2}}{|z-\tau_k|} + 2h \sum_{k=M}^{\infty} \frac{e^{-\tau_k^2}}{|z-\tau_k|} \right), \end{aligned} \tag{3.91}$$

where M is the smallest integer $\geq N+1$ such that $\tau_M > \theta x$.

For the first summation we have

$$\begin{aligned} 2h \sum_{k=N+1}^{M-1} \frac{e^{-t_k^2}}{|z-\tau_k|} &\leq \frac{1}{(1-\theta)x} \left(2h \sum_{k=N+1}^{\infty} e^{-\tau_k^2} \right) \\ &\leq \frac{1}{(1-\theta)x} \left(2h e^{-\tau_{N+1}^2} + 2h \sum_{k=N+2}^{\infty} e^{-\tau_k^2} \right) \\ &\leq \frac{1}{(1-\theta)x} \left(2h e^{-\tau_{N+1}^2} + 2 \int_{\tau_{N+1}}^{\infty} e^{-t^2}\, dt \right) \\ &\leq \frac{1}{(1-\theta)x} \left(\frac{1+2h\tau_{N+1}}{\tau_{N+1}} \right) e^{-\tau_{N+1}^2}. \end{aligned} \tag{3.92}$$

To arrive at the last line we have used the property (3.82).

For the second summation we have that

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

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

$$2h\sum_{k=M}^{\infty}\frac{e^{-\tau_k^2}}{|z-\tau_k|} \leq \frac{4}{h}\left(\frac{1+2h\tau_{N+1}}{\theta x}\right)e^{-\tau_{N+1}^2}. \tag{3.93}$$

We have, using $\dfrac{1}{\sqrt{x^2+\tau_{N+1}^2}} \leq \dfrac{1}{\tau_{N+1}}$ and (3.91), (3.92) and (3.93), that

$$|T_N(h,0)| \leq \frac{\sqrt{2}(1+2h\tau_{N+1})}{\pi\tau_{N+1}}\left[\frac{1}{(1-\theta)\,\tau_{N+1}}+\frac{4}{h\theta}\right]e^{-\tau_{N+1}^2}.$$

Choose θ such that

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

i.e.

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

Then we have that

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

Similarly, we have, using $\dfrac{x}{\sqrt{x^2+\tau_{N+1}^2}} \leq 1$ and (3.91), (3.92) and (3.93), that

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

□

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* $|z-t_k| \geq h/4$ *for* $k = N+1, N+2, \ldots$ *and* $z = x+iy$ *with* $0 \leq y < x$ *. Then, for* $h > 0$*,*

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

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

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

$$|T_N(h,1/2)| \leq \frac{\sqrt{2}x}{\pi\sqrt{x^2+t_{N+1}^2}} \left(2h \sum_{k=N+1}^{M^*-1} \frac{e^{-t_k^2}}{|z-t_k|} + 2h \sum_{k=M^*}^{\infty} \frac{e^{-t_k^2}}{|z-t_k|} \right), \tag{3.98}$$

where M^* is the smallest integer $\geq N+1$ such that $t_{M^*} > \theta^* x$.

For the first summation we have (similar to (3.92) in Proposition 3.3.5)

$$2h \sum_{k=N+1}^{M^*-1} \frac{e^{-t_k^2}}{|z-t_k|} \leq \frac{1}{(1-\theta^*)x} \left(\frac{1+2ht_{N+1}}{t_{N+1}} \right) e^{-t_{N+1}^2}. \tag{3.99}$$

For the second summation we can easily show (similar to Proposition 3.3.5) that

$$2h \sum_{k=M^*}^{\infty} \frac{e^{-t_k^2}}{|z-t_k|} \leq \frac{4}{h} \left(\frac{1+2ht_{M^*}}{\theta^* x} \right) e^{-t_{M^*}^2}.$$

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

$$2h \sum_{k=M^*}^{\infty} \frac{e^{-t_k^2}}{|z-t_k|} \leq \frac{4}{h} \left(\frac{1+2ht_{N+1}}{\theta^* x} \right) e^{-t_{N+1}^2}. \tag{3.100}$$

We have, using $\dfrac{1}{\sqrt{x^2+t_{N+1}^2}} \leq \dfrac{1}{t_{N+1}}$ and (3.98), (3.99) and (3.100), that

$$|T_N(h,1/2)| \leq \frac{\sqrt{2}(1+2ht_{N+1})}{\pi t_{N+1}} \left[\frac{1}{(1-\theta^*)\,t_{N+1}} + \frac{4}{h\theta^*} \right] e^{-t_{N+1}^2}.$$

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

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

Similarly, we have, using $\dfrac{x}{\sqrt{x^2+t_{N+1}^2}} \leq 1$ and (3.98), (3.99) and (3.100), that

$$x|T_N(h,1/2)| \leq \frac{2\sqrt{2}\,(1+2ht_{N+1})\,(h+4t_{N+1})}{\pi h t_{N+1}}\,e^{-t_{N+1}^2}. \tag{3.102}$$

□

We choose the step-size h such that the exponents of $e^{-\pi^2/h^2}$ and $e^{-\tau_{N+1}^2}$ are equal, where $\tau_{N+1} = (N+1)h$, giving

$$h := \sqrt{\frac{\pi}{N+1}}. \tag{3.103}$$

Remark 3.3.2. *We can rewrite, for $h = \sqrt{\pi/(N+1)}$, the bounds* (3.95) *and* (3.102) *as*

$$x|T_N(h,0)| \leq G_1(N) \quad \textit{and} \quad x|T_N(h,1/2)| \leq G_2(N). \tag{3.104}$$

where

$$G_1(N) := \frac{2\sqrt{2\pi}\,(1+2\pi)\,(4N+5)}{\pi^2\,e^{\pi}\sqrt{N+1}}\,e^{-\pi N}, \tag{3.105}$$

and

$$G_2(N) := \frac{2\sqrt{2\pi}\,(4N+7)\,((2\pi+1)N+3\pi+1)}{\pi^2(N+3/2)\,\sqrt{N+1}}\,e^{-\pi(N+3/2)^2/(N+1)}. \tag{3.106}$$

We can easily show, for $N \geq 1$, that

$$\frac{G_1(N)}{G_2(N)} = \frac{(1+2\pi)(4N+5)(N+3/2)\,e^{\pi/(N+1)}}{e^{\pi}(4N+7)\,((2\pi+1)N+3\pi+1)}\,e^{2\pi} \geq 1. \tag{3.107}$$

Hence, the bounds in Proposition 3.3.5 are also bounds for $|T_N(h,1/2)|$ and $|T_N(h,1/2)|/|w(z)|$.

The following result gives a bound for $T(h,\alpha)$ in the upper half of the first quadrant.

Proposition 3.3.7. *Suppose* $\alpha = 0$ *or* $\alpha = 1/2$ *and* $z = x + \mathrm{i}y$ *with* $y \geq x \geq 0$. *Then, for* $h > 0$,

$$|T_N(h,\alpha)| \leq \frac{(1+2h\,\tau_{N+1})}{\pi\tau_{N+1}^2}\,\mathrm{e}^{-\tau_{N+1}^2}, \quad \textit{and} \tag{3.108}$$

$$\frac{|T_N(h,\alpha)|}{|w(z)|} \leq \frac{(1+2h\,\tau_{N+1})(1+2\sqrt{\pi}\,\tau_{N+1})}{\pi\tau_{N+1}^2}\,\mathrm{e}^{-\tau_{N+1}^2}. \tag{3.109}$$

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

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

Thus, we have

$$|T_N(h,\alpha)| \leq 2h \sum_{k=N+1}^{\infty} e^{-\tau_k^2}|F(\tau_k)|,$$

and, using (3.53),

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

Since

$$|z^2 - \tau_k^2|^2 = y^4 + \tau_k^4 + x^4 + 2x^2y^2 + 2\tau_k^2(y^2 - x^2) \geq y^4 + \tau_k^4,$$

$$\begin{aligned}|T_N(h,\alpha)| &\leq \frac{2\sqrt{2}hy}{\pi}\sum_{k=N+1}^{\infty}\frac{\mathrm{e}^{-\tau_k^2}}{\sqrt{y^4+\tau_k^4}}\\ &\leq \frac{\sqrt{2}y}{\pi\sqrt{y^4+\tau_{N+1}^4}}\left(2h\mathrm{e}^{-\tau_{N+1}^2} + 2\int_{\tau_{N+1}}^{\infty}\mathrm{e}^{-t^2}dt\right)\\ &\leq \frac{\sqrt{2}y\,(1+2h\,\tau_{N+1})}{\pi\tau_{N+1}\sqrt{y^4+\tau_{N+1}^4}}\,\mathrm{e}^{-\tau_{N+1}^2}.\end{aligned}$$

Moreover

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

Thus we have that

$$|T_N(h,\alpha)| \leq \frac{(1+2h\,\tau_{N+1})}{\pi\,\tau_{N+1}^2}\,\mathrm{e}^{-\tau_{N+1}^2},$$

and

$$y\,|T_N(h,\alpha)| \leq \frac{\sqrt{2}(1+2h\,\tau_{N+1})}{\pi\,\tau_{N+1}}\,\mathrm{e}^{-\tau_{N+1}^2}.$$

□

The following two results give bounds for the absolute and relative errors of approximating $w(z)$ by $I_N(h,\alpha)$ given by (3.87).

Theorem 3.3.2. *Suppose that $I_N^*(h,\alpha)$ is given by* (3.86) *and $h=\sqrt{\pi/(N+1)}$. Then, for $z=x+\mathrm{i}y$ with $0\leq y<x$, we have*

$$|w(z)-I_N^*(h,\alpha)| \;\leq\; c_N\,\mathrm{e}^{-\pi N}, \quad \textit{and} \tag{3.110}$$

$$\frac{|w(z)-I_N^*(h,\alpha)|}{|w(z)|} \;\leq\; c_N^*\,\sqrt{N+1}\,\mathrm{e}^{-\pi N}, \tag{3.111}$$

where

$$c_N := \frac{100\sqrt{2}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)} + \frac{10\sqrt{2}\,(1+2\pi)}{\pi^2\mathrm{e}^{\pi}} \tag{3.112}$$

and

$$c_N^* := \left(\frac{1+\sqrt{2}\,\pi\sqrt{N+1}}{\sqrt{N+1}}\right) c_N, \tag{3.113}$$

with β is given by (3.43). *Further, c_N and c_N^* decrease as N increases with $c_1\approx 0.63$, $c_1^*\approx 3.24$,*

$$\lim_{N\to\infty} c_N = \frac{10\sqrt{2}(1+2\pi)}{\pi^2\,e^{\pi}} \approx 0.45 \quad \textit{and} \quad \lim_{N\to\infty} c_N^* = \frac{20(1+2\pi)}{\pi\,e^{\pi}} \approx 2.0. \tag{3.114}$$

Proof. Applying Proposition 3.3.4 for $h=\sqrt{\pi/(N+1)}$ yields that

$$\begin{aligned}|w(z)-I^*(h,\alpha)| &\leq \left(\frac{40\sqrt{2}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}\right)\mathrm{e}^{-\pi N}\\ &\leq \left(\frac{100\sqrt{2}\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}\right)\mathrm{e}^{-\pi N},\end{aligned} \tag{3.115}$$

and

$$\begin{aligned}\frac{|w(z)-I^*(h,\alpha)|}{|w(z)|} &\leq \frac{20\sqrt{2}\,(4+5\sqrt{2}\pi\sqrt{N+1})\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}\,\mathrm{e}^{-\pi N}\\ &\leq \frac{100\sqrt{2}\,(1+\sqrt{2}\pi\sqrt{N+1})\left(1+2\sqrt{\pi}\,\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi\mathrm{e}^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}\,\mathrm{e}^{-\pi N}.\end{aligned} \tag{3.116}$$

Similarly, we find, using Proposition 3.3.5 for $h=\sqrt{\pi/(N+1)}$, that

$$|T_N(h,\alpha)| \leq \frac{2\sqrt{2}\,(1+2\pi)\,(5+4N)}{\pi^2\mathrm{e}^{\pi}(N+1)}\,\mathrm{e}^{-\pi N} \tag{3.117}$$

$$\leq \frac{10\sqrt{2}\,(1+2\pi)}{\pi^2\mathrm{e}^{\pi}}\,\mathrm{e}^{-\pi N}, \tag{3.118}$$

and

$$\frac{|T_N(h,\alpha)|}{|w(z)|} \leq \frac{10\sqrt{2}\left(1+\sqrt{2}\pi\sqrt{N+1}\right)(1+2\pi)}{\pi^2\mathrm{e}^{\pi}}\,\mathrm{e}^{-\pi N}. \tag{3.119}$$

