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 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.

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 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 [21] (see also [9, 8]) described wave propagation via the integral

Date: September 25, 2013.

Key words and phrases. Rayleigh-Sommerfeld integral, Fresnel approximation, Fraunhofer approximation, approximation by Gaussians, unequally spaced fast Fourier transform, quadratures for band-limited functions.

The authors are with the Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO 80309, USA. E-mail: lewisrd@colorado.edu; beylkin@colorado.edu (corresponding author); lucas.monzon@colorado.edu. This research was partially supported by NSF grants DMS-1009951, DGE-0801680, DMS-0602284, and DOE/ORNL grant 4000038129.
Proc R Soc A 469: 20130323, http://dx.doi.org/10.1098/rspa.2013.0323 .

$$u(\mathbf{x},z) = -\frac{1}{2\pi}\int_{\mathbb{R}^2} f(\mathbf{y})\,\frac{\partial}{\partial z}\left(\frac{e^{i2\pi R}}{R}\right)d\mathbf{y}, \quad z>0, \tag{1.1}$$

where $R = \sqrt{z^2 + \|\mathbf{x}-\mathbf{y}\|^2}$. Given the field $u(\mathbf{y},0) = f(\mathbf{y})$ in the plane $z=0$, (1.1) 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 in this intermediate region, the standard practice is to replace the kernel 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 (1.1) 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) [11, 3, 19], 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 significantly increases the size of the output region where the evaluation of (1.1) is accurate.

The Fresnel approximation is an important, often-used, and frequently-referenced, approximation to the propagated field. While our method can be viewed as a generalization of the Fresnel approximation, the two approximations are not directly related. The Fresnel approximation replaces the Rayleigh-Sommerfeld kernel by a single Gaussian with a purely imaginary exponent. In our method, for any user-specified target accuracy, we use a nonlinear algorithm to approximate the kernel as a short linear combination of Gaussians with complex-valued exponents. In both the Fresnel approximation and our method, the use of Gaussians leads to a fast algorithm to apply the approximate kernel. However, accuracy control in our approach relies on analysis-based approximations rather than directly on analytic formulae. Equally as important as the approximation of the kernel with respect to accuracy control is a careful discretization of the resulting integrals using quadratures for band-limited functions (see §2.2).

Given the known limitations of the Fresnel approximation (see, e.g., [22, 30, 12]), many researchers have sought methods to improve its accuracy, e.g., by constructing series expansions of the propagated field where the Fresnel approximation appears as the first term in the series [13, 1]. Unfortunately (to the best of our knowledge) such an approach does not lead to a fast algorithm with controlled accuracy. The expansions derived in these papers can be used in a limited number of cases if the boundary data are known analytically. However, if the boundary data are provided

numerically (e.g., measured, or produced by a computational procedure such as phase recovery), such analytic expansions can yield only a limited accuracy. We further comment on this topic in §C.4 of the online supplement.

The need for an accurate propagation algorithm arises in areas such as computational holography [14], optical component design [20], and antenna design [2]. A particularly interesting application area is X-ray diffraction microscopy [23], 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 (see §§A and B of the online supplement).

The paper is organized as follows. The necessary mathematical preliminaries are reviewed in §2. We describe our new algorithm in §3, then discuss its region of validity in §4. In §5 we provide several numerical examples, then summarize our results in §6. By introducing this new algorithm, we hope to stimulate accuracy improvements in computational optical systems by essentially eliminating numerical errors.

2. Preliminaries

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,

$$\left(\Delta + k^2\right) u = 0, \tag{2.1}$$

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 (1.1) yields the solution $u(\mathbf{x}, z)$ of the Dirichlet problem for (2.1) in the half-space $z > 0$ that satisfies the Sommerfeld radiation condition [28, 29],

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

Given the boundary data $u(\mathbf{x},0)=f(\mathbf{x})$, we rewrite (1.1) as

$$u(\mathbf{x},z)=\int_{\mathbb{R}^2} f(\mathbf{y})\,K_z(\|\mathbf{x}-\mathbf{y}\|)\,d\mathbf{y}, \tag{2.2}$$

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

$$K_z(r)=\frac{e^{i2\pi z\sqrt{1+(r/z)^2}}}{iz}\left(\frac{1}{1+(r/z)^2}+\frac{i}{2\pi z\left(1+(r/z)^2\right)^{\frac{3}{2}}}\right),\quad r\geq 0. \tag{2.3}$$

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 (2.2) in the Fourier domain as

$$u(\mathbf{x},z)=\int_{\mathbb{R}^2}\widehat{f}(\mathbf{p})\,\widehat{K}_z(\|\mathbf{p}\|)\,e^{i2\pi\mathbf{x}\cdot\mathbf{p}}\,d\mathbf{p}, \tag{2.4}$$

where the Fourier transform of the Rayleigh-Sommerfeld kernel (cf. [24] and references therein) is given by

$$\widehat{K}_z(\rho)=e^{i2\pi z\sqrt{1-\rho^2}},\quad \rho\geq 0. \tag{2.5}$$

Our goal is to evaluate (2.2) accurately 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 (2.2) or (2.4) impractical. In §3 we will show how to approximate (2.3) 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 and for very large values of z, the problem may be solved using asymptotic methods (see §§A and B of the online supplement).

Remark 1. Given the normal derivative of the boundary data

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

Lord Rayleigh's formula for the Neumann problem reads

$$u(\mathbf{x},z)=-\frac{1}{2\pi}\int_{\mathbb{R}^2} g(\mathbf{y})\,\frac{e^{i2\pi R}}{R}\,d\mathbf{y},\quad R=\left(z^2+\|\mathbf{x}-\mathbf{y}\|^2\right)^{\frac{1}{2}},\ z>0. \tag{2.6}$$

With minor modifications, our approach is also applicable to evaluating (2.6).

