Scalar Scattering Theory and Physics-inspired Optimization for Computational Imaging

by

Subeen Pang
B.Sc., Seoul National University (2018)
M.S., Massachusetts Institute of Technology (2021)

Submitted to the Department of Mechanical Engineering
and Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mechanical Engineering and Computation
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2024

Authored by: Subeen Pang
Department of Mechanical Engineering
and Center for Computational Science and Engineering
May 16, 2024

Certified by: George Barbastathis
Professor of Mechanical Engineering
Thesis Supervisor

Accepted by: Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Chairman, Department Committee on Graduate Theses

Accepted by: Youssef M. Marzouk
Professor of Aeronautics and Astronautics
Co-Director, Center for Computational Science and Engineering

Scalar Scattering Theory and Physics-inspired Optimization for Computational Imaging

by

Subeen Pang

Submitted to the Department of Mechanical Engineering
and Center for Computational Science and Engineering
on May 16, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering and Computation

Abstract

This thesis explores the realm of computational imaging, focusing on the critical problems of phase retrieval and optical scattering—essential for accurately extracting physical information from photons. It aims to enhance the understanding and computational efficiency of existing models by addressing the fundamental challenges encountered due to diffraction effects, multiple scattering, and noise. Specifically, the thesis proposes improvements and comprehensive analyses of models related to phase retrieval, such as the Transport-of-Intensity Equation (TIE), and optical scattering approximations, including the Lippmann-Schwinger Equation (LSE).

For phase retrieval, this work introduces mathematical approaches to reduce the TIE's sensitivity to experimental conditions and provides a quantitative comparison with other methods to clarify its applicability. It also explores the adjoint method for solving the TIE, which significantly enhances numerical stability, and discusses the analytical relationship between non-paraxial formulations and conventional phase retrieval methods, deepening our understanding of the field.

In the domain of optical scattering, where information in photons is further encoded via complex light-matter interactions, this thesis examines several models derived from the scalar wave equation such as the LSE, the Born series, and the beam propagation method. It provides a direct and quantitative analysis of their relationships and numerical stability, highlighting the strengths and weaknesses of these models in various experimental contexts, which has not been discussed thoroughly in previous studies.

Additionally, the thesis tackles the computational challenges associated with the LSE by proposing numerical strategies and integrating neural networks as a learnable regularization. This approach aims to reduce computational demands while maintaining generalizability across different scattering objects.

Overall, this work contributes to the field of computational imaging by offering a deeper understanding of phase retrieval and optical scattering models, alongside presenting solutions to overcome their limitations. It sets the stage for further theoretical

analysis and practical applications in physics, where accurate information retrieval from photons is crucial.

Thesis Supervisor: George Barbastathis
Title: Professor of Mechanical Engineering

Acknowledgments

I would like to extend my heartfelt gratitude to everyone at MIT who supported me throughout my graduate studies, making this journey both enjoyable and intellectually stimulating.

First and foremost, I am deeply indebted to Professor George Barbastathis for his unwavering guidance, insightful advice, and invaluable discussions that significantly shaped my research. I am also profoundly grateful to all the members of the 3D Optical Systems Group for their encouragement and for continually challenging me to grow as a scientist and engineer.

To my family—my parents and my younger brother—I owe a special thanks for their constant support, understanding, and encouragement. Their unwavering belief in me has been a source of strength throughout this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In many branches of physics, photons have been considered as one of the most important carriers of physical information. In very small scales, we use so-called X-rays to obtain information on the atomic structure or composition of materials. On the other hand, even in larger scale, we use infrared light sources to retrieve hints on the universe far from our planet. This is because valuable information about physical characteristics of objects is encoded in photons via complex light-matter interactions, which we often refer to as the optical scattering.

The main objective of computational imaging is to facilitate the retrieval of the physical information from photons in quantitative ways. Fig. (1-1) is a simple description on computational imaging, which consists of four parts: light sources and corresponding illuminations, scattering with objects, propagation in space toward detectors, and lastly measurements at detectors. Contrary to the simple structure of the description, we face multiple difficult problems for the accurate retrieval of information in photons. One of the well-known problems is the phase retrieval problem, estimating the phase part ϕ of an electromagnetic wave ψ. To be more specific, note that what is usually collected at detectors is the intensity of such wave and thus, the phase data is lost. However, it carries indispensable information on characterizing a physical system, which leads to extensive studies in many branches of physics regard-

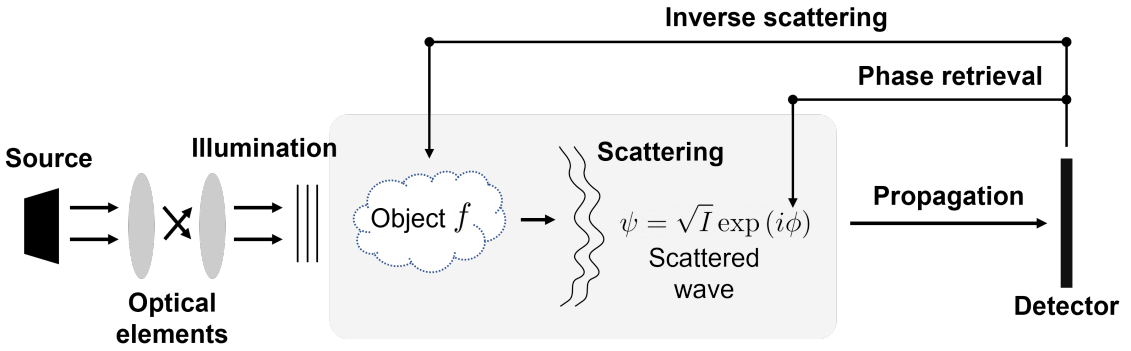

Figure 1-1: Simplified overview on the computational imaging. In computational imaging, the main objective is to retrieve physical information on objects of our interest from optical measurements using computational methods. Core components in computational imaging are: photon sources and corresponding illumination, optical wave ψ with an intensity I and a phase ϕ that originates from an interaction between an illumination and an object, and a detector for optical measurements.

ing the problem. For example, the phase part is required to estimate the scattering cross-section and corresponding atomic structure in X-ray crystallography, and without an appropriate phase retrieval process, the estimation is limited to objects under special assumptions [28].

