Nonlinear Approximations in Filter Design and Wave Propagation

by

Ryan D. Lewis

B.S., University of Colorado, Boulder, 2003

M.S., University of Colorado, Boulder, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2013

This thesis entitled:
Nonlinear Approximations in Filter Design and Wave Propagation
written by Ryan D. Lewis
has been approved for the Department of Applied Mathematics

Gregory Beylkin

Bradley Alpert

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Lewis, Ryan D. (Ph.D., Applied Mathematics)

Nonlinear Approximations in Filter Design and Wave Propagation

Thesis directed by Prof. Gregory Beylkin

This thesis has two parts. In both parts we use nonlinear approximations to obtain accurate solutions to problems where traditional numerical approaches rapidly become computationally infeasible.

The first part describes a systematic method for designing highly accurate and efficient infinite impulse response (IIR) and finite impulse response (FIR) filters given their specifications. In our approach, we first meet the specifications by constructing an IIR filter, without requiring the filter to be causal, and possibly with a large number of poles. We then construct, for any given accuracy, an optimal IIR version of such filter. Finally, also for any given accuracy, we convert the IIR filter to an efficient FIR filter cascade. In this FIR approximation, the non-causal part of the IIR filter only introduces an additional delay. Because our IIR construction does not have to enforce causality, the filters we design are more efficient than filters designed by existing methods.

The second part describes a fast algorithm to propagate, for any desired accuracy, a time-harmonic electromagnetic field between two planes separated by free space. The analytic formulation of this problem (circa 1897) requires the evaluation of the Rayleigh-Sommerfeld integral. If the distance between the planes is small, this integral can be accurately evaluated in the Fourier domain; if the distance is large, it can be accurately approximated by asymptotic methods. The computational difficulties arise in the intermediate region where, in order to obtain an accurate solution, it is necessary to apply the oscillatory Rayleigh-Sommerfeld kernel as is. In our approach, we accurately approximate the kernel by a short sum of Gaussians with complex exponents and then efficiently apply the result to input data using the unequally spaced fast Fourier transform. The resulting algorithm has the same computational complexity as methods based on the Fresnel approximation. We demonstrate that while the Fresnel approximation may provide adequate accuracy near the optical axis, the accuracy deteriorates significantly away from the optical axis. In contrast, our method maintains controlled accuracy throughout the entire computational domain.

Dedication

For my father, Kelly D. Lewis, who taught me to have an inquisitive mind and to enjoy the process of discovery.

Acknowledgements

I am grateful to my advisor, Gregory Beylkin, who has been generous with his knowledge, friendship, and patience during the preparation of this thesis. I also want to acknowledge the past and present members of our research group, especially Lucas Monzón, Matt Reynolds, and Dave Biagioni. Brad Alpert from NIST provided many useful suggestions to improve the content and quality of this document.

Finally, thanks to my family for their support and encouragement, and especially to my wife, Ling, and my parents, Kelly and Vernita.

Contents

Chapter

1 Introduction 1

1.1 Filter Design 1

1.1.1 Historical Introduction to Digital Filter Design 1

1.1.2 Preliminaries 3

1.1.3 Contribution of this Thesis 9

1.2 Light Propagation 10

1.2.1 Contribution of this Thesis 14

2 On the Design of Highly Accurate and Efficient IIR and FIR Filters 15

2.1 Introduction 16

2.2 Preliminaries 17

2.2.1 Designing IIR Filters From a Desired Impulse Response 17

2.2.2 Reduction of the Number of Poles 22

2.2.3 Efficient FIR Approximation of IIR Filters 24

2.3 Filling the Gaps 26

2.4 Filter Implementations 33

2.5 Design Examples 36

2.5.1 Frequency Selective Filters 36

2.5.2 Quadrature Mirror Filters 37

2.6 Conclusion 40

3 Fast and Accurate Propagation of Coherent Light 43

3.1 Introduction 44

3.2 Preliminaries 46

3.2.1 The Rayleigh-Sommerfeld Formula 46

3.2.2 Slepian Functions 48

3.2.3 Band-Limiting the Boundary Data 50

3.2.4 Approximation of Functions by Linear Combinations of Exponentials and Gaussians 51

3.2.5 Decompositions of Low-Rank Matrices 52

3.2.6 The Approximations of Fresnel and Fraunhofer 52

3.3 A New Algorithm for Fast and Accurate Light Propagation 53

3.3.1 Approximation of the Kernel with Controlled Error 54

3.3.2 Discretization of Integrals 55

3.3.3 Rapid Evaluation of the Field 56

3.3.4 Computational Cost 60

3.4 Size of the Output Region 60

3.5 Numerical Examples 62

3.5.1 A Gaussian Beam 62

3.5.2 Focusing Waves and the Fresnel Approximation 64

3.5.3 Relationship Between Computational Cost and Propagation Distance 69

3.6 Conclusions 70

3.7 Appendix A: Accurate Propagation in the Fourier Domain 71

3.8 Appendix B: A Comment on the Fraunhofer Approximation 73

3.9 Appendix C: Algorithm for Approximation by Exponential Sums 75

Bibliography 77

Appendix

A Technical Details Concerning Light Propagation 82

A.1 Proof of Lemma 18 82

A.2 Rigorous Estimates Relating to Computational Complexity 83

A.2.1 Number of Terms Needed to Approximate the Kernel 84

A.2.2 Number of Input Samples 85

A.2.3 Number of Terms Needed to Approximate the Tensors 86

List of Tables

Table

2.1 Poles and weights of the lowpass filter H and the number of factors required for each pole in $\widetilde{H}$ 30

2.2 Poles and weights of the "staircase" filter H, and the number of factors required for each pole in $\widetilde{H}$ 37

2.3 Poles, weights, and number of factors required in the FIR approximation $\widetilde{E}_{80}$ of the IIR QMF E_{80} 41

3.1 Computational cost as a function of propagation distance 70

List of Figures

Figure

2.1 Frequency response of the lowpass filters H_d, H, and $\widetilde{H}$ 30

2.2 Poles of the sub-optimal lowpass filter H_d and the equivalent near-optimal filter H 31

2.3 The "staircase" filter H_d and the approximation H 38

2.4 Poles of the sub-optimal "staircase" filter H_d and equivalent near-optimal filter H 38

2.5 Approximate QMF $\widetilde{E}_{80}$ and its perfect reconstruction error 41

3.1 Propagation of boundary data with a Gaussian profile 64

3.2 Propagation of a spherical wave restricted to a square aperture and converging to a point on the optical axis 66

3.3 Comparison of the magnitude of the field evaluated near the focal point on the x-axis 67

3.4 Comparison of the magnitude of the field for a focal point 5° off the optical axis computed by our algorithm and by the Fresnel approximation 68

3.5 Comparison of the error of our method and that of the Fresnel approximation 68

Chapter 1

Introduction

1.1 Filter Design

1.1.1 Historical Introduction to Digital Filter Design

In order to place Chapter 2 (see also [BLM12]) in its proper context, we give a brief historical account of digital filter design methods. In the mid-1960s Charles Rader and his collaborators at Lincoln Labs, including Bernard Gold and Joseph Levin, used discrete signal processing techniques, and recursive filters in particular, to simulate analog telecommunications equipment (vocoders) that could be modeled by a system of ordinary differential equations [Rad06].[1] That experience, and the encouragement of Allen Oppenheim from the Massachusetts Institute of Technology, persuaded Gold and Rader to publish what is considered to be the first monograph devoted to discrete signal processing in 1969 [GR69]. In that book, Gold and Rader describe methods to design digital recursive filters by adapting existing design algorithms for analog filters. The resulting filters have an infinite impulse response, and are therefore known as IIR filters. It is interesting to note that the filter design

[1] Discrete signal processing far predates the digital computer. For example, Sir Isaac Newton described finite difference methods for operating on sequences in his **Principia** in 1687 [New87, Book III]. In the 1750s, Jean d'Alembert, Leonard Euler, Daniel Bernoulli, and Joseph-Louis Lagrange used what today we would call a trigonometric series, a notion intimately connected with discrete signal processing, to study the motion of a vibrating string. Famously, this did not prevent the judges at the Paris Academy, including Lagrange, from giving a lukewarm reception to Jean Baptiste Joseph Fourier's 1807 paper [Fou07] suggesting that an arbitrary function could be decomposed into such a trigonometric series. Two year earlier, in an 1805 unpublished manuscript, Carl Friedrich Gauss derived a version of what is now known as the fast Fourier transform to process samples of the position of the asteroid Pallas [HJB84].

problem they considered was a reformulation of an optimal uniform rational approximation problem solved by E. I. Zolotarjov in 1877 [Zol77].

An important contribution to the development of digital filters occurred in 1971 when Thomas Parks of Rice University, and his student James McClellan, published their famous Parks-McClellan algorithm for the optimal design of polynomial (i.e., non-recursive) digital filters [PM72]. Such filters have a finite impulse response and are therefore known as FIR filters. In effect, their filter design method specialized the so-called Remez algorithm for optimal uniform polynomial approximation to the filter design problem (see, e.g., [Rem69]). Later, Larry Rabiner from Bell Labs conducted extensive tests on the algorithm, and consequently it is often called the Parks-McClellan-Rabiner algorithm [MP05]. This polynomial approximation approach marked a major departure from the rational approximation approach of Gold and Rader in [GR69].

Since 1971, many new algorithms to construct FIR and IIR filters have appeared, but it is probably correct to say that the filters most commonly used in applications are FIR filters designed using the Parks-McClellan-Rabiner algorithm. At least one reason for this is the availability of a reliable and efficient computer implementation of the algorithm [Dig79]. A second reason, perhaps equally important, is that the trade-offs engendered by certain practical considerations—viz., the requirements of causality, stability, and a linear phase—often make FIR filters appear more desirable than IIR filters (see §1.1.3).

In Chapter 2 we first show that the practical design constraints mentioned above (causality, stability, and a linear phase) can be relaxed in an essential way. This eliminates the factors that had made IIR filters appear less desirable than FIR filters, and therefore significantly alters the trade-offs between the two types of filters. Next, in situations where an FIR filter is desirable, we show how to convert an IIR filter into an equivalent and efficient FIR filter. By combining these observations with a recently-discovered algorithm for near-optimal rational approximation [BM05], we obtain a new filter design algorithm and use it to construct FIR filters that are considerably more computationally efficient than their

equivalent "optimal" Parks-McClellan-Rabiner counterparts.

1.1.2 Preliminaries

This section may be skipped by readers already familiar with discrete signal processing. We collect results and terminology used in Chapter 2. The material presented here may be found in any introductory discrete signal processing textbook (e.g., [OS89]).

A discrete signal $x \in \ell_\infty(\mathbb{Z})$, $x(n) \in \mathbb{C}$, for $n = \ldots, -1, 0, 1, \ldots$, is a bounded, bi-infinite sequence of complex numbers. From now on, we will will drop the term "discrete" and simply write "signal." We will also refer to the index n as "time," although this is not really correct since x may represent, e.g., a spatial process. For our purposes, we define a filter T, $T : \ell_\infty \to \ell_\infty$, as a bounded, linear, shift-invariant operator defined on the space of signals, i.e., there exists some $M \geq 0$ such that for all $x, y \in \ell_\infty$, $\alpha, \beta \in \mathbb{C}$, and $k \in \mathbb{Z}$,

$$\|Tx\|_\infty \leq M \|x\|_\infty \quad \text{(bounded)}$$

$$T(\alpha x + \beta y) = \alpha Tx + \beta Ty \quad \text{(linear)}$$

$$Tx(\cdot - k) = (Tx)(\cdot - k) \quad \text{(shift-invariant)}.$$

Although there are other interesting classes of operators, the class of bounded, linear, shift-invariant operators is by far the most common, and is the only class we consider here. As is well known, the action of all such operators on a signal x may be expressed as a convolution,

$$Tx(n) = \sum_{k=-\infty}^{\infty} h(k)\, x(n-k), \tag{1.1}$$

for some bi-infinite complex sequence $h \in \ell_1$, which is called the impulse response of the operator T. For the remainder of this section, we will call such an operator T a filter.

The goal of a filter design problem is to construct a filter T possessing certain desired properties. Generally, these properties are of two distinct classes:

(1) Properties concerning the action T has on input signals. For example, we may require the Fourier transform (defined below) of the sequence h associated with the filter to

possess certain desirable characteristics, such as attenuating some frequencies while leaving other frequencies undisturbed.

(2) Properties concerning the computational cost of applying T to an input signal x.

Often, these two classes of properties conflict, and it is necessary to strike a balance between accuracy (i.e., properties of the first class) and efficiency (i.e., properties of the second class). Optimizing this balance is one of the most important aspects of typical filter design problems.

1.1.2.1 The z-Transform and the Discrete-Time Fourier Transform

Given a complex bi-infinite sequence g, we define its z-transform by the formal series

$$G(z) = \sum_{n=-\infty}^{\infty} g(n) z^{-n}, \quad z \in \mathbb{C}. \tag{1.2}$$

An important special case of the z-transform is the discrete-time Fourier transform, defined by the formal series

$$G\left(e^{i\omega}\right) = \sum_{n=-\infty}^{\infty} g(n) e^{-i\omega n}, \quad \omega \in \mathbb{R}, \tag{1.3}$$

which is simply $G(z)$ evaluated on the unit circle. When there is no risk of confusion between the continuous-time Fourier transform of a function and the discrete-time Fourier transform of a sequence, we will follow convention and simply call the 2π-periodic function $G(e^{i\omega})$ the Fourier transform of the sequence g.

Let us now consider the convergence of (1.2). If g is the impulse response of a filter G, then (1.2) converges absolutely for all z on the unit circle (since we have stipulated that G operates on ℓ_∞ and therefore $g \in \ell_1$).[2] Because $g \in \ell_1$, we may recover the impulse response g from its Fourier transform by the inversion formula

$$g(n) = \frac{1}{2\pi} \int_0^{2\pi} G\left(e^{i\omega}\right) e^{i\omega n} \, d\omega. \tag{1.4}$$

[2] In certain applications, it is more appropriate to consider signals as elements of ℓ_2, and therefore the impulse responses of bounded, linear, shift-invariant operators on such signals are also elements of ℓ_2.

It is common to refer to the filter as "the filter G," to the sequence g as "the impulse response of G," to the function $G(z)$ as "the transfer function of G," and to the Fourier transform $G(e^{i\omega})$ as "the frequency response of G."

The convergence of (1.2) and (1.3) is more complicated if $g \notin \ell_1$. If we now assume that g is a signal, then $g \in \ell_\infty$ and its Fourier transform (1.3) is a generalized function (for a suitable space of test functions—the precise choice does not matter here). We will never attempt to compute the value of a Fourier transform of a sequence $g \in \ell_\infty$, and so we will not need the full machinery of generalized functions. Therefore, we omit further details and simply state two facts that are sufficient for our needs: if $g \in \ell_\infty$, then

(1) the sequence g may be recovered from its Fourier transform, the generalized function $G(e^{i\omega})$, by a proper interpretation of (1.4), and

(2) if $g \in \ell_\infty$ is a signal and $h \in \ell_1$ is the impulse response of a filter, and $G(e^{i\omega})$ and $H(e^{i\omega})$ are their respective Fourier transforms, then the inverse Fourier transform of the product $G(e^{i\omega})\,H(e^{i\omega})$ is the convolution

$$\sum_{k=-\infty}^{\infty} g(n-k)\,h(k), \quad n \in \mathbb{Z},$$

which is simply the result of applying the filter H to the signal g.

1.1.2.2 Classes of Filters

Given a filter H with impulse response h, where by definition, $h \in \ell_1$, we say that H is a finite impulse response (FIR) filter if h has a finite number of nonzero elements. Otherwise, H is an infinite impulse response (IIR) filter. The distinction is important: if H is FIR, then we can apply it to an input sequence x by a direct evaluation of the convolution (1.1). On the other hand, if H is IIR, then we must use a more sophisticated method to apply it to input sequences. Observing that $h \in \ell_1$ implies that $|h(n)| \to 0$ as $|n| \to \infty$, it may appear tempting to truncate the infinite sum (1.1) once the elements of h become

sufficiently small, essentially discarding the small elements of h to construct a new FIR filter $\widetilde{H}$ with impulse response $\widetilde{h}$. However, a direct implementation of this strategy is generally not advisable, since the decay of h will be quite slow when $H\left(e^{i\omega}\right)$ contains sharp transitions or is highly peaked, as is often the case in applications. In such cases, either the truncation error resulting from discarding the small elements of h will be unacceptably large, so that $\widetilde{H}$ is a poor approximation of H, or so many elements must be retained that the calculation of the convolution (1.1) using $\widetilde{h}$ will be computationally prohibitive. We note that one component of our new filter design method may be interpreted as a technique to truncate the impulse response of an IIR filter in such a way that the resulting FIR filter is both accurate and computationally efficient.

Let us distinguish between two other important classes of filters. We say that a filter H is causal if $h\left(n\right) = 0$ for all $n < 0$, and noncausal otherwise. If H is causal, then the output signal y that results from applying H to an input signal x may be written as

$$y\left(n\right) = Hx\left(n\right) = \sum_{k=0}^{\infty} x\left(n-k\right) h\left(k\right),$$

so that the n-th element of y only depends on elements of x that have arrived by time n. In many applications, such as telecommunications, causality is essential, though there exist some applications, e.g., those involving spatial data, where noncausal filters are acceptable. Observe that if H is simply a right-sided sequence, meaning that there exists some N such that $h\left(n\right) = 0$ for all $n < N$, then if H is noncausal (i.e., $N < 0$) we can obtain a causal filter H' from H by setting $h'\left(n\right) = h\left(n-N\right)$. It is straightforward to verify that the z-transforms of H and H' are related by $H'\left(z\right) = z^{-N} H\left(z\right)$. Such a transformation is called "introducing a pure delay." Clearly, any FIR filter can be transformed into a causal FIR filter by the introduction of such a delay.

Finally, we define the concept of a linear phase filter. If we factor the Fourier transform of a filter H as

$$H\left(e^{i\omega}\right) = A\left(\omega\right) e^{i\phi(\omega)},$$

where $A(\omega)$ is a real-valued function and $\phi(\omega)$ is a real-valued continuous function on the interval $\omega \in (-\pi, \pi]$, then we say that H has a linear phase if $\phi(\omega) = -\alpha\omega + \beta$, for some $\alpha, \beta \in \mathbb{R}$, and a nonlinear phase otherwise. (In fact, we have given the condition for a generalized linear phase, which is often more useful in practice.) If a filter has a nonlinear phase, then it is dispersive. Dispersion is undesirable in many applications, and it is therefore common to require a filter to have a linear, or at least approximately linear, phase. For example, the presence of dispersion in a telecommunications system will distort voice quality and must therefore be controlled.

1.1.2.3 Rational Filters

We now restrict our attention to filters whose z-transforms are rational functions of z, continuous on the unit circle,

$$H(z) = \frac{P(z)}{Q(z)} = \frac{\sum_{k=0}^{N_P} p_k z^{-k}}{\sum_{k=0}^{N_Q} q_k z^{-k}}, \qquad Q\left(e^{i\omega}\right) \neq 0. \tag{1.5}$$

This form of filter is by far the most common, though other useful forms exist (e.g., splines are not of this form). The roots of P are called zeros of the filter H, and the roots of Q are called the poles of H. We assume for simplicity that $N_P = N_Q$, that all the roots of Q are distinct, and that P and Q have no common factors. Expand (1.5) in partial fractions to obtain

$$H(z) = c_0 + \sum_{k=1}^{N_Q} \frac{s_k}{1 - \gamma_k / z}.$$

By assumption, $|\gamma_k| \neq 1$ since Q does not vanish on the unit circle, so we may partition the poles γ_k into the sets γ_k^{in}, $k = 1, \ldots, M^{\text{in}}$, and γ_k^{out}, $k = 1, \ldots, M^{\text{out}}$, where $\left|\gamma_k^{\text{in}}\right| < 1$ and $|\gamma_k^{\text{out}}| > 1$. We now multiply terms involving poles outside the unit circle by z/z and rewrite them as

$$\frac{s_k^{\text{out}}}{1 - \gamma_k^{\text{out}} / z} = s_k^{\text{out}} + \frac{-s_k^{\text{out}}}{1 - z/\gamma_k^{\text{out}}}.$$

This transformation introduces an extra pole-zero pair at the origin, but does not disturb the value of $H(z)$ on the unit circle. With this rearrangement, the rational filter H may be

written as

$$H(z) = c_0' + \sum_{k=1}^{M^{\text{in}}} \frac{w_k^{\text{in}}}{1 - \gamma_k^{\text{in}}/z} + \sum_{m=1}^{M^{\text{out}}} \frac{w_k^{\text{out}}}{1 - z/\gamma_k^{\text{out}}}, \tag{1.6}$$

where $c_0' = c_0 + s_1^{\text{out}} + \cdots s_{M^{\text{out}}}^{\text{out}}$ and $w_k^{\text{out}} = -s_k^{\text{out}}$. The impulse response of H is

$$h(n) = \begin{cases} c_0' + \sum_{k=1}^{M^{\text{in}}} w_k^{\text{in}} + \sum_{k=1}^{M^{\text{out}}} w_k^{\text{out}}, & \text{if } n = 0, \\ \sum_{k=1}^{M^{\text{in}}} w_k^{\text{in}} \left(\gamma_k^{\text{in}}\right)^n, & \text{if } n > 0, \\ \sum_{k=1}^{M^{\text{out}}} w_k^{\text{out}} \left(\gamma_k^{\text{out}}\right)^n, & \text{if } n < 0. \end{cases} \tag{1.7}$$

We see that H is, in general, a noncausal IIR filter and that its impulse response h is a linear combination of decaying exponentials. The filters we study in Chapter 2 have the form (1.6).

Let us apply the rational filter H in (1.5) to an input signal x to obtain the output signal y,

$$y(n) = Hx(n) = \sum_{k=-\infty}^{\infty} x(k)\, h(n-k),$$

and then compute the z-transform of both sides of this equation, yielding

$$Y(z) = X(z) \frac{P(z)}{Q(z)}$$

or, for z on the unit circle,

$$Y(z)\, Q(z) = X(z)\, P(z), \quad |z| = 1, \tag{1.8}$$

where it is understood that X and Y are generalized functions. Apply the formula for the inverse Fourier transform (1.4) to (1.8) to obtain

$$\sum_{k=0}^{N_Q} q_k y(n-k) = \sum_{k=0}^{N_P} p_k x(n-k), \quad n \in \mathbb{Z}, \tag{1.9}$$

which expresses the output y as the solution of an order N_Q constant coefficient difference equation. This formula explains the interest in rational filters: with an appropriate set of initial conditions, we may run the recursion forwards (or occasionally backwards) in time to efficiently compute the output y resulting from applying an IIR filter to an input signal x—provided the recursion is stable in the direction we select.

Finally, observe that if $Q(z) = 1$, then (1.9) simplifies to

$$y(n) = \sum_{k=0}^{N_P} p_k x(n-k),$$

so that H is simply a polynomial, and is therefore a causal FIR filter.

1.1.3 Contribution of this Thesis

From the preceding discussion in §1.1.2, it is clear that IIR and FIR filter design requires constructing either a rational or polynomial function $H(z)$, respectively, to satisfy certain design constraints. It would seem that IIR filters have a clear advantage, since it is well known that a complicated function can be approximated by a rational function using far fewer terms than would be required by an equivalent approximating polynomial. On the other hand, a rational filter by definition has an infinite impulse response, and so it would appear that we must use the recursive formula (1.9) to apply such a filter to input signals. Assuming that we want to to compute the recursion in the forward direction, then for this recurrence to be stable we must require that all the roots of Q lie inside the unit circle. But this restriction has a severe side effect: such a filter cannot have a linear phase! In fact, it can be shown that a necessary condition for a filter to have a linear phase is that its impulse response $h(n)$ be symmetric about some value $n = s$, which is clearly impossible if all the roots of Q lie in the unit circle (cf. (1.7)). (Technically, the point of symmetry s need not be an integer—observe that (1.4) is defined for all real n.) Moreover, classical results from approximation theory show that the best uniform rational approximation to many important functions requires poles both inside and outside the unit circle. For example, in 1877 Zolotarjov [Zol77] used the Jacobi elliptic functions to construct the best uniform rational approximation on the interval $x \in [-1, 1]$ to the function

$$f(x) = \begin{cases} 1, & \text{if } |x| < t - \epsilon, \\ \\ 0, & \text{if } |x| > t + \epsilon, \end{cases}$$

where the parameters t and ϵ satisfy $t < 1$ and $\epsilon < \min(t, 1-t)$, and the behavior of f in the intervals $|x| \in [t-\epsilon, t+\epsilon]$ is unspecified. The resulting optimal rational approximation of this function, which corresponds to a lowpass filter with linear phase, is known explicitly and requires poles both inside and outside the unit circle.