Since

$$|w(z)-I_N^*(h,\alpha)| \leq |w(z)-I^*(h,\alpha)|+|T_N(h,\alpha)|,$$

the first bound follows by combining (3.115) and (3.118). Similarly and since

$$\frac{|w(z)-I_N^*(h,\alpha)|}{|w(z)|} \leq \frac{|w(z)-I^*(h,\alpha)|}{|w(z)|}+\frac{|T_N(h,\alpha)|}{|w(z)|},$$

the second bound follows by combining (3.116) and (3.119). □

Theorem 3.3.3. *Suppose that $I_N(h,\alpha)$ is given by* (3.87) *and $h=\sqrt{\pi/(N+1)}$. Then for $z=x+\mathrm{i}y$ with $y\geq\max(x,\pi/h)$, we have*

$$|w(z)-I_N(h,\alpha)| \leq b_N\,\mathrm{e}^{-\pi N}, \quad \textit{and} \tag{3.120}$$

$$\frac{|w(z)-I_N(h,\alpha)|}{|w(z)|} \leq b_N^*\sqrt{N+1}\,\mathrm{e}^{-\pi N}, \tag{3.121}$$

where

$$b_N := \frac{40\sqrt{2}\left(1+2\sqrt{\pi}\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi e^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}+\frac{2e^{1/4}}{e^{\pi}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}+\frac{(1+2\pi)}{\pi^2(N+1)\mathrm{e}^{\pi}}, \tag{3.122}$$

and

$$b_N^* := \left(\frac{1+2\pi\sqrt{N+1}}{\sqrt{N+1}}\right)b_N, \tag{3.123}$$

with β is given by (3.43). *Further, b_N and b_N^* decrease as N increases with $b_1\approx 0.20$, $b_1^*\approx 0.84$,*

$$\lim_{N\to\infty} b_N = \frac{2e^{1/4}}{e^{\pi}} \approx 0.11 \quad \textit{and} \quad \lim_{N\to\infty} b_N^* = \frac{4\pi e^{1/4}}{e^{\pi}} \approx 0.70. \tag{3.124}$$

Proof. Applying Proposition 3.3.3 for $h=\sqrt{\pi/(N+1)}$ yields that

$$|w(z)-I(h,\alpha)| \leq \left(\frac{40\sqrt{2}\left(1+2\sqrt{\pi}\mathrm{e}^{-\beta\pi(N+1)}\right)}{9\pi e^{\pi}\sqrt{N+1}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}+\frac{2e^{1/4}}{e^{\pi}\left(1-\mathrm{e}^{-2\pi(N+1)}\right)}\right)e^{-\pi N}, \tag{3.125}$$

and

$$\frac{|w(z)-I(h,\alpha)|}{|w(z)|} \leq (1+5\sqrt{2}\pi\sqrt{N+1})|w(z)-I_N(h,1/2)|. \tag{3.126}$$

Similarly, we obtain by applying Proposition 3.3.6 for $h=\sqrt{\pi/(N+1)}$ that

$$|T_N(h,\alpha)| \leq \frac{1+2\pi}{\pi^2\mathrm{e}^{\pi}(N+1)}\,\mathrm{e}^{-\pi N} \tag{3.127}$$

and

$$\frac{|T_N(h,\alpha)|}{|w(z)|} \leq (1+2\pi\sqrt{N+1})\,|T_N(h,\alpha)|. \tag{3.128}$$

Since

$$|w(z) - I_N(h,\alpha)| \le |w(z) - I(h,\alpha)| + |T_N(h,\alpha)|,$$

and

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

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 $0 < x \le y < \pi/h$, that*

$$|w(z) - I_N^*(h,\alpha)| \le b_N \, \mathrm{e}^{-\pi N} \tag{3.129}$$

and

$$\frac{|w(z) - I_N^*(h,\alpha)|}{|w(z)|} \le b_N^* \sqrt{N+1}\, \mathrm{e}^{-\pi N}, \tag{3.130}$$

where b_N and 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 and 3.3.3), and that explores the accuracy and efficiency of our approximation $w_N(z)$ given by (3.21) in comparison with the approximations (3.8), (3.11) and (3.16). The numerical calculations in Figures 3.1 are implemented for $z = 10^p e^{i\theta}$, with $p = -6(0.06)6$ and $\theta = 0(\pi/400)\pi/2$ giving in total 40401 values (these are the same test values used in Weideman [62]).

In Figure 3.1 we compute the maximum values of the absolute and relative errors in our approximation (3.21) to $w(z)$, implemented using the *Matlab* code in Listing A.3, and the approximation (3.8) from Weideman [62], implemented using the *Matlab* code in Table 1 [62], as a function of N. The exact value of $w(z)$ is computed for our approximation by $w_{20}(z)$, and for Weideman's approximation with $N = 40$ in the formula (3.8). From Figure 3.1 we read off that:

(i) the exponential convergence predicted by the bounds in Theorem 3.2.1 is achieved;

(ii) the approximation w_N achieves, with N as small as 11, maximum absolute and relative errors which are $\le 10^{-15}$.

(iii) the approximation w_N, with $N \leq 14$, 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 $|z| \to 0$, 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 $\alpha = 2.75$ 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 $z = 10^p e^{i\theta}$, with $p = -6(0.0006)6$ and $\theta = 0(\pi/400)\pi/2$ 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 $\alpha = 2.75$ and $M = 5$, and at least as accurate as Weideman's approximation (3.8) with $N = 40$. We can read off of the timing comparisons that the approximation $w_N(z)$ and its code `w(z,N)` is significantly more efficient, for the stated range of values of z, than the approximation (3.8) and its code `cef(z,40)` in Table 1 of Weideman [62]; and at the same time these timings confirm the efficiency of the approximation (3.16) and the higher efficiency (in addition to its high accuracy) of the approximation (3.11).

Algorithm	Maximum absolute error	Computation time in seconds
`w(z,11)`	1.11×10^{-15}	0.64
`cef(z,40)`	1.30×10^{-15}	1.46
`comperf(z)`	5.53×10^{-10}	0.90
`Faddeyeva_v2(z,13)`	3.92×10^{-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).

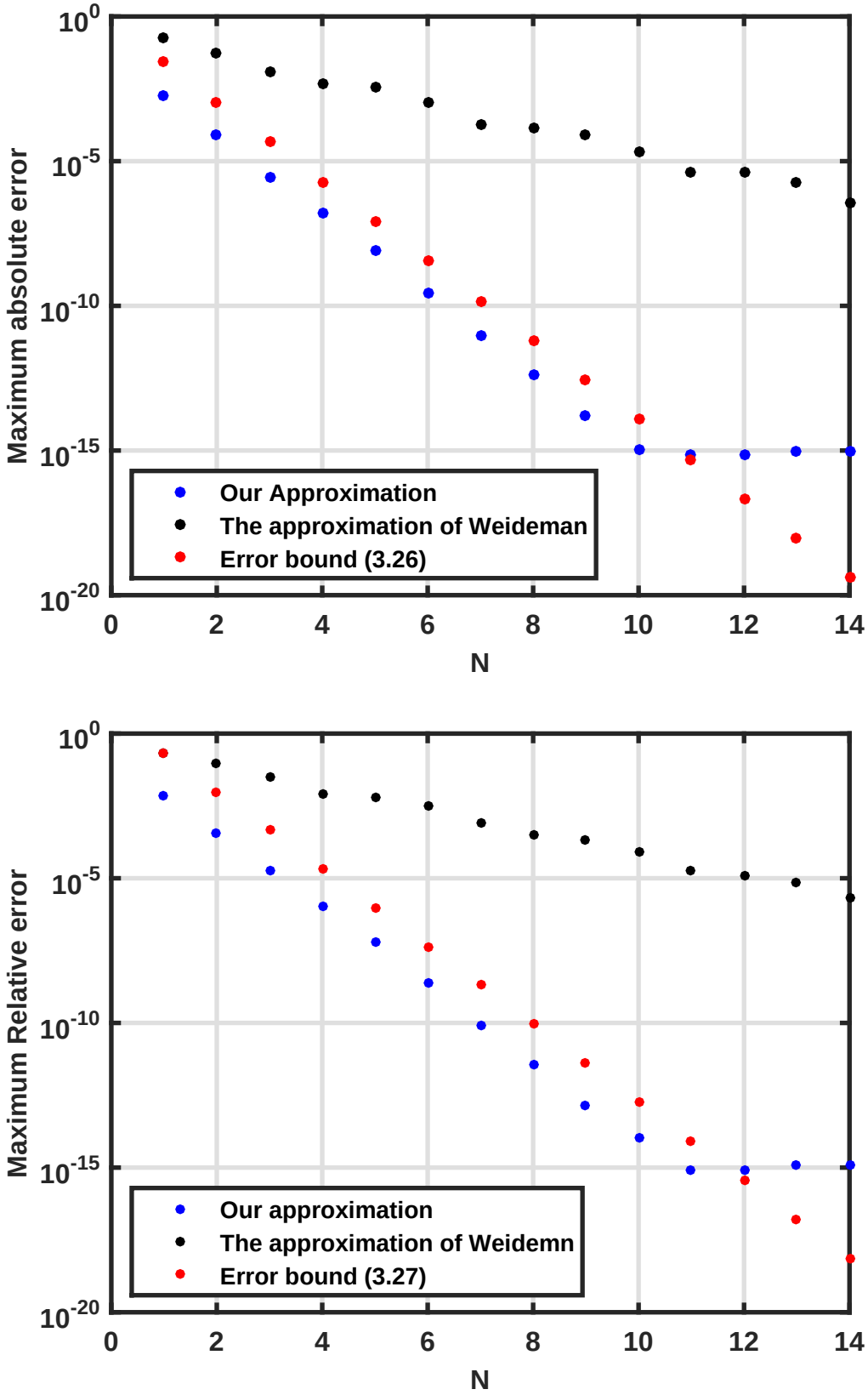

Fig. 3.1 Accuracy of our approximation (3.21) and its error bounds in Theorem 3.2.1, as a function of N, in comparison with Weideman's approximation (3.8).

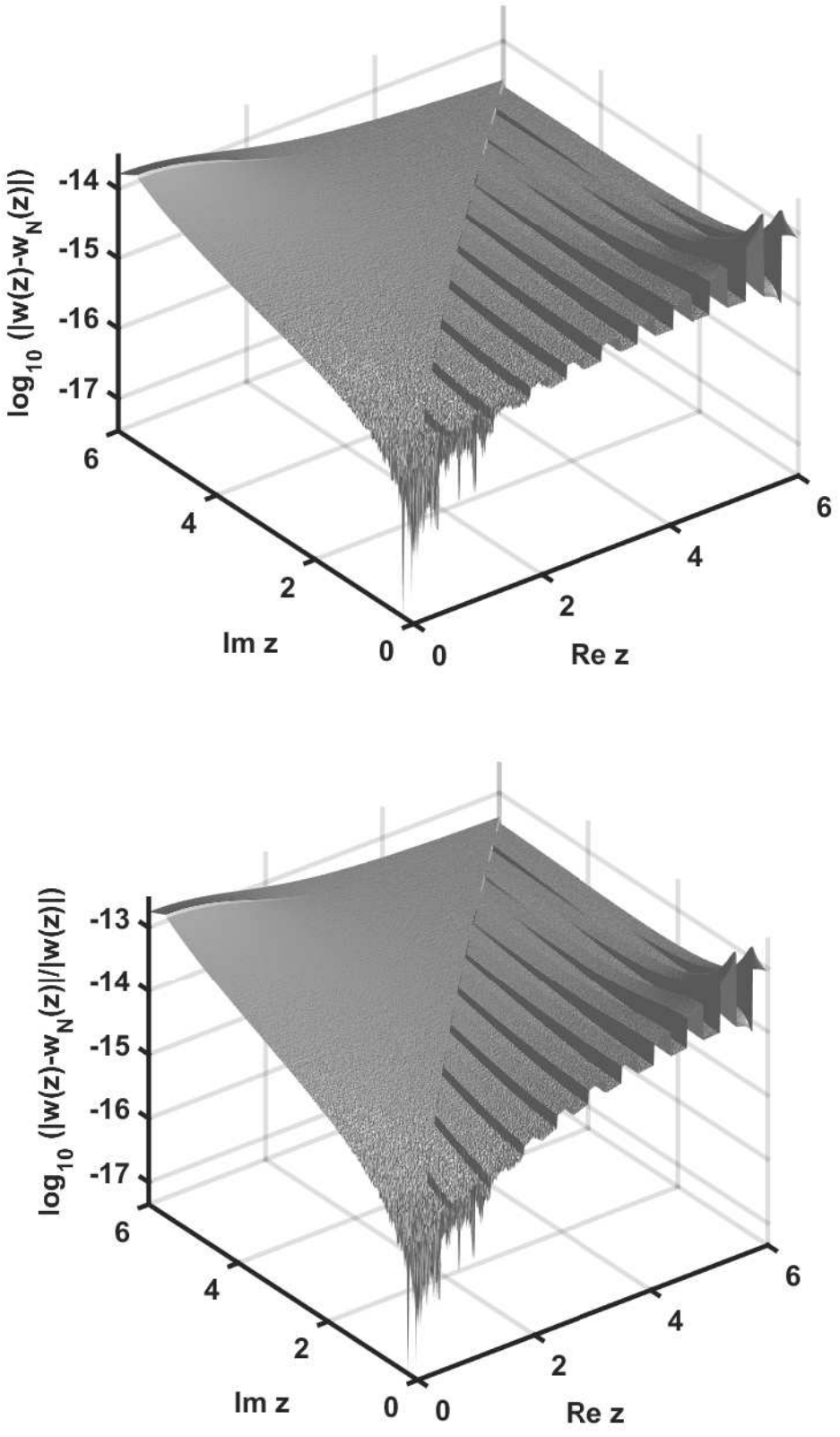

Fig. 3.2 The surfaces of the absolute (top) and relative (bottom) errors of the approximation $w_N(z)$ given by (3.21) with $N = 9$, where the exact value of $w(z)$ is computed by $w_{20}(z)$.

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 BEMs are discretisations of boundary integral equations (BIEs), and the kernel of these BIEs is given in terms of the acoustic Green's function for an impedance half-plane, so that computing each matrix element in the BEM discretisations requires computation of this Green's function; in other words, the computation of sound propagation from a coherent line source above a homogeneous impedance plane. Thus, efficient and accurate calculation of the solution to this problem is of great interest for a wide range of acoustics and outdoor noise control applications.