Oftentimes, the connection between ψ and physical properties of objects is not straightforward, implying that the retrieval of the phase information is not sufficient. This is because of the complex optical scattering noted above and one should look for scattering models to have such connection. For instance, signals collected in space telescopes suffer from dusts and gravitational effects, which should be *decrypted* to unveil the physics of stars from photons. In computational imaging, such decryption is conducted by, first, analyzing the forward scattering representing the mapping from an object to ψ, and subsequently the inverse scattering that one tries to approximate an object from ψ.

These problems, the phase retrieval and the inverse scattering, would constitute the most fundamental and important parts in computational imaging. In particular, when the characteristic length scale of objects ranges from a few to thousands wavelengths and the scalar wave approximation applies, diffraction effects and multiple scattering inside objects become important, which may add another layer of diffi-

culty in computational imaging [79]. Accordingly, there have been numerous studies proposing computational models on how to take such difficulties into account. However, some of such models still require improvements, or even if they are shown to be applicable, there are cases for which quantitative analysis should be conducted further for better understanding on models. Based on phenomenal models in the phase retrieval and the optical scattering, this thesis proposes concrete analysis and possible improvements for such models, alleviating well-known difficulties in computational imaging. It should be emphasized that, as illustrated above, the phase retrieval and optical scattering problems have also been crucial parts in many branches of physics involving imaging. It is expected that techniques developed in this work can be directly applied on various research, facilitating more accurate characterization of physical systems.

1.2 Thesis overview

As illustrated in the previous section, this thesis is dedicated to proposing computational improvements and better analysis on various models regarding the phase retrieval and the approximation of optical scattering. For the phase retrieval, the transport-of-intensity equation (TIE) is considered. As TIE is known to sensitive to experimental conditions, a few mathematical ways to alleviate such problems are proposed. In addition, it is quantitatively compared to relevant methods with different assumptions on the wave propagation for better understanding on its applicability. Furthermore, when it comes to optical scattering, several different models are also discussed. Such models have different properties and these properties are known to have significant influence on our estimation on objects. Yet, there has not been detailed and quantitative analysis on their relationship. Subsequently, this thesis proposes concrete description on the relationship. Finally, based on the description, a model called the Lippmann-Schwinger equation (LSE) is discussed for its accuracy but limited applicability due to the heavy computational burden. Consequently, this thesis suggests numerical tricks and neural networks to mitigate the computational

problems regarding the LSE.

In Chapter 2, TIE is proposed as a model for the phase retrieval. Compared to other phase retrieval models, origins of its sensitivity on experimental conditions are reviewed. In turn, a new idea to interpret and solve TIE is proposed through the adjoint method. It is shown that this idea can greatly improve the numerical stability of TIE. Furthermore, another formulation that is non-paraxial but closely related with the proposed idea is considered, and the analytical relationship between these methods is discussed for better understanding on various phase retrieval methods.

In Chapter 3, three models, in the scalar scattering theory are discussed. They are all originated from the scalar wave equation, but the direct and quantitative relationship between them has not been proposed in detail. Starting from the LSE, other famous models such as the Born series and the beam propagation method (BPM) are derived. In addition, the reason that they exhibit different numerical stability is discussed. Compared to LSE, BPM has distinct strengths and weaknesses, whose applicability on different experimental situations are presented in analytical ways.

In Chapter 4, numerical strategies to alleviate heavy computational requirements in the LSE are shown. The most significant component is to leverage formulations proposed in the proximal gradient descent where regularziations can be applied to facilitate the convergence of the LSE. In turn, it is proposed to use neural networks as a learnable regularization as appropriate forms of regularizations for such problem are not known. In addition, the structure of neural networks is carefully designed to consume far less amount of computational power compared to classical approaches and to promote generalizability on various scattering objects.

Altogether, this thesis show quantitative descriptions on famous models in computational imaging, providing more fundamental understanding on advantages and disadvantages regarding them. Furthermore, this work presents mathematical ideas on significantly alleviating disadvantages in phenomenal models that have not been resolved very well. While it is not easy to convey an unified description on all existing models in the phase retrieval and the optical scattering, it is expected that such mod-

els can still be regarded as extensions to models discussed in this work, promoting theoretical analysis on such models.

Chapter 2

Noise robust phase retrieval via the transport-of-intensity equation with adjoint method

2.1 Introduction

As introduced in Chapter 1, the transport-of-intensity equation (TIE) has been one of the important models in the phase retrieval. It is an equation that relates intensity derivatives to information on the phase of a wave:

$$-k_0 n_b \frac{\partial I(\mathbf{x}, z)}{\partial z} = \nabla_{\mathbf{x}} \cdot \Big(I(\mathbf{x}, z) \nabla_{\mathbf{x}} \phi(\mathbf{x}, z) \Big), \tag{2.1}$$

where $k_0 = 2\pi/\lambda$, n_b is the refractive index of background, λ is the vacuum wavelength, and $\nabla^2_{\mathbf{x}}$ denotes the Laplacian applied on the lateral dimensions $\mathbf{x}$, z is the optical axis. Here, I and ϕ denote the intensity and the phase part of a wave ψ respectively, i.e.

$$\psi(\mathbf{x}, z) = \sqrt{I(\mathbf{x}, z)} e^{i\phi(\mathbf{x}, z)}. \tag{2.2}$$

Compared to other phase retrieval algorithms, TIE has several advantages. In principle, the phase ϕ can be obtained by just collecting multiple intensities at

different defocused positions z. Hence, experiments can be conducted in a non-interferometric way; there is no need for a reference beam, making the experimental setting very simple. Furthermore, unlike methods like [16], TIE does not require many measurements, though collecting more measurements may alleviate the ill-posedness from noise. Compared to famous methods such as the Gerchberg-Saxton algorithm, TIE is not based on iterative projections, so the numerical convergence is always guaranteed. Lastly, if necessary, TIE can even be applied on incoherent cases [80]. Accordingly, TIE has been utilized in many phase retrieval problems, due to its simple setting and deterministic convergence.

However, TIE also suffers from serious drawbacks. Such drawbacks originate from inherent assumptions behind TIE and such assumptions are strongly coupled with the solution process, resulting in various numerical instabilities. In this chapter, some theoretical bottlenecks in TIE are reviewed and a new interpretation on the equation is proposed to mitigate drawbacks of the equation.