Roughly speaking, this state of affairs could be summarized as follows: many of the results from the theory of optimal rational approximation could not be directly applied to filter design problems because, typically, the resulting filters would contain poles outside the unit circle, and therefore could not be applied to input signals in a stable manner. If a rational filter were designed with all its poles inside the unit circle, then it would usually be sub-optimal and could not posses a linear phase. The algorithm we describe in Chapter 2 alters this state of affairs. Our key observation is that terms in the partial fraction expansion (1.6) corresponding to poles outside the unit circle may be accurately transformed into polynomials. Although these polynomials have a high degree, they are computationally efficient because the number of operations needed to apply them to input signals is proportional to the logarithm of their degree. Therefore, we are free to design IIR filters with poles both inside and outside the unit circle. We then use this new freedom to adapt recent results in near-optimal rational approximation [BM05] to the filter design problem. Since we do not constrain the poles of our IIR filters to be inside the unit circle, we obtain a simple method for constructing linear phase filters if the specifications so require. The result is a new method to construct highly accurate and efficient digital filters.

1.2 Light Propagation

In Chapter 3 of this thesis (see also [LBM13]), we describe an algorithm to propagate, for any user-specified accuracy, a time-harmonic electromagnetic field between two parallel planes separated by a linear, isotropic, and homogeneous medium. As is well known, in such

a medium Maxwell's equations simplify to the scalar Helmholtz equation,

$$\left(\Delta + k^2\right) u = 0, \tag{1.10}$$

where the wavenumber $k = 2\pi/\lambda$, λ is the wavelength, and $u(\mathbf{x}, z)$ is the complex amplitude of one component of the vector-valued electric field at a point $(\mathbf{x}, z) \in \mathbb{R}^3$. For simplicity, we measure all distances in wavelengths and therefore set $\lambda = 1$.

In 1897 Lord Rayleigh described a formula for a solution u of (1.10). Given the boundary data

$$u(\mathbf{x}, 0) = f(\mathbf{x}),$$

Rayleigh's formula expresses u as the convolution

$$u(\mathbf{x}, z) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} f(\mathbf{y}) \frac{\partial}{\partial z} \left(\frac{e^{i2\pi R}}{R} \right) d\mathbf{y}, \quad z > 0, \tag{1.11}$$

where

$$R = \sqrt{z^2 + \|\mathbf{x} - \mathbf{y}\|^2}$$

[Ray97, Bou54, BWB99]. This formula is surprising because (1.10) is a second order PDE, yet the solution to the boundary value problem appears only to depend on the value of u on the boundary, and not on its normal derivative. In fact, (1.10) has two solutions—one solution consisting of outgoing waves, and the other solution consisting of incoming waves. In many physical situations, including the one we study here, the incoming solution may be rejected on physical grounds. It turns out that Rayleigh's formula (1.11) gives the outgoing solution, which was demonstrated in 1912 by Arnold Sommerfeld when he introduced his celebrated radiation condition. This condition states that a necessary and sufficient condition for a solution u of (1.10) to be an outgoing solution is that it satisfy [Som12, Som49]

$$\lim_{s \to \infty} s \left(\frac{\partial u}{\partial s} - i2\pi u \right) = 0, \quad \text{where } s = \|(\mathbf{x}, z)\| \text{ and } z > 0.$$

It can be shown that Rayleigh's solution (1.11) satisfies this condition, and therefore (1.11) is known as the Rayleigh-Sommerfeld formula.

There are only a few examples of boundary data f where (1.11) can be evaluated analytically, so it is usually necessary to employ numerical methods to compute u. It turns out to be difficult to construct a numerical scheme to accurately and efficiently evaluate (1.11) in a large region of space. To see why, we rewrite (1.11) as

$$u(\mathbf{x}, z) = \int_{\mathbb{R}^2} f(\mathbf{y})\, K_z(\|\mathbf{x}-\mathbf{y}\|)\, d\mathbf{y}, \tag{1.12}$$

where the Rayleigh-Sommerfeld kernel $K_z(r)$ is given by

$$K_z(r) = \frac{e^{i2\pi z\sqrt{1+(r/z)^2}}}{iz}\left(\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z\left(1+(r/z)^2\right)^{\frac{3}{2}}}\right), \quad r \geq 0. \tag{1.13}$$

Denoting the Fourier transform of the boundary data as

$$\widehat{f}(\mathbf{p}) = \int_{\mathbb{R}^2} f(\mathbf{x})\, e^{-i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{x},$$

we write (1.12) via an integral in the Fourier domain,

$$u(\mathbf{x}, z) = \int_{\mathbb{R}^2} \widehat{f}(\mathbf{p})\, \widehat{K}_z(\|\mathbf{p}\|)\, e^{i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p}, \tag{1.14}$$

where the Fourier transform of the Rayleigh-Sommerfeld kernel is given by (see [She67] and references therein)

$$\widehat{K}_z(\rho) = e^{i2\pi z\sqrt{1-\rho^2}}, \quad \rho \geq 0. \tag{1.15}$$

It is clear that $K_z(r)$ is a highly oscillatory function of r when z is small and that $\widehat{K}_z(\rho)$ is a highly oscillatory function of ρ when z is large. For many physically interesting choices of the distance z in the intermediate region, $K_z(r)$ and $\widehat{K}_z(\rho)$ are both highly oscillatory, making the direct numerical computation of u using either (1.12) or (1.14) impractical.

Let us mention a few existing numerical schemes to compute u. One popular method [Syp95] uses (1.14) to propagate the field via calculations in the Fourier domain. To eliminate the rapid oscillations of the kernel $\widehat{K}_z$, the total propagation distance z is divided into M smaller steps, $z = z_1 + \cdots + z_M$, and a zero-padding operation is performed in the spatial

domain between each propagation step. This has the effect of attenuating the rapidly-oscillating portions of the solution so that the computational problem remains tractable, but it has the numerical side effect of introducing $M-1$ artificial apertures into the problem. Each of these apertures introduces diffraction artifacts that combine to significantly degrade the final accuracy of the computed solution. An alternative Fourier propagation method [MS09] does the propagation in a single step, but first uses a heuristic argument based on geometric optics to lowpass filter the boundary data prior to propagation. This method resolves many of the computational difficulties associated with the oscillatory kernel $\widehat{K}_z$; however, because of the lowpass filtering operation, it offers only limited accuracy.

A different class of methods attempt to compute u using (1.12), i.e., by evaluating an integral in the spatial domain. The most straightforward of these is to replace the Rayleigh-Sommerfeld kernel K_z by the Fresnel approximation,

$$K_z\left(\|\mathbf{x}-\mathbf{y}\|\right) \approx F_z\left(\mathbf{x},\mathbf{y}\right) = \frac{e^{i2\pi z}e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz}e^{i\frac{\pi}{z}\|\mathbf{y}\|^2}e^{-i\frac{2\pi}{z}\mathbf{x}\cdot\mathbf{y}},$$

so that (1.12) becomes

$$u\left(\mathbf{x},z\right) \approx \frac{e^{i2\pi z}e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz}\int\limits_{\mathbb{R}^2} f\left(\mathbf{y}\right)e^{i\frac{\pi}{z}\|\mathbf{y}\|^2}e^{-i\frac{2\pi}{z}\mathbf{x}\cdot\mathbf{y}}\,d\mathbf{y}, \tag{1.16}$$

which can be evaluated using the discrete Fourier transform (see, e.g., [Goo05]). Unfortunately, the Fresnel approximation is only accurate for points $\mathbf{x}$ close to the optical axis, and the accuracy of (1.16) degrades rapidly as $\mathbf{x}$ moves away from the optical axis. In [SW06], Shen and Wang describe a method to discretize the integral (1.12) using a quadrature formula and then evaluate the resulting summation using the discrete Fourier transform. However, for their method to maintain accuracy, many samples of the boundary data f are required if the desired output region is not very close to the optical axis, so the computational cost of their method is prohibitive when seeking an accurate solution in a large output region.

1.2.1 Contribution of this Thesis

In Chapter 3 we present a fast algorithm to evaluate, for any user-specified accuracy, the Rayleigh-Sommerfeld integral (1.12) in the spatial domain. In our approach, for a given accuracy $\epsilon > 0$, we approximate the kernel K_z by a short sum of Gaussians with complex exponents. The resulting approximate kernel is then efficiently applied to input data using the unequally spaced fast Fourier transform [DR93, Bey95b].

Our approach may be viewed as a generalization of the Fresnel approximation. While the Fresnel approximation replaces the Rayleigh-Sommerfeld kernel with a single Gaussian with a purely imaginary exponent, we use a nonlinear algorithm to approximate the kernel, for any user-selected accuracy, as a short linear combination of Gaussians with complex exponents.

We demonstrate that while the Fresnel approximation, as it is currently used, may provide adequate accuracy near the optical axis, the accuracy deteriorates significantly away from the optical axis. In contrast, our method maintains controlled accuracy throughout the entire computational domain of interest.

Chapter 2

On the Design of Highly Accurate and Efficient IIR and FIR Filters

This chapter contains a reprint of

[BLM12] G. Beylkin, R. D. Lewis, and L. Monzón, **On the design of highly accurate and efficient IIR and FIR filters**, IEEE Trans. Signal Process. **60** (2012), no. 8, 4045–4054.

ON THE DESIGN OF HIGHLY ACCURATE AND EFFICIENT IIR AND FIR FILTERS

GREGORY BEYLKIN, RYAN D. LEWIS, AND LUCAS MONZÓN

Abstract. We describe a systematic method for designing highly accurate and efficient infinite impulse response (IIR) and finite impulse response (FIR) filters given their specifications. In our approach, we first meet the specifications by constructing an IIR filter with, possibly, a large number of poles. We then construct, for any given accuracy, an optimal IIR version of such filter (with a minimal number of poles). Finally, also for any given accuracy, we convert the IIR filter to an efficient FIR filter cascade (either serial or parallel). Since in this FIR approximation the non-causal part of the IIR filter only introduces an additional delay (as a function of the desired accuracy), our IIR construction does not have to enforce causality. Thus, we obtain a simple method for constructing linear phase filters if the specifications so require. All of these procedures are accomplished via robust, fast algorithms. We provide several illustrative examples of our method.

2.1 Introduction

In his 2006 paper "The Rise and Fall of Recursive Digital Filters," [Rad06] Rader gives a brief history of filter design methods. He describes how the perceived pros and cons of recursive and non-recursive filters changed over time as new design and implementation techniques were discovered. The goal of our paper is to offer an addendum to this history by providing a new systematic method of designing both types of filters. Our approach is based on a combination of several approximation algorithms and a few observations. We cite algorithms for constructing near-optimal rational approximations [BM05, BM09], a new high accuracy reduction algorithm [HB12], and a somewhat obscure short note [Bey95a]. Our key observation is that it is relatively easy to construct an accurate but sub-optimal (with a large number of poles) rational filter that satisfies the design criteria. We describe an effective approach for the sub-optimal construction well suited for the optimization algorithm. We then rely on robust nonlinear algorithms for optimal rational approximation to minimize the number of poles for a desired accuracy.

We first construct an infinite impulse response (IIR) filter that satisfies the design criteria without attempting to make the design optimal. We next find an equivalent (vis-à-vis the specifications) IIR filter with a near-minimal number of poles. We then convert, for any given accuracy, the IIR filter to an efficient finite impulse response (FIR) filter. It is well known that approximating a rational function with a polynomial for a set accuracy may require a polynomial of high degree. Despite the high degree of our FIR filter, its implementation cost is low and requires only $\mathcal{O}\left(\log \epsilon^{-1}\right)$ operations, where ϵ is the desired accuracy. This efficiency is achieved by expressing the FIR filter as a cascade (either serial or parallel) where each factor is computationally inexpensive. Importantly for the many applications that require linear phase filters, we may easily design IIR filters with exact linear phase. In our method, the non-causal part of the IIR filter results in a finite delay in the FIR approximation that does not disturb the phase of the filter.

The combination of these design steps leads to a robust, nearly automatic, method for filter design. We believe that our approach contributes to the state-of-the-art of filter design as summarized in the conclusion of Rader's paper.

2.2 Preliminaries

In this section, we introduce notation and present the algorithms used in our filter design method. Given a filter, we identify its impulse response $h(n)$ with its z-transform,

$$H(z) = \sum_{n=-\infty}^{\infty} h(n) z^{-n}, \tag{2.1}$$

where the sum in (2.1) converges on the unit circle. To recall, if $h(n)$ contains only a finite number of nonzero terms, then $H(z)$ is an FIR filter. Otherwise, $H(z)$ is an IIR filter. If $h(n) = 0$ for all $n < 0$ then $H(z)$ is causal (and non-causal otherwise).

We introduce two filter design algorithms whose origins may be traced to the work of Adamjan, Arov, and Krein (AAK theory) [AAK68a, AAK68b, AAK71]. The algorithm in §2.2.1 is often adequate but may require extended precision arithmetic for intermediate computations. The reduction algorithm in §2.2.2 (see [BM05, BM10]) is significantly more efficient and its new version in [HB12] achieves high accuracy using only the standard double precision arithmetic. Finally, following [Bey95a], we describe an algorithm to convert IIR filters to efficient FIR filters while maintaining arbitrary finite accuracy.

2.2.1 Designing IIR Filters From a Desired Impulse Response

Our first algorithm constructs an IIR filter $H(z)$ whose impulse response $h(n)$ agrees with some desired impulse response $h_d(n)$, up to some finite but arbitrary accuracy $\epsilon > 0$ over a certain range of the index $n \in \mathbb{Z}$.

Our solution makes use of an algorithm in [BM05, BM10]. Given a sequence

$$h_d(n), \quad 1 \le n \le 2N + 1$$

and a target accuracy $\epsilon > 0$, we determine the optimal (minimal) number of nodes γ_m and weights w_m such that

$$\left| h_d(n) - \sum_{m=1}^{M} w_m \gamma_m^n \right| < \epsilon, \quad 1 \le n \le 2N+1. \tag{2.2}$$

We now describe the steps of the algorithm to obtain this approximation.

Algorithm 1:

- Build the $N+1 \times N+1$ Hankel matrix

$$\mathbf{H}_{k\ell} = h_d(k+\ell+1), \quad k, \ell \in [0, N]. \tag{2.3}$$

- Find a vector $\mathbf{u} = (u_0, \ldots, u_N)^T$ satisfying

$$\mathbf{H}\mathbf{u} = \sigma \overline{\mathbf{u}}, \tag{2.4}$$

  with positive σ close to the target accuracy ϵ, where $\overline{\mathbf{u}} = (\overline{u}_0, \ldots, \overline{u}_N)^T$ denotes the element-wise complex conjugate of the vector $\mathbf{u}$. A problem of this form is known as a con-eigenvalue problem (see, e.g., [HJ90, §4.6]), $\mathbf{u}$ is a con-eigenvector, and σ is a con-eigenvalue. In our case, $\mathbf{H}$ is a Hankel matrix and hence symmetric; the existence of a solution $(\sigma, \mathbf{u})$ follows from Takagi's factorization (see, e.g., [BM05, pp. 22]), as does the fact that we may take σ to be a singular value of $\mathbf{H}$ and $\mathbf{u}$ to be a specific singular vector.

- Given singular values $\sigma_0 \ge \sigma_1 \ge \ldots \ge \sigma_N$, we select a sufficiently small σ_M, which determines the accuracy of approximation, and the corresponding singular vector $\mathbf{u} = (u_0, \ldots, u_N)^T$.

- Compute the roots γ_m of the con-eigenpolynomial $u(z) = \sum_{n=0}^{N} u_n z^n$ whose coefficients are the entries of the vector $\mathbf{u}$ from the previous step.

- Obtain the weights w_m by solving the least-squares Vandermonde system

$$\sum_{m=1}^{N} w_m \gamma_m^n = h_d(n), \quad 1 \le n \le 2N+1. \tag{2.5}$$

Typically, only M weights w_m have absolute value larger than the target accuracy ϵ. We then retain only those nodes γ_m that correspond to the significant weights and solve the corresponding Vandermonde system (2.5) again. For cases of practical interest in digital filtering, the sequence $h_d(n)$ exhibits decay as n becomes large. As a result, the nodes of interest lie inside the unit disk, $|\gamma_m| < 1$.

Remark 1.

- Typically, singular values decay rapidly so the number of terms M in the approximation (2.2) satisfies $M = \mathcal{O}(\log \epsilon^{-1})$.

- To approximate a sequence

$$h_d(n), \quad -2N - 1 \leq n \leq -1,$$

by a sum

$$h_d(n) \approx \sum_{m=1}^{M} w_m \gamma_{m,}^n \quad -2N - 1 \leq n \leq -1,$$

we simply re-index $n \mapsto -n$ and use Algorithm 1. In this case the nodes γ_m lie outside the unit disk, $|\gamma_m| > 1$, provided $h_d(n)$ decays as n becomes large and negative.

- We note that we may formulate this algorithm in terms of the singular value decomposition (SVD) without invoking the con-eigenvalue problem. However, Algorithm 3, which may be derived from Algorithm 1, requires this formulation. For a detailed analysis we refer to [BM05, BM10].

- The nodes γ_m turn out to be the poles of the transfer function $H(z)$ that we construct via the next algorithm.

Let us now describe how to use Algorithm 1 to solve a filter design problem. Given a desired impulse response $h_d(n)$ for $n \in [-N_2, N_1]$ and target accuracy $\epsilon > 0$, construct an IIR filter

$H(z)$ with a (nearly) minimal number of poles whose impulse response $h(n)$ satisfies

$$|h_d(n) - h(n)| < \epsilon, \quad n \in [-N_2, N_1]. \tag{2.6}$$

In the special case that $N_2 = 0$, $H(z)$ is a causal filter.

Algorithm 2:

- Determine poles γ_m^{in} and weights w_m^{in} such that

$$\left| h_d(n) - \sum_{m=1}^{M^{\text{in}}} w_m^{\text{in}} \left(\gamma_m^{\text{in}}\right)^n \right| < \epsilon, \quad 1 \le n \le N_1,$$

  where $\left|\gamma_m^{\text{in}}\right| < 1$ using Algorithm 1.

- Determine poles γ_m^{out} and weights w_m^{out} such that

$$\left| h_d(n) - \sum_{m=1}^{M^{\text{out}}} w_m^{\text{out}} \left(\gamma_m^{\text{out}}\right)^n \right| < \epsilon, \quad -N_2 \le n \le -1,$$

  where $|\gamma_m^{\text{out}}| > 1$ again using Algorithm 1.

- Compute the constant w_0 as

$$w_0 = h_d(0) - \sum_{m=1}^{M^{\text{in}}} w_m^{\text{in}} - \sum_{m=1}^{M^{\text{out}}} w_m^{\text{out}}.$$

- The resulting IIR filter $H(z)$, with impulse response $h(n)$, has $M^{\text{in}} + M^{\text{out}}$ poles and is given by

$$H(z) = w_0 + \sum_{m=1}^{M^{\text{in}}} \frac{w_m^{\text{in}}}{1 - \gamma_m^{\text{in}}/z} + \sum_{m=1}^{M^{\text{out}}} \frac{w_m^{\text{out}}}{1 - z/\gamma_m^{\text{out}}}. \tag{2.7}$$

It may not be immediately obvious why this algorithm should work. Indeed, it is rather surprising that the poles of an optimal IIR filter are related to the roots of a con-eigenpolynomial of a Hankel matrix constructed from the filter's impulse response. The theory underlying our method may be found in [BM09] and traced back to the work of Adamjan, Arov, and Krein (AAK theory) [AAK68a, AAK68b, AAK71]. In this sense, our algorithm is related to algorithms in [GST83] and [CPC92]. But while those algorithms suggest that the input

sequence $h_d(n)$ be windowed in some fashion—thereby modifying (perhaps substantially) the desired frequency response $H_d(e^{j\omega})$—ours does not. Also, our algorithm leads to a way to reduce the number of poles in a sub-optimal IIR filter, which we describe below. First, let us make a few remarks about typical filter design problems.

Remark 2. In many cases of practical interest, some type of symmetry exists between $h_d(n)$ and $h_d(-n)$. In such cases a corresponding symmetry is induced between poles inside and outside the unit disk and their corresponding weights. For example, it is quite common for the impulse response to be real and symmetric,

$$h_d(n) \in \mathbb{R} \quad \text{and} \quad h_d(-n) = h_d(n),$$

in which case it is not difficult to show that poles appear at conjugate-reciprocal locations and the corresponding weights are complex conjugates, so that with a suitable reordering

$$M^{\text{in}} = M^{\text{out}}, \; w_m^{\text{in}} = \overline{w}_m^{\text{out}}, \; \text{and } \gamma_m^{\text{in}} = 1/\overline{\gamma}_m^{\text{out}} \; .$$

Additionally, poles inside the unit disk appear in conjugate pairs, so that for each $m \in [1, M^{\text{in}}]$, either both w_m^{in} and γ_m^{in} are real, or there exists a $m' \in [1, M^{\text{in}}]$ such that

$$w_m^{\text{in}} = \overline{w}_{m'}^{\text{in}} \quad \text{and} \quad \gamma_m^{\text{in}} = \overline{\gamma}_{m'}^{\text{in}}.$$

If such symmetries are present, then it is not necessary to approximate the negative half of the sequence $h_d(n)$. Instead, we approximate only the positive half, which gives us the poles and corresponding weights inside the unit disk, and then use the appropriate symmetry relations to obtain the poles and weights outside the unit disk.

Remark 3. If we are given a frequency response $H_d(e^{j\omega})$, we may use or design an appropriate quadrature rule to compute the impulse response $h_d(n)$. We need to compute a sufficient number of terms so that h_d has decayed to a level substantially smaller than ϵ for both negative and positive indices. This may lead to a rather large matrix $\mathbf{H}$; the algorithm we describe next in combination with the construction in §2.3 allows us to avoid computing with large matrices.

2.2.2 Reduction of the Number of Poles

The filter design algorithm in §2.2.1 is simple to implement and produces excellent filters. As input, it requires a portion of the desired impulse response, $h_d(n)$. For the output filter $H(z)$ to be satisfactory—i.e., for $|H_d(e^{j\omega}) - H(e^{j\omega})|$ to be less than the target accuracy ϵ—the portion of $h_d(n)$ provided as input should have decayed to a level smaller than ϵ. If $H_d(e^{j\omega})$ contains sharp transitions or is highly peaked, then the sequence $h_d(n)$ decays slowly, resulting in a large Hankel matrix $\mathbf{H}$ in (2.3). Computing the SVD of this matrix can be time consuming and may require extended precision arithmetic. In this section we present an alternative approach: by reducing the number of poles in a sub-optimal (but easy to obtain) IIR filter, we bypass a costly SVD.

In §2.3 we demonstrate how to obtain a sub-optimal (with a large number of poles) IIR filter satisfying a particular set of filter design requirements. We now describe an algorithm that takes such a sub-optimal filter as input and produces a near-optimal filter as output. We write the sub-optimal filter as

$$T_0(z) + H_0(z) = T_0(z) + \sum_{m=1}^{M_0^{\text{in}}} \frac{s_m^{\text{in}}}{1 - p_m^{\text{in}}/z} + \sum_{m=1}^{M_0^{\text{out}}} \frac{s_m^{\text{out}}}{1 - z/p_m^{\text{out}}},$$

where $|p_m^{\text{in}}| < 1$ and $|p_m^{\text{out}}| > 1$. We separate the Laurent polynomial $T_0(z)$ in the filter description to make $H_0(z)$ a proper rational function. A Laurent polynomial $T_0(z)$ is a finite linear combination of positive and negative integer powers of z. The important property is that the poles of $T_0(z)$ (if any) be located at the origin. In many cases $T_0(z)$ is simply a constant; for example, in (2.7) $T_0(z) = w_0$. We also assume that the poles of $H_0(z)$ are simple.