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 and his collaborators in [27, 17, 18, 25, 26] 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 [4]; 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 [4] 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 [4] that, similar to the classical case, the eigenvalues of this operator,

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

quantify the fraction of energy (L^2-norm) of ψ_j 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 [4]), for a user-specified accuracy ϵ, we select a linear subspace of the eigenfunctions, $\operatorname{span}\{\psi_j\}_{j=0}^J$, 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 §2.3 below.

Identifying this subspace allows us to accurately evaluate integrals involving the boundary data. Following [4, 5] (see also [33]), we have

Theorem 2. *Let $W = \left[-\frac{w}{2}, \frac{w}{2}\right]^2$ be a square output window and $A = \left[-\frac{a}{2}, \frac{a}{2}\right]^2$ be the spatial domain. Then for all functions $f \in \operatorname{span}\{\psi_j\}_{j=0}^J$ and for any target accuracy ϵ, we can use the algorithms in [4, 5] to obtain 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*

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

The number of quadrature nodes required to obtain accuracy ϵ depends on the space-bandwidth product $(a+w)c$ as $M = \mathcal{O}\left((a+w)c \log \epsilon^{-1}\right)$. These quadratures are known as generalized Gaussian quadratures for band-limited functions.

2.2.1. *The Unequally Spaced Fast Fourier Transform.* We need to evaluate trigonometric sums of the form

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

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 [11, 3, 19]) with computational complexity $\mathcal{O}\left(N^2 \log N + M^2 \log^2 \epsilon^{-1}\right)$.

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 (2.2) outside of A may be neglected,

$$\int\limits_{\mathbf{x}\notin A} |f(\mathbf{x})|^2 \, d\mathbf{x} \leq \epsilon^2 \|f\|_2^2 . \tag{2.7}$$

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 evaluate (2.2) accurately. It follows from (2.5) 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,

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

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

$$c_e = \sqrt{1 + \left(\frac{\log \epsilon^{-1}}{2\pi z}\right)^2}. \tag{2.8}$$

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,

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

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 the function 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.

2.4. **Approximation of Functions by Linear Combinations of Exponentials and Gaussians.** We use an algorithm in [6] (see also [7]) to approximate, for a target accuracy ϵ, a smooth function $f(x)$ by a nearly optimal linear combination of Gaussians. Since the algorithm in [6] 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 a (nearly) minimal number of complex-valued weights w_ℓ and exponents η_ℓ, such that

$$\left| f(x) - \sum_{\ell=1}^{L} w_\ell e^{-\eta_\ell x^2} \right| \leq \epsilon, \quad x \in [0,1]. \tag{2.9}$$

For completeness, we recall this algorithm for approximation by exponentials in §C.1 of the online supplement.

For the functions $f(x)$ considered in this paper, the number of terms L in approximation (2.9) satisfies $L = \mathcal{O}\left(\log \epsilon^{-1}\right)$. This behavior is typical and occurs for a wide variety of functions encountered in applications.

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 [10, 16]. The computational complexity of these algorithms is $\mathcal{O}\left(MN \log k + (M+N)\, k^2\right)$ (cf. $\mathcal{O}(MNk)$ for the direct approach utilizing a rank-revealing QR factorization).

2.6. **The Approximations of Fresnel and Fraunhofer.** Our method of approximating the kernel (2.3) 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 (2.3)

$$\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z\left(1+(r/z)^2\right)^{\frac{3}{2}}} \approx 1 \tag{2.10}$$

and

$$e^{i2\pi z\sqrt{1+(r/z)^2}} \approx e^{i2\pi z} e^{i\frac{\pi}{z}r^2}. \tag{2.11}$$

The Fresnel approximation uses this approximate kernel in place of the Rayleigh-Sommerfeld kernel in (2.2), yielding

$$\begin{aligned} u(\mathbf{x},z) &\approx \frac{e^{i2\pi z}}{iz} \int_{\mathbb{R}^2} f(\mathbf{y})\, e^{i\frac{\pi}{z}\|\mathbf{x}-\mathbf{y}\|^2}\, d\mathbf{y} \\ &= \frac{e^{i2\pi z} e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz} \int_{\mathbb{R}^2} f(\mathbf{y})\, e^{i\frac{\pi}{z}\|\mathbf{y}\|^2} e^{-i\frac{2\pi}{z}\mathbf{x}\cdot\mathbf{y}}\, d\mathbf{y} \end{aligned} \tag{2.12}$$

(see, e.g., [15, §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 (2.11) is especially deleterious—see §5.3). In §§4 and 5 we demonstrate that the accuracy of the Fresnel approximation rapidly deteriorates away from the optical axis.

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 (2.12), leads to the Fraunhofer (sometimes called far-field) approximation

$$u(\mathbf{x}, z) \approx \frac{e^{i2\pi z} e^{i\frac{\pi}{z}\|\mathbf{x}\|^2}}{iz} \widehat{f}\left(\frac{\mathbf{x}}{z}\right) \tag{2.13}$$

(see, e.g., [15, §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. A more accurate approximation in the far field is given by

$$u(\mathbf{x}, z) \sim -\frac{ize^{i2\pi\sqrt{z^2+\|\mathbf{x}\|^2}}}{z^2+\|\mathbf{x}\|^2} \widehat{f}\left(\frac{\mathbf{x}}{\sqrt{z^2+\|\mathbf{x}\|^2}}\right), \quad z \to \infty, \tag{2.14}$$

which may be evaluated via the USFFT. We further discuss the Fraunhofer approximation and derive (2.14) in §B of the online supplement.

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 §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 (2.2), we need to compute

$$u(\mathbf{x}, z) = \int_A f(\mathbf{y}) K_z(\|\mathbf{x}-\mathbf{y}\|)\, d\mathbf{y}, \quad \mathbf{x} \in W, \tag{3.1}$$

where K_z is the Rayleigh-Sommerfeld kernel (2.3). Our algorithm comprises three steps. First, in §3.1, we accurately approximate the Rayleigh-Sommerfeld kernel by a linear combination of Gaussians using the algorithm briefly described in §2.4. Second, in §3.2, we use the resulting approximation in (3.1) and accurately discretize the ensuing integrals via the generalized Gaussian quadratures for band-limited functions from Theorem 2. Finally, in §3.3, we use the algorithms referred to in §2.5 for computing the SVDs of low-rank matrices to rearrange the resulting sums for rapid and accurate evaluation via the USFFT (see §2.2.1).