We adopt here the same notations used originally in [14] to introduce the problem and its equations. In the cartesian coordinate system $Oxyz$, we choose the mono-frequency line source to be parallel to the z-axis. The homogenous impedance plane is chosen to be $y = 0$, and the homogenous, stationary fluid half-space is $y > 0$ as shown in Figure 4.1. The problem is two-dimensional in the Oxy plane and the acoustic field is independent of z. The position of the source is $\mathbf{r_0} = (x_0, y_0)$, the position of the image of the source is $\mathbf{r'_0} = (x_0, -y_0)$, the position of the receiver is $\mathbf{r} = (x, y)$ and the angle of incidence is θ_0 with $0 \leq \theta_0 \leq \pi/2$ and $\gamma := \cos\theta_0 = (y + y_0)/d'$. Let $d = |\mathbf{r} - \mathbf{r_0}|$ be the distance from the source to the receiver,

$d' = |\mathbf{r} - \mathbf{r}_0'|$ 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

$$\nabla^2 G_\beta(\mathbf{r}, \mathbf{r_0}) + k^2 G_\beta(\mathbf{r}, \mathbf{r_0}) = \delta(\mathbf{r} - \mathbf{r_0}), \quad y > 0; \tag{4.1}$$

(ii) the **impedance boundary condition**, that is

$$\frac{\partial}{\partial y} G_\beta(\mathbf{r}, \mathbf{r_0}) + \mathrm{i}k\beta G_\beta(\mathbf{r}, \mathbf{r_0}) = 0, \quad y = 0; \tag{4.2}$$

(iii) the **Sommerfeld radiation condition**, that is

$$\frac{\partial}{\partial r} G_\beta(\mathbf{r}, \mathbf{r_0}) - \mathrm{i}k\beta G_\beta(\mathbf{r}, \mathbf{r_0}) = \mathrm{o}(r^{-1/2}), \quad G_\beta(\mathbf{r}, \mathbf{r_0}) = \mathrm{O}(r^{1/2}), \tag{4.3}$$

uniformly in $\theta \in (0, \pi)$ as $r \to \infty$, where (r, θ) are the polar coordinates of $\mathbf{r}$.

Note that ∇^2 is the Laplace operator, $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$, and δ is the Dirac delta function.

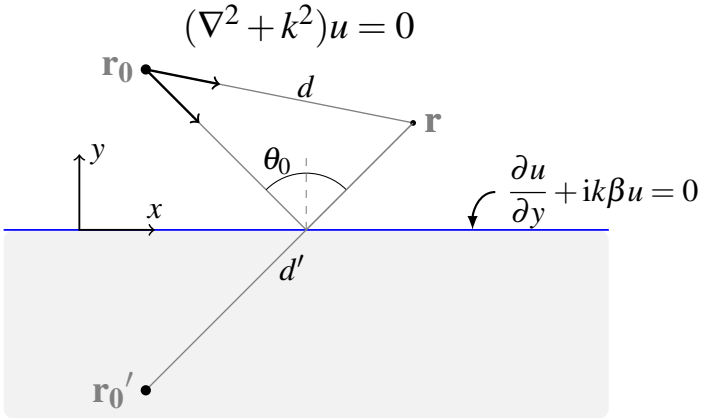

Fig. 4.1 The positions of the source ($\mathbf{r_0}$) and the receiver ($\mathbf{r}$) above the homogeneous impedance plane. The cross-section is in the plane perpendicular to the line source.

It is well known (see [14]) that the solution for a rigid surface boundary, i.e. for $\beta = 0$, is given by the methods of images as

$$G_0(\mathbf{r},\mathbf{r_0}) = -\frac{\mathrm{i}}{4}\left(H_0^{(1)}(kd) + H_0^{(1)}(kd')\right), \tag{4.4}$$

where $H_0^{(1)}$ is the Hankel function of the first kind of order 0. For the general case of an energy-absorbing boundary with $\mathrm{Re}(\beta) > 0$, the solution $G_\beta(\mathbf{r},\mathbf{r_0})$ is given by

$$G_\beta(\mathbf{r},\mathbf{r_0}) = G_0(\mathbf{r},\mathbf{r_0}) + P_\beta(\mathbf{r},\mathbf{r_0}). \tag{4.5}$$

Substituting for $G_\beta(\mathbf{r},\mathbf{r_0})$ in equations (4.1)–(4.3) shows that $P_\beta(\mathbf{r},\mathbf{r_0})$ satisfies the Sommerfeld radiation condition (4.3) and the following equations:

$$\nabla^2 P_\beta(\mathbf{r},\mathbf{r_0}) + k^2 P_\beta(\mathbf{r},\mathbf{r_0}) = 0, \quad y > 0 \tag{4.6}$$

and

$$\frac{\partial}{\partial y} P_\beta(\mathbf{r},\mathbf{r_0}) + \mathrm{i}k\beta P_\beta(\mathbf{r},\mathbf{r_0}) = -\mathrm{i}k\beta G_0(\mathbf{r},\mathbf{r_0}), \quad y = 0. \tag{4.7}$$

Applying a Fourier transform operator to the previous equations converts them into an ordinary differential equation with boundary conditions which can be solved to get an expression for the Fourier transform of P_β. Using the inverse Fourier transform we find that [14]

$$P_\beta(\mathbf{r},\mathbf{r_0}) = \frac{\mathrm{i}\beta}{2\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{\mathrm{i}k\phi(s)}}{\sqrt{1-s^2}(\sqrt{1-s^2}+\beta)}\, ds, \tag{4.8}$$

where $\mathrm{Re}\sqrt{1-s^2} \geq 0$, $\mathrm{Im}\sqrt{1-s^2} \geq 0$ and

$$\phi(s) := (y+y_0)\sqrt{1-s^2} - (x-x_0)s. \tag{4.9}$$

The integral representation (4.8) is not suitable for numerical quadrature due to the oscillatory behavior of the integrand. The substitution $s = \sin\theta$ simplifies the integrand and removes its branch point singularities at $s = \pm 1$, allowing us to rewrite P_β as

$$P_\beta(\mathbf{r},\mathbf{r_0}) = \frac{i\beta}{2\pi}\int_L \frac{e^{i\rho\cos(\theta-\theta_0)}}{\cos\theta+\beta}\, d\theta, \tag{4.10}$$

where L is the path in the θ-plane from $-\pi/2 + i\infty \to -\pi/2 \to \pi/2 \to \pi/2 - i\infty$. The integrand in (4.10) is highly oscillatory for large $\rho = kd'$. To remove this oscillation we

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

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

where

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

with

$$F(t) := -\frac{\beta + \gamma(1+\mathrm{i}t^2)}{\sqrt{t^2-2\mathrm{i}}(t^2-z_1^2)(t^2-z_2^2)}, \quad -\frac{\pi}{2} < \arg\sqrt{t^2-2\mathrm{i}} < \frac{\pi}{2}, \tag{4.13}$$

$$z_1 := \sqrt{\mathrm{i}a_+}, \quad -\frac{\pi}{4} < \arg\sqrt{ia_+} < \frac{3\pi}{4}, \tag{4.14}$$

$$z_2 := \sqrt{\mathrm{i}a_-}, \quad 0 < \arg\sqrt{ia_-} < \frac{\pi}{2}, \tag{4.15}$$

$$a_\pm := 1 + \beta\gamma \mp \sqrt{1-\beta^2}\sqrt{1-\gamma^2}, \quad \mathrm{Re}\sqrt{1-\beta^2} \geq 0, \tag{4.16}$$

and

$$P_\beta^{(s)} := \frac{\beta \mathrm{e}^{\mathrm{i}\rho}}{\pi} \frac{\pi \mathrm{e}^{-\mathrm{i}\rho a_+}}{2\sqrt{1-\beta^2}}\, \delta_s, \tag{4.17}$$

where

$$\delta_s := \begin{cases} 2, & \mathrm{Im}\,\beta < 0, \mathrm{Re}\,a_+ < 0 \\ 1, & \mathrm{Im}\,\beta < 0, \mathrm{Re}\,a_+ = 0, \\ 0, & \text{otherwise} \end{cases} \tag{4.18}$$

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 (4.1)–(4.3) (and the corresponding 3D version) is of major interest in the literature (e.g. Thomasson [58, 59], Kawai *et al.* [35], Filippi [21], Habault [26] and Nédélec *et al.* [19]). In particular, methods for computing $G_\beta(\mathbf{r},\mathbf{r_0})$ can be found in Chandler-Wilde and Hothersall [14], La Porte [38], O'Neil *et al.* [47], and in Nédélec *et al.* [19].

We find in Chandler-Wilde and Hothersall [14] an efficient scheme for computing $P_\beta(\mathbf{r},\mathbf{r_0})$. This proposed approximation is shown in [14], using numerical and theoretical calculations, to be accurate and efficient for a wide range of applications. This approximation

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 $P_\beta(r, r_0)$ is derived and used in [14]:

$$P_\beta(\mathbf{r}, \mathbf{r_0}) = \frac{\beta\, e^{i\rho}}{\pi} \int_0^\infty t^{-1/2} e^{-\rho t} f(t)\, dt, \quad \mathrm{Im}(\beta) > 0 \quad \text{or} \quad \mathrm{Re}(a_+) > 0, \tag{4.19}$$

and

$$P_\beta(\mathbf{r}, \mathbf{r_0}) = \frac{\beta\, e^{i\rho}}{\pi} \left(\int_0^\infty t^{-1/2} e^{-\rho t} g(t)\, dt + \frac{w(\sqrt{i\rho a_+})}{2(1-\beta^2)^{1/2}} \right), \quad \beta \neq 1, \tag{4.20}$$

where w is the Faddeeva function given by (3.3),

$$f(t) := F(\sqrt{t}) = -\frac{(\beta + \gamma(1+\mathrm{i}t))}{\sqrt{t-2\mathrm{i}}(t-ia_+)(t-ia_-)}, \quad \mathrm{Re}\sqrt{t-2i} > 0, \tag{4.21}$$

with F given by (4.13) and $a_\pm$ by (4.16), and

$$g(t) := f(t) - \frac{e^{-i\pi/4}\sqrt{a_+}}{2(1-\beta^2)^{1/2}(t-ia_+)}, \quad \mathrm{Re}\sqrt{a_+} \geq 0, \tag{4.22}$$

with $\mathrm{Re}(1-\beta^2)^{1/2} \geq 0$. The proposed approximations are

$$P^{(1)}_{n,m} := \frac{\beta\, e^{i\rho}}{\pi\sqrt{\pi}} \sum_{j=1}^{m} w_{j,n} f(x_{j,n}/\rho), \tag{4.23}$$

and

$$P^{(2)}_{n,m} := \frac{\beta\, e^{i\rho}}{\pi\sqrt{\pi}} \sum_{j=1}^{m} w_{j,n} g(x_{j,n}/\rho) + \frac{w(\sqrt{i\rho a_+})}{2(1-\beta^2)^{1/2}}, \tag{4.24}$$

where, for $n \in \mathbb{N}$ and $1 \leq m \leq n$, $P^{(1)}_{n,m}$ and $P^{(2)}_{n,m}$ are n-point Gauss-Lauguerre quadrature rule approximations applied to the integral representations (4.19) and (4.20), respectively, but then neglecting the weights $w_{j,n}$ in the quadrature rule for $j > m$. Chandler-Wilde and Hothersall [14] proposed to approximate P_β by

$$P_{n,m} := \begin{cases} P^{(1)}_{n,m}, & |1-\beta| < 0.1, \\ P^{(2)}_{n,m}, & |1-\beta| \geq 0.1. \end{cases} \tag{4.25}$$

In view of applications to outdoor sound propagation, the excess attenuation which is defined as

$$EA := -20\log_{10}|G_\beta(\mathbf{r},\mathbf{r_0})/\{(-i/4)H_0^{(1)}(\rho)\}|$$

is of interest and the error in the value of EA is bounded in [14] by

$$11\times 10^{EA/20}\,E_{n,m}\quad dB,$$

where $E_{n,m}$ is the error in the approximation (4.25) defined in [14] as

$$E_{n,m} := |P_\beta(\mathbf{r},\mathbf{r_0}) - P_{n,m}|/|(-i/4)H_0^{(1)}(\rho)|. \tag{4.26}$$

Chandler-Wilde and Hothersall [14] proved, for $\rho \geq 14\pi$ and $|\beta| \leq 1$, that

$$E_{40,22} \leq 9.2\times 10^{-11}|\beta|, \tag{4.27}$$

and they showed, using numerical calculations and the bound (4.27), that the approximation $P_{40,22}$ is accurate for $0 \leq \gamma \leq 1$, $|\beta| \leq 1$ and $\rho \geq 0.5$.