Given a target accuracy ϵ, we find a filter $H(z)$ of the form

$$H(z) = \sum_{m=1}^{M^{\text{in}}} \frac{w_m^{\text{in}}}{1 - \gamma_m^{\text{in}}/z} + \sum_{m=1}^{M^{\text{out}}} \frac{w_m^{\text{out}}}{1 - z/\gamma_m^{\text{out}}},$$

such that

$$\left|H_0(e^{j\omega}) - H(e^{j\omega})\right| < \epsilon,$$

with $M^{\text{in}} < M_0^{\text{in}}$ and $M^{\text{out}} < M_0^{\text{out}}$. This process, which we call **reduction**, is performed separately on the poles inside and outside the unit disk. Let us describe the procedure for reducing the interior poles; the procedure for reducing exterior poles is completely analogous. For simplicity of notation, we drop superscripts and let $s_m = s_m^{\text{in}}$, $p_m = p_m^{\text{in}}$, and $M_0 = M_0^{\text{in}}$.

Algorithm 3:

- Write each weight s_m in polar form, $s_m = \rho_m e^{j\theta_m}$, and compute the square roots $c_m = \rho_m^{\frac{1}{2}} e^{j\frac{\theta_m}{2}}$.

- Construct the $M_0 \times M_0$ positive definite matrix $\mathbf{A}$, where

$$\mathbf{A}_{mn} = \frac{c_m \overline{c}_n}{1 - p_m \overline{p}_n}.$$

- Find a vector $\mathbf{u} = (u_1, \ldots, u_{M_0})^T$ satisfying the con-eigenproblem

$$\mathbf{A}\mathbf{u} = \sigma \overline{\mathbf{u}}, \tag{2.8}$$

  with positive $\sigma = \sigma_M$ close to the target accuracy ϵ, where the con-eigenvalues are ordered, $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_{M_0-1}$. The matrix $\mathbf{A}$ is not necessarily symmetric, so the con-eigenvalue σ need not be a singular value of $\mathbf{A}$, but it may be shown that σ^2 is an eigenvalue of $\overline{\mathbf{A}}\mathbf{A}$ [HJ90, §4.6].

- Use the elements of the con-eigenvector $\mathbf{u} = \mathbf{u}_M$ from the previous step to build the con-eigenfunction $u(z)$,

$$u(z) = \frac{1}{\sigma} \sum_{m=1}^{M_0} \frac{\overline{s}_m u_m}{1 - \overline{p}_m z}.$$

  AAK theory guarantees that $u(z)$ has exactly M roots $\gamma_1, \gamma_2, \ldots, \gamma_M$ inside the unit disk.

- Obtain the weights $w_1, w_2, \ldots, w_M$ as the unique solution of the $M \times M$ linear system

$$\sum_{m=1}^{M} \frac{1}{1 - \gamma_m \overline{\gamma}_n} w_m = \sum_{m=1}^{M_0} \frac{1}{1 - p_m \overline{\gamma}_n} s_m.$$

The resulting IIR filter

$$H(z) = \sum_{m=1}^{M} \frac{w_m}{1 - \gamma_m / z}$$

is near-optimal and satisfies

$$\left| \sum_{m=1}^{M_0} \frac{s_m}{1 - p_m / z} - \sum_{m=1}^{M} \frac{w_m}{1 - \gamma_m / z} \right| < k\epsilon, \quad |z| = 1,$$

for $k \approx 1$.

We do not derive this algorithm here (see [HB12] for details) and note that it can be obtained from the discussion in [BM05, §6] or justified using results from AAK theory [AAK68a, AAK68b, AAK71]. We note that several significant improvements to this algorithm which use the Cauchy structure of **A** appear in [HB12]. The key improvements in [HB12] are the speed of the algorithm and a **relative** accuracy of computed con-eigenvalues resulting in accurate computations using the standard double precision arithmetic.

2.2.3 Efficient FIR Approximation of IIR Filters

In many situations the straightforward recursive realization of an IIR filter may be inconvenient. For example, an IIR filter with linear phase requires poles both inside and outside the unit disk. The data must then be accessed in reverse-time order to obtain a stable recursive realization. Also, recursive realizations implemented using fixed-point arithmetic may allow errors to accumulate, potentially reducing the filter accuracy to an unacceptable level. For these reasons, it is desirable to find FIR approximations of IIR filters.

The traditional approach to this problem uses some optimization criterion to find a fixed-length FIR filter (see, e.g., [KBG92]); efficiency is obtained by requesting a short filter. Instead, we use the approach in [Bey95a]: we specify the target accuracy ϵ, but do not fix the order of the FIR filter. We obtain a factored FIR filter where each factor is particularly simple, resulting in an efficient cascade realization. We briefly present this approximation method and refer to [Bey95a] for the details.

In our method, the problem of finding an FIR filter amounts to approximating a rational function with a polynomial for some prescribed accuracy ϵ. The construction is based on the simple identity

$$\frac{1}{1-z} = \prod_{n=0}^{\infty} \left(1 + z^{2^n}\right), \quad |z| < 1. \tag{2.9}$$

We adapt the approach in [Bey95a] to IIR filters expressed as partial fractions as constructed by Algorithms 2 and 3. The following Lemma shows how to approximate a single term in the partial fraction expansion (2.7) by an FIR filter.

Lemma 4. *Let γ, w be complex-valued with $|\gamma| < 1$, and let N be a positive integer. Then, for both causal and anti-causal partial fractions, we have the bound*

$$\left| \frac{w}{1-\gamma/z} - w \prod_{n=0}^{N} \left[1 + \left(\frac{\gamma}{z}\right)^{2^n}\right] \right| \leq |w| \, \frac{|\gamma|^{2^{N+1}}}{1-|\gamma|} \tag{2.10}$$

and

$$\left| \frac{w}{1-\gamma z} - w \prod_{n=0}^{N} \left[1 + (\gamma z)^{2^n}\right] \right| \leq |w| \, \frac{|\gamma|^{2^{N+1}}}{1-|\gamma|} \tag{2.11}$$

for all $|z| = 1$.

Proof. From (2.9) it follows that

$$\prod_{n=0}^{N} \left[1 + (\gamma z)^{2^n}\right] = \frac{1-(\gamma z)^{2^{N+1}}}{1-\gamma z} = \sum_{k=0}^{2^{N+1}-1} (\gamma z)^k, \tag{2.12}$$

for $|z| = 1$. Apply (2.12) to the identity

$$\frac{w}{1-\gamma z} = w \sum_{k=0}^{\infty} (\gamma z)^k$$

to obtain (2.11). The proof of (2.10) is identical. □

Even though the sum on the right hand side of (2.12) contains 2^{N+1} terms, the sum is represented by only $N+1$ factors in the product on the left. Given the desired accuracy ϵ, inequalities (2.10) and (2.11) show that the number of factors $N+1$ depends only sub-logarithmically on ϵ^{-1}. The next proposition shows how to approximate the entire IIR filter (2.7) by an FIR filter with a bounded absolute error. We omit the proof since it is an immediate consequence of Lemma 4.

Proposition 5. *Given an IIR filter in the form* (2.7)*, define the FIR filter*

$$\widetilde{H}(z) = w_0 + \sum_{m=1}^{M^{in}} w_m^{in} \prod_{n=0}^{N_m^{in}} \left[1 + \left(\frac{\gamma_m^{in}}{z}\right)^{2^n}\right] + \sum_{m=1}^{M^{out}} w_m^{out} \prod_{n=0}^{N_m^{out}} \left[1 + \left(\frac{z}{\gamma_m^{out}}\right)^{2^n}\right], \tag{2.13}$$

where N_m^{in}, $m = 1, 2, \ldots, M^{in}$ *and* N_m^{out}, $m = 1, 2, \ldots, M^{out}$ *satisfy*

$$\sum_{m=1}^{M^{in}} \left|w_m^{in}\right| \frac{\left|\gamma_m^{in}\right|^{2^{N_m^{in}+1}}}{1 - \left|\gamma_m^{in}\right|} + \sum_{m=1}^{M^{out}} \left|w_m^{out}\right| \frac{\left|\gamma_m^{out}\right|^{-2^{N_m^{out}+1}}}{1 - \left|\gamma_m^{out}\right|^{-1}} < \epsilon. \tag{2.14}$$

Then the FIR approximation $\widetilde{H}(z)$ *in* (2.13) *satisfies*

$$\left|H(z) - \widetilde{H}(z)\right| < \epsilon, \quad |z| = 1.$$

Remark 6. If the IIR filter $H(z)$ is non-causal, then the FIR filter (2.13) is also non-causal (viz., $\widetilde{H}(z)$ contains positive powers of z). The highest positive power of z that appears in $\widetilde{H}(z)$ depends on the desired accuracy and determines the non-causal delay associated with the FIR filter. By introducing a pure delay, $z^{-2^{N_{max}+1}+1}\widetilde{H}(z)$, where $N_{\max} = \max\left\{N_1^{\text{out}}, N_2^{\text{out}}, \ldots, N_{M_{\text{out}}}^{\text{out}}\right\}$, we obtain a causal FIR filter. Hence, both causal and non-causal IIR filters yield efficient causal FIR approximations.

2.3 Filling the Gaps

We now combine the algorithms of §2.2 to produce a systematic method for designing near-optimal filters with which we create remarkable filters not obtainable (as far as we know) by other techniques. We describe lowpass filter design as a model problem. Although in this case it may be possible to obtain an equivalent design by other means, this example allows us to compare with filters designed using alternative methods. However, for the more complicated filter design problems addressed in §2.5, we are not aware of alternative constructions with comparable efficiency.

Our method comprises three steps. We first create a sub-optimal IIR filter to satisfy the design criteria. Next, we use the reduction algorithm of §2.2.2 to find an equivalent (vis-à-vis the filter specifications) near-optimal IIR approximation of this filter. Finally, we

use the FIR approximation algorithm of §2.2.3 to obtain an efficient FIR filter. Each step in this process introduces some approximation error, so we will allocate a portion of the total allowable error, as given in the filter specifications, to each of the three steps.

Consider the following lowpass filter specification:

$$\begin{aligned} \left|H(e^{j\omega}) - 1\right| &< 10^{-4}, \quad |\omega| < \frac{80}{140} \\ \left|H(e^{j\omega})\right| &< 10^{-4}, \quad |\omega| > \frac{81}{140}, \end{aligned} \tag{2.15}$$

where $\omega \in (-\pi, \pi)$. The combination of a relatively wide passband and a narrow transition region make this a challenging problem. For example, a multirate approach utilizing decimate-by-two stages would offer only marginal improvement over a single stage approach since decimation could only be performed twice. Furthermore, the passband error specification requires the phase $\arg H(e^{j\omega})$ to be nearly zero throughout the passband. Such a requirement, equivalent to requesting approximately linear phase, is challenging for many IIR filter design techniques.

A straightforward method of using the algorithms of §2.2 to obtain an IIR filter is to begin with the piecewise linear function $H_p(e^{j\omega})$, where

$$H_p(e^{j\omega}) = \begin{cases} 1, & \text{if } |\omega| < \frac{80}{140} \\ 81 - 140\,|\omega|\,, & \text{if } |\omega| \in [\frac{80}{140}, \frac{81}{140}] \\ 0, & \text{if } |\omega| > \frac{81}{140}. \end{cases}$$

However, approximating this function allocates too many poles to the sharp corners of the transition region. Instead, we will follow an approach inspired by Butterworth digital filter design (see, e.g., [PB87, §7.2]) and begin with an an infinitely differentiable rational function that is optimally flat in the passband and the stopband. We define the function $F(w)$ by

$$F(w) = F(w; \delta, N) = \frac{1}{1 + \left(\frac{w}{\delta}\right)^{4N}}, \tag{2.16}$$

where $\delta > 0$ and N is a positive integer parameter to be specified later. $F(w)$ is infinitely

differentiable on the real axis of the w-plane, and we associate the real axis with analog frequency.

For real w, the function $F(w)$ has the partial fraction expansion

$$F(w) = 2\,\mathcal{R}\mathrm{e} \sum_{n=0}^{2N-1} \frac{r}{1-\gamma_n w},$$

where $r = (4N)^{-1}$ and

$$\gamma_n = \delta^{-1} e^{j\pi \frac{2n+1}{4N}}, \quad n = 0, 1, \ldots, 2N-1.$$

Applying the Möbius transform

$$w = \alpha(z) = j\frac{1-z}{1+z}, \tag{2.17}$$

we map the unit disk $|z| < 1$ onto the upper half plane $\mathcal{I}\mathrm{m}\, w > 0$, and obtain the IIR filter

$$H_d(z) = F\left(\alpha(z)\right) = c + 2\,\mathcal{R}\mathrm{e} \sum_{n=0}^{2N-1} \frac{\overline{s}_n}{1-\overline{p}_n z},$$

where

$$p_n = \overline{\left(\frac{\gamma_n - j}{\gamma_n + j}\right)} \quad \text{and} \quad s_n = \overline{\left(\frac{2jr\gamma_n}{\gamma_n^2 + 1}\right)}$$

for $n = 0, 1, \ldots, 2N-1$, and

$$c = 2\,\mathcal{R}\mathrm{e} \sum_{n=0}^{2N-1} \frac{jr}{j-\gamma_n}.$$

For $|z| = 1$ on the unit circle, we write $H_d(z)$ as

$$H_d(z) = c + \sum_{n=0}^{2N-1} \frac{s_n}{1-p_n/z} + \frac{\overline{s}_n}{1-\overline{p}_n z} \tag{2.18}$$

describing a lowpass non-causal IIR filter with linear phase (in fact, $H_d(e^{j\omega})$ is real and nonnegative). The filter consists of $4N$ poles appearing as points with conjugate-reciprocal symmetry. The factor of 4 in the denominator of (2.16) was chosen to produce this 4-fold symmetry. Observe that (2.18) is in the proper form for the reduction algorithm of §2.2.2, a fact we will use momentarily.

We now choose δ and N to obtain our preliminary sub-optimal IIR filter $H_d(z)$. In choosing these parameters we only concern ourselves with the accuracy of the approximation. Setting $\delta = 0.295686$ and $N = 393$ produces a filter with a maximum error of 3.3×10^{-5} in both the passband and stopband. This filter—which has 1572 poles—is obviously far from optimal. We now apply to $H_d(z)$ the reduction algorithm from §2.2.2 to obtain a near-optimal IIR filter $H(z)$. For the con-eigenvalue controlling the approximation error in (2.8), we select $\sigma \approx 3.7 \times 10^{-5}$. After applying the algorithm, the resulting IIR filter has only 30 poles, 15 inside the unit disk and 15 (conjugate-reciprocal poles) outside the unit disk. Like $H_d(z)$, the frequency response of $H(z)$ is real-valued. It has a maximum error of 8.5×10^{-5} in the passband and stopband.

As a final step, we use the approach in §2.2.3 to obtain an FIR approximation $\widetilde{H}(z)$ of $H(z)$. We construct the filter $\widetilde{H}(z)$ in the form (2.13), where we expand each pole so that the error of approximation in (2.14) does not exceed 1.5×10^{-5}. The resulting FIR filter satisfies the filter specifications (2.15), has linear phase, and its implementation requires 312 real additions and 161 real multiplications per output sample (we discuss the operation count in §2.4). For comparison, the FIR filter that satisfies (2.15) designed by the Parks-McClellan-Rabiner (PMR) algorithm [Dig79, §5.1] requires 4057 taps and needs 4056 real additions and 2029 real multiplications per output sample. Alternatively, if we were to use the PMR algorithm to produce a filter with the same passband and stopband requiring 161 multiplications per sample, the resulting filter would achieve a stopband attenuation of only 0.21, compared with 10^{-4} for our filter.

The frequency responses of $H_d(z)$, $H(z)$, and $\widetilde{H}(z)$ are shown in Fig. 2.1. The poles of the sub-optimal filter $H_d(z)$ and the poles of the near-optimal filter $H(z)$ are displayed in Fig. 2.2, where only poles inside the unit disk are shown. The poles of $H(z)$ inside the upper half of the unit disk are listed in Table 2.1. The table also lists how many factors each pole requires in its FIR approximation $\widetilde{H}(z)$.

A different approach to efficient FIR filter design is to decompose the frequency range

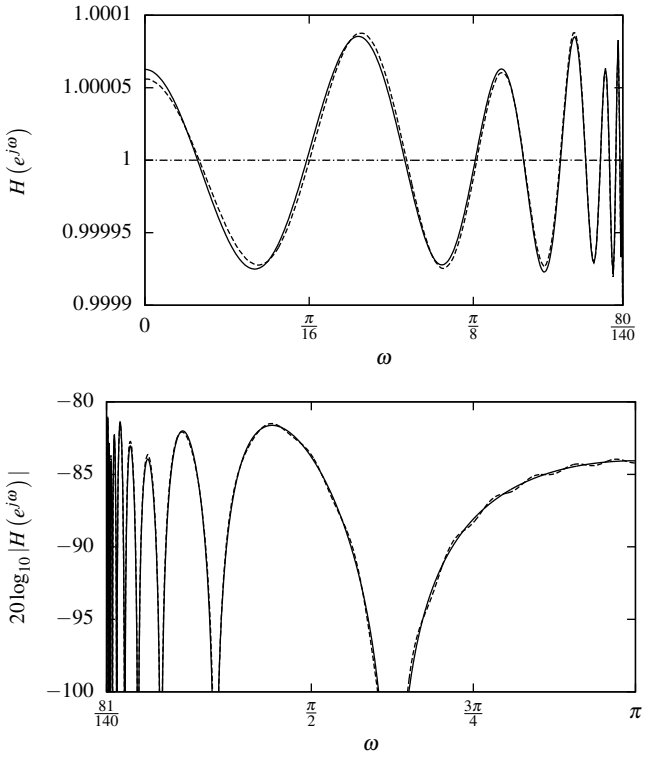

Figure 2.1: Frequency response of the lowpass filters H_d (dash-dot line), H (solid line), and $\widetilde{H}$ (dashed line) in the passband (top) and the stopband (bottom). $\widetilde{H}$ is an excellent approximation of H, making their graphs almost indistinguishable.

Table 2.1: Poles and weights of the lowpass filter H in §2.3, and the number of factors required for each pole in $\widetilde{H}$. The constant term is $w_0 = -0.18305$.

Pole z_m	Weight w_m	Factors
$0.83828 + 0.54323j$	6.1855e-7 − 3.5136e-4j	14
$0.83610 + 0.54180j$	3.6497e-6 − 5.5533e-4j	12
$0.83131 + 0.53862j$	2.5683e-5 − 1.4427e-3j	11
$0.81892 + 0.53003j$	1.7411e-4 − 3.8056e-3j	9
$0.78836 + 0.50650j$	1.2072e-3 − 9.9390e-3j	8
$0.72188 + 0.44108j$	7.9194e-3 − 2.4407e-2j	7
$0.62320 + 0.27635j$	4.0707e-2 − 4.1560e-2j	5
0.57310	8.2954e-2	5

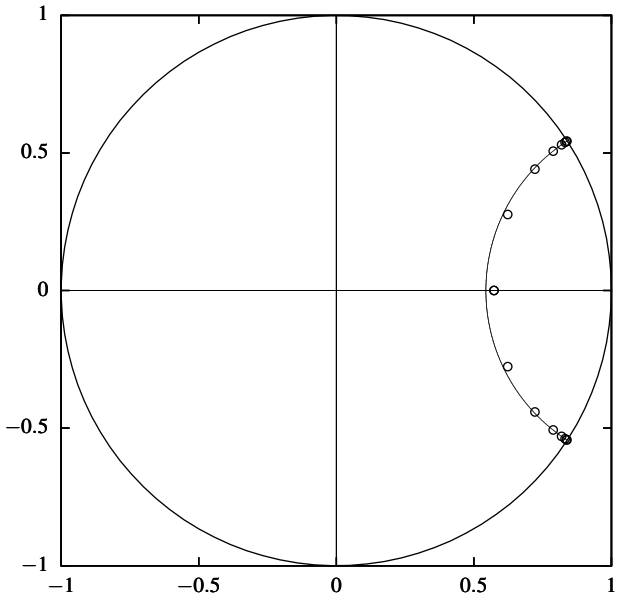

Figure 2.2: Poles of the sub-optimal lowpass filter H_d (small dots) and the equivalent near-optimal filter H (open circles). The poles of H_d are so closely spaced that they appear to form a solid arc.

$\omega \in (-\pi, \pi)$ into subbands and design an efficient FIR filter for each subband [MMS93]. Efficiency is generally obtained by designing sparse FIR filters. Although our method is entirely different, the structure of the resulting FIR filters in (2.13) also has a subband interpretation. Each pole γ_m within the unit disk represents a subband: the subband is centered at the pole's argument $\arg \gamma_m$, and its bandwidth depends on the pole's proximity to the unit circle. The formula (2.9) yields an efficient FIR filter for each subband. Thus, one may view our near-optimal IIR filters as near-optimal subband decompositions of desired frequency responses.

A few remarks are in order.

Remark 7. Many existing IIR design algorithms (see [CJ82] for an early example or [TCHR01] for a more recent one) contain a computationally expensive step to ensure that the IIR filter is causal (i.e., all poles lie within the unit disk), which obviously precludes a filter with linear phase. Our FIR approximation algorithm shows that such restrictions are not necessary, since a non-causal IIR filter may be efficiently approximated by an FIR filter to any desired accuracy. The emphasis, then, should be on minimizing the number of poles rather than ensuring that all poles lie within the unit disk.

Remark 8. Approximation by splines is another excellent method of producing sub-optimal IIR filters. They are especially useful for producing more complicated frequency responses. Splines have accurate and efficient rational approximations, so it is easy to obtain a sub-optimal IIR filter $H_d(z)$ given a sequence of spline coefficients. We ensure our reduced filters $H(z)$ are efficient by choosing splines of sufficiently high degree, so they have many continuous derivatives. Finally, we note that the spline expansion coefficients may be obtained rapidly using the algorithm in [BC02] and [JBB10, Appx.], which makes use of the Fast Fourier Transform.

Remark 9. By combining the "building block" function $F(w; \delta, N)$ with the standard frequency transformations used to construct IIR filters from the classical analog filters (see,

e.g., [PB87]), our method generates the common frequency selective filters (viz., lowpass, bandpass, highpass, and bandstop). We demonstrate further uses of $F(w;\delta,N)$ in §2.5.

Remark 10. Many desired impulse responses, such as those requiring linear phase, are two-sided; they decay to the left and right of a central maximum. Conceptually (and sometimes numerically) a causal IIR approximation is obtained by windowing, truncating, and shifting the desired two-sided impulse response to the right, which (in most cases) puts the maximum amplitude significantly to the right of the origin. Directly applying this approach, as in [GST83, CPC92], produces serious numerical difficulties effectively precluding the optimality implied by the underlying AAK theory. For example, to obtain the same quality approximation as in our approach, a causal IIR filter must have a similar number of accurate impulse response coefficients as in the causal FIR filter $z^{-2^{N_{max}+1}+1}\widetilde{H}(z)$ (see Remark 6). Thus, for high accuracy, the central peak of the impulse response would be located far away from the origin. This would cause serious numerical difficulties in approximating such sequence by an efficient causal IIR filter. This may explain why examples in the literature for such approximations only deal with low accuracy filters.

2.4 Filter Implementations

There are several possible implementations of our FIR approximations of IIR filters. The choice depends on the implementation medium (hardware vs. software), on the purpose of the filter, and on the filter itself. For example, if we have an IIR filter of the form

$$H(z) = \frac{P(z)}{Q(z)}, \tag{2.19}$$

then [Bey95a] shows how to replace $1/Q(z)$ by a cascade of FIR factors, where the application of each factor requires only a single multiplication and addition. Alternatively, we may begin with an IIR filter expressed in partial fractions,

$$H(z) = w_0 + \sum_{m=1}^{M^{\text{in}}} \frac{w_m^{\text{in}}}{1-\gamma_m^{\text{in}}/z} + \sum_{m=1}^{M^{\text{out}}} \frac{w_m^{\text{out}}}{1-z/\gamma_m^{\text{out}}},$$

which is the form produced by the algorithms in SS2.2.1 and 2.2.2. One implementation path is to rewrite this filter in the form (2.19), then realize the FIR approximation as a single cascade. An alternative is to approximate each term in the partial fraction expansion separately, obtaining an FIR approximation of the form (2.13). In this way, each pole may be applied in parallel. Such an implementation is especially advantageous for software-based realizations given the current prevalence of multiprocessors. We will discuss this type of parallel realization in some detail, then conclude with several remarks about other implementation considerations.