2.2 Inherent assumptions in TIE and its comparison to propagation models

Though TIE has been mainly used for the phase retrieval because of its simplicity, it can also be thought as a model that estimates the wave propagation under the scalar wave approximation. Accordingly, it is closely related to the scalar Helmholtz equation,

$$\nabla^2\psi(\boldsymbol{r}) + (k_0 n(\boldsymbol{r}))^2\psi(\boldsymbol{r}) = 0 \tag{2.3}$$

where n is the refractive index of objects and we denote $\boldsymbol{r} = (\mathbf{x}, z)$. A special solution to Eq. (2.3) under uniform media is the so-called the paraxial wave form:

$$\psi(\mathbf{r}) \Rightarrow \psi(\mathbf{x}, z)e^{ik_0 n_b z}. \tag{2.4}$$

where $\psi(\mathbf{x}, z)$ is a slowly-varying wavefront in the lateral dimensions. Substituting the paraxial form to Eq. (2.3) results in the paraxial wave equation:

$$\nabla_{\mathbf{x}}^2 \psi(\mathbf{x}, z) + 2ik_0 n_b \frac{\partial \psi(\mathbf{x}, z)}{\partial z} + \mathcal{O}\left(\left\| \frac{\partial \psi^2(\mathbf{x}, z)}{\partial z^2} \right\|\right) = 0. \tag{2.5}$$

Here, $\nabla_{\mathbf{x}}^2$ denotes the Laplacian applied on the lateral dimensions.

To derive TIE, we multiply Eq. (2.5) by $\psi^\star(\mathbf{x}, z)$ and take the complex conjugate:

$$\begin{aligned} \psi^\star \left[\nabla_{\mathbf{x}}^2 \psi\right] + 2ik_0 n_b \psi^\star \frac{\partial \psi}{\partial z} = 0 \\ \psi \left[\nabla_{\mathbf{x}}^2 \psi^\star\right] + 2ik_0 n_b \psi \frac{\partial \psi^\star}{\partial z} = 0, \end{aligned} \tag{2.6}$$

where $\star$ denotes the complex conjugate. If the two equations in Eq. (2.6) are added, taking the real part of the paraxial equation, we can obtain the TIE [97]. Considering the Poynting vector, it can be shown that the TIE in Eq. (2.1) represents the conservation of energy during the propagation [80].

TIE can be connected to other well-known propagation models such as the angular spectrum propagation and the Fresnel propagation via Eqs. (2.3) and (2.5). For example, the solution to Eq. (2.3) in the free space can also be expressed by the angular spectrum propagation:

$$\psi(\mathbf{x}, \Delta z) = \hat{\mathcal{F}}^\dagger \mathcal{H}_{\Delta z} \hat{\mathcal{F}} \psi(\mathbf{x}, z) \tag{2.7}$$

where the notation $\dagger$ represents the adjoint. The operators $\hat{\mathcal{F}}$ and $\mathcal{H}_{\Delta z}$ represent the Fourier transform and the angular spectrum kernel corresponding to a propagation by Δz, respectively. More specifically,

$$\begin{aligned} \hat{\mathcal{F}} : \psi(\mathbf{x}, z) &\to \int d\mathbf{x}\ \psi(\mathbf{x}, z) e^{-2\pi i \langle \mathbf{u}, \mathbf{x} \rangle} \\ \mathcal{H}_{\Delta z} : \psi(\mathbf{u}, z) &\to e^{i\Delta z \sqrt{(k_0 n_b)^2 - (2\pi)^2 \langle \mathbf{u}, \mathbf{u} \rangle}} \psi(\mathbf{u}, z) \end{aligned} \tag{2.8}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. On the other hand, the Fresnel propagation is the

paraxial version of the angular spectrum propagation:

$$\begin{aligned}\psi(\mathbf{x}, z+\Delta z) &= \hat{\mathcal{F}}^{\dagger}\mathcal{H}_{\Delta z}\hat{\mathcal{F}}\psi(\mathbf{x}, z) \\ &\approx \hat{\mathcal{F}}^{\dagger}\left[e^{ik_0 n_b \Delta z} e^{-i\pi\lambda/n_b \Delta z\langle \mathbf{u},\mathbf{u}\rangle + \mathcal{O}(u^4)}\right]\hat{\mathcal{F}}\psi(\mathbf{x}, z) \\ &\overset{\Delta z\to 0}{\approx} \hat{\mathcal{F}}^{\dagger}\left[1 - i\pi\frac{\lambda}{n_b}\Delta z\,\langle \mathbf{u},\mathbf{u}\rangle + \mathcal{O}\left((\Delta z)^2 u^4\right)\right]\hat{\mathcal{F}}\psi(\mathbf{x}, z) \\ &\approx \psi(\mathbf{x}, z) + i\frac{\Delta z}{2k_0 n_b}\nabla_{\mathbf{x}}^2\psi(\mathbf{x}, z),\end{aligned} \tag{2.9}$$

where the term $e^{-i\pi\lambda/n_b\Delta z\langle \mathbf{u},\mathbf{u}\rangle}$ corresponds to the Fresnel kernel in the Fourier space. Note that Eq. (2.9) is the finite-difference version of Eq. (2.5). In the third line in Eq. (2.9), the constant phase term $e^{ik_0 n_b \Delta z}$ is dropped for the simplicity. This corresponds to an additional term $2(k_0 n_b)^2\psi(\mathbf{x}, z)$ to Eq. (2.5),

$$\nabla_{\mathbf{x}}^2\psi(\mathbf{x}, z) + 2ik_0 n_b\frac{\partial\psi(\mathbf{x}, z)}{\partial z} + 2(k_0 n_b)^2\psi(\mathbf{x}, z) + \mathcal{O}\left(\left\|\frac{\partial\psi^2(\mathbf{x}, z)}{\partial z^2}\right\|\right) = 0, \tag{2.10}$$

which some of the original studies on the TIE such as [97] use for the derivation of the TIE, instead of Eq. (2.9). To explicitly show the relationship between the terms, perform the Taylor expansion on $e^{ik_0 n_b \Delta z}$ as well and drop $(\Delta z)^2$ dependencies, which gives

$$\psi(\mathbf{x}, \Delta z) \approx \psi(\mathbf{x}, z) + i\frac{\Delta z}{2k_0 n_b}\nabla_{\mathbf{x}}^2\psi(\mathbf{x}, z) + ik_0 n_b\Delta z\psi(\mathbf{x}, z). \tag{2.11}$$

This equation reveals the relationship. In the following sections, the $\mathbf{x}$ subscript in ∇ is dropped for the notational simplicity, unless this can cause mathematical confusion.