La Porte [38] proposed an approximation of $P_\beta(r,r_0)$ based on the modified trapezium rule (1.18) (with $\alpha = 0$) applied to the integral representation (4.12). The proposed approximation is

$$P_\beta^{h,N,H} := \frac{\beta e^{i\rho}}{\pi}\left(I_N^*(h,0) + \frac{\pi\,\mathrm{e}^{-\mathrm{i}\rho a_+}}{2\sqrt{1-\beta^2}}\,\delta_s\right) \tag{4.28}$$

$$= \frac{\beta e^{i\rho}}{\pi}\left[hF(0) + 2h\sum_{k=1}^{N} e^{-\rho k^2h^2}F(kh) + C(h,0) + \frac{\pi\,\mathrm{e}^{-\mathrm{i}\rho a_+}}{2\sqrt{1-\beta^2}}\,\delta_s\right], \tag{4.29}$$

where

$$C(h,0) := \frac{\pi}{2\sqrt{1-\beta^2}}\left[\frac{2\mathbf{H}(H-y_2)\,e^{-i\rho a_-}}{1-e^{-2i\pi z_2/h}}\,V + \frac{\mathbf{H}(H-|y_1|)\,e^{-i\rho a_+}}{1-e^{-2i\pi z_1/h}}\,\delta_+^{(1)}\right], \tag{4.30}$$

$$V := \frac{\beta\sqrt{1-\gamma^2} - \gamma\sqrt{1-\beta^2}}{\sqrt{ia_-}\sqrt{i(a_- - 2)}},\quad \mathrm{Re}\sqrt{\cdot} \geq 0, \tag{4.31}$$

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

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

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

$$M_H(F) := \sup_{x\in\mathrm{R},|y|=H} |F(x+iy)| \le \frac{|\beta|+\gamma}{\sqrt{1-H^2}}\,\widehat{M}, \tag{4.34}$$

where

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

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 $0.5 \le \rho \le 8.54$, $0 \le \gamma \le 1$ and $0.1 \le |\beta| \le 1$.

This chapter of the thesis builds on the work of La Porte [38] but extends this work significantly. The main issues with the approximation $P_\beta^{h,N,H}$ 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 $\rho > 0$, $0 \le \gamma \le 1$ and β with $\mathrm{Re}(\beta) > 0$; 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 $P_\beta(r, r_0)$, for $0 \le \beta \le 1$, based on the following representation for $P_\beta(\mathbf{r}, \mathbf{r_0})$:

$$P_\beta(\mathbf{r}, \mathbf{r_0}) = I_1 + I_2,$$

where

$$I_1 := \frac{ik\beta}{2\pi} \int_0^1 H_0^{(1)}(k|\mathbf{r} - \widetilde{\mathbf{r}}_\mathbf{0}|) e^{ik\beta\eta}\, d\eta, \quad \widetilde{\mathbf{r}}_\mathbf{0} = (x_0, -(y_0 + \eta)),$$

and

$$I_2 := \frac{ik\beta}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-\sqrt{\lambda^2 - k^2}(y+y_0)}\, e^{-\sqrt{\lambda^2-k^2} - ik\beta}}{\sqrt{\lambda^2 - k^2}\left(\sqrt{\lambda^2 - k^2} - ik\beta\right)} e^{i\lambda(x - x_0)}\, d\lambda.$$

The computation scheme of O'Neil *et al.* [47] is based on approximating the integral I_1 by a 16th order Gauss–Legendre quadrature rule on the dyadic subintervals $[0, 2^{-m}]$, $[2^{-m}, 2^{-m+1}]$, ..., $[2^{-1}, 1]$ (giving in total $16m$ quadrature points), and approximating the integral I_2 by a truncated trapezium rule in the variable $t \in [-t_{max}, t_{max}]$ along the contour $\lambda = t - i\tanh(t)$ with $t_{max} = |k| + 20$ and the number of quadrature points of the order $\mathrm{O}(k + \alpha)$ with $\alpha = k\beta$. It is clear here that we have two sources of error in this scheme corresponding to the two quadrature rules and there are no theoretical results in O'Neil *et al.* [47] on the errors in these suggested approximations, in contrast to the work for other approximations [14], [38] and this thesis. In the numerical calculations they provide the maximum achieved accuracy doesn't exceed 10^{-11}, while the same accuracy is reached with only 22 quadrature points using the scheme in [14], and the same accuracy with only 11 quadrature points of a (modified) trapezium rule using the proposed scheme in this thesis.

The rest of this chapter will be as follows. Section 4.2 gives a summary of the main results; §4.3 is the main section which contains the derivations of our approximation based on the truncated modified trapezium rule (1.22) and error bounds for this approximation; and finally §4.4 demonstrates, using numerical calculations, the accuracy of our approximation in comparison with the approximations of [14] and [38].

4.2 Summary of the main results

The main contributions of this chapter are: (i) to derive, based on the truncated modified trapezium rule (1.23), the approximation $P_{\beta,N}$ (given by (4.36) below) to P_β; (ii) to prove rigorous and uniform error bound on $|P_\beta - P_{\beta,N}|$; and (iii) to carry out systematic numerical

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

Let $P_\beta(\mathbf{r},\mathbf{r_0})$ be given by equations (4.11)–(4.18) and $H := \min(0.9,\widetilde{A}_N)$ with $\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)}$, and recall that $\mathbf{H}$ is given by (4.32). Then we propose, for $\rho > 0$, $0 \le \gamma \le 1$ and β in the half-plane $\mathrm{Re}(\beta) > 0$, cut from 1 to $+\infty$ along the real axis[1], to approximate P_β by

$$P_{\beta,N} := \frac{\beta\, e^{i\rho}}{\pi}\left(\widehat{I}_N + \frac{\pi\,\mathrm{e}^{-\mathrm{i}\rho a_+}}{2\sqrt{1-\beta^2}}\,\delta_s\right), \tag{4.36}$$

where δ_s is given by (4.18) and

$$\widehat{I}_N := \begin{cases} I_N^*(h,0), & \text{if } |\phi(|x_1|/h) - 1/2| \le 1/4\,, \\ \\ I_N^*(h,1/2), & \text{otherwise}\,, \end{cases} \tag{4.37}$$

with the step-size h given by Remark 4.3.3 and ϕ is defined by (3.7),

$$I_N^*(h,0) = \frac{\beta\mathrm{e}^{\mathrm{i}\rho}}{\pi}\left[hF(0) + 2h\sum_{k=1}^{N} e^{-\tau_k^2}F(\tau_k) + C(h,0)\right], \tag{4.38}$$

and

$$I_N^*(h,1/2) := \frac{\beta\mathrm{e}^{\mathrm{i}\rho}}{\pi}\left[2h\sum_{k=0}^{N} e^{-t_k^2}F(t_k) + C(h,1/2)\right], \tag{4.39}$$

with $\tau_k := kh$, $t_k := (k+1/2)h$ and F given by (4.13),

$$C(h,0) := \frac{\pi}{2\sqrt{1-\beta^2}}\left[\frac{-2\mathbf{H}\,(H-y_2)\,e^{-i\rho a_-}}{1-e^{-2i\pi z_2/h}}\,\Omega + \frac{\mathbf{H}\,(H-|y_1|)\,e^{-i\rho a_+}}{1-e^{-2i\pi z_1/h}}\,\delta_+^{(1)}\right],$$

and

$$C(h,1/2) := \frac{\pi}{2\sqrt{1-\beta^2}}\left[\frac{-2\mathbf{H}\,(H-y_2)\,e^{-i\rho a_-}}{1+e^{-2i\pi z_2/h}}\,\Omega + \frac{\mathbf{H}\,(H-|y_1|)\,e^{-i\rho a_+}}{1+e^{-2i\pi z_1/h}}\,\delta_+^{(2)}\right],$$

[1] This half-plane with the cut from 1 to $+\infty$ is referred to through this chapter as the "cut half-plane" as in [14].

where

$$\Omega := \begin{cases} +1, & \text{if} \quad \beta\sqrt{1-\gamma^2} - \gamma\sqrt{1-\beta^2} = \sqrt{ia_-}\sqrt{i(a_- - 2)}, \\ -1, & \text{otherwise}, \end{cases} \tag{4.40}$$

$$\delta_+^{(1)} := \begin{cases} 2\,e^{-2i\pi z_1/h}, & y_1 < 0\,, \\ 1 + e^{-2i\pi z_1/h}, & y_1 = 0\,, \\ 2, & y_1 > 0, \end{cases} \tag{4.41}$$

and

$$\delta_+^{(2)} := \begin{cases} -2\,e^{-2i\pi z_1/h}, & y_1 < 0\,, \\ 1 - e^{-2i\pi z_1/h}, & y_1 = 0\,, \\ 2, & y_1 > 0. \end{cases} \tag{4.42}$$

The main error estimate that we prove is

Theorem 4.2.1. *Let h be given by Remark 4.3.3 and $H := \min(0.9, \widetilde{A}_N)$ with $\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)}$. Then, for $\rho > 0$, $0 \le \gamma \le 1$ and β in the cut half-plane,*

$$|P_\beta - P_{\beta,N}| \le \begin{cases} \dfrac{|\beta|}{\pi}\,\Psi_N\,e^{-\sqrt{3}\,\pi\,(N+1)/2}, & \text{if} \quad \widetilde{A}_N \le 0.9, \\ \dfrac{|\beta|}{\pi}\,(N+1)^{1/3}\,\Upsilon_N\,e^{-1.5\rho^{1/3}H^{2/3}(N+1)^{2/3}}, & \text{otherwise}, \end{cases} \tag{4.43}$$

where

$$\Psi_N := C_N + \frac{2\pi}{|1-\beta^2|^{1/2}}, \tag{4.44}$$

$$\Upsilon_N := \widetilde{C}_N + \frac{2\pi}{(N+1)^{1/3}\,|1-\beta^2|^{1/2}}, \tag{4.45}$$

with

$$C_N := (|\beta|+1)\left[\frac{384\sqrt{10}\,(4|\beta|+7)(1+4\sqrt{\pi\rho})\rho^{3/2}}{\pi^{3/2}(N+1)^2\left(1-e^{-2\pi(N+1)/\sqrt{3}}\right)}+20\left(1+\frac{1}{\widetilde{A}_N}\right)\right], \quad (4.46)$$

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

$$\widetilde{K}_N := 8K_N\left(1+\frac{\rho^{1/3}}{a\,\pi^{1/3}H^{1/3}(N+1)^{1/3}}\right), \quad (4.47)$$

where K_N is given by (4.120)*. Further, for fixed β and ρ, C_N and $\widetilde{C}_N$ decrease as N increases with*

$$\lim_{N\to\infty} C_N = 20(|\beta|+1) \quad \textit{and} \quad \lim_{N\to\infty}\widetilde{C}_N = 16(|\beta|+1)\pi^{-2/3}H^{-2/3}\rho^{-1/3}. \quad (4.48)$$

4.3 The proposed approximation and its error bounds

In this section we derive the approximation (4.36) to $P_\beta(r,r_0)$. Using (4.11), (4.12) and (4.17) we can rewrite P_β as

$$P_\beta(\mathbf{r},\mathbf{r_0}) = \frac{\beta \mathrm{e}^{\mathrm{i}\rho}}{\pi}\left(I+\frac{\pi\,\mathrm{e}^{-\mathrm{i}\rho a_+}}{2\sqrt{1-\beta^2}}\,\delta_s\right). \quad (4.49)$$

where

$$I = \int_{-\infty}^{\infty} e^{-\rho t^2}F(t)\,dt, \quad (4.50)$$

and F is given by (4.13). Note that $e^{-\rho t^2}F(t)$ is meromorphic for $|\mathrm{Im}(t)|<1$ with simple poles at $t=\pm z_1$ and $t=\pm z_2$. Let $R_1 := \mathrm{Res}\left(e^{-\rho t^2}F(t), z_1\right)$ and $R_2 := \mathrm{Res}\left(e^{-\rho t^2}F(t), z_2\right)$, then we can show that

$$R_1 = \frac{i e^{-i\rho a_+}(\beta+\gamma(1-a_+))}{2\sqrt{ia_+}\sqrt{i(a_+-2)}(a_+-a_-)}, \quad -\frac{3\pi}{4} < \arg\sqrt{i(a_+-2)} < \frac{\pi}{4}, \quad (4.51)$$

$$R_2 = \frac{i e^{-i\rho a_-}(\beta+\gamma(1-a_-))}{2\sqrt{ia_-}\sqrt{i(a_--2)}(a_--a_+)}, \quad -\frac{\pi}{2} < \arg\sqrt{i(a_--2)} < \frac{\pi}{2}, \quad (4.52)$$

$$\mathrm{Res}\left(e^{-\rho t^2}F(t), -z_1\right) = -R_1 \quad \text{and} \quad \mathrm{Res}\left(e^{-\rho t^2}F(t), -z_2\right) = -R_2. \quad (4.53)$$

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

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

$$R_2 = \frac{ie^{-i\rho a_-}}{4\sqrt{1-\beta^2}}\,\Omega, \tag{4.55}$$

where

$$\Omega := \begin{cases} +1, & \text{if} \quad \beta\sqrt{1-\gamma^2} - \gamma\sqrt{1-\beta^2} = \sqrt{ia_-}\sqrt{i(a_--2)}, \\ -1, & \text{otherwise}. \end{cases} \tag{4.56}$$

Proof. Using (4.51) we have that

$$e^{-i\rho a_+}R_1 = \frac{i(\beta+\gamma(1-a_+))}{2\sqrt{ia_+}\sqrt{i(a_+-2)}(a_+-a_-)}, \tag{4.57}$$

and using the following relations from [14]

$$(\beta+\gamma(1-a_\pm))^2 = -a_\pm(a_\pm-2)(1-\gamma^2) \tag{4.58}$$

and

$$(a_\pm - a_\mp)^2 = 4(1-\beta^2)(1-\gamma^2), \tag{4.59}$$

we find that

$$e^{-2i\rho a_+}R_1^2 = \frac{(\beta+\gamma(1-a_+))^2}{4a_+(a_+-2)(a_+-a_-)^2} \tag{4.60}$$

$$= \frac{-(1-\gamma^2)}{4(a_+-a_-)^2} \tag{4.61}$$

$$= -\frac{1}{16(1-\beta^2)}, \tag{4.62}$$

which implies that

$$e^{-i\rho a_+}R_1 = \pm\frac{i}{4\sqrt{1-\beta^2}}. \tag{4.63}$$

The branch cuts of $\sqrt{ia_+}$ and $\sqrt{i(a_+-2)}$ given by (4.14) and (4.51), respectively, ensure that the two expressions (4.57) and (4.63), considered as functions of β, are analytic in the cut half-plane and agree to within a change of sign. To determine the correct choice of sign it is sufficient to evaluate the two expressions for $\beta=\gamma=0.5$. This gives that

$$R_1 = -\frac{\mathrm{i}\,\mathrm{e}^{-\mathrm{i}\rho a_+}}{4\sqrt{1-\beta^2}}. \tag{4.64}$$