As mentioned in §2.2.3 and demonstrated by the lowpass filter $H(z)$ constructed in §2.3, the poles of IIR filters with real valued, even- or odd-symmetric impulse responses either have non-zero imaginary part and appear with 4-fold symmetry (conjugate-reciprocal pairs inside and outside the unit disk) or are purely real and have 2-fold reciprocal symmetry. A similar symmetry exists for the weights. With this in mind, we pick representative poles inside the upper half of the unit disk or on the real axis, and write our IIR filter as

$$H(z) = w_0 + \sum_{m=1}^{M^{\text{real}}} \frac{a_m}{1 - p_m/z} + \frac{a_m}{1 - p_m z} + \sum_{m=1}^{M^{\text{cpx}}} \frac{\alpha_m}{1 - \rho_m/z} + \frac{\overline{\alpha}_m}{1 - \overline{\rho}_m/z} + \frac{\alpha_m}{1 - \rho_m z} + \frac{\overline{\alpha}_m}{1 - \overline{\rho}_m z}, \quad (2.20)$$

where w_0, a_m, p_m are real and $|p_m| < 1$; α_m, ρ_m are complex, $|\rho_m| < 1$ and $\mathcal{I}\mathrm{m}\,\rho_m > 0$. For the real terms in the first sum, we use (2.9) to write

$$\frac{a}{1 - p/z} + \frac{a}{1 - pz} = \left[2a - ap(z + z^{-1})\right] \prod_{n=0}^{\infty} \left[1 + p^{2^{n+1}} + p^{2^n}\left(z^{2^n} + z^{-2^n}\right)\right].$$

The infinite product may be truncated with bounded error using Proposition 5. For the complex terms in the second sum, we write

$$\frac{\alpha}{1 - \rho/z} + \frac{\overline{\alpha}}{1 - \overline{\rho}/z} + \frac{\alpha}{1 - \rho z} + \frac{\overline{\alpha}}{1 - \overline{\rho} z} = \left[b_0 + b_1(z + z^{-1}) + b_2(z^2 + z^{-2})\right] \times \prod_{n=0}^{\infty} \left[c_{0,n} + c_{1,n}\left(z^{2^n} + z^{-2^n}\right) + c_{2,n}\left(z^{2^{n+1}} + z^{-2^{n+1}}\right)\right],$$

where the real coefficients are given by

$$\begin{aligned}
b_0 &= 4\left(\mathcal{R}\mathrm{e}\,\alpha + |\rho|^2\,\mathcal{R}\mathrm{e}\,\alpha + \mathcal{R}\mathrm{e}\left(\rho^2\overline{\alpha}\right)\right)\\
b_1 &= -2\left(\mathcal{R}\mathrm{e}\left(\rho\alpha\right) + 2\,\mathcal{R}\mathrm{e}\left(\rho\overline{\alpha}\right) + |\rho|^2\,\mathcal{R}\mathrm{e}\left(\rho\overline{\alpha}\right)\right)\\
b_2 &= 2\,|\rho|^2\,\mathcal{R}\mathrm{e}\,\alpha\\
c_{0,n} &= 1 + \left(2\,\mathcal{R}\mathrm{e}\left(\rho^{2^n}\right)\right)^2 + |\rho|^{2^{n+2}}\\
c_{1,n} &= 2\,\mathcal{R}\mathrm{e}\left(\rho^{2^n}\right)\left(1 + |\rho|^{2^{n+1}}\right)\\
c_{2,n} &= |\rho|^{2^{n+1}}\,.
\end{aligned}$$

Just as for the real poles, Proposition 5 may be used to truncate the infinite product.

In this way, each real pole is approximated as a cascade where each factor requires 2 additions and 1 multiplication (we factor out the terms $1 + p^{2^{n+1}}$), followed by a factor requiring 2 additions and 2 multiplications. Each complex pole is approximated as a cascade with factors requiring 4 real additions and 2 real multiplications (where we factor out the terms $c_{0,n}$) followed by a factor requiring 4 real additions and 3 real multiplications. Let N_m^{real} denote the number of factors needed to approximate the real pole p_m, and N_m^{cpx} denote the number of factors needed for the complex pole ρ_m (see (2.13)). Then the total computational cost of a parallel implementation is

$$\begin{aligned}
\#\mathrm{Adds} &= 4N^{\mathrm{cpx}} + 5M^{\mathrm{cpx}} + 2N^{\mathrm{real}} + 3M^{\mathrm{real}}\\
\#\mathrm{Mults.} &= 2N^{\mathrm{cpx}} + 3M^{\mathrm{cpx}} + N^{\mathrm{real}} + 2M^{\mathrm{real}} + 1,
\end{aligned}$$

where

$$N^{\mathrm{cpx}} = \sum_{m=1}^{M^{\mathrm{cpx}}} N_m^{\mathrm{cpx}} \quad \text{and} \quad N^{\mathrm{real}} = \sum_{m=1}^{M^{\mathrm{real}}} N_m^{\mathrm{real}},$$

which includes the costs of the constant term w_0 and combining the output of each parallel component. We emphasize that these are real additions and multiplications, even though the poles and weights are generally complex.

We conclude this section with a series of remarks.

Remark 11. For software-based realizations, the parallel structure of our FIR approximations is simple to implement and yields fast codes. For hardware-based realizations, the serial cascade structure in [Bey95a] may also be considered.

Remark 12. A non-causal filter lacking a symmetric impulse response does not possess symmetry of poles inside and outside the unit disk. In this situation, the poles inside the unit disk may be applied using the standard recursive equations and the poles outside the unit disk using an appropriate FIR approximation.

Remark 13. Lowpass filters used in downsampling applications, such as digital tuning or sigma/delta A/D conversion, are an important special case. These filters are characterized by a narrow passband, narrow transition band, and tight error tolerances. Lowpass FIR filters designed by our method are especially convenient in these situations. Since factors in the cascade have terms z^{2^n}, we can apply a factor then decimate by two prior to applying the next factor. This approach dramatically reduces the memory and number of arithmetic operations required to implement the filter. We note that strategically interlacing decimation and filtering stages has been used with great success in the field of multirate signal processing (see [CR83] and references therein).

2.5 Design Examples

2.5.1 Frequency Selective Filters

We now turn to a more complicated sub-optimal filter, and thereby obtain a near-optimal filter that could not easily be obtained by other means. Let us consider a "staircase" filter $H_d(z)$ constructed by using the Möbius transform (2.17) together with the function

$$\frac{1}{2}F\left(w;\frac{3}{8},25\right)+\frac{1}{2}F\left(w;\frac{5}{8},25\right),$$

where $F(w;\delta,N)$ is defined in (2.16) (see Fig. 2.3). The sub-optimal filter $H_d(z)$ is a real-valued IIR filter with 200 poles. We reduce their number by choosing a con-eigenvalue of

Table 2.2: Poles and weights of the "staircase" filter H in §2.5.1, and the number of factors required for each pole in $\widetilde{H}$. The constant term is $w_0 = 0.60057$.

Pole z_m	Weight w_m	Factors
$0.73729 + 0.64330j$	2.6451e-5 $- 3.7663$e-$3j$	9
$-0.21254 + 0.94461j$	7.2305e-6 $- 5.5944$e-$3j$	8
$0.67809 + 0.59327j$	-4.3178e-4 $- 1.1810$e-$2j$	7
$-0.18571 + 0.83540j$	6.8318e-4 $- 1.7861$e-$2j$	6
$0.44780 + 0.42670j$	-1.3985e-2 $- 7.0914$e-$2j$	4
-8.7442e-$2 + 0.42165j$	-2.9205e-2 $- 0.16703j$	4
-0.50671	-1.4759e-2	3

$\sigma \approx 4.6 \times 10^{-4}$ in Algorithm 3 to obtain a new filter $H(z)$ with 26 poles, of which 13 are inside the unit disk and 13 are outside. The approximation error $|H_d(e^{j\omega}) - H(e^{j\omega})|$ is shown in Fig. 2.3. The error displays almost exact equioscillation, consistent with our claim that IIR filters produced by our method are near-optimal. The pre- and post-reduction pole locations are shown in Fig. 2.4, where only poles inside the unit disk are displayed. The pole pattern is complicated enough that it is not clear how one would produce these poles by other means. The poles, weights, and number of factors required in an FIR approximation with error less than 10^{-3} are shown in Table 2.2 (only poles inside the upper part of the unit disk are listed).

2.5.2 Quadrature Mirror Filters

Our approach allows us to widen the range of useful properties in the design of FIR Quadrature Mirror Filters (QMFs). The perfect reconstruction condition requires the low-pass filter of the QMF pair to satisfy

$$H(z)H(z^{-1}) + H(-z)H(-z^{-1}) = 1. \tag{2.21}$$

Such filters give rise to filter banks, and, with simple additional constraints, to orthonormal wavelet bases. Filter banks provide methods for efficiently applying operators to signals, in particular, operators that in the standard representation result in very long filters, such as fractional derivatives or the Hilbert transform (see, e.g., [Bey92]). Filter banks have proven

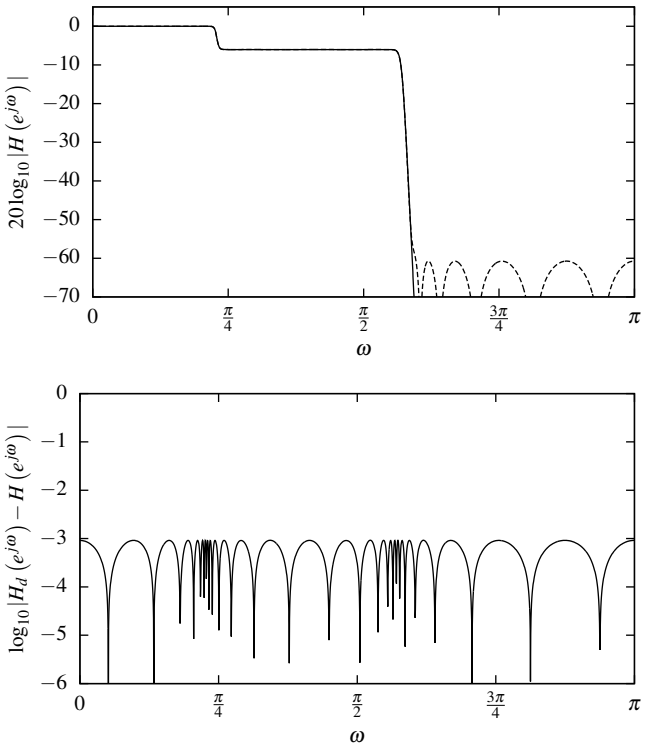

Figure 2.3: The "staircase" filter H_d (solid line) and the approximation H (dashed line) (top). The equioscillation approximation error (bottom) shows that H is near-optimal.

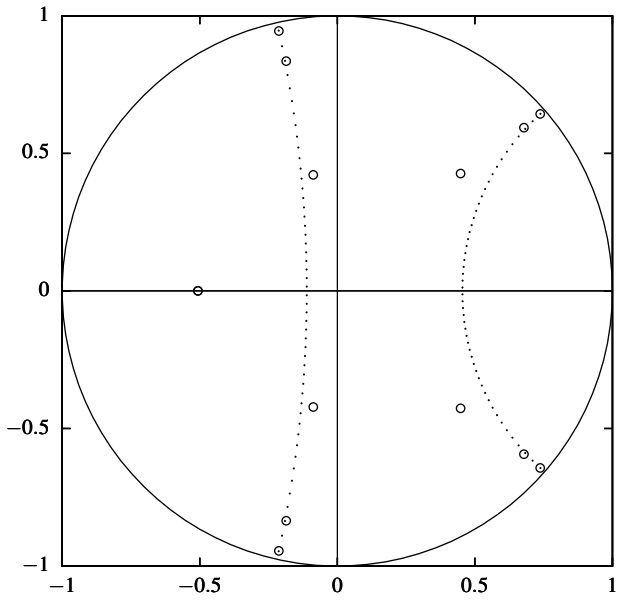

Figure 2.4: Poles of the sub-optimal "staircase" filter H_d (small dots) and equivalent near-optimal filter H (open circles).

useful for applications in signal processing, numerical analysis, and data compression (see, e.g., [JMR01]).

Depending on the application, we may request different properties of the filter (2.21). Algebraically, many of these properties are interrelated and several are mutually exclusive. For example, no FIR QMF can be symmetric but nothing prevents the design of symmetric IIR QMFs. We note that many such restrictions on properties of QMFs are fragile; i.e., for any finite accuracy these restrictions disappear, and we use this fact as a tool for the design of approximate QMFs with the desired properties. Some examples may be found in [Bey95a] and here we construct approximate IIR and FIR QMFs that are symmetric (i.e., have linear phase), efficient, and have attractive flatness and subband isolation properties.

In [Mon99] a particularly interesting family of symmetric IIR QMFs is introduced,

$$E_{4N}(z) = \frac{(1+z)^{2N}\left((1+z)^{2N} + (-1)^N\sqrt{2}(1-z)^{2N}\right)}{(1+z)^{4N} + (1-z)^{4N} + (-1)^N\sqrt{2}(1-z^2)^{2N}}, \tag{2.22}$$

where the positive integer parameter N simultaneously controls the flatness of the passband and stopband and the width of the transition region. It may be that the value N required to achieve a sufficiently narrow transition band results in a filter that is excessively flat. We show how to use our method to obtain an efficient FIR approximation of the original QMF that retains the desired sharpness but gains efficiency by reducing the excessive flatness. An example of such a QMF frequency response is illustrated in Fig. 2.5.

The filter flatness is controlled by the root of order $2N$ at $z = -1$ of $E_{4N}(z)$. To obtain a more efficient, but less flat, IIR filter, we factor our a portion of this high-order root and apply the reduction algorithm from §2.2.2 to the remaining terms. Observing that $E_{4N}(z)$ is real-valued on the unit circle, we select some integer $S < N$ (which controls the flatness of the new filter) and rewrite $E_{4N}(z)$ as

$$E_{4N}(z) = \left(\frac{1+z}{2}\right)^S \left(\frac{1+z^{-1}}{2}\right)^S \left[c + \sum_{n=1}^{2N} \frac{s_n}{1-p_n/z} + \frac{\overline{s}_n}{1-\overline{p}_n z}\right]. \tag{2.23}$$

We may now reduce the expression in brackets and construct an FIR approximation of the

result. For example, we choose $N = 20$, yielding an IIR filter with 40 poles inside the unit disk and 40 poles outside. This filter has an appealingly narrow transition band, but the passband is flatter than may be required for many filter bank applications. We select $S = 3$ in (2.23) and apply the reduction algorithm of §2.2.2 to obtain a new IIR filter with only 17 poles inside the unit disk and 17 outside. Finally, we use the FIR approximation algorithm of §2.2.3 to obtain an FIR filter $\widetilde{E}_{80}(z)$ that approximates $E_{80}(z)$ on the unit circle with an error bounded by 10^{-8}. Because $\widetilde{E}_{80}(z)$ approximates $E_{80}(z)$ so closely, it (approximately) inherits the same properties as $E_{80}(z)$. In particular, $\widetilde{E}_{80}(z)$ is symmetric and approximately satisfies the perfect reconstruction condition (2.21) with an error that does not exceed 10^{-8} on the unit circle. It is also very flat because of the root of order 6 at $z = -1$. The approximate QMF $\widetilde{E}_{80}(z)$ and the perfect reconstruction error

$$\widetilde{E}_{80}(z)\widetilde{E}_{80}(z^{-1}) + \widetilde{E}_{80}(-z)\widetilde{E}_{80}(-z^{-1}) - 1$$

are shown in Fig. 2.5. The poles, weights, and number of factors in the FIR approximation are shown in Table 2.3 (only the poles in the upper half of the unit disk are listed).

Remark 14. We note that the order of the zero, $2S$, yields an approximate interpolating property for the filter bank [Dau93, MBH99]. Directly constructing FIR QMFs with this property leads to the so-called Coiflets (design of which which is difficult, see, e.g., [Dau93, MBH99]) and the resulting filters cannot be symmetric. In comparison, our construction is simple and provides additional properties.

2.6 Conclusion

We have described a new method of designing accurate and efficient IIR and FIR filters. Our method has several advantages. First, the FIR filters it produces are more efficient than FIR filters constructed by other methods, when such constructions are even possible. Second, many properties (such as symmetric filters satisfying the perfect reconstruction condition) can only be obtained by IIR filters. Our method produces FIR filters that, with any finite

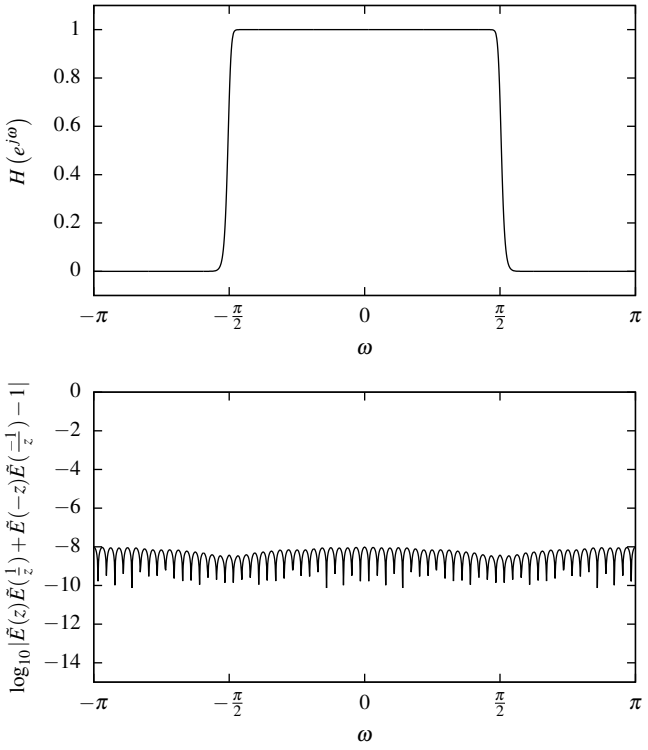

Figure 2.5: Approximate QMF $\widetilde{E}_{80}$ (top) and its perfect reconstruction error (bottom).

Table 2.3: Poles, weights, and number of factors required in the FIR approximation $\widetilde{E}_{80}$ of the IIR QMF E_{80} in §2.5.2. The constant term is $w_0 = -1.24533870508$.

Pole z_m	Weight w_m	Factors
$1.614962028091702\text{e-}8 + 0.9427643190768825j$	$0.1156203491819011 - 8.058245848823418\text{e-}2j$	9
$1.523267607493497\text{e-}8 + 0.9063478708933764j$	$-6.602258562520147\text{e-}2 - 0.1235199582810649j$	8
$9.452755533145547\text{e-}4 + 0.8038059179307812j$	$0.1435760298523706 - 5.972837054722412\text{e-}3j$	7
$2.385100189342754\text{e-}3 + 0.7758181038788966j$	$-8.466685108153214\text{e-}3 - 0.1405348047690257j$	7
$-4.287984710863575\text{e-}3 + 0.6711924419976365j$	$0.1673367117011427 - 4.488479171905594\text{e-}2j$	6
$1.654221356678399\text{e-}2 + 0.5487007195130383j$	$0.1691065050219905 + 4.884707929391815\text{e-}2j$	5
$3.826740048904385\text{e-}2 + 0.4009109376978119j$	$9.267944965286919\text{e-}2 + 0.1068308245943145j$	5
$4.944326086546647\text{e-}2 + 0.2253118851038638j$	$8.840369128305046\text{e-}3 + 6.662132497247104\text{e-}2j$	4
0.5130884721438124	$-1.582530456670174\text{e-}6$	4

accuracy, approximately possess these properties. Third, our filters have a straightforward parallel implementation. Finally, by approximating IIR filters with FIR filters, we can consider IIR filters with properties, such as linear phase, not obtainable by causal IIR filters.

Chapter 3

Fast and Accurate Propagation of Coherent Light

This chapter contains a preprint of

> [LBM13] R. D. Lewis, G. Beylkin, and L. Monzón, **Fast and accurate propagation of coherent light**, 2013, to be submitted.

FAST AND ACCURATE PROPAGATION OF COHERENT LIGHT

RYAN D. LEWIS, GREGORY BEYLKIN, AND LUCAS MONZÓN

Abstract. We describe a fast algorithm to propagate, for any user-specified accuracy, a time-harmonic electromagnetic field between two parallel planes separated by a linear, isotropic, and homogeneous medium. The analytic formulation of this problem (circa 1897) requires the evaluation of the so-called Rayleigh-Sommerfeld integral. If the distance between the planes is small, this integral can be accurately evaluated in the Fourier domain; if the distance is very large, it can be accurately approximated by asymptotic methods. In the large intermediate region of practical interest, where the oscillatory Rayleigh-Sommerfeld kernel must be applied directly, current numerical methods can be highly inaccurate without indicating this fact to the user. In particular, we demonstrate that while the often-used Fresnel approximation may yield adequate accuracy near the optical axis, the accuracy deteriorates significantly away from the axis. In our approach, for any user-specified accuracy $\epsilon > 0$, we approximate the kernel by a short sum of Gaussians with complex-valued exponents and then efficiently apply the result to the input data using the

unequally spaced fast Fourier transform. The resulting algorithm has computational complexity $\mathcal{O}\left(N^2 \log N \log^2 \epsilon^{-1} + M^2 \log^4 \epsilon^{-1}\right)$, where we evaluate the solution on an $N \times N$ grid of output points given an $M \times M$ grid of input samples. Our algorithm maintains its accuracy throughout the entire computational domain.

3.1 Introduction

A measurement system can be no more accurate than the least accurate of its constituent parts. A critical part of many computational optical systems is a numerical algorithm to propagate a time-harmonic electromagnetic field between two parallel planes separated by a linear, isotropic, and homogeneous medium. Within the experimental community, it is well understood that algorithms used for this purpose give approximate solutions. However, virtually none of the current algorithms provide a mechanism to control or estimate their error and, for this reason, may generate inaccurate results without indicating this fact to the user. This state of affairs is somewhat surprising since one might expect that in the computer age, of all the sources of error in an optical system, numerical error ought to be the easiest to eliminate.

At the end of 19th century, Lord Rayleigh [Ray97] described wave propagation via the integral

$$u(\mathbf{x}, z) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} f(\mathbf{y}) \frac{\partial}{\partial z} \left(\frac{e^{i2\pi R}}{R} \right) d\mathbf{y}, \quad z > 0, \tag{3.1}$$

where $R = \sqrt{z^2 + \|\mathbf{x} - \mathbf{y}\|^2}$ (see also [Bou54, BWB99]). Given the field $f(\mathbf{y})$ in the plane $z = 0$, (3.1) describes the field $u(\mathbf{x}, z)$, $z > 0$, that satisfies the Sommerfeld radiation condition. Expressing all distances in wavelengths, we note that if the propagation distance is small, then the kernel of this integral operator is highly oscillatory, but the computation can then proceed in an accurate manner in the Fourier domain. On the other hand, if the distance is very large, then application of this kernel asymptotically reduces to a scaled

Fourier transform. The computational difficulties arise in the intermediate region where, in order to obtain an accurate solution, it is necessary to apply this oscillatory kernel as is. Currently, the standard practice is to replace the kernel in this intermediate region by its Fresnel approximation. We show that this approximation yields only limited accuracy even near the optical axis, and that the accuracy deteriorates significantly away from the optical axis. Perhaps what is most troubling is that the accuracy of approximation is not controlled.

In this paper we present a fast algorithm to evaluate the Rayleigh-Sommerfeld integral (3.1) with any user-specified accuracy. We approximate the kernel by a short sum of Gaussians with complex-valued exponents. The number of terms in our approximation is nearly minimal for a given accuracy ϵ. The resulting approximate kernel is then efficiently applied to input data using the unequally spaced fast Fourier transform (USFFT) [DR93, Bey95b], yielding an algorithm of computational complexity $\mathcal{O}\left(N^2 \log N \log^2 \epsilon^{-1} + M^2 \log^4 \epsilon^{-1}\right)$, where we evaluate the solution on an $N \times N$ grid of output points given a grid of $M \times M$ input samples, the same order of complexity as algorithms based on the Fresnel approximation. Our approach also increases significantly the size of the output region where the evaluation of (3.1) is accurate.