Consequently, it is clear how the angular spectrum propagation, the Fresnel propagation, and the TIE are related to each other. Particularly, the Fresnel propagation and the TIE both stem from the same paraxial approximation. In addition, TIE can be regarded as a convenient version of the Fresnel propagation by deliberately ignoring information on the imaginary part of the wave, making itself be easily applicable to non-interferometric experiments. When it comes to their behaviors, they will converge to the same solution under the paraxial approximation, but the angular spectrum propagation is the only model that can give correct propagation estima-

tions when such approximation is no longer valid. In practice, however, the accurate estimation of the wave propagation can be a subtle problem, as it would depend on conditions such as the sampling frequency, numerical aperture, and aberrations [74, 105]. Under the paraxial approximation, it has been extensively suggested that the deviation between paraxial and non-paraxial methods becomes negligible [35, 105], while the high-order frequency terms in non-paraxial methods might lead to unexpected side effects compared to paraxial methods. Hence in the following sections, the performance of TIE is compared with the angular spectrum to demonstrate its validity, but at the same time, it should be noted that the angular spectrum can be an overkill and may have numerical side effects in practice.

2.3 Numerical instabilities in using TIE

Based on previous studies on TIE, the typical experimental procedure can be summarized as follows.

- Collect intensities from different positions (defocused intensities) on the optical axis, i.e. $I_i = I(\mathbf{x}, z_i)$, $i \in \{1, 2, \cdots N\}$.
- Estimate the intensity derivative at the position of interest (usually at the in-focus position) using the measurements I_i.
- Solve for ϕ in Eq. (2.1), which is the phase at the position of interest. The equation can be solved either by directly inverting differential operators $\nabla\cdot$ and ∇, or in a variational way [8].

During the procedure, one can face a few numerical instabilities. First, estimating the intensity derivative from measurements can become a non-trivial problem under noise. For example, using the first-order central difference,

$$\frac{\partial I}{\partial z} = \frac{I(\mathbf{x}, \Delta z) - I(\mathbf{x}, -\Delta z)}{2\Delta z} + \mathcal{O}\big((\Delta z)^2 + \varepsilon/\Delta z\big) \tag{2.12}$$

where ε represents noise. Here, we face a dilemma:

- If we use very small Δz, the error term $\varepsilon/\Delta z$ dominates, and we should be very careful about the precision on defocus positions.
- If we use large Δz, $(\Delta z)^2$ becomes no longer negligible. Likewise, our assumption in Eq. (2.9) for the paraxial wave equation is violated.

In addition, the combination of the divergence and the gradient operators on the right hand side of Eq. (2.1) exhibits numerically unfavorable behaviors. Introducing an auxilliary function ξ [97] such that

$$\nabla\xi(\mathbf{x}, z) \equiv I(\mathbf{x}, z)\nabla\phi(\mathbf{x}, z), \tag{2.13}$$

TIE can be reformulated as

$$\begin{aligned} \frac{\partial I}{\partial z} &= -\frac{1}{k_0 n_b}\nabla^2\xi \\ \nabla^2\phi &= \nabla\cdot\left[\frac{1}{I}\nabla\xi\right]. \end{aligned} \tag{2.14}$$

Subsequently, the direct solution to ϕ can be expressed by using Fourier transform:

$$\phi = \frac{n_b k_0}{(2\pi)^2}\hat{\mathcal{F}}^\dagger\frac{[\mathbf{u}\cdot]}{\langle\mathbf{u},\mathbf{u}\rangle}\hat{\mathcal{F}}\left(I^{-1}\right)\hat{\mathcal{F}}^\dagger\frac{\mathbf{u}}{\langle\mathbf{u},\mathbf{u}\rangle}\hat{\mathcal{F}}\frac{\partial I}{\partial z}, \tag{2.15}$$

where with slight abuse of notation, $[\mathbf{u}\cdot]$ means the dot product with the frequency components in the lateral dimensions. For convenience, if is further assumed that I is nearly uniform at the position of interest,

$$\phi = \frac{n_b k_0}{(2\pi)^2 I}\hat{\mathcal{F}}^\dagger\frac{1}{\langle\mathbf{u},\mathbf{u}\rangle}\hat{\mathcal{F}}\frac{\partial I}{\partial z}. \tag{2.16}$$

The kernel $\langle\mathbf{u},\mathbf{u}\rangle^{-1}$ has a singularity at the origin and significantly penalizes high frequency information from $\frac{\partial I}{\partial z}$. It is well-known that due to the singularity at the origin, the low frequency part of noise in $\frac{\partial I}{\partial z}$ can be amplified, resulting in cloud-like artifacts in the estimation of ϕ. On the other hand, due to the high-frequency penalty, TIE often has difficulty in recovering sharp features. In particular, such behavior is exacerbated as we use large Δz. Of course, poor $\frac{\partial I}{\partial z}$ estimation under large Δz can be one cause of such behavior, but in addition, note that the quadratic penalty

$\langle \mathbf{u}, \mathbf{u} \rangle^{-1}$ originates from the paraxial approximation in Eq. (2.9). As we increase Δz, $\mathcal{O}((\Delta z)^2)$ terms are no longer negligible and the propagation kernel becomes not quadratic but oscillatory, implying that the deviation between the TIE and the angular spectrum propagation becomes large especially in the high frequency regime. In other words, the slowly-varying envelope approximation becomes invalid and high-frequency components of the wavefront would not be well estimated. Previous studies like [8] discuss the behavior with the transfer function.

In summary, TIE has useful advantages over other phase retrieval methods, but due to inherent limitations in its kernel and the choice of Δz, it is prone to producing infeasible estimations. In what follows, a new technique is proposed to alleviate the numerical problems regarding the TIE reviewed in this section, which can greatly help the phase retrieval process.