3.1. Approximation of the Kernel with Controlled Error.

The key observation behind the Fresnel approximation is that the phase of the kernel (2.3) is approximately quadratic, cf. (2.11), 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 $0 \le r \le (a+w)/\sqrt{2}$. We demonstrate how to obtain, for any user-specified accuracy $\epsilon_K > 0$, an approximation $\widetilde{K}_z(r)$ such that

$$\left|K_z(r) - \widetilde{K}_z(r)\right| \le \frac{\epsilon_K}{z}, \quad r \in \left[0, \frac{a+w}{\sqrt{2}}\right]. \tag{3.2}$$

We emphasize that in (3.2) 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

$$A_z(r) = \left(\frac{1}{1+(r/z)^2} + \frac{i}{2\pi z \left(1+(r/z)^2\right)^{\frac{3}{2}}} \right) e^{i2\pi z\left(\sqrt{1+(r/z)^2}-1-\frac{1}{2}(r/z)^2\right)}. \tag{3.3}$$

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 §2.4 to compute, for a desired accuracy $\epsilon_K > 0$, complex-valued weights w_ℓ and exponents η_ℓ such that

$$\left| A_z(r) - \sum_{\ell=1}^{L} w_\ell e^{-\eta_\ell r^2} \right| \le \epsilon_K, \quad r \in \left[0, \frac{a+w}{\sqrt{2}}\right], \tag{3.4}$$

leading to the approximation

$$\widetilde{K}_z(r) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell e^{-\left(\eta_\ell - i\frac{\pi}{z}\right)r^2}, \tag{3.5}$$

satisfying (3.2). We define $\widetilde{u}(\mathbf{x}, z)$ to be the result of using the approximate kernel $\widetilde{K}_z(r)$ in (3.1),

$$\widetilde{u}(\mathbf{x}, z) = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \int_A f(\mathbf{y})\, e^{-\left(\eta_\ell - i\frac{\pi}{z}\right)\|\mathbf{x}-\mathbf{y}\|^2}\, d\mathbf{y}. \tag{3.6}$$

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

Proposition 3. *Let $\widetilde{u}$ be the function defined in* (3.6)*, with weights w_ℓ and exponents η_ℓ, $\ell = 1, \ldots, L$, as in* (3.4)*. Then*

$$|u(\mathbf{x}, z) - \widetilde{u}(\mathbf{x}, z)| \le \frac{\epsilon_K \|f\|_1}{z}, \quad \mathbf{x} \in \left[-\frac{w}{2}, \frac{w}{2}\right]^2, \tag{3.7}$$

where the field $u(\mathbf{x}, z)$ is given by (3.1)*.*

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 (3.4), we rearrange (3.6) as

$$\begin{aligned} \widetilde{u}(\mathbf{x}, z) = {} & \frac{e^{i2\pi z}}{iz} \\ & \times \sum_{\ell=1}^{L} w_\ell e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{x}\|^2} \int_A f(\mathbf{y})\, e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{y}\|^2} e^{2\alpha_\ell \mathbf{x}\cdot\mathbf{y}} e^{i2\beta_\ell \mathbf{x}\cdot\mathbf{y}}\, d\mathbf{y}. \end{aligned} \tag{3.8}$$

A straightforward estimate of the bandlimit of the integrands (see §C.3.2 of the online supplement) 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. Using the bandlimit c', we discretize the integrals in (3.8), for a desired accuracy ϵ_Q, using the quadratures from Theorem 2.

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 $\tau_m \tau_{m'}$. We select an $N \times N$ grid of output locations $\mathbf{x}_{nn'} = (x_n, x_{n'}) \in W$, $n, n' = 1, \ldots, N$. We then apply the quadrature from Theorem 2 to the integrals in (3.8) and obtain an approximation to the output field at the desired locations as

$$u_{nn'} = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \sum_{m,m'=1}^{M} \tau_m \tau_{m'} \mathbf{T}^{(\ell)}_{nn'mm'} f\left(\mathbf{y}_{mm'}\right) e^{i2\beta_\ell \mathbf{x}_{nn'} \cdot \mathbf{y}_{mm'}}. \tag{3.9}$$

In (3.9) the $N \times N \times M \times M$ fourth-order tensors $\mathbf{T}^{(\ell)}$, $\ell = 1, \ldots, L$, are given by

$$\mathbf{T}^{(\ell)}_{nn'mm'} = e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{x}_{nn'}\|^2} e^{-(\alpha_\ell + i\beta_\ell)\|\mathbf{y}_{mm'}\|^2} \mathbf{S}^{(\ell)}_{nm} \mathbf{S}^{(\ell)}_{n'm'}, \tag{3.10}$$

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

$$\mathbf{S}^{(\ell)}_{nm} = e^{2\alpha_\ell x_n y_m}, \tag{3.11}$$

where $n = 1, \ldots, N$ and $m = 1, \ldots, M$. From Theorem 2 we obtain the bound

$$|\widetilde{u}\left(\mathbf{x}_{nn'}, z\right) - u_{nn'}| \leq \frac{\epsilon_Q \left\|f\right\|_1}{z}, \tag{3.12}$$

where $\widetilde{u}\left(\mathbf{x}_{nn'}, z\right)$ is given by (3.6) and $u_{nn'}$ by (3.9).

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 (2.12) via either the FFT or the USFFT. In our approach, the exponents in approximation (3.5) 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 (3.9).