Our approximation of the kernel may be viewed as a generalization of the Fresnel approximation. The Fresnel approximation replaces the Rayleigh-Sommerfeld kernel with a single Gaussian with a purely imaginary exponent, whereas we use a nonlinear algorithm to approximate the kernel as a short linear combination of Gaussians with complex-valued exponents (for any user-specified accuracy).

The need for an accurate propagation algorithm arises in areas such as computational holography, optical component design, and antenna design. A particularly interesting application area is X-ray diffraction microscopy, and related techniques, where one attempts to form an image of a microscopic sample from measurements of the magnitude of its diffraction pattern. These inverse problems are usually solved by iterative methods that include a light propagation algorithm. Therefore, the accuracy of the propagation algorithm ultimately

limits the accuracy of the reconstructed image. The speed of a propagation algorithm is obviously also of critical importance for applications employing iterative methods.

The numerical algorithms that we use are designed to yield any user-specified accuracy. This includes controlled accuracy in the rapid computation of integrals. The methods that we employ for this purpose (specifically the USFFT and generalized Gaussian quadratures for band-limited functions) can significantly improve the performance and accuracy of even the standard methods for light propagation, as we observe in Appendices A and B (§§3.7 and 3.8).

The paper is organized as follows. We begin by reviewing the necessary mathematical preliminaries in §3.2. We describe our new algorithm in §3.3, then discuss its region of validity in §3.4. In §3.5 we provide several numerical examples, then summarize our results in §3.6. By introducing this new algorithm, we hope to stimulate accuracy improvements in computational optical systems by essentially eliminating numerical errors.

3.2 Preliminaries

3.2.1 The Rayleigh-Sommerfeld Formula

The behavior of a time-harmonic electromagnetic field in a linear, isotropic, and homogeneous medium is described by the scalar Helmholtz equation,

$$\left(\Delta + k^2\right) u = 0, \tag{3.2}$$

where the wavenumber $k = 2\pi/\lambda$, λ is the wavelength, and $u(x, y, z)$ is the complex amplitude of one component of the vector-valued electric field at a point $(x, y, z) \in \mathbb{R}^3$. We may consider each component of the field separately since their governing equations decouple in an isotropic homogeneous medium, allowing us to work with the scalar form of the Helmholtz equation instead of its vector form.

It is convenient to associate one coordinate of the three-dimensional Cartesian system with the optical axis—we choose the z-coordinate for this purpose, and will often represent a

point $(x, y, z_0) \in \mathbb{R}^3$ as $(\mathbf{x}, z_0)$, where $\mathbf{x} \in \mathbb{R}^2$ lies in the plane $z = z_0$ transverse to the optical axis. We find it natural to measure distances in the units of wavelengths and therefore, for the remainder of this paper, set the wavenumber $k = 2\pi$.

The Rayleigh-Sommerfeld integral (3.1) yields the solution $u(\mathbf{x}, z)$ of the Dirichlet problem (3.2) in the half-space $z > 0$ that satisfies the Sommerfeld radiation condition [Som12, Som49],

$$\lim_{s\to\infty} s\left(\frac{\partial u}{\partial s} - i2\pi u\right) = 0, \quad \text{where } s = \|(\mathbf{x}, z)\| \text{ and } z > 0.$$

Given the boundary data $u(\mathbf{x}, 0) = f(\mathbf{x})$, we rewrite (3.1) as

$$u(\mathbf{x}, z) = \int_{\mathbb{R}^2} f(\mathbf{y})\, K_z(\|\mathbf{x} - \mathbf{y}\|)\, d\mathbf{y}, \tag{3.3}$$

where the Rayleigh-Sommerfeld kernel $K_z(r)$ is given by

$$K_z(r) = \frac{e^{i2\pi z\sqrt{1+(r/z)^2}}}{iz}\left(\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z\left(1+(r/z)^2\right)^{\frac{3}{2}}}\right), \quad r \geq 0. \tag{3.4}$$

Denoting the Fourier transform of the boundary data as

$$\widehat{f}(\mathbf{p}) = \int_{\mathbb{R}^2} f(\mathbf{x})\, e^{-i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{x},$$

we write (3.3) in the Fourier domain as

$$u(\mathbf{x}, z) = \int_{\mathbb{R}^2} \widehat{f}(\mathbf{p})\, \widehat{K}_z(\|\mathbf{p}\|)\, e^{i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p}, \tag{3.5}$$

where the Fourier transform of the Rayleigh-Sommerfeld kernel (cf. [She67] and references therein) is given by

$$\widehat{K}_z(\rho) = e^{i2\pi z\sqrt{1-\rho^2}}, \quad \rho \geq 0. \tag{3.6}$$

Our goal is to evaluate (3.3) in such a way that the computational cost does not increase with the distance z. It is clear that the spatial kernel $K_z(r)$ is a highly oscillatory function of r when z is small, and that the Fourier domain kernel $\widehat{K}_z(\rho)$ is a highly oscillatory

function of ρ when z is large. For many physically interesting choices of the distance z in the intermediate region, $K_z(r)$ and $\widehat{K}_z(\rho)$ are both highly oscillatory, making the direct numerical computation of u using either (3.3) or (3.5) impractical. In §3.3 we will show how to approximate (3.4) with controlled error and then describe a fast and accurate algorithm to apply the resulting approximate Green's function to boundary data. Our algorithm mainly addresses the propagation problem for intermediate and large values of z—for small values of z, it is well known that the problem may be solved using Fourier methods, which we discuss briefly in Appendix A (§3.7), and for very large values of z, the problem may be solved using asymptotic methods, as discussed in Appendix B (§3.8).

Remark 15. Given the normal derivative of the boundary data

$$\frac{\partial}{\partial z} u(\mathbf{x}, z)\Big|_{z=0} = g(\mathbf{x}),$$

Rayleigh's formula for the Neumann problem reads

$$u(\mathbf{x}, z) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} g(\mathbf{y}) \frac{e^{i2\pi R}}{R} \, d\mathbf{y}, \quad z > 0. \tag{3.7}$$

With minor modifications, the approach of this paper is also applicable to evaluating (3.7).

3.2.2 Slepian Functions

All physically realistic fields must eventually decay in space and, at the same time, are essentially band-limited in the Fourier domain. An appropriate mathematical description of such fields was initiated by Slepian et al. in [SP61, LP61, LP62, Sle64, Sle78] by considering a space-limiting and band-limiting integral operator and using its eigenfunctions to identify a class of functions that have controlled concentration in both the space and the Fourier domains. Slepian et al. showed that this integral operator commutes with the differential operator of classical mathematical physics describing the prolate spheroidal wave functions, i.e., both operators share the same eigenfunctions.

For our purposes, we use eigenfunctions with controlled concentration in a square in the spatial domain and band-limited to a disk in the Fourier domain. The construction of such eigenfunctions is described in [BKM07]; it differs from the traditional construction since there is no differential operator available in this case.

Denoting a square in the spatial domain by $A = \left[-\frac{a}{2}, \frac{a}{2}\right]^2$ and selecting a disk of radius c in the Fourier domain, following [BKM07] let us define the space-limiting and band-limiting operator $\mathcal{Q} : L^2(A) \to L^2(A)$,

$$\mathcal{Q}[f](\mathbf{x}) = \int_A f(\mathbf{y}) \frac{cJ_1(2\pi c \|\mathbf{x}-\mathbf{y}\|)}{\|\mathbf{x}-\mathbf{y}\|} d\mathbf{y},$$

where J_1 is the first order Bessel function of the first kind. It is shown in [BKM07] that, similar to the classical case, the eigenvalues of this operator,

$$\mathcal{Q}\psi_j = \mu_j \psi_j, \quad j = 0, 1, \ldots,$$

allow us to quantify the proportion of the L^2-norm of the corresponding eigenfunctions outside of A,

$$1 - \mu_j^2 = \frac{\int_{\mathbb{R}^2 \setminus A} |\psi_j(\mathbf{x})|^2 \, d\mathbf{x}}{\int_{\mathbb{R}^2} |\psi_j(\mathbf{x})|^2 \, d\mathbf{x}}.$$

The eigenvalues satisfy $0 < \mu_j < 1$ and we order them in decreasing order, $\mu_0 > \mu_1 \geq \mu_2 \geq \cdots > 0$. Since they have a sharp transition from being nearly one to being nearly zero (see [BKM07]), for a user-specified accuracy ϵ, we select a linear subspace of the eigenfunctions, $\text{span}\{\psi_j\}_{j=0}^{J}$, with corresponding eigenvalues $\mu_j \geq 1 - \epsilon$. Given boundary data f, we project f onto this subspace, where the choice of parameters, i.e., the domain A and the bandlimit c, is described in §3.2.3 below.

Identifying this subspace allows us to accurately evaluate integrals involving the boundary data. Following [BKM07, BM02] (see also [XRY01]), we have

Theorem 16. *Let $W = \left[-\frac{w}{2}, \frac{w}{2}\right]^2$ be a square output window and fix the positive integer J. Then for any target accuracy ϵ there is a (nearly optimal) tensor product grid of quadrature*

nodes $\mathbf{y}_{mm'} = (y_m, y_{m'}) \in A$, $m, m' = 1, \ldots, M$, *and corresponding weights* $\tau_m \tau_{m'} > 0$ *so that for all functions* $f \in \operatorname{span}\{\psi_j\}_{j=0}^{J}$, *we have*

$$\left| \int_A f(\mathbf{y})\, e^{i\mathbf{x}\cdot\mathbf{y}}\, d\mathbf{y} - \sum_{m,m'=1}^{M} \tau_m \tau_{m'} f(\mathbf{y}_{mm'})\, e^{i\mathbf{x}\cdot\mathbf{y}_{mm'}} \right| \leq \epsilon \|f\|_1, \quad \mathbf{x} \in W.$$

These quadratures are known as generalized Gaussian quadratures for band-limited functions.

3.2.2.1 The Unequally Spaced Fast Fourier Transform

We need to evaluate trigonometric sums of the form

$$\sum_{m,m'=1}^{M} \tau_m \tau_{m'} f(\mathbf{y}_{mm'})\, e^{i\mathbf{x}\cdot\mathbf{y}_{mm'}}$$

at output points $\mathbf{x}_{nn'} = (x_n, x_{n'})$, where $n, n' = 1, \ldots, N$. Such sums can be evaluated rapidly, for any user-specified accuracy ϵ, using the USFFT (see [DR93, Bey95b, LG05]) with computational complexity $\mathcal{O}\left(N^2 \log N + M^2 \log^2 \epsilon^{-1}\right)$.

3.2.3 Band-Limiting the Boundary Data

For a given accuracy ϵ, there exists some square region $A = A(\epsilon) = \left[-\frac{a}{2}, \frac{a}{2}\right]^2$ such that the values of the boundary data f in (3.3) outside of A may be neglected,

$$\int_{\mathbf{x}\notin A} |f(\mathbf{x})|^2\, d\mathbf{x} \leq \epsilon^2 \|f\|_2^2. \tag{3.8}$$

In this paper, we refer to the region A where the field is concentrated as an aperture.

Let us determine the highest spatial frequency c that must be propagated in order to accurately evaluate (3.3). It follows from (3.6) that evanescent waves corresponding to spatial frequencies above $\rho = \|\mathbf{p}\| > 1$ are attenuated exponentially fast as a function of the propagation distance z. This implies that, for a given distance z and accuracy ϵ, there exists some bandlimit $c_e > 1$ such that frequencies greater than c_e may be neglected,

$$\left| u(\mathbf{x}, z) - \int_{\|\mathbf{p}\|\leq c_e} \widehat{f}(\mathbf{p})\, \widehat{K}_z(\|\mathbf{p}\|)\, e^{i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p} \right| \leq \epsilon \|f\|_2.$$

A good estimate of this bandlimit is obtained by setting $e^{-2\pi z\sqrt{c_e^2-1}} = \epsilon$ so that

$$c_e = \sqrt{1 + \left(\frac{\log \epsilon^{-1}}{2\pi z}\right)^2}. \tag{3.9}$$

It may happen that the boundary data f has a bandlimit much larger than c_e. In such cases, we set $c_f = 2c_e$ and replace f by its band-limited version,

$$\tilde{f}(\mathbf{x}) = \int\limits_{\|\mathbf{p}\| \leq 2c_e} \widehat{f}(\mathbf{p})\, h(\|\mathbf{p}\|)\, e^{i2\pi \mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p},$$

where the window function $h(\rho)$ satisfies $|h(\rho) - 1| \leq \epsilon$ for $0 \leq \rho \leq c_e$ and drops smoothly to zero in the interval $\rho \in (c_e, 2c_e]$. The function $\widetilde{f}$ will be band-limited to the disk of radius c_f and concentrated in a square aperture $\widetilde{A}$ that is somewhat larger than the original aperture A. This spreading can be controlled by an appropriate choice of h—one convenient choice is a linear combination of shifted Gaussians. We use this new, larger, aperture in place of the original aperture and therefore set $A = \widetilde{A}$. It may also happen that the bandlimit c_f of boundary data is known a priori and is less than c_e, so it is not necessary to propagate spatial frequencies with magnitudes $\rho = \|\mathbf{p}\| \in [c_f, c_e]$. In either case, we set the highest spatial frequency that must be propagated to $c = c_f$, where c_f is defined as just described.

3.2.4 Approximation of Functions by Linear Combinations of Exponentials and Gaussians

We use an algorithm in [BM05] (see also [BM10]) to approximate, for a target accuracy ϵ, a smooth function $f(x)$ by a nearly optimal linear combination of Gaussians. Since the algorithm in [BM05] finds a nearly minimal number of exponential terms, we apply it to the function $g(t) = f\left(\sqrt{t}\right)$. Changing variables back, $t \mapsto x^2$, yields an approximation by Gaussians with the (nearly) minimal number of complex-valued weights w_ℓ and exponents η_ℓ, such that

$$\left| f(x) - \sum_{\ell=1}^{L} w_\ell e^{-\eta_\ell x^2} \right| \leq \epsilon, \quad x \in [0,1]. \tag{3.10}$$

For completeness, we recall this algorithm for approximation by exponentials in Appendix C (§3.9).

For the functions $f(x)$ considered in this paper, the number of terms L in approximation (3.10) satisfies $L = \mathcal{O}(\log \epsilon^{-1})$. This behavior is typical and occurs for a wide variety of functions encountered in applications.

3.2.5 Decompositions of Low-Rank Matrices

In order to compute the singular value decomposition (SVD) of a low-rank matrix $\mathbf{S} \in \mathbb{C}^{N\times M}$, where $\mathbf{S}$ has numerical rank k for a given accuracy ϵ, we use algorithms described in [CGMR05, HMT11]. The computational complexity of these algorithms is $\mathcal{O}(MN \log k + (M+N)k^2)$ (cf. with $\mathcal{O}(MNk)$ for the direct approach utilizing a rank-revealing QR factorization).

3.2.6 The Approximations of Fresnel and Fraunhofer

Our method of approximating the kernel (3.4) resembles the approach that leads to the Fresnel approximation, which we now recall. If the propagation distance is significantly larger than both the spatial extent of the input field and the desired output region, so that $r = \|\mathbf{x}-\mathbf{y}\| < z$, it is common to use this assumption to make the (rather dramatic) approximations in (3.4)

$$\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z\left(1+(r/z)^2\right)^{\frac{3}{2}}} \approx 1 \tag{3.11}$$

and

$$e^{i2\pi z\sqrt{1+(r/z)^2}} \approx e^{i2\pi z} e^{i\frac{\pi}{z}r^2}. \tag{3.12}$$

The Fresnel approximation uses this approximate kernel in place of the Rayleigh-Sommerfeld kernel in (3.3), yielding

$$\begin{aligned} u(\mathbf{x}, z) &\approx \frac{e^{i2\pi z}}{iz} \int_{\mathbb{R}^2} f(\mathbf{y})\, e^{i\frac{\pi}{z}\|\mathbf{x}-\mathbf{y}\|^2}\, d\mathbf{y} \\ &= \frac{e^{i2\pi z} e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz} \int_{\mathbb{R}^2} f(\mathbf{y})\, e^{i\frac{\pi}{z}\|\mathbf{y}\|^2} e^{-i\frac{2\pi}{z}\mathbf{x}\cdot\mathbf{y}}\, d\mathbf{y} \end{aligned} \tag{3.13}$$

(see, e.g., [Goo05, §4.2]). Since the latter integral can be computed using the fast Fourier transform (FFT), this approximation is widely used despite its potentially low accuracy (it turns out that the poor approximation of the kernel's phase in (3.12) is especially deleterious—see §3.5.2).

When z is much larger than the spatial extent of $f(\mathbf{y})$, it is common to make the further approximation $e^{i\frac{\pi}{z}\|\mathbf{y}\|^2} \approx 1$, which, when used in (3.13), leads to the Fraunhofer (sometimes called far-field) approximation

$$u(\mathbf{x}, z) \approx \frac{e^{i2\pi z} e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz} \widehat{f}\left(\frac{\mathbf{x}}{z}\right) \tag{3.14}$$

(see, e.g., [Goo05, §4.3]). The Fraunhofer approximation, which relates the output field to the scaled Fourier transform of the input field, is especially common in antenna design and X-ray diffraction microscopy. In §§3.4 and 3.5 we demonstrate that the accuracy of the Fresnel approximation rapidly deteriorates away from the optical axis. We discuss the Fraunhofer approximation further in Appendix B (§3.8).

3.3 A New Algorithm for Fast and Accurate Light Propagation

In this section we describe a fast algorithm to compute, for a fixed propagation distance z and any user-specified accuracy $\epsilon > 0$, the field $u(\mathbf{x}, z)$ in a square output window $W = \left[-\frac{w}{2}, \frac{w}{2}\right]^2$. We assume that the boundary data f has already been replaced with its space-limited and band-limited version, as described in §3.2.3. Hence, f is band-limited with some bandlimit c and concentrated in a square aperture $A = \left[-\frac{a}{2}, \frac{a}{2}\right]^2$ so that, according to (3.3),

we need to compute

$$u(\mathbf{x}, z) = \int_A f(\mathbf{y})\, K_z(\|\mathbf{x} - \mathbf{y}\|)\, d\mathbf{y}, \quad \mathbf{x} \in W, \tag{3.15}$$

where K_z is the Rayleigh-Sommerfeld kernel (3.4). Our algorithm comprises three steps. First, in §3.3.1, we accurately approximate the Rayleigh-Sommerfeld kernel by a linear combination of Gaussians using the algorithm briefly described in §3.2.4. Second, in §3.3.2, we use the resulting approximation in (3.15) and accurately discretize the ensuing integrals using the generalized Gaussian quadratures for band-limited functions from Theorem 16. Finally, in §3.3.3, we use the algorithms referred to in §3.2.5 for computing the SVDs of low-rank matrices to rearrange the resulting sums for rapid and accurate evaluation via the USFFT (see §3.2.2.1).

3.3.1 Approximation of the Kernel with Controlled Error

The key observation behind the Fresnel approximation is that the phase of the kernel (3.4) is approximately quadratic, cf. (3.12), at least for small values of r/z. We also use this observation but, in addition, exploit the fact that the rest of the phase can be accommodated via an approximation with controlled error, valid throughout a large computational domain.

Due to the finite sizes of the output window W and input aperture A, it is only necessary to approximate the kernel $K_z(r)$ on the interval $0 \le r \le (a + w)/\sqrt{2}$. In fact, for any user-specified accuracy $\epsilon_K > 0$, we demonstrate how to obtain an approximation $\widetilde{K}_z(r)$ such that

$$\left| K_z(r) - \widetilde{K}_z(r) \right| \le \frac{\epsilon_K}{z}, \quad r \in \left[0, \frac{a+w}{\sqrt{2}}\right]. \tag{3.16}$$

We emphasize that in (3.16) the desired accuracy ϵ_K is scaled by the propagation distance z since the magnitude of the kernel decays like z^{-1} along the optical axis.

Inspired by the Fresnel approximation, we rewrite the kernel as

$$K_z(r) = \frac{e^{i2\pi z} e^{i\frac{\pi}{z}r^2}}{iz} A_z(r),$$

where

$$A_z(r) = \left(\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z \left(1+(r/z)^2\right)^{\frac{3}{2}}} \right) e^{i2\pi z\left(\sqrt{1+(r/z)^2}-1-\frac{1}{2}(r/z)^2\right)}. \tag{3.17}$$

Having removed the factor $e^{i\frac{\pi}{z}r^2}$ capturing most of the oscillatory behavior of the kernel, the function A_z is non-oscillatory over a large region of space. We use the algorithm in §3.2.4 to compute, for a desired accuracy $\epsilon_K > 0$, complex-valued weights w_ℓ and exponents η_ℓ such that

$$\left| A_z(r) - \sum_{\ell=1}^{L} w_\ell e^{-\eta_\ell r^2} \right| \leq \epsilon_K, \quad r \in \left[0, \frac{a+w}{\sqrt{2}}\right], \tag{3.18}$$

leading to the approximation

$$\widetilde{K}_z(r) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell e^{-\left(\eta_\ell - i\frac{\pi}{z}\right)r^2}, \tag{3.19}$$

satisfying (3.16). We define $\widetilde{u}(\mathbf{x}, z)$ to be the result of using the approximate kernel $\widetilde{K}_z(r)$ in (3.15),

$$\widetilde{u}(\mathbf{x}, z) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \int_A f(\mathbf{y})\, e^{-\left(\eta_\ell - i\frac{\pi}{z}\right)\|\mathbf{x}-\mathbf{y}\|^2}\, d\mathbf{y}. \tag{3.20}$$

The following proposition bounds the absolute error of the approximation and is an immediate consequence of the preceding discussion.

Proposition 17. *Let $\widetilde{u}$ be the function defined in* (3.20)*, with weights w_ℓ and exponents η_ℓ, $\ell = 1, \ldots, L$, as in* (3.18)*. Then*

$$|u(\mathbf{x}, z) - \widetilde{u}(\mathbf{x}, z)| \leq \frac{\epsilon_K \|f\|_1}{z}, \quad \mathbf{x} \in \left[-\frac{w}{2}, \frac{w}{2}\right]^2, \tag{3.21}$$

where the field $u(\mathbf{x}, z)$ is given by (3.15)*.*

3.3.2 Discretization of Integrals

Letting $\alpha_\ell = \mathcal{R}\mathrm{e}\,\eta_\ell$ and $\beta_\ell = \mathcal{I}\mathrm{m}\,\eta_\ell - \frac{\pi}{z}$, where η_ℓ, $\ell = 1, \ldots, L$, are as in (3.18), we rearrange (3.20) as

$$\widetilde{u}(\mathbf{x}, z) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{x}\|^2} \int_A f(\mathbf{y})\, e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{y}\|^2} e^{2\alpha_\ell \mathbf{x}\cdot\mathbf{y}} e^{i2\beta_\ell \mathbf{x}\cdot\mathbf{y}}\, d\mathbf{y}. \tag{3.22}$$

A straightforward estimate of the bandlimit of the integrands (see §A.2) may be bounded (for each term, independently of ℓ) by

$$c' = c + \frac{a^2}{2\sqrt{2}z} + \frac{\pi a w}{\sqrt{2}z},$$

where c is the bandlimit of the input function f. Given bandlimit c', we discretize the integrals in (3.22), for a desired accuracy ϵ_Q, using the quadratures from Theorem 16. Let $\mathbf{y}_{mm'} = (y_m, y_{m'}) \in A$, $m, m' = 1, \ldots, M$, be the $M \times M$ tensor product grid of quadrature nodes with the corresponding quadrature weights $\tau_m \tau_{m'}$.