Using (4.16) and (4.52), we have that

$$\begin{aligned} R_2 &= \frac{ie^{-i\rho a_-}(\beta+\gamma(1-a_-))}{2\sqrt{ia_-}\sqrt{i(a_--2)}(a_--a_+)} && (4.65)\\ &= \frac{ie^{-i\rho a_-}(\beta-\beta\gamma^2-\gamma\sqrt{1-\beta^2}\sqrt{1-\gamma^2})}{4\sqrt{ia_-}\sqrt{i(a_--2)}\sqrt{1-\beta^2}\sqrt{1-\gamma^2}} && (4.66)\\ &= \frac{ie^{-i\rho a_-}}{4\sqrt{1-\beta^2}}\left(\frac{\beta\sqrt{1-\gamma^2}-\gamma\sqrt{1-\beta^2}}{\sqrt{ia_-}\sqrt{i(a_--2)}}\right). && (4.67)\end{aligned}$$

We can easily show, using (4.16), that

$$\left(\beta\sqrt{1-\gamma^2}-\gamma\sqrt{1-\beta^2}\right)^2 = -a_-(a_--2) \tag{4.68}$$

so that

$$\beta\sqrt{1-\gamma^2}-\gamma\sqrt{1-\beta^2} = \pm\sqrt{ia_-}\sqrt{i(a_--2)}, \tag{4.69}$$

and the second result follows. □

Applying the modified trapezium rule (1.18) to the integral I given by (4.50) yields that

$$I^*(h,\alpha) := I(h,\alpha)+C(h,\alpha), \tag{4.70}$$

where

$$I(h,\alpha) = \sum_{k\in\mathbb{Z}} e^{-\rho(k-\alpha)^2h^2}F((k-\alpha)h), \tag{4.71}$$

F is given by (4.13), and, using (1.16), Remark 1.2.2 and Lemma 4.3.1,

$$C(h,\alpha) := \frac{\pi}{2\sqrt{1-\beta^2}}\left[\frac{-2\mathbf{H}(H-y_2)\,e^{-i\rho a_-}}{1-e^{-2i\pi(\alpha+z_2/h)}}\,\Omega+\frac{\mathbf{H}(H-|y_1|)\,e^{-i\rho a_+}}{1-e^{-2i\pi(\alpha+z_1/h)}}\,\delta_+\right] \tag{4.72}$$

where Ω is given by (4.56),

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

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 $\alpha = 0$ or $\alpha = 1/2$,

$$E^*(h,\alpha) := I - I^*(h,\alpha),$$

where I and $I^*(h,\alpha)$ 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*

$$|E^*(h,\alpha)| \le \frac{\Delta(H)\, e^{\rho H^2/4 - \pi H/h}}{1 - e^{-\pi H/h}}, \tag{4.75}$$

where

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

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 $\varepsilon \in (0, H/4)$ and consider the case $|H - |y_1|| \ge \varepsilon$ and $|H - y_2| \ge \varepsilon$. Then, using Proposition 1.2.4, we have that

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

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

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

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

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

where, for $j=1,2$,

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

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

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

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

$$|E^*(h,\alpha)| \le \left|\int_{\Gamma_H^*} e^{-\rho z^2} F(z)(1-g(z))dz\right| + \sum_{j=1}^{2}\left|\int_{\gamma_j} e^{-\rho z^2} F(z)(1-g(z))dz\right| + |C(h,\alpha)|. \tag{4.81}$$

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

$$\begin{aligned} |C(h,\alpha)| &\le \frac{\pi}{2|1-\beta^2|^{1/2}}\left(\frac{2e^{-2\pi|y_1|/h}}{1-e^{-2\pi|y_1|/h}} + \frac{2e^{-2\pi y_2/h}}{1-e^{-2\pi y_2/h}}\right) \\ &\le \frac{\pi}{2|1-\beta^2|^{1/2}}\left(\frac{4e^{-2\pi(H-4\varepsilon)/h}}{1-e^{-2\pi(H-4\varepsilon)/h}}\right) \\ &= \frac{2\pi e^{-2\pi(H-4\varepsilon)/h}}{|1-\beta^2|^{1/2}(1-e^{-2\pi(H-4\varepsilon)/h})}. \end{aligned} \tag{4.82}$$

For the integral over the path Γ_H^*, we have that

$$\left|\int_{\Gamma_H^*} e^{-\rho z^2} F(z)(1-g(z))dz\right| \le \frac{2\sqrt{\pi}M_H(F)}{\sqrt{\rho}(1-e^{-2\pi H/h})}\, e^{\rho H^2 - 2\pi H/h} \tag{4.83}$$

and $M_H(F)$ is bounded by (4.78).

For $z \in \gamma_j$ with $j = 1,2$, we have

$$\begin{aligned} |z^2 - 2i| &\geq 0.1(1 + H - 4\varepsilon) \geq 0.1 && (4.84)\\ |z^2 - z_1^2| &\geq \varepsilon\,(|y_1| + H - 4\varepsilon) \geq \varepsilon\,(H - 4\varepsilon), && (4.85)\\ |z^2 - z_2^2| &\geq \varepsilon\,(y_2 + H - 4\varepsilon) \geq \varepsilon\,(H - 4\varepsilon), && (4.86)\\ |\beta + \gamma(1 + iz^2)| &\leq (|\beta| + \gamma)(1 + (1 + z_j)^2) \leq (|\beta| + 1)(7|\beta| + 4). && (4.87) \end{aligned}$$

Thus we have that

$$\left| \int_{\gamma_j} e^{-\rho z^2} F(z)(1 - g(z))dz \right| \leq \frac{2\sqrt{10}(|\beta| + 1)(7|\beta| + 4)\, e^{-\frac{\pi^2}{\rho h^2}}}{\varepsilon^2 (H - 4\varepsilon)^2 (1 - e^{-2\pi(H - 4\varepsilon)/h})} \left| \int_{\gamma_j} e^{-\rho Z^2} \right|, \tag{4.88}$$

where $Z := z - i\frac{\pi}{\rho h} = X + iY$. Note that $\left| e^{-\rho Z^2} \right| = e^{\rho K}$, where

$$K = Y^2 - X^2 \leq \left(y - \frac{\pi}{\rho h} \right)^2. \tag{4.89}$$

Since $H - 4\varepsilon \leq y \leq H$ and $H = \min\left(0.9, \frac{\pi}{\rho h}\right)$, we have

$$K \leq \left(\frac{\pi}{\rho h} - H + 4\varepsilon \right)^2. \tag{4.90}$$

Thus,

$$\left| \int_{\gamma_j} e^{-\rho Z^2} \right| \leq 2\pi\, e^{\rho \left(\frac{\pi}{\rho h} - H + 4\varepsilon \right)^2}, \tag{4.91}$$

and

$$\left| \int_{\gamma_j} e^{-\rho z^2} F(z)(1 - g(z))dz \right| \leq \frac{4\pi\sqrt{10}(|\beta| + 1)(7|\beta| + 4)}{\varepsilon^2 (H - 4\varepsilon)^2 (1 - e^{-2\pi(H - 4\varepsilon)/h})}\, e^{P_2}, \tag{4.92}$$

where $P_2 = \rho \left(\frac{\pi}{\rho h} - H + 4\varepsilon \right)^2 - \frac{\pi^2}{\rho h^2}$.

Note that, for $\varepsilon > 0$ and $H := \min\left(0.9, \frac{\pi}{\rho h}\right)$, we can easily show that

$$\rho H^2 - 2\pi H/h = \rho \left(H - \frac{\pi}{\rho h} \right)^2 - \frac{\pi^2}{\rho h^2} \leq P_2. \tag{4.93}$$

Choosing ε to maximise $\varepsilon^2 (H - 4\varepsilon)^2$ gives that $\varepsilon = H/8$. For this choice of ε we have

$$M_H(F) \leq \frac{256\sqrt{10}(|\beta| + 1)(7 + 4|\beta|)}{H^4},$$

and

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

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 $T_N(h,\alpha)$ as defined in (1.24) for $\alpha = 0$ or $\alpha = 1/2$. 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 $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ are given by (4.14) and (4.15), respectively, then we have the following result.

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

$$|F(\tau_k)| \leq \frac{8}{h}(|\beta|+1)\left(1+\frac{1}{\tau_{N+1}}\right). \quad (4.95)$$

Proof. For $\tau_k = kh$ with $k = N+1, N+2, \ldots$, we have

$$\begin{aligned}
|\beta+\gamma(1+i\tau_k^2)| &\leq (|\beta|+1)|1+i\tau_k^2| = (|\beta|+1)\sqrt{1+\tau_k^4}, && (4.96)\\
\sqrt{|\tau_k^2-2i|} &= \sqrt[4]{1+\tau_k^4} && (4.97)\\
|\tau_k^2 - z_1^2| &= |\tau_k - z_1||\tau_k + z_1| \geq \frac{h}{4}(\tau_k + |x_1|), && (4.98)\\
|\tau_k^2 - z_2^2| &= |\tau_k - z_2||\tau_k + z_2| \geq |\tau_k - y_2|(\tau_k + x_2). && (4.99)
\end{aligned}$$

Since $\mathrm{Re}(a_-) \geq 1$, we have, for $z_2 = \sqrt{ia_-} = x_2 + iy_2$, that $y_2 \geq \dfrac{1}{2x_2} > 0$ and

$$|\tau_k^2 - z_2^2| \geq \frac{1}{2x_2}(\tau_k + x_2). \quad (4.100)$$

Combining the above inequalities, we find that

$$|F(\tau_k)| \leq \frac{8x_2(|\beta|+1)\sqrt{1+\tau_k^4}}{h\sqrt[4]{1+\tau_k^4}(\tau_k+|x_1|)(\tau_k+x_2)} \tag{4.101}$$

$$\leq \frac{8(|\beta|+1)\sqrt[4]{1+\tau_k^4}}{h(\tau_k+|x_1|)} \tag{4.102}$$

$$\leq \frac{8(|\beta|+1)(1+\tau_k)}{h(\tau_k+|x_1|)}, \tag{4.103}$$

where the last line comes from

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

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

$$|F(\tau_k)| \leq \frac{8}{h}(|\beta|+1) \times \begin{cases} \dfrac{1+\tau_{N+1}}{|x_1|+\tau_{N+1}}, & \text{if } |x_1| \leq 1\ , \\ \\ 1, & \text{otherwise}\ , \end{cases} \tag{4.104}$$

but

$$\frac{1+\tau_{N+1}}{|x_1|+\tau_{N+1}} \leq \left(1+\frac{1}{\tau_{N+1}}\right),$$

and hence the result follows. □

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

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

we have

$$|T_N(h,0)| \leq \frac{8\,(|\beta|+1)(1+2h\rho\,\tau_{N+1})}{h\rho\,\tau_{N+1}}\left(1+\frac{1}{\tau_{N+1}}\right)\,e^{-\rho\,\tau_{N+1}^2}. \tag{4.105}$$

Proof. Using Lemma 4.3.2 we find that

$$\begin{aligned}
|T_N(h,0)| &\leq \frac{8M_N(|\beta|+1)}{h}\left(2h\sum_{k=N+1}^{\infty}e^{-\rho\tau_k^2}\right)\\
&= \frac{8M_N(|\beta|+1)}{h}\left(2he^{-\rho\tau_{N+1}^2}+2h\sum_{k=N+2}^{\infty}e^{-\rho\tau_k^2}\right)\\
&\leq \frac{8M_N(|\beta|+1)}{h}\left(2he^{-\rho\tau_{N+1}^2}+2\int_{\tau_{N+1}}^{\infty}e^{-\rho t^2}\,dt\right)\\
&\leq \frac{8M_N(|\beta|+1)}{h}\left(2he^{-\rho\tau_{N+1}^2}+\frac{e^{-\rho\tau_{N+1}^2}}{\rho\,\tau_{N+1}}\right)\\
&= \frac{8M_N(|\beta|+1)(1+2h\rho\,\tau_{N+1})}{h\rho\,\tau_{N+1}}\;e^{-\rho\tau_{N+1}^2}.
\end{aligned}$$

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

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

□

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

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

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

$$\left(1+\frac{1}{t_{N+1}}\right)\leq\left(1+\frac{1}{\tau_{N+1}}\right),$$

and hence we have that

$$|T_N(h,1/2)|\leq\frac{8(|\beta|+1)(1+2h\rho\,\tau_{N+1})}{h\rho\,\tau_{N+1}}\left(1+\frac{1}{\tau_{N+1}}\right)\;e^{-\rho\tau_{N+1}^2}. \tag{4.108}$$

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 $\rho>0$, $H:=\min(0.9,\pi/(\rho h))$ and $\tau_{N+1}=(N+1)h$ with $N\in\mathbb{N}$, we define two possible choices, h_N^* and 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

bounds (4.75) and (4.105) on the discretisation and truncation errors. We will set $h = h_N^*$ if the value of H given by (4.109) is ≤ 0.9, otherwise we will use $h = h_N$.