2.4 Interpreting TIE as an ordinary differential equation coupled with transport-of-phase equation

To mitigate the major problems in TIE discussed in the last section, one may require an idea that can satisfy the following conditions:

- Avoid estimating $\frac{\partial I}{\partial z}$ explicitly from (possibly noisy) measurements.
- Adopt arbitrarily small Δz for the validity of the Fresnel approximation, without worrying about the precision on the measurement positions.
- Leverage information from multiple measurements; it is an inefficiency in the information-theory sense if such information is all fused into $\frac{\partial I}{\partial z}$.

In this section, a new interpretation on the TIE is proposed to tackle the aforementioned problems. That is to treat the TIE as an ordinary differential equation (ODE) with respect to z:

$$-k_0 n_b \frac{d\boldsymbol{I}(z)}{dz} = \nabla \cdot \Big(\boldsymbol{I}(z) \nabla \boldsymbol{\phi}(z) \Big), \tag{2.17}$$

where the boldfaced functions corresponds to the discretized version of the original functions on the Cartesian coordinates, i.e.

$$\boldsymbol{I}(z) \equiv \left(\cdots, \underbrace{I(x_i, y_j, z)}_{[i + N_y(j-1)]^{\text{th}} \text{ element of a vector}}, \cdots \right)^{\mathsf{T}}, \tag{2.18}$$

where N_y is the number of discretization points along the y axis. Subsequently, we consider the following procedure:

- Starting from our initial guess on $\boldsymbol{I}$ and $\boldsymbol{\phi}$ at some point, integrate Eq. (2.17). This requires information on $\boldsymbol{\phi}$ over the optical axis, assume that we have such information for now.
- Compute a loss function between estimated $\boldsymbol{I}$'s and the measurements at the designated defocus positions.
- Evaluate the gradient of the loss function with respect to $\boldsymbol{\phi}(z^\circ)$ where z° corresponds to the position of interest at which one wants to retrieve the phase.

During the procedure, $\frac{\partial I}{\partial z}$ is not directly computed from measurements. Rather, it is only approximated based on current estimations on $\boldsymbol{I}$ and $\boldsymbol{\phi}$, which can even be regularized. Moreover, the choice on Δz is no longer a critical problem; Δz can be set arbitrarily small, or the adaptive step size can be leveraged based on the current slope (variance of $\boldsymbol{I}$ over z). Lastly, all individual measurements contribute to the loss function, not just fused into $\frac{\partial I}{\partial z}$, resulting in estimations on $\boldsymbol{I}$ and $\boldsymbol{\phi}$ being adequately penalized based on measurements.

However, solving Eq. (2.17) has non-trivial problems. First, as mentioned above, any information on $\boldsymbol{\phi}$ on arbitrary positions on the optical axis is not generally available in experimental conditions of TIE, while it is necessary to integrate the equation. Second, the computation of the gradient of the loss function may not look straightforward. Intuitively, one has to imagine a *flow of gradients* over z in the opposite direction to the integration (e.g. backpropagation with layers where each layer can correspond to an integration step).

For the first problem regarding the ODE interpretation, note that TIE is the real part of the paraxial wave equation. From Eq. (2.6), the imaginary part of the wave equation can also be derived, leading to the transport-of-phase equation (TPE) [113, 97, 114]:

$$2k_0 n_b I^2 \frac{\partial \phi}{\partial z} = \frac{1}{2} I \nabla^2 I - \frac{1}{4} (\nabla I)^2 - I^2 (\nabla \phi)^2 + k_0 n_b I^2. \tag{2.19}$$

Subsequently, a system of coupled ODEs consisting of TIE and TPE can be presented:

$$\begin{aligned} \boldsymbol{\varphi} = \boldsymbol{I} \oplus \boldsymbol{\phi} \quad &\Longleftrightarrow \quad \frac{d\boldsymbol{\varphi}}{dz} = \frac{d\boldsymbol{I}}{dz} \oplus \frac{d\boldsymbol{\phi}}{dz} \\ -k_0 n_b \frac{d\boldsymbol{I}(z)}{dz} &= \nabla \cdot \Big(\boldsymbol{I}(z) \nabla \boldsymbol{\phi}(z) \Big) \\ 2k_0 n_b \boldsymbol{I}^2 \frac{d\boldsymbol{\phi}}{dz} &= \frac{1}{2} \boldsymbol{I} \nabla^2 \boldsymbol{I} - \frac{1}{4} (\nabla \boldsymbol{I})^2 - \boldsymbol{I}^2 (\nabla \boldsymbol{\phi})^2 + k_0 n_b \boldsymbol{I}^2. \end{aligned} \tag{2.20}$$

where $\oplus$ represents the concatenation of two vectors. Starting from the position of interest over the optical axis and corresponding estimations on $\boldsymbol{\varphi}$, Eq. (2.20) can be integrated, and the loss function is evaluated. In this manner, information on the phase propagation is no longer ignored; all information contained in the original wave equation is leveraged. While the theoretical existence of TPE has been mentioned in previous studies [113, 114], it is of limited usage or often requires serious assumptions as it is difficult to detect how the phase of a wave varies in the propagation in conventional imaging systems. Furthermore, this makes overall experimental settings complex, which reduces one of the main advantages in TIE. By considering TIE and TPE as coupled ODEs, such limitations are remediated.

For the concrete review on the solution process regarding the coupled TIE and TPE, computational methods on the gradient of the loss function in an ODE-constrained problem should also be discussed. Here, an ODE-constrained problem implies a minimization of a loss function whose estimations are constrained to be related to a solution of an ODE. Appendix A illustrates a method called the adjoint method [87, 18], evaluating the gradient of the loss function for general equality-constrained optimization problems. Intuitively, the adjoint method tells that the core compo-

nent of such evaluation is an equation on the adjoint of an operator that constitutes equality constraints. In physics, adjoints or physical operators usually imply a reversal of domains; the adjoint of the projection in computational tomography is the backprojection, and that of the wave propagation is the backpropagation, etc. In a similar manner, the adjoint method results in another ODE that should be solved in the opposite direction to an original ODE, which is described in Eq. (A.14).