Let us consider an $N \times N$ grid of user-selected output locations $\mathbf{x}_{nn'} = (x_n, x_{n'}) \in W$, $n, n' = 1, \ldots, N$. We apply the quadrature from Theorem 16 to the integrals in (3.22) and obtain an approximation to the output field at the desired locations as

$$u_{nn'} = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \sum_{m,m'=1}^{M} \tau_m \tau_{m'} \mathbf{T}^{(\ell)}_{nn'mm'} f\left(\mathbf{y}_{mm'}\right) e^{i2\beta_\ell \mathbf{x}_{nn'} \cdot \mathbf{y}_{mm'}}. \tag{3.23}$$

In (3.23) the $N \times N \times M \times M$ fourth-order tensors $\mathbf{T}^{(\ell)}$, $\ell = 1, \ldots, L$, are given by

$$\mathbf{T}^{(\ell)}_{nn'mm'} = e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{x}_{nn'}\|^2} e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{y}_{mm'}\|^2} \mathbf{S}^{(\ell)}_{nm} \mathbf{S}^{(\ell)}_{n'm'}, \tag{3.24}$$

where $n, n' = 1, \ldots, N$ and $m, m' = 1, \ldots, M$, and the $N \times M$ second-order tensors (matrices) $\mathbf{S}^{(\ell)}$, $\ell = 1, \ldots, L$, are given by

$$\mathbf{S}^{(\ell)}_{nm} = e^{2\alpha_\ell x_n y_m}, \tag{3.25}$$

where $n = 1, \ldots, N$ and $m = 1, \ldots, M$. From Theorem 16 we obtain the bound

$$\left|\widetilde{u}\left(\mathbf{x}_{nn'}, z\right) - u_{nn'}\right| \leq \frac{\epsilon_Q \left\|f\right\|_1}{z}, \tag{3.26}$$

where $\widetilde{u}\left(\mathbf{x}_{nn'}, z\right)$ is given by (3.20) and $u_{nn'}$ is given by (3.23).

3.3.3 Rapid Evaluation of the Field

In the Fresnel approximation of the kernel, the exponent in the quadratic phase factor is purely imaginary, making it easy to compute (3.13) via either the FFT or the USFFT.

In our approach, the exponents in approximation (3.19) are complex-valued, although the magnitude of their real parts is small relative to the aperture and output window sizes (we describe below how to ensure that this is the case). This observation allows us to develop a fast algorithm to evaluate (3.23).

We want to evaluate the inner summations in (3.23) rapidly using the USFFT. Towards this end, we look for an approximation of $\mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'}$ in a form where the output indices n, n' are split from the input indices m, m'. As the first step, we use the SVD to write the matrices in (3.25) as a sum of outer products,

$$\mathbf{S}^{(\ell)}_{nm} = \sum_{q=1}^{\min(M,N)} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq}, \tag{3.27}$$

where the singular values $\sigma^{(\ell)}_1 \geq \sigma^{(\ell)}_2 \geq \cdots \geq 0$ are arranged in decreasing order and the columns of matrices $\mathbf{U}^{(\ell)}$ and $\mathbf{V}^{(\ell)}$ are orthonormal. By properly selecting parameters as described below in §3.3.4, we ensure that the $N \times M$ matrices $\mathbf{S}^{(\ell)}$ have a low numerical rank (typically less than 25). We then use the algorithms described in §3.2.5 to rapidly compute these SVDs and apply the result to approximate $\mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'}$ by a low-separation-rank tensor with indices n, n' split from the indices m, m'. The error estimate is provided by (see §A.1 for the proof)

Lemma 18. *Let $\sigma^{(\ell)}_q$, $\mathbf{U}^{(\ell)}_{nq}$, and $\mathbf{V}^{(\ell)}_{mq}$, where $\ell = 1, \ldots, L$, $q = 1, \ldots, \min(M, N)$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$, be as in (3.27). For a desired accuracy $\epsilon_R > 0$, let $I^{(\ell)}$, $\ell = 1, \ldots, L$, be the smallest integer such that*

$$\sum_{q=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_q \leq \epsilon_R.$$

Then for $\ell = 1, \ldots, L$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$, we have the approximations

$$\left| \mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'} - \sum_{q,s=1}^{I^{(\ell)}} \sigma^{(\ell)}_q \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{nq}\mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{mq}\mathbf{V}^{(\ell)}_{m's} \right| \leq \epsilon_R 2e^{\frac{|\alpha_\ell|}{2}aw}.$$

Using Lemma 18, we approximate $\mathbf{T}^{(\ell)}$ in (3.24) as $\sum_{r=1}^{R^{(\ell)}} \mathbf{P}^{(\ell)}_{nn'r}\mathbf{Q}^{(\ell)}_{mm'r}$, with $\ell = 1, \ldots, L$, $n, n' = 1, \ldots, N$, and $m, m' = 1, \ldots, M$, where we have re-indexed the resulting double summation using a single index and, also, have collected terms that depend on the output coordinate $\mathbf{x}_{nn'}$ as the $N \times N \times R^{(\ell)}$ tensors $\mathbf{P}^{(\ell)}$ and terms that depend on the input coordinate $\mathbf{y}_{mm'}$ as the $M \times M \times R^{(\ell)}$ tensors $\mathbf{Q}^{(\ell)}$. Lemma 18 implies that

$$\left| \mathbf{T}^{(\ell)}_{nn'mm'} - \sum_{r=1}^{R^{(\ell)}} \mathbf{P}^{(\ell)}_{nn'r}\mathbf{Q}^{(\ell)}_{mm'r} \right| \leq \epsilon_R 2 e^{\frac{|\alpha_\ell|}{2}\left(a^2+w^2+aw\right)}. \tag{3.28}$$

We define $\widetilde{u}_{nn'}$ to be the result of using approximation (3.28) in (3.23),

$$\widetilde{u}_{nn'} = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \sum_{r=1}^{R^{(\ell)}} \mathbf{P}^{(\ell)}_{nn'r} \sum_{m,m'=1}^{M} \tau_m \tau_{m'} \mathbf{Q}^{(\ell)}_{mm'r} f\left(\mathbf{y}_{mm'}\right) e^{i2\beta_\ell \mathbf{x}_{nn'} \cdot \mathbf{y}_{mm'}}. \tag{3.29}$$

It follows from (3.28) that

$$|u_{nn'} - \widetilde{u}_{nn'}| \leq \frac{\epsilon_R \left\|f\right\|_1}{z} 2 \sum_{\ell=1}^{L} |w_\ell| \, e^{\frac{|\alpha_\ell|}{2}\left(a^2+w^2+aw\right)}, \tag{3.30}$$

where $u_{nn'}$ is given by (3.23) and we estimated

$$\sum_{m,m'=1}^{M} \tau_m \tau_{m'} \left|f\left(\mathbf{y}_{mm'}\right)\right| \approx \left\|f\right\|_1.$$

From the bounds given in §A.2, we have

$$2 \sum_{\ell=1}^{L} |w_\ell| \, e^{\frac{|\alpha_\ell|}{2}\left(a^2+w^2+aw\right)} \leq b,$$

where b is a small constant that we incorporate into ϵ_R so that the bound (3.30) becomes

$$|u_{nn'} - \widetilde{u}_{nn'}| \leq \frac{\epsilon_R \left\|f\right\|_1}{z}. \tag{3.31}$$

Combining the error bounds (3.21), (3.26), and (3.31), we obtain

Theorem 19. *The error of computing the field u from* (3.15) *using* (3.29) *is bounded by*

$$\left|u\left(\mathbf{x}_{nn'}, z\right) - \widetilde{u}_{nn'}\right| \leq \frac{\left(\epsilon_K + \epsilon_Q + \epsilon_R\right) \left\|f\right\|_1}{z}. \tag{3.32}$$

Formula (3.29) allows us to compute the field $\widetilde{u}_{nn'}$ rapidly. We first apply $\mathbf{Q}^{(\ell)}_{mm'r}$ as a pre-factor to the input samples $f(\mathbf{y}_{mm'})$, then compute the inner sums using the USFFT, and finally apply $\mathbf{P}^{(\ell)}_{nn'r}$ to the result as a post-factor.

In our presentation, we used three different accuracies, ϵ_K, ϵ_Q, and ϵ_R, in the three steps of deriving the final approximation of the field in (3.29) in order to emphasize these as separate steps. In practice, we choose these accuracies to be the same, and set $\epsilon_K = \epsilon_Q = \epsilon_R = \epsilon/3$ to achieve the final accuracy of ϵ.

Remark 20. It is not necessary for the aperture and output window to be square. Indeed, the USFFT allows us to place the output coordinates at arbitrary locations in the output window. We have used a tensor product grid here for simplicity—with minor modifications, our algorithm may be used to compute the field anywhere in the output window with the same computational cost. The input aperture may also have any shape, provided that accurate quadrature rules are used to discretize the integrals in (3.22). We note that near optimal quadratures for circular apertures are described in [BKM07].

Remark 21. **Simplifications for separable boundary data.** As with the Fresnel approximation, our approach simplifies in the case of boundary data that are separable in Cartesian or polar coordinates. For example, suppose the function f is separable in Cartesian coordinates, viz.,

$$f(\mathbf{x}) = f(x_1, x_2) = \sum_{s=1}^{S} f_1^{(s)}(x_1)\, f_2^{(s)}(x_2) \tag{3.33}$$

for some functions $f_1^{(s)}$ and $f_2^{(s)}$, $s = 1, \ldots, S$. In such cases the application of the approximate kernel (3.19) simplifies to the calculation of several one-dimensional USFFTs. Substitute (3.33) into (3.22) and rearrange to obtain an approximation for the field u in a separated form,

$$\widetilde{u}(x_1, x_2, z) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \sum_{s=1}^{S} u_1^{(\ell,s)}(x_1)\, u_2^{(\ell,s)}(x_2),$$

where the functions $u_1^{(\ell,s)}$ and $u_2^{(\ell,s)}$, $\ell = 1, \ldots, L$, $s = 1, \ldots, S$, are obtained in a manner completely analogous to the method described above except that they may be evaluated by

one-dimensional integrals. We obtain similar formulae if the boundary data are concentrated in a disk and separable in polar coordinates.

3.3.4 Computational Cost

It can be shown (see §A.2) that the number of terms in (3.18) may be estimated as $L = \mathcal{O}\left(\gamma^4 \log \epsilon^{-1}\right)$, where

$$\gamma = \frac{a+w}{\sqrt{2} z^{\frac{3}{4}}}. \tag{3.34}$$

In order to control the number of terms L, we restrict the parameter γ by the empirically-determined constant

$$\gamma \leq 2.62. \tag{3.35}$$

This, in turn, limits the domain where our approximation is valid, although this domain is significantly larger than that of the Fresnel approximation. We discuss this further in §3.4. Moreover, this bound also implies that the ratio $|\alpha_\ell| / (aw)$ is small, causing the matrices $\mathbf{S}^{(\ell)}$ in (3.25) to have a low numerical rank.

We now estimate the computational cost of our algorithm. The cost of evaluating (3.29) depends on the number of USFFTs that must be computed, viz., $R = R^{(1)} + \cdots + R^{(L)}$, estimated as $R = \mathcal{O}\left(\log^2 \epsilon^{-1}\right)$, and on the cost of each USFFT (see §3.2.2.1). Hence, the overall computational cost of our algorithm is $\mathcal{O}\left(N^2 \log N \log^2 \epsilon^{-1} + M^2 \log^4 \epsilon^{-1}\right)$. For actual computing times see §3.5.3.

3.4 Size of the Output Region

In §3.3.4 we ensured that our algorithm is efficient by requiring γ from (3.34) to satisfy (3.35). The practical impact of this requirement is to establish a relationship between the input aperture side-length a, propagation distance z, and output window side-length w. In particular, for a fixed aperture size and propagation distance, the largest output window

that our algorithm can accommodate is

$$w_{\max} = 3.71 \times z^{\frac{3}{4}} - a, \tag{3.36}$$

provided that this number is positive. If it is negative, then the propagation distance is small with respect to the aperture size—in such cases, the propagation problem under consideration should be treated in the Fourier domain or using near-field methods. We address this case further in Appendix A (§3.7).

Using the same reasoning, we also define the quantity $z_{\min}$ as

$$z_{\min} = 0.174 \times a^{\frac{4}{3}}, \tag{3.37}$$

which, for a fixed aperture size a, gives the minimum propagation distance before our algorithm can be used.

Let us find analogues of (3.36) and (3.37) for the Fresnel approximation (3.13). Recall that the only mechanism to control the error when using the Fresnel approximation is to restrict the size of the output region. We first determine the analogue of (3.36), that is, for a given accuracy ϵ, let us find $w'_{\max}$, the largest possible output window where the Fresnel approximation is guaranteed to achieve accuracy ϵ. Since the Fresnel approximation replaces the phase of the Rayleigh-Sommerfeld kernel (3.4) with $e^{i2\pi z}e^{i\frac{\pi}{z}r^2}$, we find, for a desired accuracy ϵ, the maximum value of $r'_{\max}$ such that

$$\left| e^{2\pi z\sqrt{1+(r/z)^2}} - e^{i\left(2\pi z + \frac{\pi}{z}r^2\right)} \right| \leq \epsilon, \quad r \in [0, r'_{\max}],$$

viz., $r'_{\max} \approx \sqrt{2}\left(\frac{\epsilon}{\pi}\right)^{\frac{1}{4}} z^{\frac{3}{4}}$, so that, for a square aperture with side-length a, the largest possible square output window has side-length

$$w'_{\max} \approx 2\left(\frac{\epsilon}{\pi}\right)^{\frac{1}{4}} z^{\frac{3}{4}} - a,$$

an analogue of (3.36) for the Fresnel approximation. The analogue of (3.37) for the Fresnel approximation is

$$z'_{\min} \approx \left(\frac{\pi}{16\epsilon}\right)^{\frac{1}{3}} a^{\frac{4}{3}},$$

which gives, for a desired accuracy ϵ and aperture size a, the minimum propagation distance required before the Fresnel approximation can be used.

To illustrate the difference between $w_{\max}$ and $z_{\min}$ for our method and $w'_{\max}$ and $z'_{\min}$ for the Fresnel approximation, let us choose $\epsilon = 10^{-3}$, so that $w'_{\max} = 0.267 \times z^{\frac{3}{4}} - a$. If $a = 5000$ wavelengths, then after propagating $z = 5 \times 10^6$ wavelengths, we find that

$$\frac{w_{\max}}{w'_{\max}} \approx 16.7,$$

so the largest side-length of our output window is approximately 17 times larger than that of the Fresnel approximation. If the propagation distance is only $z = 250,000$ wavelengths, then $w_{\max} \approx 36,480$ wavelengths while $w'_{\max}$ for the Fresnel approximation is negative, implying that 3-digit accuracy of the Fresnel approximation cannot be guaranteed in **any** output window. In fact, for this accuracy, the minimum propagation distance for the Fresnel approximation is $z'_{\max} \approx 497,000$ wavelengths, compared with $z_{\min} \approx 14,880$ for our method.

If we choose the accuracy threshold to be $\epsilon = 10^{-6}$, then the minimum propagation distance for the Fresnel approximation increases to $z'_{\min} \approx 5 \times 10^6$ wavelengths, whereas the minimum distance for our method does not depend on the desired accuracy, and therefore remains unchanged at $z_{\min} \approx 14,880$.

3.5 Numerical Examples

3.5.1 A Gaussian Beam

To demonstrate the accuracy of our algorithm, we choose boundary data that allows the field to be accurately computed by an alternative approach. For this purpose, we select the boundary data with a Gaussian profile given by

$$f(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|^2}{\sigma^2}}, \tag{3.38}$$

where σ determines the width of the beam measured in the units of wavelengths. It can be shown that the propagating (i.e., non-evanescent) portion of the field is given by (cf. (3.5))

$$\begin{aligned} u_p(\mathbf{x}, z) &= \int\limits_{\|\mathbf{p}\| \le 1} \widehat{f}(\mathbf{p})\, \widehat{K}_z(\|\mathbf{p}\|)\, e^{i2\pi \mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p} \\ &= 4\sqrt{\frac{\pi}{2}} \sum_{k=0}^{\infty} i^k f_k\, j_k \left(2\pi z \sqrt{1 + (\|\mathbf{x}\|/z)^2}\right) \overline{P}_k \left(\left(1 + (\|\mathbf{x}\|/z)^2\right)^{-\frac{1}{2}}\right), \end{aligned} \tag{3.39}$$

where j_k is the k-th order spherical Bessel function of the first kind,

$$\overline{P}_k(s) = \sqrt{(2k+1)/2} P_k(s)$$

is the normalized k-th degree Legendre polynomial, and the coefficients f_k are defined as

$$f_k = \pi\sqrt{2\pi}\sigma^2 \int_0^1 s e^{-\pi^2\sigma^2\left(1-s^2\right)} \overline{P}_k(s)\, ds. \tag{3.40}$$

These coefficients decay rapidly once k is sufficiently large, so that we may truncate the sum in (3.39) to obtain a simple formula to compute the non-evanescent portion of the field to any desired accuracy. The error committed by neglecting the evanescent waves may be bounded by

$$\begin{aligned} |u_e(\mathbf{x}, z)| &= \left| \int\limits_{\|\mathbf{p}\| > 1} \widehat{f}(\mathbf{p})\, e^{-2\pi z\sqrt{\|\mathbf{p}\|^2 - 1}} e^{i2\pi \mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p} \right| \\ &\le 2(\pi\sigma)^2 e^{-(\pi\sigma)^2} \int_1^{\infty} \rho e^{-2\pi z\sqrt{\rho^2-1}}\, d\rho. \end{aligned} \tag{3.41}$$

Provided that (3.41) is less than the accuracy sought, we may disregard the evanescent portion of the field entirely and regard (3.39) as a formula to compute the field generated by the boundary data (3.38) for the desired accuracy.

In our example, we choose $\sigma = 5$ wavelengths, a square aperture of size $a = 50$ wavelengths, a propagation distance of $z = 1000$ wavelengths, a square output window of size $w = 450$ wavelengths, and a desired accuracy of $\epsilon = 10^{-6}$. We then use our algorithm to evaluate the (axially-symmetric) field at $N = 256$ points along the x-axis using $M \times M =$

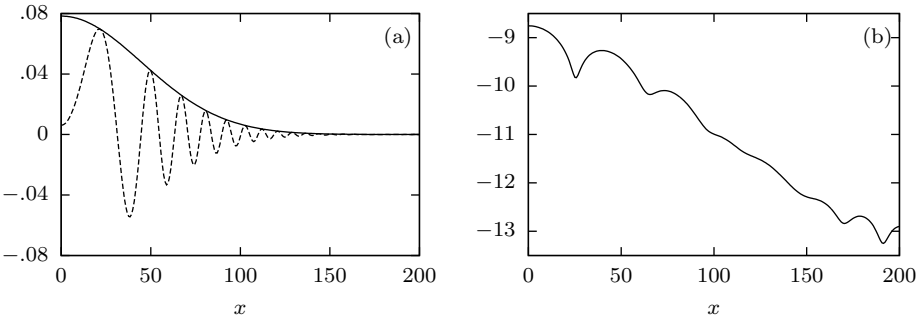

Figure 3.1: Propagation of boundary data with a Gaussian profile. The field magnitude, $|u(x,0)|$, (solid line) and its real part, $\mathcal{R}\mathrm{e}\, u(x,0)$ (dashed line) evaluated along the positive x-axis (a) and the attained accuracy $\log_{10}|u(x,0)-\widetilde{u}(x,0)|$ (b).

512×512 input samples. With this choice of parameters, the number of terms needed to approximate the kernel is $L = 8$, and the number of USFFTs required to evaluate the field is $R^{(1)} + \cdots + R^{(8)} = 52$.

To determine the accuracy of the result, we first compute (3.41) and find that the evanescent part of the solution is undetectable, viz., $|u_e(\mathbf{x}, 1000)| \leq 8.7 \times 10^{-113}$ for all $\mathbf{x}$. We also find that the coefficients (3.40) decay to $|f_k| \leq 10^{-15}$ once $k \geq 200$, so we truncate the sum (3.39) after 200 terms and use it to determine the accuracy of our algorithm. We display the results in Figure 3.1 and note that the obtained accuracy is better than the accuracy goal 10^{-6} (the bound in Lemma 18 is not tight).

3.5.2 Focusing Waves and the Fresnel Approximation

Next we compare the field computed by our algorithm to that obtained via the Fresnel approximation by considering the boundary data

$$f(\mathbf{x}) = \begin{cases} e^{-i2\pi\sqrt{z_0^2+\|\mathbf{x}-\mathbf{r}_0\|^2}}, & \text{if } \mathbf{x} \in \left[-\frac{a}{2}, \frac{a}{2}\right]^2, \\ 0, & \text{otherwise,} \end{cases}$$

representing a spherical wave restricted to a square aperture and converging to the point $(\mathbf{r}_0, z_0)$. In Figure 3.2 we show the magnitude of the resulting field, $|u(\mathbf{x}, z_0)|$, in the plane $z =$

z_0 transverse to the optical axis and containing the focal point, for the choice of parameters $\mathbf{r}_0 = (0,0)$, $z_0 = 100,000$ wavelengths, and $a = 2500$ wavelengths—as expected, the field magnitude is approximately a scaled version of the function $|\mathrm{sinc}\,(x)\,\mathrm{sinc}\,(y)|$.

Now let us move the focal point away from the optical axis. We fix the propagation distance to $z_0 = 100,000$ wavelengths and set the focal point to

$$\mathbf{r}_0 = (z_0 \sin\theta, 0)\,,$$

where θ is the angle between the optical axis and the ray from the origin to the focal point $(\mathbf{r}_0, z_0)$. We select accuracy $\epsilon = 10^{-3}$ and compare, for several values of θ in the range 0 to 5 degrees, the field computed by our algorithm, $\widetilde{u}\,(\mathbf{x}, z_0)$, and the field computed by the Fresnel approximation, $u_f\,(\mathbf{x}, z_0)$, near the focal point $\mathbf{x} = \mathbf{r}_0$. Results displayed in Figure 3.3 demonstrate that the accuracy of the Fresnel approximation deteriorates rapidly as the focal point moves away from the optical axis. We also display the diffraction pattern computed by our algorithm and the pattern computed by the Fresnel approximation for $\theta = 5^\circ$ in Figure 3.4. The diffraction pattern obtained by the Fresnel approximation is both shifted and blurred when compared to the correct pattern.

We compare the two methods in a different manner in Figure 3.5, where we plot the error of each method at the focal point, i.e., $|u\,(\mathbf{r}_0, z_0) - \widetilde{u}\,(\mathbf{r}_0, z_0)|$ and $|u\,(\mathbf{r}_0, z_0) - u_f\,(\mathbf{r}_0, z_0)|$, as a function of the angle θ (we determined the true value $u\,(\mathbf{r}_0, z_0)$ by direct numerical integration). Our method maintains its accuracy for all $\theta \in [0, 5]$ degrees, while the Fresnel approximation is accurate to approximately 3 digits for $\theta = 0^\circ$ but has essentially no accurate digits for $\theta > 4^\circ$.

Remark 22. An often-cited paper on numerical light propagation [Syp95] (see also [SPG03]) claims that the Fresnel approximation produces accurate results at angles up to 18 degrees off the optical axis. Our example demonstrates that this claim is unsustainable.

Remark 23. From Figure 3.3, it may appear tempting to attempt to "correct" the Fresnel approximation by introducing a change of variable $\mathbf{x} \mapsto g\,(\mathbf{x})$, where the function $g : \mathbb{R}^2 \to \mathbb{R}^2$

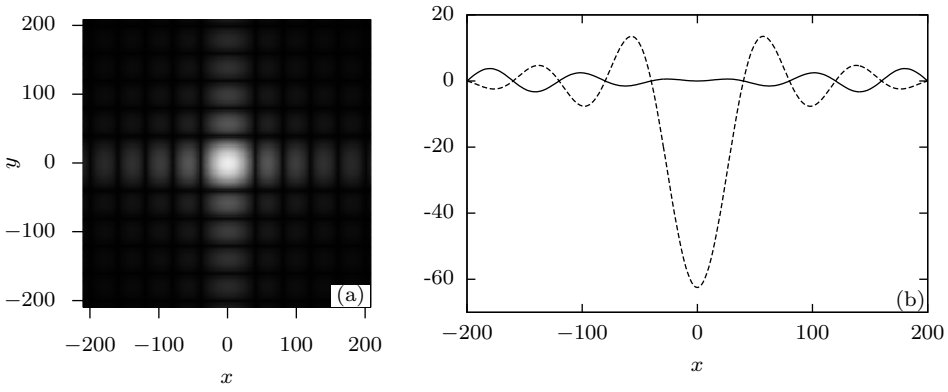

Figure 3.2: Propagation of a spherical wave restricted to a square aperture and converging to a point on the optical axis. As expected, the diffraction pattern is approximately a scaled version of the function $|\mathrm{sinc}\,(x)\,\mathrm{sinc}\,(y)|$. We display the magnitude of the field, $|u\,(x,y)|$, (a), and the real and the imaginary parts, $\mathcal{R}\mathrm{e}\,u\,(x,0)$ (solid line) and $\mathcal{I}\mathrm{m}\,u\,(x,0)$ (dashed line) of the field on the x-axis (b).