(i) Let h_N^* be such that

$$H = \frac{\pi}{\rho h_N^*} \quad \text{and} \quad \frac{1}{4}\rho H^2 - \frac{\pi}{h_N^*} H = -\rho (N+1)^2 (h_N^*)^2, \tag{4.109}$$

then

$$h_N^* := \sqrt{\frac{\sqrt{3}\,\pi}{2\rho (N+1)}}. \tag{4.110}$$

(ii) Let h_N be such that,

$$H = 0.9 \quad \text{and} \quad \frac{1}{4}\rho H^2 - \frac{\pi}{h_N} H = -\rho (N+1)^2 (h_N)^2, \tag{4.111}$$

then h_N is the zero of the cubic polynomial $P(h)$ defined by

$$P(h) := 4\rho\,(N+1)^2\, h^3 + \rho H^2\, h - 4\pi H.$$

Since $P(0) = -4\pi H < 0$ and $P^{'}(h) = 12\rho\,(N+1)^2\, h^2 + \rho H^2 > 0$, h_N is the unique real zero of this polynomial. We can express the equation $P(h_N) = 0$ as

$$a^3 + 3ba - 1 = 0, \tag{4.112}$$

where

$$h_N^{'} = \left(\frac{\pi H}{\rho\,(N+1)^2}\right)^{1/3}, \quad a = \frac{h_N}{h_N^{'}} \quad \text{and} \quad b = \frac{H^2}{12(N+1)^2\,(h_N^{'})^2}. \tag{4.113}$$

Using the well-known formula to solve the cubic equation (4.112) for a, we find

$$a = \sqrt[3]{\frac{1}{2} + \sqrt{\frac{1}{4} + b^3}} + \sqrt[3]{\frac{1}{2} - \sqrt{\frac{1}{4} + b^3}} \tag{4.114}$$

The following lemma (see La Porte [38, Proposition 2.3.17]) gives lower and upper bounds for a.

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

Remark 4.3.3. *Let $H := \min(0.9, \widetilde{A}_N)$ with $\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)}$, and set*

$$h := \begin{cases} h_N^* := \sqrt{\dfrac{\sqrt{3}\,\pi}{2\rho\,(N+1)}}, & \widetilde{A}_N \leq 0.9 \\ \\ h_N = a\left(\dfrac{\pi H}{\rho\,(N+1)^2}\right)^{1/3}, & \textit{otherwise}, \end{cases} \tag{4.116}$$

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

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

The following result bounds the expression

$$\frac{1 + 2h\rho\, t_{N+1}}{h\rho\, t_{N+1}}$$

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 $\rho > 0$, $N \in \mathbb{N}$ and $\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)}$, and h be given as in Remark 4.3.3. Then, for $\tau_{N+1} = (N+1)h$,*

$$\frac{1 + 2h\rho\,\tau_{N+1}}{\rho h\,\tau_{N+1}} \leq \begin{cases} \dfrac{5}{2}, & \textit{if} \quad \widetilde{A}_N \leq 0.9, \\ \\ (N+1)^{1/3}\, K_N, & \textit{otherwise}, \end{cases} \tag{4.119}$$

where

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

Proof. For $h = h_N^* = \sqrt{\dfrac{\sqrt{3}\,\pi}{2\rho(N+1)}}$, we have

$$\frac{1+2\rho\, h\,\tau_{N+1}}{\rho\, h\,\tau_{N+1}} = 2 + \frac{1}{\rho(N+1)(h_N^*)^2} = 2 + \frac{2}{\sqrt{3}\pi} \leq 2.5.$$

For $h = h_N = a\left(\dfrac{\pi H}{\rho\,(N+1)^2}\right)^{1/3}$, we have that

$$\frac{1+2h\rho\,\tau_{N+1}}{h\rho\,\tau_{N+1}} = 2 + \frac{(N+1)^{1/3}}{\rho^{1/3}(\pi H)^{2/3}\,a^2}.$$

Using the lower bound in Lemma 4.3.3, we see that

$$a^2 \geq \frac{1}{(1+3b)^2} \geq \frac{1}{2+18b^2},$$

and hence we have

$$\begin{aligned}\frac{1+2h\rho\,\tau_{N+1}}{h\rho\,\tau_{N+1}} &\leq 2 + \frac{2(N+1)^{1/3}}{\rho^{1/3}(\pi H)^{2/3}} + \frac{18b^2(N+1)^{1/3}}{\rho^{1/3}(\pi H)^{2/3}} \\ &= 2 + \frac{2(N+1)^{1/3}}{\rho^{1/3}(\pi H)^{2/3}} + \frac{\rho H^2}{8\pi^2(N+1)},\end{aligned}$$

where the last line comes from using (4.118) with some simplifications, and the second bound follows. □

4.3.4 Bounding the total error

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

$$E_N^*(h,\alpha) = I - I_N^*(h,\alpha),$$

where I is given by (4.50) and $I_N^*(h,\alpha)$ is the truncation of $I^*(h,\alpha)$ given by (4.70). Note that

$$E_N^*(h,\alpha) = E^*(h,\alpha) + T_N(h,\alpha),$$

and hence the following result gives upper bound to $E_N^*(h,\alpha)$ using Propositions 4.3.1 and 4.3.2 and Lemma 4.3.4.

Theorem 4.3.5. *Let $\rho > 0$, $H := \min(0.9, \widetilde{A}_N)$ with $\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)}$ and h is given by Remark 4.3.3. Then, for $0 \le \gamma \le 1$, β in the cut half-plane and $\alpha = 0, 1/2$,*

$$|E_N^*(h,\alpha)| \le \begin{cases} \Psi_N\, e^{-\sqrt{3}\pi(N+1)/2}, & \text{if} \quad \widetilde{A}_N \le 0.9, \\ \\ (N+1)^{1/3}\Upsilon_N\, e^{-1.5\rho^{1/3}H^{2/3}(N+1)^{2/3}}, & \text{otherwise}, \end{cases}$$

where

$$\Psi_N := C_N + \frac{2\pi}{|1-\beta^2|^{1/2}}, \tag{4.121}$$

$$\Upsilon_N := \widetilde{C}_N + \frac{2\pi}{(N+1)^{1/3}|1-\beta^2|^{1/2}}, \tag{4.122}$$

with

$$C_N := (|\beta|+1)\left[\frac{384\sqrt{10}(4|\beta|+7)(1+4\sqrt{\pi\rho})\rho^{3/2}}{\pi^{3/2}(N+1)^2\left(1-e^{-2\pi(N+1)/\sqrt{3}}\right)} + 20\left(1+\frac{1}{\widetilde{A}_N}\right)\right], \tag{4.123}$$

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

$$\widetilde{K}_N := 8K_N\left(1+\frac{\rho^{1/3}}{a\,\pi^{1/3}H^{1/3}(N+1)^{1/3}}\right), \tag{4.124}$$

where K_N is given by (4.120)*. Further, for fixed β and ρ, C_N and $\widetilde{C}_N$ decrease as N increases with*

$$\lim_{N\to\infty} C_N = 20(|\beta|+1) \quad \textit{and} \quad \lim_{N\to\infty} \widetilde{C}_N = 16(|\beta|+1)\pi^{-2/3}H^{-2/3}\rho^{-1/3}. \tag{4.125}$$

Proof. We consider first the case when $H = \widetilde{A}_N$ and $h = h_N^* = \sqrt{\dfrac{\sqrt{3}\,\pi}{2\rho(N+1)}}$. Using (4.75) for these values with some simplifications gives that

$$|E^*(h,\alpha)| \le \left(\frac{384\sqrt{10}(|\beta|+1)(7+4|\beta|)(1+4\sqrt{\pi\rho})\rho^{3/2}}{\pi^{3/2}\,(N+1)^2\left(1-e^{-2\pi(N+1)/\sqrt{3}}\right)} + \frac{2\pi}{|1-\beta^2|^{1/2}}\right) e^{-\sqrt{3}\pi(N+1)/2}. \tag{4.126}$$

Using Proposition 4.3.2 and Lemma 4.3.4 for the same values gives that

$$|T_N(h,\alpha)| \le 20(|\beta|+1)\left(1+\frac{1}{\widetilde{A}_N}\right) e^{-\sqrt{3}\pi(N+1)/2}, \tag{4.127}$$

and hence the first bound follows.

Now we consider the case $H = 0.9$ and $h = h_N = a\left(\dfrac{\pi H}{\rho\,(N+1)^2}\right)^{1/3}$. Using (4.75) we find that

$$|E^*(h,\alpha)| \le \left(\frac{782\sqrt{10\pi}(1+4\sqrt{\pi\rho})(|\beta|+1)(7+4|\beta|)}{\sqrt{\rho}\,\left(1-e^{-0.9\pi h_N}\right)} + \frac{2\pi}{|1-\beta^2|^{1/2}}\right) e^{\rho H^2/4-\pi H/h_N}.$$

Using Proposition 4.3.2 and Lemma 4.3.4 for these values we have

$$|T_N(h,\alpha)| \le (|\beta|+1)\,\widetilde{K}_N\,(N+1)^{1/3}\,e^{\rho H^2/4-\pi H/h_N}, \tag{4.128}$$

where $\widetilde{K}_N = 8K_N\left(1+\dfrac{\rho^{1/3}}{a\,\pi^{1/3}H^{1/3}(N+1)^{1/3}}\right)$ and K_N is given by (4.120).

For $h_N = a\left(\dfrac{\pi H}{\rho\,(N+1)^2}\right)^{1/3}$ and using Lemma 4.3.3, we see that

$$\begin{aligned}
\frac{\rho H^2}{4} - \frac{\pi H}{h_N} &= \frac{\rho H^2}{4} - \frac{\rho^{1/3}\pi^{2/3}H^{2/3}(N+1)^{2/3}}{a} \\
&\le \frac{\rho H^2}{4} - \rho^{1/3}\pi^{2/3}H^{2/3}(N+1)^{2/3} - \rho^{1/3}\pi^{2/3}H^{2/3}(N+1)^{2/3}\,b \\
&= \frac{\rho H^2}{4} - \rho^{1/3}\pi^{2/3}H^{2/3}(N+1)^{2/3} - \frac{\rho H^2}{12} \\
&= \frac{1}{6}\rho H^2 - \rho^{1/3}\pi^{2/3}H^{2/3}(N+1)^{2/3} \\
&= \rho^{1/3}\left(\frac{1}{6}\rho^{2/3}H^2 - \pi^{2/3}H^{2/3}(N+1)^{2/3}\right).
\end{aligned}$$

Note

$$\widetilde{A}_N := \sqrt{2\pi(N+1)/(\sqrt{3}\rho)} > 0.9 \quad \text{implies that} \quad \rho < \frac{2\pi(N+1)}{\sqrt{3}H^2},$$

with $H = 0.9$. Hence, we see that

$$\begin{aligned}
\frac{\rho H^2}{4} - \frac{\pi H}{h_N} &\le \rho^{1/3}H^{2/3}(N+1)^{2/3}\left(\frac{1}{6}\left(\frac{2\pi}{\sqrt{3}}\right)^{2/3} - \pi^{2/3}\right) \\
&< -1.5\,\rho^{1/3}H^{2/3}(N+1)^{2/3},
\end{aligned}$$

and the second bound follows. □

4.4 Numerical results

In this section we show numerical calculations that illustrate and confirm theoretical results (Theorem 4.3.5), and that explores the accuracy and efficiency of our approximation

$P_{\beta,N}$ given by (4.36) in comparison with the approximations (4.25) and (4.28). Systematic numerical calculations are implemented for $\theta_0 = 0^o(10^o)90^o$, $|\beta| = 0.1(0.1)0.999$ and $\arg(\beta) = -89^o(8.9^o)89^o$, 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 $P_{\beta,N}$ given by (4.36) by $P^{(2)}_N$. We do not have access to exact values for P_β and so using different accurate approximations to P_β:

(i) Our approximation $P_{\beta,N}$ 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* code [9], supplied to us by the author.

(iii) The approximation (4.28) in La Porte [38] with $N = 100$, implemented in *Matlab* by the author of this thesis.

In Table 4.1, for the considered parameter values of γ and β, we compute the maximum of

$$E_{approx} := |P_{100,100} - P_{40,22}|/|(-i/4)H_0^{(1)}(\rho)|\,; \tag{4.129}$$

$$E^{(1)} := |P^{(1)}_{100} - P_{40,22}|/|(-i/4)H_0^{(1)}(\rho)|\,; \tag{4.130}$$

$$E^{(2)} := |P^{(2)}_{100} - P_{40,22}|/|(-i/4)H_0^{(1)}(\rho)|. \tag{4.131}$$

The calculations in Table 4.1 confirm the high accuracy of $P_{40,22}$ in Chandler-Wilde and Hothersall [14] and show that $P^{(1)}_{100}$ in La Porte [38] is as accurate as $P_{100,100}$ in Chandler-Wilde and Hothersall [14]; and demonstrate that our approximation $P^{(2)}_{100}$ is at least as accurate as the other two approximations. Additionally, it is also clear that the three approximations achieve an acceptable accuracy for small values of ρ.