In order to evaluate the inner summations in (3.9) rapidly, 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 (3.11) as a sum of outer products,

$$\mathbf{S}^{(\ell)}_{nm} = \sum_{q=1}^{\min(M,N)} \sigma^{(\ell)}_q \mathbf{U}^{(\ell)}_{nq} \mathbf{V}^{(\ell)}_{mq}, \tag{3.13}$$

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 §3.4, 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 §2.5 to compute these SVDs rapidly 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 §C.2 of the online supplement for a proof)

Lemma 4. *Let $\sigma_q^{(\ell)}$, $\mathbf{U}_{nq}^{(\ell)}$, and $\mathbf{V}_{mq}^{(\ell)}$, where $\ell = 1, \ldots, L$, $q = 1, \ldots, \min(M, N)$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$, be as in (3.13). For a desired accuracy $\epsilon_R > 0$, let $I^{(\ell)}$, $\ell = 1, \ldots, L$, be the smallest integer such that*

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

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

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

Using Lemma 4, we approximate $\mathbf{T}^{(\ell)}$ in (3.10) as $\sum_{r=1}^{R^{(\ell)}} \mathbf{P}_{nn'r}^{(\ell)} \mathbf{Q}_{mm'r}^{(\ell)}$, 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 4 implies that

$$\left| \mathbf{T}_{nn'mm'}^{(\ell)} - \sum_{r=1}^{R^{(\ell)}} \mathbf{P}_{nn'r}^{(\ell)} \mathbf{Q}_{mm'r}^{(\ell)} \right| \leq \epsilon_R 2 e^{\frac{|\alpha_\ell|}{2}(a^2+w^2+aw)}. \tag{3.14}$$

We define $\widetilde{u}_{nn'}$ to be the result of using approximation (3.14) in (3.9),

$$\widetilde{u}_{nn'} = \frac{e^{i2\pi z}}{iz} \sum_{\ell=1}^{L} w_\ell \sum_{r=1}^{R^{(\ell)}} \mathbf{P}_{nn'r}^{(\ell)} \sum_{m,m'=1}^{M} \tau_m \tau_{m'} \mathbf{Q}_{mm'r}^{(\ell)} f\left(\mathbf{y}_{mm'}\right) e^{i2\beta_\ell \mathbf{x}_{nn'} \cdot \mathbf{y}_{mm'}}. \tag{3.15}$$

It follows from (3.14) and the estimate

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

that

$$|u_{nn'} - \widetilde{u}_{nn'}| \leq \frac{\epsilon_R \|f\|_1}{z} 2 \sum_{\ell=1}^{L} |w_\ell| \, e^{\frac{|\alpha_\ell|}{2}(a^2+w^2+aw)}, \tag{3.16}$$

where $u_{nn'}$ is given by (3.9). We also have

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

where b is a small constant (this can be shown using the techniques from §C.3 of the online supplement). Incorporating b into ϵ_R, the bound (3.16) becomes

$$|u_{nn'} - \widetilde{u}_{nn'}| \leq \frac{\epsilon_R \|f\|_1}{z}. \tag{3.17}$$

Combining the error bounds (3.7), (3.12), and (3.17), we obtain

Theorem 5. *The error of computing the field u from* (3.1) *using* (3.15) *is bounded by*

$$|u(\mathbf{x}_{nn'}, z) - \widetilde{u}_{nn'}| \leq \frac{(\epsilon_K + \epsilon_Q + \epsilon_R) \|f\|_1}{z}. \tag{3.18}$$

The expression for $\widetilde{u}_{nn'}$ in (3.15) allows us to evaluate the field 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 the three steps of deriving the final approximation of the field (3.15), we used three different accuracies, ϵ_K, ϵ_Q, and ϵ_R, 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 ϵ.

Remark 6. 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. 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 (3.8). We note that near optimal quadratures for circular apertures are described in [4].

Remark 7. *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.,

$$f(\mathbf{x}) = f(x_1, x_2) = \sum_{s=1}^{S} f_1^{(s)}(x_1) f_2^{(s)}(x_2) \tag{3.19}$$

for some functions $f_1^{(s)}$ and $f_2^{(s)}$, $s = 1, \ldots, S$. In such cases the application of the approximate kernel (3.5) simplifies to the evaluation of several one-dimensional USFFTs. Substitute (3.19) into (3.8) 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 $u_2^{(\ell,s)}$, $\ell = 1, \ldots, L$, $s = 1, \ldots, S$, can be evaluated by one-dimensional integrals. We obtain similar formulae if the boundary data are concentrated in a disk and separable in polar coordinates.

3.4. **Computational Cost.** The number of terms in (3.4) may be estimated as $L = \mathcal{O}\left(\gamma^4 \log \epsilon^{-1}\right)$, where

$$\gamma = \frac{a + w}{\sqrt{2} z^{\frac{3}{4}}} \tag{3.20}$$

(see §C.3.1 of the online supplement). In order to control the number of terms L, we restrict the parameter γ by the empirically-determined constant

$$\gamma \leq 2.62. \tag{3.21}$$

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 §4. This bound also implies that the ratio $|\alpha_\ell| / (aw)$ is small, leading to a low numerical rank of the matrices $\mathbf{S}^{(\ell)}$ in (3.11).

The cost of evaluating (3.15) depends on the number of USFFTs, $R = R^{(1)} + \cdots + R^{(L)}$, which is estimated as $R = \mathcal{O}\left(\log^2 \epsilon^{-1}\right)$. 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 §5.4.

4. Size of the Output Region

In §3.4 we ensured that our algorithm is efficient by requiring γ from (3.20) to satisfy (3.21). 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

$$w_{\max} = 3.71 \times z^{\frac{3}{4}} - a, \tag{4.1}$$

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 (see §A of the online supplement).