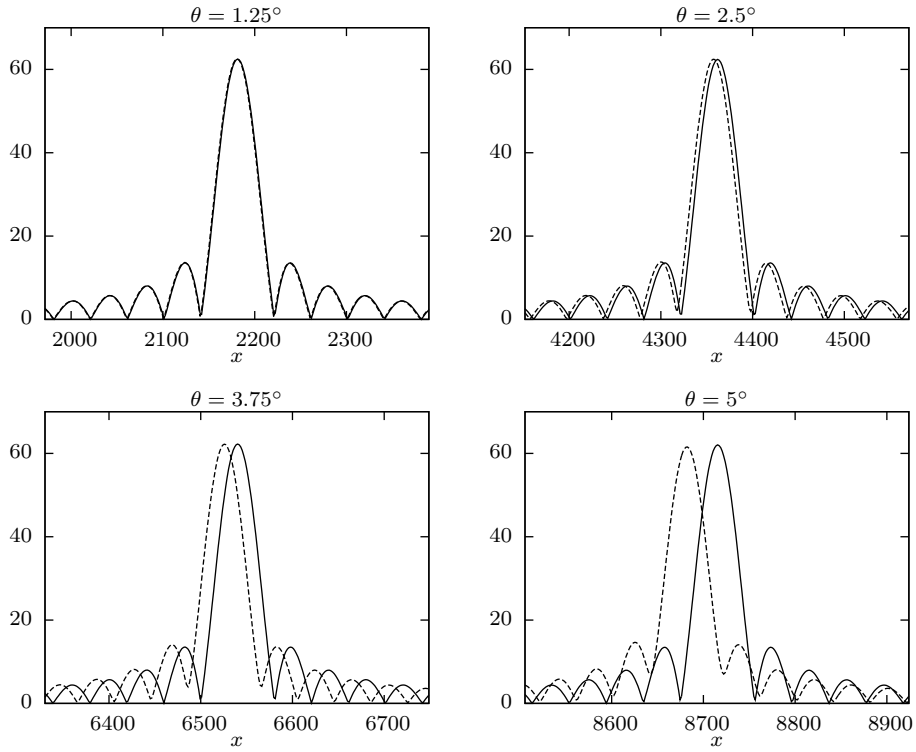

Figure 3.3: Comparison of the magnitude of the field evaluated near the focal point on the x-axis. We display the magnitude $|u(x, 0, z_0)|$, computed by our algorithm (solid line, correct to 3 digits) and by the Fresnel approximation (dashed line), as the focal point of a converging spherical wave moves away from the optical axis. The Fresnel approximation incorrectly computes both the position and the shape of the focal spot, e.g., compare the nulls between the main lobe and first side lobes in the bottom-right plot.

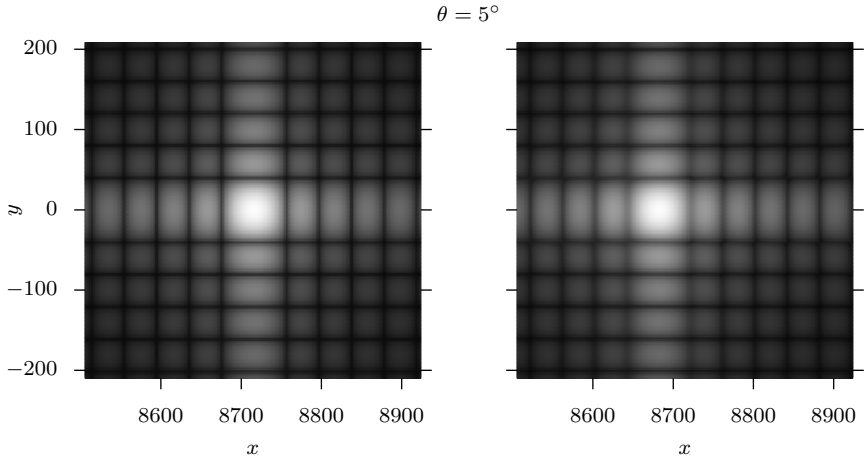

Figure 3.4: Comparison of the magnitude of the field for a focal point 5° off the optical axis computed by our algorithm correct to 3 digits (left), and by the Fresnel approximation (right). To enhance contrast, we plot the square root of the magnitude, $|u(x,y)|^{1/2}$. The Fresnel approximation shifts the location of the focal spot, and blurs the boundaries between the mainlobe and sidelobes. See also the bottom-right plot in Figure 3.3.

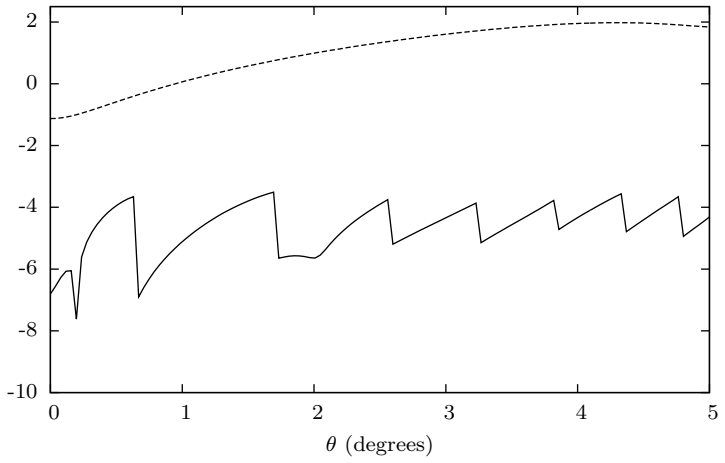

Figure 3.5: Comparison of the error of our method and that of the Fresnel approximation. We display the error of our method, $\log_{10}|u(\mathbf{r}_0,z_0)-\widetilde{u}(\mathbf{r}_0,z_0)|$ (solid line), and the error of the Fresnel approximation, $\log_{10}|u(\mathbf{r}_0,z_0)-u_f(\mathbf{r}_0,z_0)|$ (dashed line), at the focal point $(\mathbf{r}_0,z_0)$ of a converging spherical wave. We note that as the angle θ increases, additional terms are added to approximation (3.19), improving accuracy by about 1.5 digits each time and giving the solid line a "sawtooth" shape.

would be selected with the goal of rescaling the field computed by the Fresnel approximation, $u_f\left(g\left(\mathbf{x}\right), z\right)$, to more closely match the true field, $u\left(\mathbf{x}, z\right)$. In effect, the strategy would be to rescale the x-axis for the dashed lines in Figure 3.3 to better align the peaks of the solid and dashed lines. Unfortunately, our example shows that this approach could not succeed because the Fresnel approximation incorrectly computes the shape of the focal spot, in addition to its position (compare the nulls between the main lobe and side lobes in the bottom-right plot in Figure 3.3).

3.5.3 Relationship Between Computational Cost and Propagation Distance

The computational cost of our algorithm depends on the number of USFFTs required in (3.29), i.e., $R = R^{(1)} + \cdots + R^{(L)}$, where L is the number of terms needed to approximate the kernel in (3.19). As it turns out, R decreases with increasing z, which is expected since the application of the Rayleigh-Sommerfeld kernel asymptotically reduces to a single scaled Fourier transform as $z \to \infty$ (see Appendix B (§3.8)). On the other hand, for smaller values of z the field changes rapidly, and many USFFTs are required to accurately compute the field in these computationally-challenging regions.

Let us fix the aperture size $a = 2000$ wavelengths, the desired accuracy $\epsilon = 10^{-3}$, and set the number of input samples and output samples to $M \times M = N \times N = 512 \times 512$. We now examine the dependence of R on the propagation distance z for two different choices of output window size:

(1) a fixed output window of size $w = 10,000$ wavelengths, and

(2) a variable output window $w = w_{\max}$, where $w_{\max}$ is defined in (3.36) and is the largest output window that our method can accommodate for a given propagation distance z.

In Table 3.1, for several propagation distances, we give the number of terms, L, needed to approximate the kernel and the number of USFFTs, R, required to compute the field for

	$w = 10,000$			$w = w_{\max}$			
z	L	R	Time [s]	w	L	R	Time [s]
$50,000$	9	486	371	$10,389$	10	577	439
$100,000$	5	132	101	$18,836$	10	440	331
$250,000$	3	40	31.4	$39,426$	10	306	224
$1,000,000$	2	16	13.6	$115,170$	10	189	143
$10,000,000$	1	3	3.46	$696,895$	10	112	85.2

Table 3.1: Computational cost as a function of propagation distance. We show the dependence of the number of terms, L, and of the number of needed USFFTs, R, on the propagation distance, z, as well as the actual computing time. The center section corresponds to the fixed window size $w = 10,000$ wavelengths and the right section to the largest possible window, $w = w_{\max}$, where $w_{\max}$ is defined in (3.36).

these two choices of output window size.

In Table 3.1 we also provide timing results for a MATLAB-based implementation of the algorithm. These timings were obtained on a laptop computer with a 2.1 GHz AMD N950 processor and 8 GB of RAM. No effort was made made to optimize the code, and we expect that a careful implementation of the algorithm will be significantly faster. We also note that all USFFTs in the evaluation of (3.29) may be computed in parallel, so that the total computational time can be reduced substantially on a multiprocessor computer system.

3.6 Conclusions

We have described a fast algorithm for the propagation of coherent light between parallel planes separated by a linear, isotropic, and homogeneous medium. Our algorithm achieves any user-specified accuracy, in contrast to existing algorithms. As a consequence, our algorithm can rapidly and accurately compute the field in non-paraxial regions, i.e., regions far from the optical axis. Importantly for practical applications, the computational complexity of our algorithm is proportional to that of the FFT.

Conceptually, our approach may be viewed as a generalization of the Fresnel approximation. While the Fresnel approximation replaces the Rayleigh-Sommerfeld kernel with a single Gaussian with a purely imaginary exponent, we use a nonlinear algorithm to approx-

imate the kernel, for any user-specified accuracy, as a short linear combination of Gaussians with complex-valued exponents. We describe an algorithm to rapidly apply this approximate kernel to input data. The result is a fast algorithm that can achieve any user-specified accuracy over a large computational domain.

3.7 Appendix A: Accurate Propagation in the Fourier Domain

We observed earlier in §3.2.1 that if the propagation distance z is small, then the propagation problem may be accurately and efficiently solved in the Fourier domain. This is not a new observation—see, e.g., [SW06, MS09]—so our objective here is simply to demonstrate how this propagation method (the so-called angular spectrum method) can be implemented using the quadratures from Theorem 16 to ensure any user-specified accuracy.

Given boundary data f and a propagation distance z, the angular spectrum method amounts to the numerical evaluation of the integral

$$u(\mathbf{x}) = \int_{\|\mathbf{p}\| \leq c} \widehat{f}(\mathbf{p}) \, \widehat{K}_z(\|\mathbf{p}\|) \, e^{i2\pi \mathbf{x}\cdot\mathbf{p}} \, d\mathbf{p}, \tag{3.42}$$

where $\widehat{K}_z$ is given in (3.6) and

$$\widehat{f}(\mathbf{p}) = \int_A f(\mathbf{y}) \, e^{-i2\pi \mathbf{y}\cdot\mathbf{p}} \, d\mathbf{y} \tag{3.43}$$

is the Fourier transform of the boundary data. In (3.42) and (3.43) the boundary data has already been replaced by its space-limited and band-limited version, as described in §3.2.3, so that the function f is concentrated in a square aperture $A = \left[-\frac{a}{2}, \frac{a}{2}\right]^2$ and is band-limited to a disk of radius c for some user-specified accuracy ϵ.

The numerical evaluation of (3.42) amounts to solving a quadrature problem. That is, for accuracy ϵ, we seek quadrature nodes $\mathbf{p}_\ell$, $\ell = 1, \ldots, L$, and associated weights ω_ℓ such that for all $\mathbf{x} \in \left[-\frac{w}{2}, \frac{w}{2}\right]^2$,

$$\left| \int_{\|\mathbf{p}\| \leq c} \widehat{f}(\mathbf{p}) \, \widehat{K}_z(\|\mathbf{p}\|) \, e^{i2\pi \mathbf{x}\cdot\mathbf{p}} \, d\mathbf{p} - \sum_{\ell=1}^{L} \omega_\ell \widehat{f}(\mathbf{p}_\ell) \, \widehat{K}_z(\|\mathbf{p}_\ell\|) \, e^{i2\pi \mathbf{x}\cdot\mathbf{p}_\ell} \right| \leq \frac{\epsilon \, \|f\|_1}{z}.$$

The construction of such quadrature rules for two-dimensional integrals of space-limited and band-limited functions, computed over disks and accurate for any user-specified accuracy ϵ, was described in [BKM07]. The resulting near-optimal quadrature nodes lie on a polar grid, so that

$$\mathbf{p}_\ell = \mathbf{p}_{mn} = \rho_m \left(\cos\phi_{mn}, \sin\phi_{mn}\right)$$

for some radial nodes ρ_m, $m = 1, \ldots, M$, and angular nodes ϕ_{mn}, $m = 1, \ldots, M$ and $n = 1, \ldots, N_m$. The total number of nodes $L = L\left(c', \epsilon\right)$ is nearly optimal and depends only weakly on the accuracy ϵ and quadratically on the bandlimit c' of the integrand. We estimate c' as $c' = c\left(a + w\right)/\sqrt{2} + c_K$, where c_K is the bandlimit of $\widehat{K}_z$ in the domain of integration. Estimating the shortest period of oscillation of $\widehat{K}_z$ in the disk of radius c yields

$$c_K \approx \begin{cases} \frac{z}{\left(c^2 z^2 + 2z\sqrt{1-c^2} - 1\right)^{\frac{1}{2}} - cz}, & \text{if } c < 1 \\ z^2 + z\sqrt{z^2 - 1}, & \text{if } c \geq 1. \end{cases} \tag{3.44}$$

Once we specify the location of the desired output samples $\mathbf{x}_n \in \left[-\frac{w}{2}, \frac{w}{2}\right]^2$, $n = 1, \ldots, N$, we compute u rapidly by evaluating

$$\widetilde{u}\left(\mathbf{x}_n, z\right) = \sum_{\ell=1}^{L} \omega_\ell \widehat{f}\left(\mathbf{p}_\ell\right) \widehat{K}_z\left(\left\|\mathbf{p}_\ell\right\|\right) e^{i2\pi \mathbf{x}_n \cdot \mathbf{p}_\ell}, \tag{3.45}$$

which requires a single USFFT.

The evaluation of (3.45) requires values of $\widehat{f}$ at the quadrature nodes $\mathbf{p}_\ell$. In most cases, $\widehat{f}$ is not known explicitly and must be computed by numerically evaluating (3.43). This is exactly the quadrature problem addressed in Theorem 16, and we obtain a tensor product grid of quadrature nodes $\mathbf{y}_{jj'} = (y_j, y_{j'})$, $j, j' = 1, \ldots, J$, and associated weights $\tau_j \tau_{j'}$ such that for each $\mathbf{p}_\ell$, $\ell = 1, \ldots, L$,

$$\left| \int_A f\left(\mathbf{y}\right) e^{-i2\pi \mathbf{y} \cdot \mathbf{p}_\ell} \, d\mathbf{y} - \sum_{j,j'=1}^{J} \tau_j \tau_{j'} f\left(\mathbf{y}_{jj'}\right) e^{-i2\pi \mathbf{y}_{jj'} \cdot \mathbf{p}_\ell} \right| \leq \epsilon \left\|f\right\|_1 . \tag{3.46}$$

Thus, $\widehat{f}\left(\mathbf{p}_\ell\right)$ may be computed accurately and rapidly with one USFFT.

Taken together, formulae (3.45) and (3.46) allow us to evaluate the field with any user-specified accuracy with only two USFFTs. This method is simple and effective, provided that the bandlimit c_K in (3.44), which increases with the distance z, is moderate. If c_K is so large that the number of quadrature nodes required to accurately compute (3.45) makes its evaluation infeasible, then the propagation problem should be treated in the spatial domain, using the method we describe in §3.3.

Remark 24. The dependence of the bandlimit c_K in (3.44) on z is nearly linear provided that $c < 1$, but it depends almost quadratically on z if $c \geq 1$. In the latter case, evanescent waves are present in the solution, and very fine sampling is required to accurately propagate them. We note that in virtually all cases of practical interest, the contribution of evanescent waves may be neglected entirely after only a few wavelengths, so in practice, we may assume that $c < 1$ unless z is small, and therefore the bandlimit $c = z^2 + z\sqrt{z^2 - 1}$ is not impractically large.

3.8 Appendix B: A Comment on the Fraunhofer Approximation

If the propagation distance z is large with respect to the sizes of the input aperture and the output window, it is common to estimate the field $u(\mathbf{x}, z)$ using the Fraunhofer approximation (3.14). We note that many optics texts derive this far-field approximation by approximating the Fresnel approximation. This is unfortunate, because the result is only valid near the optical axis and, therefore, the size of the output region where the asymptotics are accurate is severely restricted. In contrast, there is a well known asymptotic approximation of solutions to the Helmholtz equation that is valid in a much larger output region, and which may be evaluated with a single USFFT. For completeness, we now recall this alternative far-field approximation and relate it to the Fraunhofer approximation (3.14).

Recall Rayleigh's integral formula (3.1) for $u(\mathbf{x}, z)$,

$$u(\mathbf{x}, z) = -\frac{1}{2\pi} \int_A f(\mathbf{y}) \frac{\partial}{\partial z} \left(\frac{e^{i2\pi R}}{R} \right) d\mathbf{y}, \tag{3.47}$$

where A is the input aperture, $f(\mathbf{y}) = u(\mathbf{y}, 0)$ is the boundary data, and $R = \sqrt{z^2 + \|\mathbf{x} - \mathbf{y}\|^2}$. Performing the indicated differentiation gives

$$\frac{\partial}{\partial z}\left(\frac{e^{i2\pi R}}{R}\right) = z e^{i2\pi R}\left(\frac{i2\pi}{R^2} - \frac{1}{R^3}\right),$$

which has the large-R asymptotic approximation

$$\frac{\partial}{\partial z}\left(\frac{e^{i2\pi R}}{R}\right) \sim \frac{i2\pi z e^{i2\pi R}}{R^2}, \quad R \to \infty. \tag{3.48}$$

We now write R as

$$R = \sqrt{z^2 + \|\mathbf{x}\|^2}\left(1 + \frac{\|\mathbf{y}\|^2}{z^2 + \|\mathbf{x}\|^2} - 2\frac{\mathbf{x}\cdot\mathbf{y}}{z^2 + \|\mathbf{x}\|^2}\right)^{\frac{1}{2}}$$

and observe that, since $\mathbf{y} \in A$ and z is much larger than the size of the aperture A, $z \gg \|\mathbf{y}\|$. Hence, we obtain the asymptotic approximation

$$\begin{aligned} R &\sim \sqrt{z^2 + \|\mathbf{x}\|^2}\left(1 - 2\frac{\mathbf{x}\cdot\mathbf{y}}{z^2 + \|\mathbf{x}\|^2}\right)^{\frac{1}{2}} \\ &\sim \sqrt{z^2 + \|\mathbf{x}\|^2}\left(1 - \frac{\mathbf{x}\cdot\mathbf{y}}{z^2 + \|\mathbf{x}\|^2}\right), \quad \mathbf{y} \in A,\ z \to \infty. \end{aligned} \tag{3.49}$$

We emphasize that we did not need to assume that $\|\mathbf{x}\|$ is small, i.e., close to the optical axis, to obtain this approximation. Following the standard procedure, we substitute (3.49) into the exponent in (3.48) and $R \sim \sqrt{z^2 + \|\mathbf{x}\|^2}$ into the denominator, and then use the resulting asymptotic Green's function in (3.47) to arrive at

$$u(\mathbf{x}, z) \sim -\frac{iz e^{i2\pi\sqrt{z^2+\|\mathbf{x}\|^2}}}{z^2 + \|\mathbf{x}\|^2}\widehat{f}\left(\frac{\mathbf{x}}{\sqrt{z^2 + \|\mathbf{x}\|^2}}\right), \quad z \to \infty. \tag{3.50}$$

The result in (3.50) relates the far-field diffraction pattern to the rescaled Fourier transform of the boundary data. We accurately and rapidly evaluate this formula using generalized Gaussian quadratures for band-limited functions from Theorem 16 and the USFFT, in exactly the same manner as in the evaluation of (3.46).

To obtain the Fraunhofer approximation, we could assume that $\|\mathbf{x}\| \ll z$ and make a further approximation by retaining one and two terms of the expansion $\sqrt{z^2 + \|\mathbf{x}\|^2} =$

$z\left(1+\frac{1}{2}\frac{\|\mathbf{x}\|^2}{z^2}+\cdots\right)$ in the denominators and exponent of (3.50), respectively, yielding the standard Fraunhofer approximation (cf. [Goo05, §4.3]),

$$u\left(\mathbf{x},z\right)\approx\frac{e^{2\pi iz}e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz}\widehat{f}\left(\frac{\mathbf{x}}{z}\right). \tag{3.51}$$

However, this additional approximation is not advisable since, unlike (3.50), the Fraunhofer approximation is only valid for points close to the the optical axis, i.e., $\|\mathbf{x}\|\ll z$. There is no computational advantage to be gained by using (3.51), with its restricted region of validity, instead of (3.50), valid for all $\mathbf{x}$ provided that z is sufficiently large, since both of these formulae may be evaluated at the same cost with a single USFFT.

3.9 Appendix C: Algorithm for Approximation by Exponential Sums

We approximate, for any user-specified accuracy ϵ, a smooth function $f\left(x\right)$, $0\leq x\leq 1$, by a linear combination of exponentials,

$$\left|f\left(x\right)-\sum_{\ell=1}^{L}w_\ell e^{-\eta_\ell x}\right|\leq\epsilon,\quad x\in\left[0,1\right], \tag{3.52}$$

where the number of complex-valued weights w_ℓ and exponents η_ℓ is nearly minimal. We obtain this representation by solving a discrete version of the approximation problem. Given $2N+1$ evenly-spaced samples of $f\left(x\right)$ and target accuracy $\epsilon>0$, we find the (nearly) minimal number of complex-valued weights w_ℓ and nodes γ_ℓ such that

$$\left|f\left(\frac{k}{2N}\right)-\sum_{\ell=1}^{L}w_\ell\gamma_\ell^k\right|\leq\epsilon,\quad 0\leq k\leq 2N. \tag{3.53}$$

We must choose the number of samples $2N+1$ large enough so that the function can be accurately reconstructed from its samples. As a result, we obtain the solution to the continuous problem (3.52) from the solution to the discrete problem (3.53) by setting $\eta_\ell=-2N\log\gamma_\ell$. We now describe the algorithm given in [BM05] (see also [BM10]) to obtain approximation (3.53).

- Build the $N+1 \times N+1$ Hankel matrix

$$\mathbf{H}_{jk} = f\left(\frac{j+k}{2N}\right), \quad j,k \in [0,N].$$

- Find a vector $\mathbf{u} = (u_0, \ldots, u_N)^T$ satisfying

$$\mathbf{H}\mathbf{u} = \sigma\overline{\mathbf{u}},$$

  with positive σ close to the target accuracy ϵ. A problem of this form is known as a con-eigenvalue problem (see, e.g., [HJ90, §4.6]), $\mathbf{u}$ is a con-eigenvector, and σ is a con-eigenvalue. In our case, $\mathbf{H}$ is a Hankel matrix and hence symmetric; the existence of a solution $(\sigma, \mathbf{u})$ follows from Takagi's factorization (see, e.g., [BM05, pp. 22]), as does the fact that we may take σ to be a singular value of $\mathbf{H}$ and $\mathbf{u}$ to be a specific singular vector.

- Given singular values $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_N$, we select a sufficiently small σ_L, which determines the accuracy of approximation, and the corresponding singular vector $\mathbf{u} = (u_0, \ldots, u_N)^T$.

- Compute the roots γ_ℓ of the con-eigenpolynomial $u(z) = \sum_{n=0}^{N} u_n z^n$ whose coefficients are the entries of the vector $\mathbf{u}$ from the previous step.

- Obtain the weights w_ℓ by solving the least-squares Vandermonde system

$$\sum_{\ell=1}^{N} w_\ell \gamma_\ell^k = f\left(\frac{k}{2N}\right), \quad 0 \leq k \leq 2N. \tag{3.54}$$

  Typically, only L weights w_ℓ have absolute value larger than the target accuracy ϵ. We then retain only those nodes γ_ℓ that correspond to the significant weights and solve the corresponding Vandermonde system (3.54) again.