In Tables 4.2 and 4.3, for the considered parameter values of γ and β, we compute the maximum of

$$E^{(3)}_N := |P_{100,100} - P^{(2)}_N|/|(-i/4)H_0^{(1)}(\rho)|\,; \tag{4.132}$$

$$E^{(4)}_N := |P^{(2)}_{100} - P^{(2)}_N|/|(-i/4)H_0^{(1)}(\rho)|. \tag{4.133}$$

It can be seen from Tables 4.2 and 4.3 that:

(i) With $N = 9$, our approximation $P_{\beta,N}$ given by (4.36) achieves a close accuracy to $P_{40,22}$ in Chandler-Wilde and Hothersall [14];

(ii) With $N = 11$, our approximation $P_{\beta,N}$ is as accurate as $P_{40,22}$;

(iii) With $N = 21$, our approximation $P_{\beta,N}$ is more accurate than $P_{40,22}$ for all the stated range of values of parameters.

The special cases when $\gamma \approx 0$ or $\beta \approx 0$ are investigated in Figures 4.2 and 4.3. It can be seen from Figures 4.2 and 4.3 that our approximation $P_{\beta,N}$ given by (4.36) is particularly accurate when $\gamma \approx 0$ and $\rho \geq 14\pi$ with $\beta = e^{i\pi/4}$. Additionally, it can be seen that $P_{\beta,N}$ is significantly more accurate for $\gamma \approx 0$ and $\beta \approx 0$ than the approximation (4.28) in La Porte [38]; and $P_{\beta,N}$ achieves, with N as small as 10, accuracy $\leq 10^{-15}$ for $\rho \geq 20$.

$\rho = kd'$	d'	E_{approx}	$E^{(1)}$	$E^{(2)}$
0.5	0.0796	5.8×10^{-4}	5.8×10^{-4}	5.8×10^{-4}
0.75	0.119	8.1×10^{-5}	8.1×10^{-5}	8.1×10^{-5}
1.125	0.179	7.1×10^{-6}	7.1×10^{-6}	7.1×10^{-6}
1.688	0.269	3.5×10^{-7}	3.5×10^{-7}	3.5×10^{-7}
2.531	0.403	8.3×10^{-9}	8.3×10^{-9}	8.3×10^{-9}
3.797	0.604	8.4×10^{-11}	8.4×10^{-11}	8.4×10^{-11}
5.695	0.906	7.0×10^{-13}	7.0×10^{-13}	7.0×10^{-13}
8.543	1.36	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}
12.814	2.039	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}
19.222	3.059	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
28.833	4.589	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
43.249	6.883	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
64.873	10.325	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
97.31	15.487	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
145.96	23.230	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
218.95	34.847	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
328.42	51.633	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
492.63	78.404	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
738.95	117.608	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}
1108.4	176.407	3.9×10^{-13}	3.9×10^{-13}	3.9×10^{-13}

Table 4.1 Maximum values of E_{approx}, $E^{(1)}$ and $E^{(2)}$ given by (4.129), (4.130) and (4.131), respectively, for $\theta_0 = 0^o(10^o)90^o$, $|\beta| = 0.1(0.1)0.999$ and $\arg(\beta) = -89^o(8.9^o)89^o$.

$\rho = kd'$	d'	E_{approx}	$E_9^{(3)}$	$E_{11}^{(3)}$	$E_{21}^{(3)}$
0.5	0.0796	5.8×10^{-4}	5.0×10^{-5}	1.7×10^{-5}	2.3×10^{-6}
0.75	0.119	8.1×10^{-5}	2.7×10^{-6}	7.1×10^{-7}	9.5×10^{-8}
1.125	0.179	7.1×10^{-6}	1.3×10^{-6}	2.8×10^{-7}	1.8×10^{-9}
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}	9.6×10^{-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}	2.0×10^{-14}
28.833	4.589	3.9×10^{-13}	1.9×10^{-11}	3.1×10^{-13}	3.5×10^{-14}
43.249	6.883	3.9×10^{-13}	4.3×10^{-12}	5.2×10^{-14}	5.2×10^{-14}
64.873	10.325	3.9×10^{-13}	1.7×10^{-11}	6.5×10^{-14}	6.1×10^{-14}
97.31	15.487	3.9×10^{-13}	1.5×10^{-11}	1.0×10^{-13}	1.0×10^{-13}
145.96	23.230	3.9×10^{-13}	7.4×10^{-12}	1.1×10^{-13}	1.0×10^{-13}
218.95	34.847	3.9×10^{-13}	1.0×10^{-11}	1.3×10^{-13}	1.4×10^{-13}
328.42	51.633	3.9×10^{-13}	4.1×10^{-12}	1.1×10^{-13}	1.0×10^{-13}
492.63	78.404	3.9×10^{-13}	3.0×10^{-12}	1.7×10^{-13}	1.7×10^{-13}
738.95	117.608	3.9×10^{-13}	2.9×10^{-12}	7.5×10^{-14}	6.9×10^{-14}
1108.4	176.407	3.9×10^{-13}	2.0×10^{-12}	6.9×10^{-14}	7.6×10^{-14}

Table 4.2 Maximum values of E_{approx} and $E_N^{(3)}$ given by (4.129) and (4.132), respectively, with $N = 9, 11, 21$, for $\theta_0 = 0^o(10^o)90^o$, $|\beta| = 0.1(0.1)0.999$ and $\arg(\beta) = -89^o(8.9^o)89^o$.

$\rho = kd'$	d'	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 $\theta_0 = 0^o(10^o)90^o$, $|\beta| = 0.1(0.1)0.999$ and $\arg(\beta) = -89^o(8.9^o)89^o$.

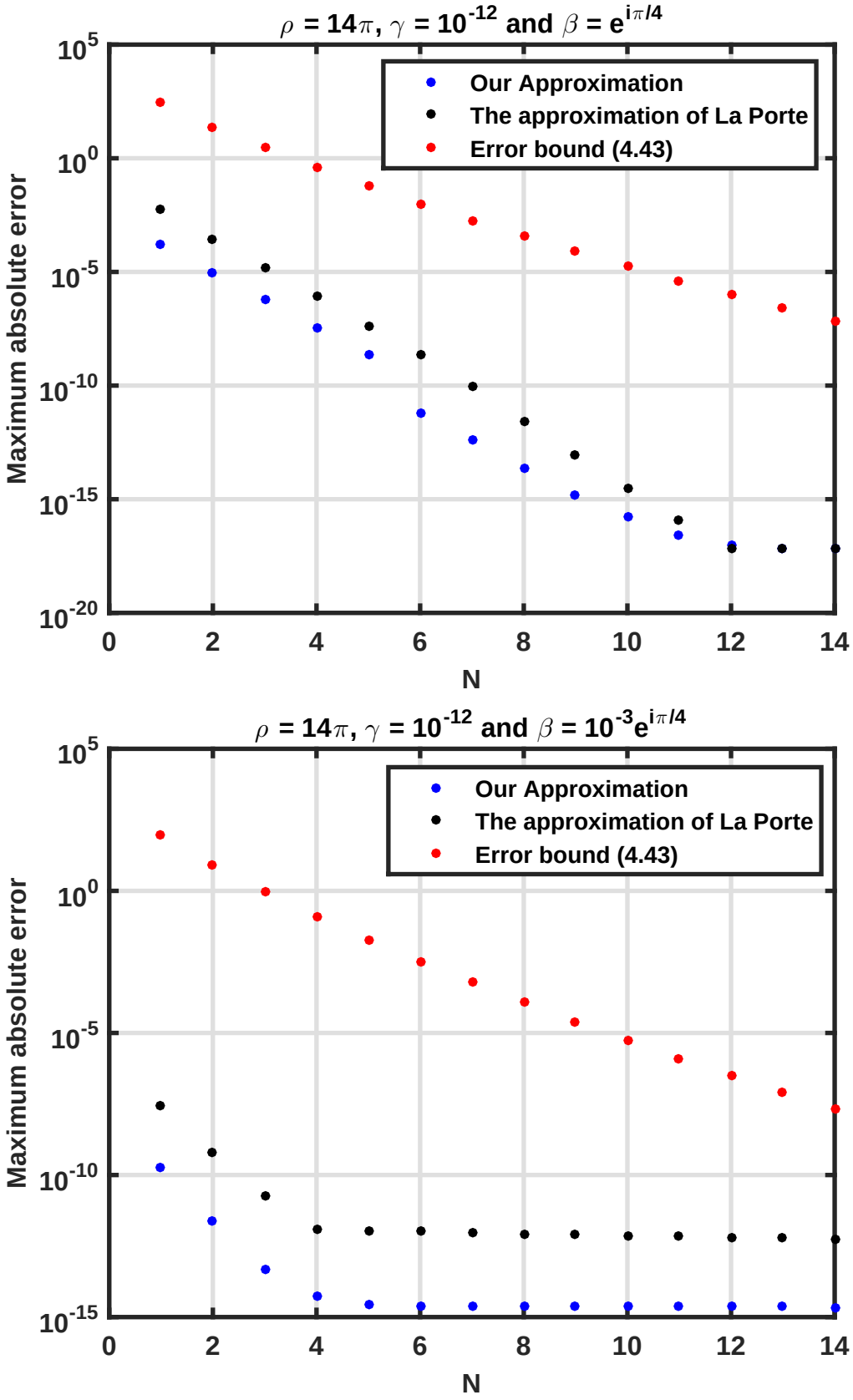

Fig. 4.2 Accuracy of our approximation (4.36) and its upper bound (4.43), as a function of N, in comparison with La Porte's approximation (4.28).

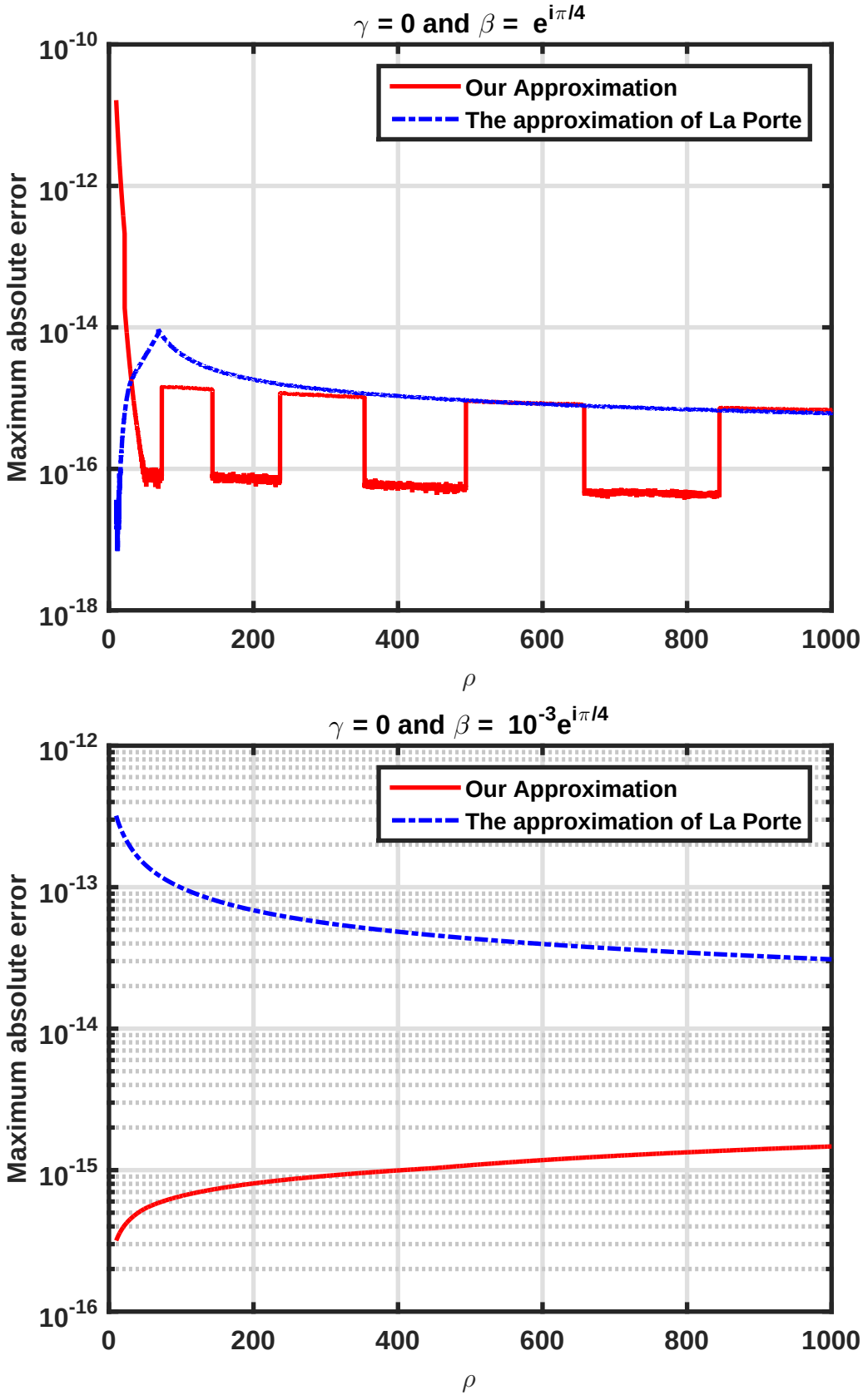

Fig. 4.3 Accuracy of our approximation (4.36), as a function of ρ, in comparison with La Porte's approximation (4.28).

Chapter 5

Concluding remarks and further work

5.1 Concluding Remarks

The objectives of this thesis were

(i) to approximate three special functions, based on the modified trapezium rule (1.18), which can be represented as integrals of the form

$$\int_{-\infty}^{\infty} e^{-\rho t^2} F(t)\, dt, \quad \rho > 0,$$

where F is an even and meromorphic function with simple poles in a strip surrounding the real line;

(ii) to prove uniform bounds on the absolute and relative errors of the proposed approximations;

(iii) and to compare, using numerical calculations, the proposed approximations with the best known ones.