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

$$z_{\min} = 0.174 \times a^{\frac{4}{3}}, \tag{4.2}$$

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

For comparison, let us find analogues of (4.1) and (4.2) for the Fresnel approximation (2.12). 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 (4.1), 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 (2.3) with $e^{i2\pi z} e^{i\frac{\pi}{z} r^2}$, we find the maximum value of $r'_{\max}$ such that

$$\left| e^{i2\pi z\sqrt{1+(r/z)^2}} - e^{i\left(2\pi z + \frac{\pi}{z} r^2\right)} \right| \leq \epsilon, \quad r \in [0, r'_{\max}]\,.$$

Using $r'_{\max} \approx \sqrt{2}\left(\frac{\epsilon}{\pi}\right)^{\frac{1}{4}} z^{\frac{3}{4}}$, we obtain an analogue of (4.1) for the Fresnel approximation,

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

giving the largest possible square output window for a square aperture with side-length a. The analogue of (4.2) for the Fresnel approximation is

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

which gives the minimum propagation distance.

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 $\epsilon = 10^{-3}$. 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 $w'_{\max}$ 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 $z_{\min} \approx 14,880$.

5. Numerical Examples

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. We select the boundary data with a Gaussian profile given by

$$f(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|^2}{\sigma^2}}, \tag{5.1}$$

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. (2.4))

$$\begin{aligned} u_p(\mathbf{x}, z) &= \int\limits_{\|\mathbf{p}\| \le 1} \widehat{f}(\mathbf{p})\, \widehat{K}_z(\|\mathbf{p}\|)\, e^{i2\pi \mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p} \\ &= 4\sqrt{\frac{\pi}{2}} \sum_{k=0}^{\infty} i^k f_k\, j_k\left(2\pi z \sqrt{1 + (\|\mathbf{x}\|/z)^2}\right) \overline{P}_k\left(\left(1 + (\|\mathbf{x}\|/z)^2\right)^{-\frac{1}{2}}\right), \end{aligned} \tag{5.2}$$

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

$$f_k = \pi \sqrt{2\pi} \sigma^2 \int_0^1 s e^{-\pi^2 \sigma^2 (1 - s^2)} \overline{P}_k(s)\, ds. \tag{5.3}$$

These coefficients decay rapidly once k is sufficiently large, so that we may truncate the sum in (5.2) to obtain a simple formula to compute the non-evanescent portion of the field to any desired accuracy. The error committed by neglecting the

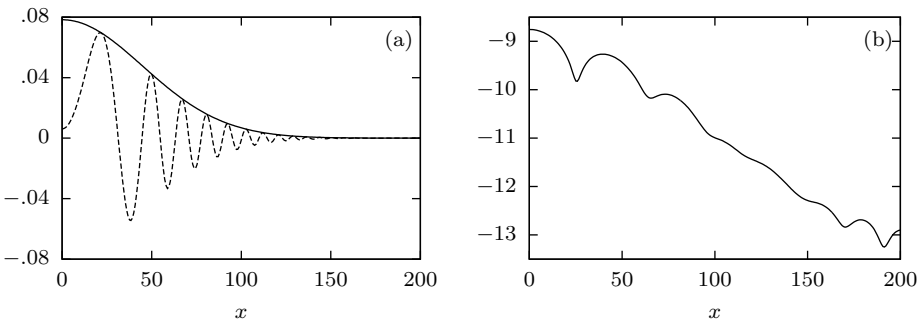

FIGURE 5.1. Propagation of boundary data with a Gaussian profile. The field magnitude, $|u(x,0)|$, (solid line) and its real part, $\mathcal{R}\mathrm{e}\, u(x,0)$ (dashed line) evaluated along the positive x-axis (a) and the attained accuracy $\log_{10}|u(x,0)-\widetilde{u}(x,0)|$ (b).

evanescent waves may be bounded by

$$|u_e(\mathbf{x},z)| = \left| \int\limits_{\|\mathbf{p}\|>1} \widehat{f}(\mathbf{p})\, e^{-2\pi z\sqrt{\|\mathbf{p}\|^2-1}} e^{i2\pi\mathbf{x}\cdot\mathbf{p}}\, d\mathbf{p} \right|$$

$$\leq 2(\pi\sigma)^2 e^{-(\pi\sigma)^2} \int_1^\infty \rho e^{-2\pi z\sqrt{\rho^2-1}}\, d\rho = \frac{\sigma^2}{2z^2} e^{-(\pi\sigma)^2}. \tag{5.4}$$

Provided that (5.4) is less than the accuracy sought, we may disregard the evanescent portion of the field entirely and regard (5.2) as a formula to compute the field generated by the boundary data (5.1) 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 $M \times M = 512 \times 512$ 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 (5.4) 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 (5.3) decay to $|f_k| \leq 10^{-15}$ once $k \geq 200$, so we truncate the sum (5.2) after 200 terms and use it to determine the accuracy of our algorithm. We display the results in Figure 5.1 and note that the obtained accuracy is better than the accuracy goal 10^{-6} (the bound in Lemma 4 is not tight).

5.2. **An Aperture in the Near Field of a Source.** We consider an aperture in the near field of a source and thereby demonstrate that our method maintains accuracy even when evanescent waves are present in the aperture field. We arrange Helmholtz point sources in the plane $z = -1$ so that the resulting field is given by (cf. (2.6))

$$u(\mathbf{x},z) = -\sum_{j=1}^{J} \varrho_j \frac{e^{i2\pi\sqrt{(z+1)^2+\|\mathbf{x}-\mathbf{r}_j\|^2}}}{2\pi\sqrt{(z+1)^2+\|\mathbf{x}-\mathbf{r}_j\|^2}},$$