2.5 Performance of the coupled TIE-TPE for phase retrieval problems

2.5.1 Numerical experiments

Evaluating the performance of the proposed method, the coupled TIE and TPE, numerical experiments are designed. In particular, previous techniques to solve the TIE are considered to see potential benefits from interpreting the TIE as an ODE. Furthermore, as illustrated in Chapter 2.2, the proposed method is further compared with non-paraxial methods for checking its theoretical validity and applicability under paraxial circumstances. To this end, four methods are tested:

- TIE, which corresponds to solving Eq. (2.15). $\frac{dI}{dz}$ is estimated by the finite difference of multiple measurements $\boldsymbol{I}_m(z)$.
- TIE-iterative, which is based on TIE, but tries to iteratively adjust the phase estimation [109].
- TIE-TPE, which is the ODE consists of TIE and TPE with a composite loss denoted as L_1:

$$L_1(\boldsymbol{\phi}(0)) = \underset{\boldsymbol{\phi}(0)}{\arg\min} \sum_i l\left(\mathcal{S}_{\boldsymbol{I}}\left[\boldsymbol{\varphi}(0) + \int_0^{z_i} dz\, \frac{d\boldsymbol{\varphi}}{dz}\right],\ \boldsymbol{I}_m(z_i)\right), \tag{2.21}$$

  where l is a loss function, $\mathcal{S}_{\boldsymbol{I}} : \boldsymbol{\varphi} \to \boldsymbol{I}$ is an intensity selection operator, where the position at which the phase should be estimated is set 0.

- Angular spectrum (AS), which is similar to TIE-TPE but we estimate the propagation of the light field with the angular spectrum method, i.e.

$$L_2(\boldsymbol{\phi}(0)) = \underset{\boldsymbol{\phi}(0)}{\arg\min} \sum_i l\left(\left\|\mathcal{P}_{z_i}\left[\sqrt{\boldsymbol{I}(0)}e^{i\boldsymbol{\phi}(0)}\right]\right\|^2,\ \boldsymbol{I}_m(z_i)\right), \tag{2.22}$$

  where $\mathcal{P}_{z_i}$ stands for the angular spectrum propagation by distance z_i.

The paraxiality of light propagation can be evaluated by the Fresnel number:

$$F = \frac{a^2}{\lambda D}, \tag{2.23}$$

where a is the characteristic size of an object, D is the propagation distance from the object. Without much loss of generality, the Fresnel number is adjusted by changing D to observe behaviors of the different phase retrieval methods under various paraxiality conditions. Specifically, setting $\lambda = 1.0$, the propagation of the light field from a square aperture at $z = 0$ with size $L = 500$ is considered. At $z = 0$, the intensity is assumed as a Gaussian beam whose standard deviation is two times the field of view. Then defocused intensities at $z = -n_z\Delta z,\ -(n_z - 1)\Delta z,\ \cdots,\ (n_z - 1)\Delta z,\ n_z\Delta z$ are collected, constituting $2n_z + 1$ images, by considering the angular spectrum propagation as a ground truth model. The ground truth phases are randomly selected from the ImageNet dataset [90]. The loss l is the mean squared error. The regularization parameter at the sinularity for TIE and TIE-iterative is chosen as 5×10^{-9}.

The ODEs in TIE-TPE are solved by using the so-called `tsit5` integration method [99], with adaptive step sizes. To obtain estimated $\boldsymbol{\varphi}$ at the designated defocused positions on the optical axis, the 4^{th} order interpolation scheme is adopted, which is implemented as the Runge-Kutta method.

2.5.2 Results

To check the validity of TIE implementations and their performance in ideal situations, a very basic experiment is considered; $n_z = 1$ and Δz is adjusted so that the

Figure 2-1: An example phase image at $z = 0$. The range of phase values is $[0, 0.4]$ in radian.

Fresnel number ranges from 1200 to 2800. In addition, it is assumed that there is no noise in measurements. For the visualization, Fig. (2-1) shows an example for the ground-truth phase image. Fig. (2-2) and Table 2.1 presents the phase estimation results. For quantitative analysis, three metrics are considered: the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), and the normalized root mean square error (NRMSE) [104, 51]. Under such ideal situations, the non-gradient methods, TIE and TIE-iterative, exhibit better performance than other methods. This can be mainly attributed to two reasons. First, at high F, TIE and TIE-iterative will not show significant discrepancy from the actual propagation of the light field, as the Fresnel approximation holds. In addition, $\left\|\frac{dI}{dz}\right\|$ is expected to be very small, which makes the gradient $\frac{\partial L_{1,2}}{\partial \boldsymbol{\phi}(0)}$ also small accordingly. This can easily stagnate the update of $\boldsymbol{\phi}(0)$. In Table 2.1, TIE and TIE-iterative seem to show better performance in overall regardless of F.

Next, the case that the intensity measurements are contaminated with noise is considered. To alleviate the difficulty in estimating the phase under high noise, multiple measurements are collected, i.e. $n_z = 3$. In this condition, based on [103], a better approximation on $\frac{dI}{dz}$ can be obtained for TIE and TIE-iterative as

$$\frac{dI}{dz} = \sum_{n=-n_z}^{n_z} \frac{b_n}{\Delta z} I_m(n\Delta z) \tag{2.24}$$

where b_n is the finite difference coefficient. As a side effect, it can be expected that

Table 2.1: The quality assessment of estimated phases from three measurements $I_m(\Delta z)$, $I_m(0)$, and $I_m(-\Delta z)$ without noise. Intensity measurements are simulated by using the angular spectrum method. Numbers presented here represent averaged metrics over 30 phase images randomly selected from the ImageNet dataset. The range of phase in each image is $[0, 0.4]$ in radian.

Metric	F	TIE	TIE-iterative	TIE-TPE	Angular spectrum
PSNR	2800	24.2696	18.6079	11.1241	11.5336
	1700	23.8639	28.9110	11.3525	11.7244
	1200	22.8106	25.5628	11.2890	11.7719
SSIM	2800	0.9061	0.6033	0.6533	0.7356
	1700	0.8712	0.8399	0.6728	0.7448
	1200	0.8237	0.7964	0.6395	0.7492
NRMSE	2800	0.1350	0.2816	0.5687	0.5532
	1700	0.1440	0.832	0.5508	0.5360
	1200	0.1586	0.1207	0.5532	0.5323

the estimation on $\frac{dI}{dz}$ becomes more stable to noise. Figs. (2-3)-(2-5) and Table 2.2 summarize estimation results. Now, TIE and TIE-iterative suffer from the well-known cloudy artifacts that originate from the low-frequency noise amplification in the Fourier space. Such amplification exists despite using Eq. (2.24) for more accurate derivative estimation and noise reduction. In contrast, results from AS and TIE-TPE contain significantly smaller amount of cloudy artifacts and maintain crisp features and the dynamic range of the original image, e.g. Fig (2-4). In other words, the figure demonstrates that the extreme sensitivity of classical TIE solvers results in not only generation of artifacts but significant deviation in terms of the dynamic range of the phase value. Such behavior is also reflected in quantitative metrics reported in Table 2.2.