We proposed approximations, building on the work of Chiarella and Reichel [15], Matta and Reichel [43] and Hunter and Regan [31], to the Fresnel integrals in Chapter 2, and the Faddeeva function in Chapter 3 and we proved bounds on the absolute and relative errors. We compared our approximations with the best available ones and we showed, using numerical calculations, that these approximations achieve absolute accuracy of 10^{-15} uniformly on the real line with $N = 12$ (the number of quadrature points) for all the Fresnel integrals, and absolute and relative accuracies of 10^{-16} with $N = 12$ uniformly in the first quadrant of the complex plane for the Faddeeva function.

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

$$\frac{\partial V}{\partial x} = -2\,\mathrm{Re}(z\,w(z)) \quad \text{and} \quad \frac{\partial V}{\partial y} = 2\,\mathrm{Im}(z\,w(z)) - \frac{2}{\sqrt{\pi}}, \tag{5.2}$$

where w is the Faddeeva function given by (3.3). One could explore the accuracy of approximating $V(x,y)$ and its derivatives using the approximation $w_N(z)$ given by (2.9) in Chapter 3.

(ii) In Chapter 4 we proved in Theorem 4.3.5 a theoretical error bound for $|P_\beta - P_{\beta,N}|$. This bound blows up algebraically when $\rho \to \infty$, but the numerical calculations in §4.4 suggest that this bound is not sharp. It is a desired improvement to prove a uniform error bound for all $\rho \geq \rho_0$, for fixed ρ_0.

(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 *Proceedings of a conference organized by the Institute of Mathematics and its Applications on Computers in mathematical research*, pages 207–235. Clarendon Press.

[13] Chandler-Wilde, S. N. and Hothersall, D. (1988b). Propagation from a line source above an impedance plane. *Proceedings of the Institute of Acoustics*, 10:523–531.

[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 I*. 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 $\int_{-\infty}^{\infty} f(x)\mathrm{e}^{-x^2}dx$. *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.

[48] Ossendrijver, M. (2016). Ancient Babylonian astronomers calculated Jupiter's position from the area under a time-velocity graph. *Science*, 351(6272):482–484.

[49] Ouis, D. (2000). Noise shielding by simple barriers: comparison between the performance of spherical and line sound sources. *Journal of Computational Acoustics*, 8(03):495–502.

[50] Poppe, G. and Wijers, C. (1990). More efficient computation of the complex error function. *ACM Transactions on Mathematical Software (TOMS)*, 16(1):38–46.

[51] Premat, E. and Gabillet, Y. (2000). A new boundary-element method for predicting outdoor sound propagation and application to the case of a sound barrier in the presence of downward refraction. *The Journal of the Acoustical Society of America*, 108(6):2775–2783.

[52] Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., et al. (1989). *Numerical recipes*, volume 3. Cambridge University Press, Cambridge.

[53] Schmelzer, T. and Trefethen, L. N. (2007). Computing the gamma function using contour integrals and rational approximations. *SIAM Journal on Numerical Analysis*, 45(2):558–571.

[54] Schreier, F. (2011). Optimized implementations of rational approximations for the Voigt and complex error function. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(6):1010–1025.

[55] Schwartz, C. (1969). Numerical integration of analytic functions. *Journal of Computational Physics*, 4(1):19–29.

[56] Schwartz, C. (2012). Numerical calculation of Bessel functions. *International Journal of Modern Physics C*, 23(12):1250084.

[57] Stenger, F. (1973). Integration formulae based on the trapezoidal formula. *IMA Journal of Applied Mathematics*, 12(1):103–114.

[58] Thomasson, S.-I. (1976). Reflection of waves from a point source by an impedance boundary. *The Journal of the Acoustical Society of America*, 59(4):780–785.

[59] Thomasson, S.-I. (1980). A powerful asymptotic solution for sound propagation above an impedance boundary. *Acta Acustica united with Acustica*, 45(2):122–125.

[60] Trefethen, L. N. and Weideman, J. (2014). The exponentially convergent trapezoidal rule. *SIAM Review*, 56(3):385–458.

[61] Turing, A. M. (1945). A method for the calculation of the zeta-function. *Proceedings of the London Mathematical Society*, 2(1):180–197.

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

[63] Zaghloul, M. R. and Ali, A. N. (2011). Algorithm 916: computing the Faddeyeva and Voigt functions. *ACM Transactions on Mathematical Software (TOMS)*, 38(2):15.

[64] Zaghloul, M. R. and Ali, A. N. (2016). Private communication.

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);
f =  (i/AN)*z.*exp(z2).*S + ez./(ez+1);
```

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.*t; t4 = t2.*t2; et2 = exp(-t2);
x2pi2 = (pi/2)*x.*x; x4 = x2pi2.*x2pi2;
a = et2(1)./(x4+t4(1)); b = t2(1)*a;
for n = 2:N
    term = et2(n)./(x4+t4(n));
    a = a + term;  b = b + t2(n)*term;
end
a = a.*x2pi2;
mx = (rootpi*AN)*x; Mx = (rootpi/AN)*x;
Chalf = 0.5*sign(mx); Shalf = Chalf;
select = abs(mx)<39;
if any(select)
    mxs = mx(select); shx = sinh(mxs); sx = sin(mxs);
    den = 0.5./(cos(mxs)+cosh(mxs));
    Chalf(select) = (shx+sx).*den;
    ssdiff = shx-sx;
    select2 = abs(mxs)<1;
    if any(select2)
        mxs = mxs(select2); mxs3 = mxs.*mxs.*mxs; mxs4 =
            mxs3.*mxs;
        ssdiff(select2) = mxs3.*(1/3 + mxs4.*(1/2520 ...
            + mxs4.*((1/19958400)+(0.001/653837184)*mxs4)));
    end
    Shalf(select) = ssdiff.*den;
end
cx2 = cos(x2pi2); sx2 = sin(x2pi2);
C = Chalf + Mx.*(a.*sx2-b.*cx2); S = Shalf - Mx.*(a.*cx2+b.*
    sx2);
```

A.2 Matlab code to compute Faddeeva function

Listing A.3 Matlab code to evaluate $w_N(z)$ given by (3.21)

```
function f = w(z,N)
h = sqrt(pi./(N+1));
AN = pi./h;
rz = real(z); rzh = rz/h; iz = imag(z);
buff = abs(rzh-floor(rzh)-0.5);
select1 = imag(z) >= max(rz,AN);
select2 =  (iz < rz) & (buff <= 0.25);
select3 = ~(select1|select2);
f = zeros(size(z));
f(select1) = w3(z(select1),N);
f(select2) = w2(z(select2),N);
f(select3) = w1(z(select3),N);
function f = w1(z,N)
a = h*((N:-1:1)+0.5); a2 = a.^2; et2 = exp(-a2); z2 = z.*z;
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);
ez = exp(-2i*AN*z); az = (2i/AN)*z;
PC1 = 2./(exp(z2).*(1+ez));
f = az.*(S0 + S1) + PC1;
    end
function f = w2(z,N)
c = h*(N:-1:1); c2 = c.^2; et3 = exp(-c2); z2 = z.*z;
S2 = et3(1)./(z2-c2(1));
for n = 2 : N
    S2 = S2 + et3(n)./(z2-c2(n));
end
ez = exp(-2i*AN*z); az = (2i/AN)*z; bz = (1i./AN)./z;
PC2 = 2./(exp(z2).*(1-ez));
f = bz + az.*S2 + PC2 ;
```

```
34 end
35 function f = w3(z,N)
36 z2 = z.*z; az = (2i/AN)*z;
37 a = h*((N:-1:1)+0.5); a2 = a.^2; et2 = exp(-a2);
38 S1 = et2(1)./(z2-a2(1));
39 for n = 2 : N
40     S1 = S1 + et2(n)./(z2-a2(n));
41 end
42 h0 = 0.5*h;
43 S0 = exp(-h0.^2)./(z2-h0.^2);
44 f = az.*(S0 + S1);
45 end
46 end
```

A.3 Matlab code to compute P_β

Listing A.4 Matlab code to compute $P_{\beta,N}$ given by (4.36)

```
function P = Pbeta(beta,gamma,rho,N)
ap = 1 + beta.*gamma - sqrt(1-beta.^2).*sqrt(1-gamma.^2);
am = 1 + beta.*gamma + sqrt(1-beta.^2).*sqrt(1-gamma.^2);
z1 = exp(1i*pi/4)*sqrt(ap); z2 = sqrt(1i*am);
AN = sqrt(2*pi*(N+1)./(sqrt(3)*rho)); H = min(0.9,AN);
V1 = beta.*sqrt(1-gamma.^2)-gamma.*sqrt(1-beta.^2);
V2 = sqrt(1i*am).*sqrt(1i*(am-2));
hN1 = sqrt(sqrt(3)*pi./(2*rho*(N+1)));
h1 = nthroot(pi*H./(rho.*(N+1).^2),3);
b = rho^(2/3).*H.^(4/3)./(12*pi^(2/3).*(N+1).^(2/3));
c = sqrt(0.25 + b.^3);
a = nthroot(0.5 + c,3)+ nthroot(0.5 - c,3); hN2 = a.*h1;
if AN <= 0.9
    h = hN1;
else
    h = hN2;
end
function d1 = d(beta)
c1 = (imag(beta)<0) & (real(ap)<0);
c2 = (imag(beta)<0) & (real(ap)==0);
c3 = ~(c1 | c2);
d1(c1)= 2; d1(c2)= 1; d1(c3)= 0;
end
function d2 = dTp(z1)
iz1 = imag(z1);
T1 = iz1 < 0; T2 = iz1 > 0; T3 = ~(T1 | T2);
e1 = exp(-2*1i*pi*z1./h);
d2(T1)= 2*e1./(1-e1); d2(T2)= 2./(1-e1); d2(T3)= (1+e1)./(1-
    e1);
end
function d3 = dMp(z1)
iz1 = imag(z1);
T1 = iz1 < 0; T2 = iz1 > 0; T3 = ~(T1 | T2);
```

```
33 e1 = exp(-2*1i*pi*z1./h);
34 d3(T1)= -2*e1(T1)./(1+e1(T1)); d3(T2)= 2./(1+e1(T2));
35 d3(T3)= (1-e1(T3))./(1+e1(T3));
36 end
37 rz1 = abs(real(z1)); rzh = rz1./h;
38 buff = abs(rzh-floor(rzh)-0.5);
39 if (buff <= 0.25);
40     P = PbetaT(beta,gamma,rho,N);
41 else
42     P = PbetaM(beta,gamma,rho,N);
43 end
44 function f1 = PbetaT(beta,gamma,rho,N)
45 Cp = exp(-1i*rho.*ap).*(dTp(z1).*heaviside(H-abs(imag(z1)))+
       d(beta)) ;
46 if V1 == V2
47     V = 1;
48 else
49     V = -1;
50 end
51 Cm = -2*V*exp(-1i*rho.*am)*heaviside(H-imag(z2))./(1-exp
       (-2*1i*pi*z2./h));
52 TC = pi*(Cp + Cm)./(2*sqrt(1-beta.^2));
53 t = h.*((N:-1:1));t2 = t.^2; et2 = - exp(-t2.*rho);
54 s1 = beta + gamma.*(1+1i*t2); s2 = sqrt(t2-2*1i);
55 s3 = t2-1i*ap; s4 = t2-1i*am;
56 S1 = et2(1).*s1(1)./(s2(1).*s3(1).*s4(1));
57 for n = 2 : N
58     S1 = S1 + et2(n).*s1(n)./(s2(n).*s3(n).*s4(n));
59 end
60 I = (beta.*exp(1i*rho)/pi).*(h./(sqrt(-2*1i)*(beta + gamma))
       + 2*h.*S1);
61 f1 = I + (beta.*exp(1i*rho)/pi).*TC;
62 end
63 function f2 = PbetaM(beta,gamma,rho,N)
64 Cp = exp(-1i*rho.*ap).*(dMp(z1).*heaviside(H-abs(imag(z1)))+
       d(beta));
```

```
65 if V1 == V2
66     V = 1;
67 else
68     V = -1;
69 end
70 Cm = -2*V*exp(-1i*rho.*am).*heaviside(H-imag(z2))./(1+exp
       (-2*1i*pi*z2./h));
71 TC = pi*(Cp + Cm)./(2*sqrt(1-beta.^2));
72 t = h.*((N:-1:1)+ 0.5); t2 = t.^2; h0 = 0.5.*h; et2 = - exp
       (-t2.*rho);
73 s1 = beta + gamma.*(1+1i*t2); s2 = sqrt(t2-2*1i);
74 s3 = t2-1i*ap; s4 = t2-1i*am;
75 S1 = et2(1).*s1(1)./(s2(1).*s3(1).*s4(1));
76 for n = 2 : N
77     S1 = S1 + et2(n).*s1(n)./(s2(n).*s3(n).*s4(n));
78 end
79 A = -(beta + gamma.*(1+1i*h0.^2)).*exp(-rho.*h0.^2);
80 B = sqrt(h0.^2 -2*1i).*(h0^2 - z1.^2).*(h0^2 - z2.^2);
81 I = (beta.*exp(1i*rho)/pi).*2*h.*(S1 + A./B);
82 f2 = I + (beta.*exp(1i*rho)/pi).*TC;
83 end
84 end
```