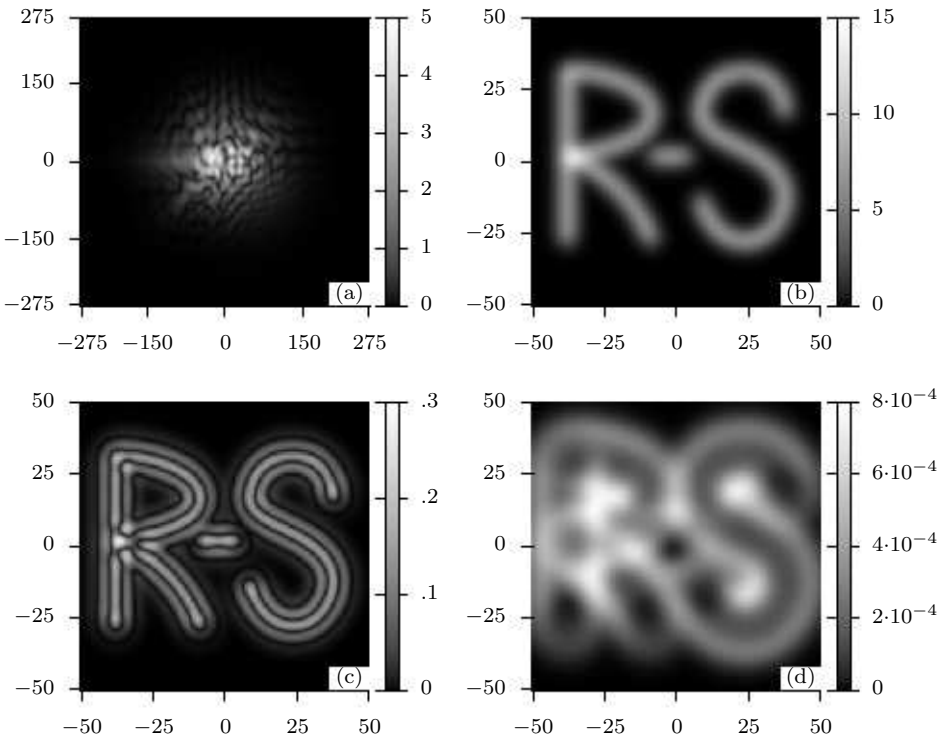

FIGURE 5.2. Comparison of the error of our method and that of the Fresnel approximation for an aperture in the near field of a source after propagating $z = 1000$ wavelengths. We display the magnitude of the input data $|f|$ (a), the magnitude of the true solution $|u|$ (b), the error of the Fresnel approximation $|u - \widetilde{u}_f|$ (c), and the error of our method $|u - \widetilde{u}|$ (d).

where $\mathbf{r}_j$ and ϱ_j, $j = 1, \ldots, J$, give the position and intensity, respectively, of each point source. The function specifying the boundary data is given by $f(\mathbf{x}) = u(\mathbf{x}, 0)$; as shown in Figure 5.2(a), to high accuracy, it is confined to a square aperture of side-length $a = 550$ wavelengths. We select propagation distance $z = 1000$ wavelengths, desired accuracy $\epsilon = 10^{-3}$, and apply our algorithm to compute the field $\widetilde{u}(\mathbf{x}, 1000)$ in a square output window of side-length $w = 100$ wavelengths. Because the aperture plane is only one wavelength from a collection of point sources, evanescent waves are present in the boundary data. In this case, we sample the boundary data on a grid of size $M \times M = 1486 \times 1486$. In Figure 5.2 we compare our approximate solution $\widetilde{u}(\mathbf{x}, 1000)$ to the true solution $u(\mathbf{x}, 1000)$ and, as requested, it is correct to slightly over 3 digits. For comparison, we also show the error of the Fresnel approximation $\widetilde{u}_f(\mathbf{x}, 1000)$, which has only about 1.5 digits of accuracy. Observe that the conditions of this numerical experiment are favorable for the Fresnel approximation, since the maximum angle that any point in the output window makes with the optical axis is only approximately 4 degrees.

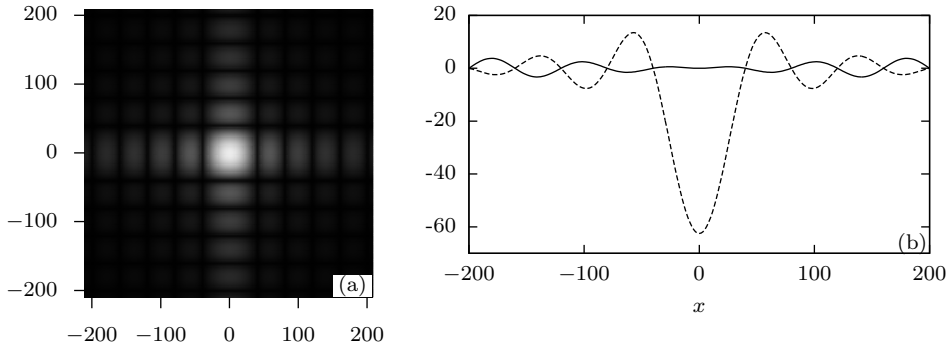

FIGURE 5.3. Propagation of a spherical wave restricted to a square aperture and converging to a point on the optical axis. As expected, the diffraction pattern is approximately a scaled version of the function $|\mathrm{sinc}\,(x_1)\,\mathrm{sinc}\,(x_2)|$. We display the magnitude of the field, $|u\,(x_1,x_2)|$, (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_1-axis (b).

5.3. **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 5.3 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 $|\mathrm{sinc}\,(x_1)\,\mathrm{sinc}\,(x_2)|$.

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, $\widetilde{u}_f\,(\mathbf{x}, z_0)$, near the focal point $\mathbf{x} = \mathbf{r}_0$. Results displayed in Figure 5.4 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 5.5. The diffraction pattern obtained by the Fresnel approximation is both shifted and blurred when compared to the correct one.