In overall, AS and TIE-TPE exhibit far better performance than TIE and TIE-iterative, unless being in a very ideal situation: F is high and intensities are measured in great accuracy. In other words, AS and TIE-TPE show comparable results to each other regardless of the inclusion of noise. This demonstrates that TIE-TPE is able to stabilize numerical problems regarding the original TIE formulation, and attains the comparable quality to AS under the paraxial condition; in fact, Table 2.2 tells that TIE-TPE can exhibit even better precision. This may be attributed to an empirical

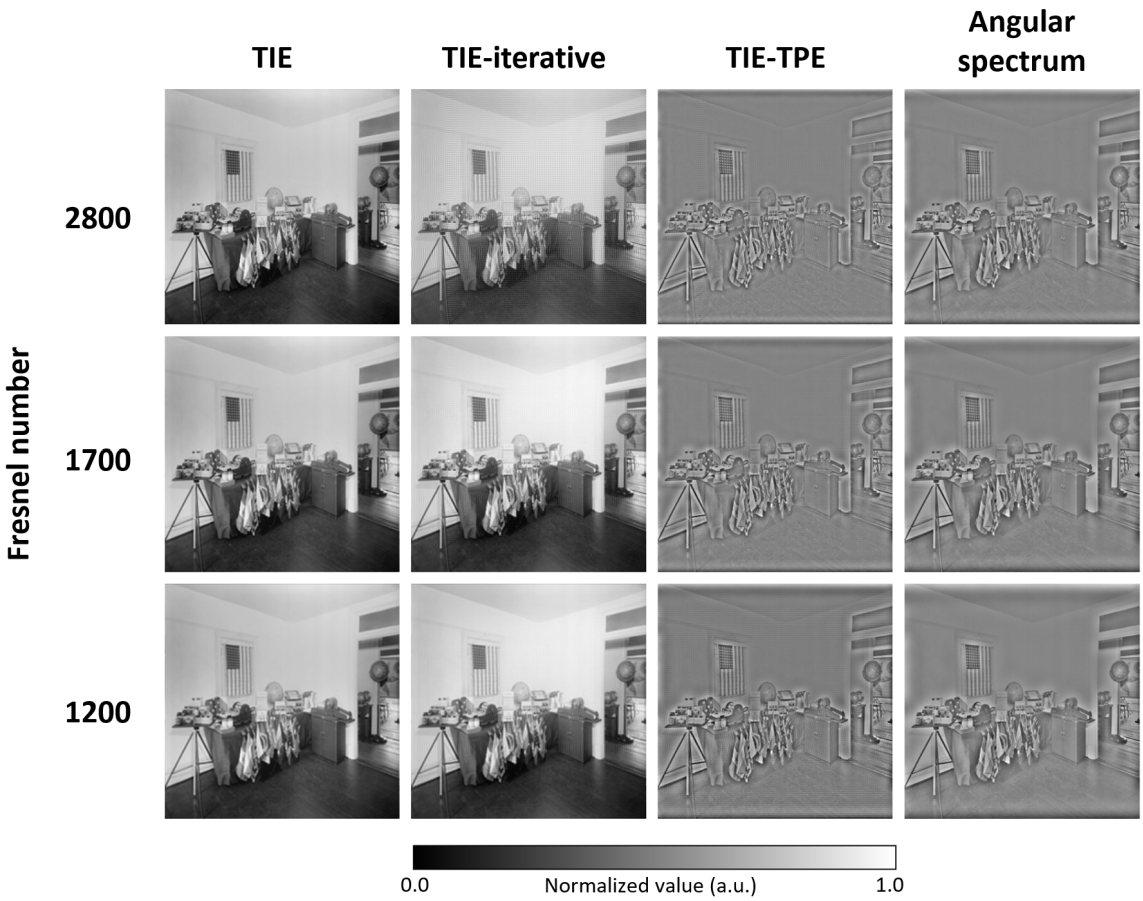

Figure 2-2: Estimated phases ($n_z = 1$) from different phase retrieval methods without noise. For qualitative visualization, each phase estimation is normalized to $[0.0, 1.0]$. For quantitative comparison of estimations from different methods presented in this figure, see Tab. 2.1. The ground truth phase image is presented in Fig. (2-1).

observation that AS tends to overestimate the maximum value of the phase in images during the experiments. Qualitatively, such behavior is also depicted in Fig. (2-4).

In addition to the experimental results, the gradient of the loss functions in AS and TIE-TPE can be analyzed further to better understand the retrieval process and corresponding mathematical properties behind each model. To start, the gradient of the loss is presented:

$$\frac{dL_{1,2}}{d\boldsymbol{\phi}(0)} = \frac{dL_{1,2}}{d\boldsymbol{I}(z)} \frac{d\boldsymbol{I}(z)}{d\boldsymbol{\phi}(0)}. \tag{2.25}$$

Table 2.2: The quality assessment of estimated phases from seven measurements $I_m(3\Delta z)$, $I_m(2\Delta z)$, $\cdots$, and $I_m(-3\Delta z)$ with noise. Intensity measurements are simulated by using the angular spectrum method. For the noise, we assume 1000 photon counts per pixel. Numbers presented here represent averaged metrics over 30 phase images randomly selected from the ImageNet dataset. The range of phase in each image is $[0, 0.4]$ in radian.