The theory underlying this algorithm may be found in [BM05] and may be traced back to the work of Adamjan, Arov, and Krein [AAK68a, AAK68b, AAK71].

Bibliography

[AAK68a] V. M. Adamjan, D. Z. Arov, and M. G. Kreĭn, Infinite Hankel matrices and generalized Carathéodory-Fejér and I. Schur problems, Funkcional. Anal. i Priložen. **2** (1968), no. 4, 1–17. MR 58 #30446

[AAK68b] ———, Infinite Hankel matrices and generalized problems of Carathéodory-Fejér and F. Riesz, Funkcional. Anal. i Priložen. **2** (1968), no. 1, 1–19. MR 38 #2591

[AAK71] ———, Analytic properties of the Schmidt pairs of a Hankel operator and the generalized Schur-Takagi problem, Math. USSR Sbornik **15** (1971), no. 1, 34–75. MR 45 #7505

[AS70] M. Abramowitz and I. A. Stegun, Handbook of mathematical functions, 9 ed., Dover Publications, 1970.

[BC02] G. Beylkin and R. Cramer, Toward multiresolution estimation and efficient representation of gravitational fields, Celestial Mechanics and Dynamical Astronomy **84** (2002), no. 1, 87–104.

[Bey92] G. Beylkin, On the representation of operators in bases of compactly supported wavelets, SIAM J. Numer. Anal. **29** (1992), no. 6, 1716–1740. MR 93k:65019

[Bey95a] ———, On factored FIR approximation of IIR filters, Appl. Comput. Harmon. Anal. **2** (1995), no. 3, 293–298. MR 1 339 807

[Bey95b] ———, On the fast Fourier transform of functions with singularities, Appl. Comput. Harmon. Anal. **2** (1995), no. 4, 363–381. MR 96i:65122

[BKM07] G. Beylkin, C. Kurcz, and L. Monzón, Grids and transforms for band-limited functions in a disk, Inverse Problems **23** (2007), no. 5, 2059–2088.

[BLM12] G. Beylkin, R. D. Lewis, and L. Monzón, On the design of highly accurate and efficient IIR and FIR filters, IEEE Trans. Signal Process. **60** (2012), no. 8, 4045–4054, http://dx.doi.org/10.1109/TSP.2012.2197397.

[BM02] G. Beylkin and L. Monzón, On generalized Gaussian quadratures for exponentials and their applications, Appl. Comput. Harmon. Anal. **12** (2002), no. 3, 332–373. MR 2003f:41048

[BM05] ———, On approximation of functions by exponential sums, Appl. Comput. Harmon. Anal. **19** (2005), no. 1, 17–48.

[BM09] ———, Nonlinear inversion of a band-limited Fourier transform, Appl. Comput. Harmon. Anal. **27** (2009), no. 3, 351–366.

[BM10] ———, Approximation of functions by exponential sums revisited, Appl. Comput. Harmon. Anal. **28** (2010), no. 2, 131–149.

[Bou54] C. J. Bouwkamp, Diffraction theory, Reports on Progress in Physics **17** (1954), no. 1, 35–100.

[BWB99] M. Born, E. Wolf, and A. B. Bhatia, Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light, 7 ed., Cambridge University Press, 1999.

[CGMR05] H. Cheng, Z. Gimbutas, P.-G. Martinsson, and V. Rokhlin, On the compression of low-rank matrices, SIAM Journal of Scientific Computing **205** (2005), no. 1, 1389–1404.

[CJ82] A. Chottera and G. Jullien, A linear programming approach to recursive digital filter design with linear phase, IEEE Trans. Circuits Syst. **29** (1982), no. 3, 139–149.

[CPC92] B. S. Chen, S. C. Peng, and B. W. Chiou, IIR filter design via optimal Hankel-norm approximation, Circuits, Devices and Systems, IEE Proceedings G **139** (1992), no. 5, 586–590.

[CR83] R. E. Crochiere and L. R. Rabiner, Multirate digital signal processing, Prentice-Hall signal processing series, Prentice-Hall, 1983.

[Dau93] I. Daubechies, Orthonormal bases of compactly supported wavelets II. Variations on a theme, SIAM J. Math. Anal. **24** (1993), no. 2, 499–519.

[Dig79] Digital Signal Processing Committee. IEEE Acoustics, Speech, and Signal Processing Society (ed.), Programs for digital signal processing, IEEE Press, 1979.

[DR93] A. Dutt and V. Rokhlin, Fast Fourier transforms for nonequispaced data, SIAM J. Sci. Comput. **14** (1993), no. 6, 1368–1393. MR 95d:65114

[Fou07] J. Fourier, Mémoire sur la propagation de la chaleur dans les corps solides, 1807, Paris.

[Goo05] J. W. Goodman, Introduction to Fourier optics, 3 ed., McGraw-Hill physical and quantum electronics series, Roberts & Co., Englewood, Colorado, 2005.

[GR69] B. Gold and C. M. Rader, Digital processing of signals, Lincoln Laboratory publications, McGraw-Hill, 1969.

[GST83] M. Gutknecht, J. Smith, and L. Trefethen, The Carathéodory-Fejér method for recursive digital filter design, IEEE Trans. Acoust. Speech Signal Process. **31** (1983), no. 6, 1417–1426.

[HB12] T. S. Haut and G. Beylkin, Fast and accurate con-eigenvalue algorithm for optimal rational approximations, SIAM J. Matrix Anal. Appl. **33** (2012), no. 4, 1101–1125, http://dx.doi.org/10.1137/110821901, see also arXiv:1012.3196 [math.NA].

[HJ90] R. A. Horn and C. R. Johnson, Matrix analysis, Cambridge University Press, Cambridge, 1990. MR 91i:15001

[HJB84] M. T. Heideman, D. H. Johnson, and C. S. Burrus, Gauss and the history of the fast Fourier transform, IEEE ASSP Magazine **1** (1984), no. 4, 14–21.

[HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, SIAM Review **53** (2011), no. 2, 217–288.

[JBB10] B. A. Jones, G. H. Born, and G. Beylkin, Comparisons of the cubed sphere gravity model with the spherical harmonics, Journal of Guidance, Control, and Dynamics **33** (2010), no. 2, 415–425.

[JMR01] S. Jaffard, Y. Meyer, and R. D. Ryan, Wavelets: tools for science & technology, revised ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. MR 2002g:00007

[KBG92] P. J. Kootsookos, R. R. Bitmead, and M. Green, The Nehari shuffle: FIR(q) filter design with guaranteed error bounds, IEEE Trans. Signal Process. **40** (1992), no. 8, 1876–1883.

[LBM13] R. D. Lewis, G. Beylkin, and L. Monzón, Fast and accurate propagation of coherent light, 2013, to be submitted.

[LG05] J-Y. Lee and L. Greengard, The type 3 nonuniform FFT and its applications, J. Comput. Phys. **206** (2005), no. 1, 1–5. MR 2135833

[LP61] H. J. Landau and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty II, Bell System Tech. J. **40** (1961), 65–84. MR 25 #4147

[LP62] ———, Prolate spheroidal wave functions, Fourier analysis and uncertainty III, Bell System Tech. J. **41** (1962), 1295–1336.

[MBH99] L. Monzón, G. Beylkin, and W. Hereman, Compactly supported wavelets based on almost interpolating and nearly linear phase filters (coiflets), Appl. Comput. Harmon. Anal. **7** (1999), no. 2, 184–210. MR 2002c:42055

[MMS93] S. K. Mitra, A. Mahalonobis, and T. Saramaki, A generalized structural subband decomposition of FIR filters and its application in efficient FIR filter design and implementation, IEEE Trans. on Circuits and Systems Part II **40** (1993), no. 6, 363–374.

[Mon99] L. Monzón, Linear phase perfect reconstruction filters and wavelets with even symmetry, arXiv:1112.5214 [math.NA] (1999).

[MP05] J. H. McClellan and T. W. Parks, A personal history of the Parks-McClellan algorithm, IEEE Signal Process. Mag. **22** (2005), no. 2, 82–86.

[MS09] K. Matsushima and T. Shimobaba, Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields, Optics Express **17** (2009), no. 22, 19662–19673.

[New87] I. Newton, Philosophiæ naturalis principia mathematica, Philosophical Transactions of the Royal Society, London, 1687.

[OS89] A. V. Oppenheim and R. W. Schafer, Discrete-time signal processing, Prentice-Hall signal processing series, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[PB87] T. W. Parks and C. S. Burrus, Digital filter design, Topics in digital signal processing, John Wiley & Sons, Inc., 1987.

[PM72] T. W. Parks and J. H. McClellan, Chebyshev approximation for nonrecursive digital filters with linear phase, IEEE Trans. Circuit Theory **CT-19** (1972), no. 2, 189–194.

[Rad06] C. M. Rader, DSP history—the rise and fall of recursive digital filters, IEEE Signal Process. Mag. **23** (2006), no. 6, 46–49.

[Ray97] Lord Rayleigh, On the passage of waves through apertures in plane screens, and allied problems, Philos. Mag. **43** (1897), 259–272.

[Rem69] E. Ya. Remez, Fundamentals of numerical methods for Chebyshev approximations, Naukova Dumka, Kiev, 1969, in Russian.

[She67] G. C. Sherman, Application of the convolution theorem to Rayleigh's integral formulas, J. Opt. Soc. Am. **57** (1967), no. 4, 546–547.

[Sle64] D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty IV. Extensions to many dimensions; generalized prolate spheroidal functions, Bell System Tech. J. **43** (1964), 3009–3057. MR 31 #5993

[Sle78] ———, Prolate spheroidal wave functions, Fourier analysis and uncertainty V. The discrete case, Bell System Tech. J. **57** (1978), 1371–1430.

[Som12] A. Sommerfeld, Die Greensche funktion der schwingungsgleichung, Jahresbericht der Deutschen Mathematiker-Vereinigung **21** (1912), 309–353.

[Som49] ———, Partial differential equations in physics, Pure and Applied Mathematics, Academic Press, New York, 1949.

[SP61] D. Slepian and H. O. Pollak, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, Bell System Tech. J. **40** (1961), 43–63. MR 25 #4146

[SPG03] M. Sypek, C. Prokopowicz, and M. Górecki, Image multiplying and high-frequency oscillations effects in the Fresnel region light propagation simulation, Optical Engineering **42** (2003), no. 11, 3158–3164.

[SW06] F. Shen and A. Wang, Fast-Fourier-transform based numerical integration method for the Rayleigh-Sommerfeld diffraction formula, Applied Optics **45** (2006), no. 6, 1102–1110.

[Syp95] M. Sypek, Light propagation in the Fresnel region. New numerical approach, Optics Communications **116** (1995), no. 1–3, 43–48.

[TCHR01] A. Tarczynski, G. D. Cain, E. Hermanowicz, and M. Rojewski, A WISE method for designing IIR filters, IEEE Trans. Signal Process. **49** (2001), no. 7, 1421–1432.

[XRY01] H. Xiao, V. Rokhlin, and N. Yarvin, Prolate spheroidal wavefunctions, quadrature and interpolation, Inverse Problems **17** (2001), no. 4, 805–838.

[Zol77] E. I. Zolotarjov, Application of the elliptic functions to the problems on the functions of the least and most deviation from zero, Zapiskah Rossijskoi Akad. Nauk., 1877, in Russian.

Appendix A

Technical Details Concerning Light Propagation

A.1 Proof of Lemma 18

Proof. Using (3.27) we have

$$\begin{aligned}
&\mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'} - \sum_{q=1}^{I^{(\ell)}} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq} \sum_{s=1}^{I^{(\ell)}} \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{m's} \\
&\qquad = \mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'} \\
&\qquad\quad - \left(\mathbf{S}^{(\ell)}_{nm} - \sum_{q=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq}\right)\left(\mathbf{S}^{(\ell)}_{n'm'} - \sum_{s=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{m's}\right) \\
&\qquad = \mathbf{S}^{(\ell)}_{nm}\left(\sum_{s=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{m's}\right) + \mathbf{S}^{(\ell)}_{n'm'}\left(\sum_{q=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq}\right) \\
&\qquad\quad - \left(\sum_{q=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq}\right)\left(\sum_{s=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{m's}\right).
\end{aligned}$$

Observing that

$$\left|\mathbf{U}^{(\ell)}_{nq}\right|, \left|\mathbf{V}^{(\ell)}_{mq}\right| \le 1 \quad \text{and} \quad \left|\mathbf{S}^{(\ell)}_{nm}\right|, \left|\mathbf{S}^{(\ell)}_{n'm'}\right| \le e^{\frac{\left|\alpha^{(\ell)}\right|}{2} aw},$$

it follows that

$$\left|\mathbf{S}^{(\ell)}_{nm}\mathbf{S}^{(\ell)}_{n'm'} - \sum_{q,s=1}^{I^{(\ell)}} \sigma^{(\ell)}_q \sigma^{(\ell)}_s \mathbf{U}^{(\ell)}_{nq}\mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{mq}\mathbf{V}^{(\ell)}_{m's}\right| \le \epsilon_R 2e^{\frac{\left|\alpha^{(\ell)}\right|}{2} aw},$$

where we have neglected the term

$$\left(\sum_{q=I^{(\ell)}+1}^{\min(M,N)} \sigma^{(\ell)}_q \left|\mathbf{U}^{(\ell)}_{nq}\mathbf{V}^{(\ell)}_{mq}\right|\right)\left(\sum_{s=I^{(\ell)}+1}^{\min(J,M)} \sigma^{(\ell)}_s \left|\mathbf{U}^{(\ell)}_{n's}\mathbf{V}^{(\ell)}_{m's}\right|\right)$$

which has size $\mathcal{O}\left(\epsilon_R^2\right)$. □

A.2 Rigorous Estimates Relating to Computational Complexity

The key step in the algorithm described in §3.3 is the construction of approximation (3.19), where, for a fixed distance z and desired accuracy ϵ_K, we approximate the Rayleigh-Sommerfeld kernel $K_z\left(r\right)$ as a linear combination of Gaussians with complex exponents. This approximation must be valid on an interval $0 \leq r \leq r_{\max}$, where, in the case of a square aperture of side-length a and square output window of side-length w, $r_{\max} = \left(a+w\right)/\sqrt{2}$. As described in §3.3.1, we obtain this approximation by removing the most-oscillatory factor $e^{i\frac{\pi}{z}r^2}$ from $K_z\left(r\right)$ then approximating the remaining function, viz. $A_z\left(r\right)$ defined in (3.17), using Gaussians with complex exponents,

$$\left| A_z\left(r\right) - \sum_{\ell=1}^{L} w_\ell e^{-\eta_\ell r^2} \right| \leq \epsilon_K, \quad 0 \leq r \leq r_{\max}.$$

Three components of this approximation ultimately determine the computational cost of our algorithm:

(1) The number of terms L.

(2) The number of input samples M^2 required to evaluate the integrals in (3.22). This depends on the maximum bandlimit of the integrands, which in turn is determined by the bandlimit of the input function and the values of the exponents η_ℓ.

(3) For each $\ell = 1, \ldots, L$, the number of terms $R^{(\ell)}$ needed in (3.28) to approximate the tensors $\mathbf{T}^{(\ell)}$ defined in (3.24).

We now provide estimates of these quantities.

A.2.1 Number of Terms Needed to Approximate the Kernel

It turns out that, for the values of r of interest here, the behavior of the function $A_z(r)$ closely resembles that of $e^{-i\frac{\pi r^4}{4z^3}}$, i.e.,

$$A_z(r) \approx e^{-i\frac{\pi r^4}{4z^3}}, \tag{A.1}$$

which comes from the Taylor series

$$\sqrt{1+\left(\frac{r}{z}\right)^2} = 1 + \frac{1}{2}\left(\frac{r}{z}\right)^2 - \frac{1}{8}\left(\frac{r}{z}\right)^4 + \mathcal{O}\left(\left(\frac{r}{z}\right)^6\right).$$

Recall that one of our goals is to construct an algorithm whose computational cost does not increase with z. Approximation (A.1) implies that the number of terms L needed to approximate A_z will depend on the ratio $\gamma = r_{\max}/z^{\frac{3}{4}}$ (and also on the desired accuracy ϵ_K). We can estimate the number of terms required using techniques similar to those in [BM10], where functions are approximated as linear combinations of complex Gaussians by manipulating their integral representations. The derivations are somewhat technical, so here we simply present the bound

$$L = L(\gamma, \epsilon) \leq \frac{2\log\epsilon_K^{-1}}{\pi} \times \frac{-B - \sqrt{B^2 - 8B\left(\log\epsilon_K + \log 2\sqrt{\pi B}\right)}}{B - \sqrt{B^2 - 8B\log\epsilon_K}}, \tag{A.2}$$

where

$$B = \frac{\pi\gamma^4}{4}. \tag{A.3}$$

The rightmost factor in (A.2) depends approximately linearly on B and only weakly on ϵ_K, so that $L = \mathcal{O}\left(\gamma^4 \log\epsilon_K^{-1}\right)$. Since the number of terms grows rapidly with γ, we require that $\gamma \leq 2.62$ to ensure that the approximation is efficient. This implies that the maximum output window is given by (3.36), and we assume that a, w, and z satisfy $\gamma \leq 2.62$ for the remainder of this section.

The same integral-based techniques that lead to (A.2) also yield the bounds

$$|\alpha_\ell| \leq \frac{B+D}{r_{\max}^2} \tag{A.4}$$

and

$$|\beta_\ell| \leq \frac{\pi}{z} + \frac{D}{r_{\max}^2}, \tag{A.5}$$

where

$$D = D\left(B, \epsilon_K\right) = \sqrt{B^2 - 8B\left(\log \epsilon_K + \log 2\sqrt{\pi B}\right)}.$$

(Recall from §3.3.2 that $\alpha_\ell = \mathcal{R}\mathrm{e}\,\eta_\ell$ and $\beta_\ell = \mathcal{I}\mathrm{m}\,\eta_\ell - \frac{\pi}{z}$, where η_ℓ, $\ell = 1, \ldots, L$, are the exponents used to approximate the Rayleigh-Sommerfeld kernel in (3.19).) We will use these bounds below to determine the number of required input samples M^2 and the number of terms $R^{(\ell)}$ in the approximations (3.28). To simplify the computations that follow, let us estimate their values. Since $\gamma \leq 2.62$, we have from (A.3) that $B \leq 37$. For $\epsilon_K = 10^{-3}$ we have $D \leq 50.1$, and for $\epsilon_K = 10^{-6}$ we have $D \leq 67.5$. Since $r_{\max} = (a+w)/\sqrt{2}$ is typically several thousand wavelengths, and often much larger, we can see that

$$|\alpha_\ell| \ll aw \quad \text{and} \quad |\beta_\ell| \approx \frac{\pi}{z}.$$

A.2.2 Number of Input Samples

The number of quadrature nodes (input samples) M^2 required in Theorem 16 to evaluate the integrals in (3.22) is determined by the bandlimits of the integrands and the desired accuracy ϵ_Q. We use the bound (A.5) to estimate the number of input samples required to accurately evaluate the integrals.

We start by rescaling the variables $\mathbf{x}$ and $\mathbf{y}$ to the unit square by defining $\mathbf{x} = \frac{w}{2}\mathbf{x}'$ and $\mathbf{y} = \frac{a}{2}\mathbf{y}'$, so that the integrals in (3.22) become

$$\frac{a^2}{4} \int\limits_{[-1,1]^2} f\left(\frac{a}{2}\mathbf{y}'\right) e^{-\frac{(\alpha_\ell + i\beta_\ell)a^2}{4}\|\mathbf{y}'\|^2} e^{\frac{\alpha_\ell aw}{4}\mathbf{x}'\cdot\mathbf{y}'} e^{i\frac{\beta_\ell aw}{2}\mathbf{x}'\cdot\mathbf{y}'}\, d\mathbf{y}', \quad \ell = 1, \ldots, L.$$

The bandlimit of the integrand is the sum of the bandlimits of each of the factors. Let c be the bandlimit of the the (rescaled) input function $f\left(\frac{a}{2}\mathbf{y}'\right)$. We now estimate the bandlimits of the other factors in the integrands.

- When estimating the bandlimit of $e^{-\frac{(\alpha_\ell + i\beta_\ell)a^2}{4}\|\mathbf{y}'\|^2}$, we may neglect the influence of α_ℓ since $|\alpha_\ell| \ll a^2$. We may also use the approximation $|\beta_\ell| \approx \frac{\pi}{z}$. Since, in the unit square, the shortest period of oscillation of the function $e^{-i\frac{\pi a^2}{4z}\|\mathbf{y}'\|^2}$ is $\sqrt{2}\left(\sqrt{1+\frac{4z}{a^2}}-1\right) \approx \frac{2\sqrt{2}z}{a^2}$, we estimate the bandlimit of this term as $c_2 = \frac{a^2}{2\sqrt{2}z}$.

- Since $|\alpha_\ell| \ll aw$, the factor $e^{\frac{\alpha_\ell a w}{4}\mathbf{x}'\cdot\mathbf{y}'}$ does not significantly impact the bandlimit of the integrand, so we neglect it completely.

- Because $\|\mathbf{x}'\| \leq \sqrt{2}$, the bandlimit of the factor $e^{i\frac{\beta_\ell a w}{2}\mathbf{x}'\cdot\mathbf{y}'}$ is $c_3 = \frac{\pi a w}{\sqrt{2}z}$, where we used the approximation $|\beta_\ell| \approx \frac{\pi}{z}$.

Thus, the bandlimit of the integrand is approximately

$$c' = c + c_2 + c_3 = c + \frac{a^2}{2\sqrt{2}z} + \frac{\pi a w}{\sqrt{2}z}. \tag{A.6}$$

From [BM02], we have that, for a desired accuracy ϵ_Q, the number of samples required to evaluate the integrals (3.22) satisfies $M^2 = \mathcal{O}\left((c')^2 \log^2 \epsilon_Q^{-1}\right)$.

A.2.3 Number of Terms Needed to Approximate the Tensors

Now let us use the bound (A.4) to estimate the number of terms required in the approximations (3.28). The tensors $\mathbf{S}^{(\ell)}$ in (3.25) are discrete approximations of the functions

$$S^{(\ell)}(x,y) = e^{2\alpha_\ell xy}, \quad x \in \left[-\frac{w}{2}, \frac{w}{2}\right],\ y \in \left[-\frac{a}{2}, \frac{a}{2}\right], \text{ and } \ell = 1, \ldots, L,$$

which have the Chebyshev expansions

$$e^{2\alpha_\ell xy} = J_0(-i\alpha_\ell a x) + 2\sum_{n=1}^{\infty} i^n J_n(-i\alpha_\ell a x)\, T_n\left(\frac{2y}{a}\right),$$

where J_n is the n-th order Bessel function of the first kind and T_n is the n-th degree Chebyshev polynomial of the first kind. For fixed x, the magnitude of the Bessel functions decay super-exponentially as $n \to \infty$. In fact, using [AS70, (9.1.62)], we have the bound

$$|J_n(-i\alpha_\ell a x)| \leq \frac{|\alpha_\ell a x|^n\, e^{|\alpha_\ell a x|}}{2^n n!}.$$

Now observe that $|\alpha_\ell a x| \leq D\frac{aw}{(a+w)^2}$, where we used (A.4) and the fact that $r_{\max} = (a+w)/\sqrt{2}$. For some desired accuracy ϵ_R, let P be the smallest integer such that $\left|J_n\left(-iD\frac{aw}{(a+w)^2}\right)\right|^2 \leq \epsilon_R$ for all $n \geq P$. Then we may estimate the number of terms $R^{(\ell)}$ in (3.28) as

$$R^{(\ell)} \leq (P+1)^2 = \mathcal{O}\left(\log \epsilon_R^{-1}\right).$$

If we assume that the output window is at least as large as the input aperture, i.e., $w \geq a$, then the argument of the Bessel function satisfies

$$\left|D\frac{aw}{(a+w)^2}\right| \leq \frac{D}{4},$$

and it is easy to verify that the numerical rank R of each matrix $\mathbf{S}^{(\ell)}$ satisfies $R \leq 19$ for $\epsilon_R = 10^{-3}$ and $R \leq 28$ for $\epsilon_R = 10^{-6}$.