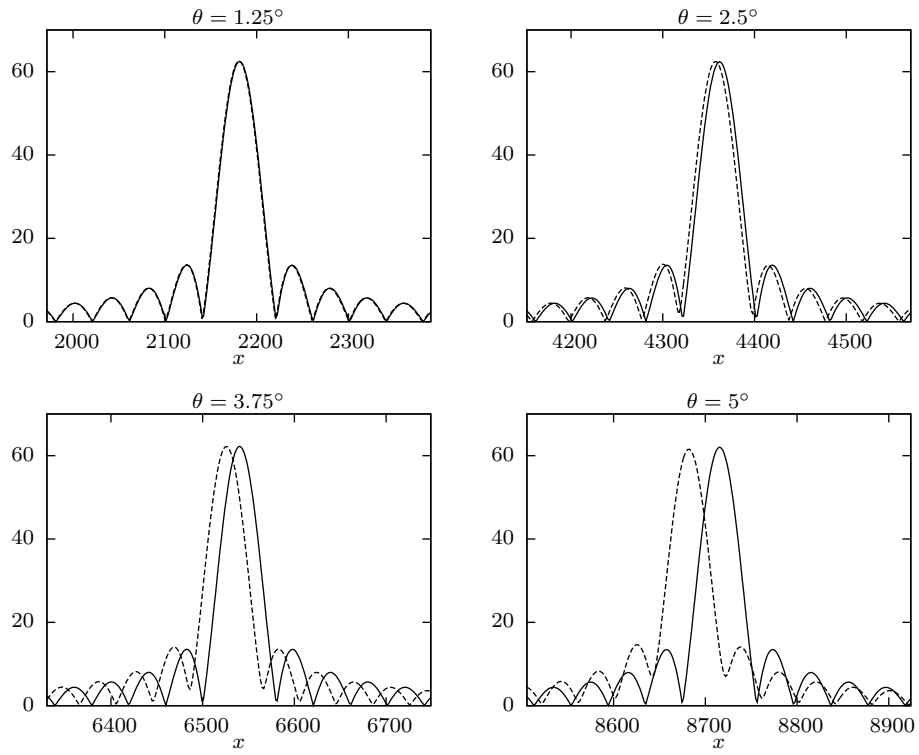

FIGURE 5.4. Comparison of the magnitude of the field evaluated near the focal point on the x_1-axis. We display the magnitude $|u(x, 0, z_0)|$, computed by our algorithm (solid line, correct to 3 digits) and by the Fresnel approximation (dashed line), as the focal point of a converging spherical wave 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.

In Figure 5.6 we compare 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 $|u(\mathbf{r}_0, z_0) - \widetilde{u}_f(\mathbf{r}_0, z_0)|$, 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 $\theta \in [0°, 5°]$, while the Fresnel approximation is accurate to approximately 3 digits for $\theta = 0°$ but has essentially no accurate digits for $\theta > 4°$. This example demonstrates that a belief that the Fresnel approximation produces accurate results at angles up to 18 degrees off the optical axis (see, e.g., [31, 32]) is not true in general.

Remark 8. From Figure 5.4, 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 $g : \mathbb{R}^2 \to \mathbb{R}^2$ would be selected with the goal of rescaling the field computed by the Fresnel approximation, $\widetilde{u}_f(g(\mathbf{x}), z)$, to more closely match the true field, $u(\mathbf{x}, z)$. In effect, the strategy would be to rescale the x-axis for the dashed lines

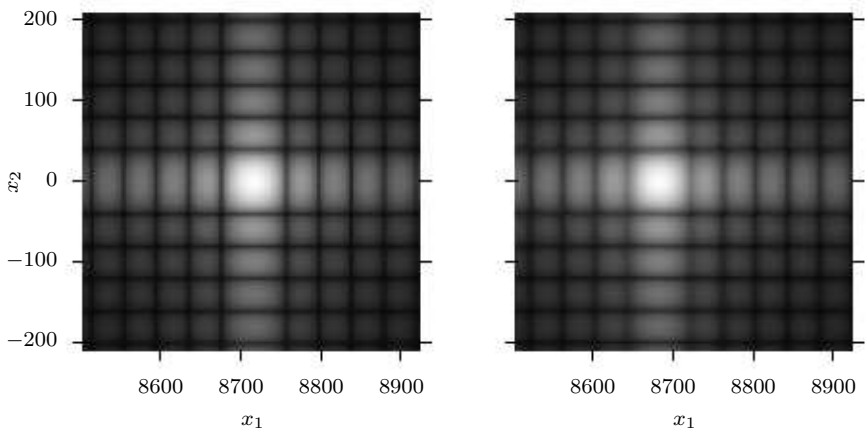

FIGURE 5.5. Comparison of the magnitude of the field for a focal point 5° off the optical axis computed by our algorithm correct to 3 digits (left), and by the Fresnel approximation (right). To enhance contrast, we plot the square root of the magnitude, $|u(x_1, x_2)|^{1/2}$. The Fresnel approximation shifts the location of the focal spot, and blurs the boundaries between the mainlobe and sidelobes. See also the bottom-right plot in Figure 5.4.

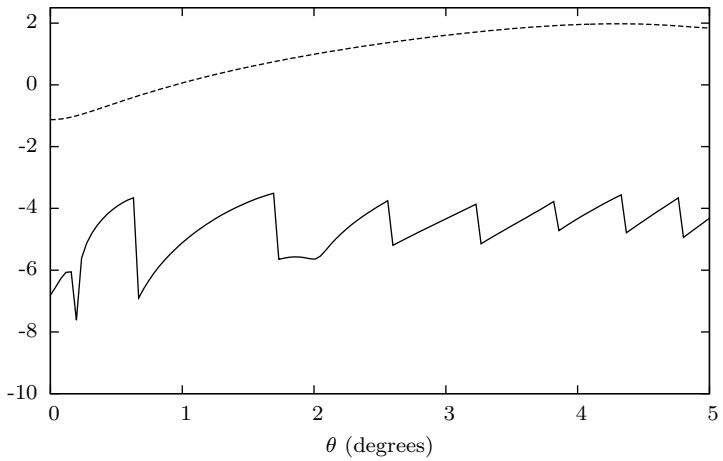

FIGURE 5.6. Comparison of the error of our method and that of the Fresnel approximation. We display the error of our method, $\log_{10}|u(\mathbf{r}_0, z_0) - \widetilde{u}(\mathbf{r}_0, z_0)|$ (solid line), and the error of the Fresnel approximation, $\log_{10}|u(\mathbf{r}_0, z_0) - \widetilde{u}_f(\mathbf{r}_0, z_0)|$ (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 (3.5), improving accuracy by about 1.5 digits each time and giving the solid line a "sawtooth" shape.

in Figure 5.4 to better align the peaks of the solid and dashed lines. Unfortunately, our example shows that 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 5.4).

5.4. **Representative Examples of Computational Cost.** The computational cost of our algorithm depends on the number of USFFTs required in (3.15), i.e., $R = R^{(1)} + \cdots + R^{(L)}$, where L is the number of terms needed to approximate the kernel in (3.5). 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 $z \to \infty$, cf. (2.14). On the other hand, for smaller values of z the field changes rapidly, and many USFFTs are required to compute the field accurately 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 (4.1) and is the largest output window that our method can accommodate for a given propagation distance z.

In Table 1, we provide the number of terms, L, needed to approximate the kernel for several propagation distances, as well as the number of USFFTs, R, required to compute the field for these two choices of output window size.

	$w = 10,000$			$w = w_{\max}$			
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 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 in seconds. 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 (4.1).

In Table 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 (3.15) may be computed in parallel, so that the total computational time can be reduced substantially on a multiprocessor computer system.

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. In contrast to current algorithms, our algorithm achieves any user-specified accuracy. As a consequence, we can rapidly and accurately compute the field in non-paraxial regions, i.e., regions far from the optical axis, with computational complexity proportional to that of the FFT. The overall result is a fast algorithm that can achieve any user-specified accuracy over a large computational domain.

Acknowledgments

We thank Dr. Bradley Alpert from NIST for providing many useful comments and suggestions.

References

[1] M. A. Alonso, A. A. Asatryan, and G. W. Forbes, *Beyond the Fresnel approximation for focused waves*, J. Opt. Soc. Am. A **16** (1999), no. 8, 1958–1969.
[2] C. A. Balanis (ed.), *Modern antenna handbook*, Wiley, 2008.
[3] G. Beylkin, *On the fast Fourier transform of functions with singularities*, Appl. Comput. Harmon. Anal. **2** (1995), no. 4, 363–381. MR 96i:65122
[4] 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.
[5] 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
[6] ———, *On approximation of functions by exponential sums*, Appl. Comput. Harmon. Anal. **19** (2005), no. 1, 17–48.
[7] ———, *Approximation of functions by exponential sums revisited*, Appl. Comput. Harmon. Anal. **28** (2010), no. 2, 131–149.
[8] 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.
[9] C. J. Bouwkamp, *Diffraction theory*, Reports on Progress in Physics **17** (1954), no. 1, 35–100.
[10] 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.
[11] A. Dutt and V. Rokhlin, *Fast Fourier transforms for nonequispaced data*, SIAM J. Sci. Comput. **14** (1993), no. 6, 1368–1393. MR 95d:65114
[12] G. W. Forbes, *Validity of the Fresnel approximation in the diffraction of collimated beams*, J. Opt. Soc. Am. A **13** (1996), no. 9, 1816–1826.
[13] G. W. Forbes, D. J. Butler, R. L. Gordon, and A. A. Asatryan, *Algebraic corrections for paraxial wave fields*, J. Opt. Soc. Am. A **14** (1997), no. 12, 3300–3315.
[14] J. W. Goodman, *Digital image formation from electronically detected holograms*, Proc. SPIE: Computerized Imaging Techniques, vol. 0010, SPIE, 1967, pp. 176–181.
[15] J. W. Goodman, *Introduction to Fourier optics*, 3 ed., McGraw-Hill physical and quantum electronics series, Roberts & Co., Englewood, Colorado, 2005.
[16] 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.
[17] 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
[18] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty III*, Bell System Tech. J. **41** (1962), 1295–1336.
[19] 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
[20] R. Piestun and J. Shamir, *Control of wave-front propagation with diffractive elements*, Optics Letters **19** (1994), no. 11, 771–773.

[21] Lord Rayleigh, *On the passage of waves through apertures in plane screens, and allied problems*, Philos. Mag. **43** (1897), 259–272.
[22] F. Schwering, *On the range of validity of Fresnel-Kirchhoff's approximation formula*, Antennas and Propagation, IRE Transactions on **10** (1962), no. 1, 99–100.
[23] D. Shapiro, P. Thibault, T. Beetz, V. Elser, M. Howells, C. Jacobsen, J. Kirz, E. Lima, H. Miao, A. M. Neiman, and D. Sayre, *Biological imaging by soft x-ray diffraction microscopy*, Proc. Natl. Acad. Sci. USA **102** (2005), no. 43, 15343–15346.
[24] G. C. Sherman, *Application of the convolution theorem to Rayleigh's integral formulas*, J. Opt. Soc. Am. **57** (1967), no. 4, 546–547.
[25] 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
[26] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty V. The discrete case*, Bell System Tech. J. **57** (1978), 1371–1430.
[27] 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
[28] A. Sommerfeld, *Die Greensche funktion der schwingungsgleichung*, Jahresbericht der Deutschen Mathematiker-Vereinigung **21** (1912), 309–353.
[29] ———, *Partial differential equations in physics*, Pure and Applied Mathematics, Academic Press, New York, 1949.
[30] W. H. Southwell, *Validity of the Fresnel approximation in the near field*, J. Opt. Soc. Am. **71** (1981), no. 1, 7–14.
[31] M. Sypek, *Light propagation in the Fresnel region. New numerical approach*, Optics Communications **116** (1995), no. 1–3, 43–48.
[32] 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.
[33] H. Xiao, V. Rokhlin, and N. Yarvin, *Prolate spheroidal wavefunctions, quadrature and interpolation*, Inverse Problems **17** (2001), no. 4, 805–838.