Metric	F	TIE	TIE-iterative	TIE-TPE	Angular spectrum
PSNR	2800	-19.1852	-20.2212	8.4090	7.2116
	1700	-15.9884	-16.7825	8.6958	6.7882
	1200	-17.1122	-19.1091	8.7023	6.3468
SSIM	2800	0.211	0.182	0.4010	0.3904
	1700	0.391	0.329	0.3403	0.2686
	1200	0.426	0.312	0.3073	0.2239
NRMSE	2800	18.3249	20.6283	0.8344	0.9152
	1700	12.7417	13.9629	0.7983	0.9626
	1200	14.4368	18.1587	0.8013	0.9958

From Eq. (A.14), we can derive the analytical expression for Eq. (2.25):

$$\begin{aligned} \frac{dL_1}{d\boldsymbol{\phi}(0)} &= -\mathcal{S}_{\boldsymbol{\phi}}\left[h^\dagger(0)\right] \\ \frac{dh}{dz} &= \left(-\frac{\partial y}{\partial \boldsymbol{\varphi}}\right)^\dagger h + \left[2\big(\boldsymbol{I}(z) - \boldsymbol{I}_m(z)\big)\delta(z - z_i) \oplus \mathbf{0}\right] \end{aligned} \tag{2.26}$$

where y represents the expression for $\frac{d\boldsymbol{\varphi}}{dz}$ in Eq. (2.20), $\mathcal{S}_{\boldsymbol{\phi}} : \boldsymbol{\varphi} \to \boldsymbol{\phi}$ is a phase selection operator, and z_i represents measurement points on the optical axis. The adjoint equation on h should be integrated from $h(n_z\Delta z) = 0$ to $h(0)$ backwards to compute the final derivative of the loss. For mathematical simplicity, it is assumed that Δz is reasonably small and $n_z = 1$. Under such condition, the intensity difference term on the right hand side of $\frac{dh}{dz}$ would only be nonzero at $z = \Delta z$, and the light field strongly satisfies the paraxial approximation. In addition, it may be also assumed that the gradient of $\boldsymbol{I}$ is negligible compared to other terms in Eq. (2.20), as often proposed in classical solutions to TIE. Then,

$$y : \boldsymbol{\varphi} \underset{\text{approximately}}{\to} \left[-\frac{1}{k_0 n_b}\boldsymbol{I}(z)\nabla^2\boldsymbol{\phi}\right] \oplus \left[\frac{2}{k_0 n_b}(\nabla\boldsymbol{\phi})^2 + \frac{1}{2}\right], \tag{2.27}$$

and

$$h(0) \approx \left(\Delta z \frac{\partial y}{\partial \boldsymbol{\varphi}}^{\dagger}\right) \left[2\big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\big) \oplus \mathbf{0}\right], \tag{2.28}$$

where

$$\frac{\partial y}{\partial \boldsymbol{\varphi}}^{\dagger} \approx \begin{pmatrix} \mathcal{D}\left(-\frac{1}{k_0 n_b}\nabla^2 \boldsymbol{\phi}\right) & -\frac{1}{k_0 n_b}\boldsymbol{I}(z)\nabla^2 \\ 0 & \frac{2}{k_0 n_b}(\nabla \boldsymbol{\phi}) \circ \nabla \end{pmatrix}^{\dagger}, \tag{2.29}$$

and $\mathcal{D}(\mathbf{v})$ represents a diagonal matrix whose diagonal elements are $\mathbf{v}$, and $\circ$ corresponds to the Hadamard product applied on each column. Subsequently, from Eq. (2.26), it can be deduced that

$$\frac{dL_1}{d\boldsymbol{\phi}(0)} \approx -\frac{2\boldsymbol{I}(\Delta z)\Delta z}{k_0 n_b}\nabla^2\big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\big). \tag{2.30}$$

Under the Fresnel approximation, Eq. (2.9), taking ∇^2 on the components (I and ϕ) of the light field would imply the degree of propagation. In other words, if the Laplacian is large, the phase difference between the current position to the next position on the optical axis is expected to be large. Hence, intuitively, the term $\nabla^2\big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\big)$ would contain some information on the difference between the estimated phase change and the real phase change near $z = 0$. Subsequently, Eq. (2.30) may be interpreted that one should adjust the phase so that such difference is minimized, which is physically plausible.

In AS, one can also derive the expression for the gradient of the loss:

$$\frac{\partial \boldsymbol{\psi}(\Delta z)}{\partial \boldsymbol{\phi}(0)} = i\mathcal{P}_{\Delta z}\left[\sqrt{\boldsymbol{I}(0)}\mathcal{D}\Big(\exp\left(i\boldsymbol{\phi}(0)\right)\Big)\right], \tag{2.31}$$

and

$$\frac{\partial L_2}{\partial \boldsymbol{\phi}(0)}^{\dagger} = 4\mathrm{Im}\left[\sqrt{\boldsymbol{I}(0)}\mathcal{D}\Big(\exp\left(i\boldsymbol{\phi}(0)\right)\Big)\mathcal{P}_{-\Delta z}\mathcal{D}\Big(\boldsymbol{\psi}^*(z)\Big)\Big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\Big)\right]. \tag{2.32}$$

The core component that appears both in Eqs. (2.30) and (2.32) is the back-propagation

of the wavefront represented by the error in intensities,

$$\nabla^2\big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\big) \Longleftrightarrow \mathcal{P}_{-\Delta z} \cdots \Big(\boldsymbol{I}(\Delta z) - \boldsymbol{I}_m(\Delta z)\Big), \tag{2.33}$$

where just different degrees of the paraxiality are assumed. At initial stages of the optimization process, it is expected that the phase estimation first tries to recover high-frequency features such as edges in measurements, as such features result in a very large Laplacian. This would be the reason that AS and TIE-TPE tends to focus on reconstructing edges in Fig (2-2), under ideal situations without noise. Subsequently, it can be deduced that such behavior, combined with the variational formulation that can suppress effects from noise, leads to better performance under the existence of noise. Eq. (2.33) also emphasizes that AS and TIE-TPE back-propagate information contained in measurements in different ways. Specifically, AS includes higher-order frequency components compared to TIE-TPE, and it may be expected that such behavior would have negative effects under noise due to its discontinuous nature. In this manner, TIE-TPE can exhibit extra numerical stability under the paraxial approximation, which can be one of the prospective reasons that TIE-TPE achieves better accuracy, presented in Table 2.2. However, additional analysis should be conducted, because other factors such as the numerical stability of ODE solvers can also be influenced in practice.

In summary, the proposed method, TIE-TPE, is successful in improving the performance of the original TIE. Prospective reasons that TIE-TPE exhibits better performance compared to relevant phase retrieval methods are also discussed in analytical ways. While further studies are required for more concrete investigation behind different phase retrieval models and their relationship to TIE-TPE, it is expected that this work can facilitate the phase retrieval problem.

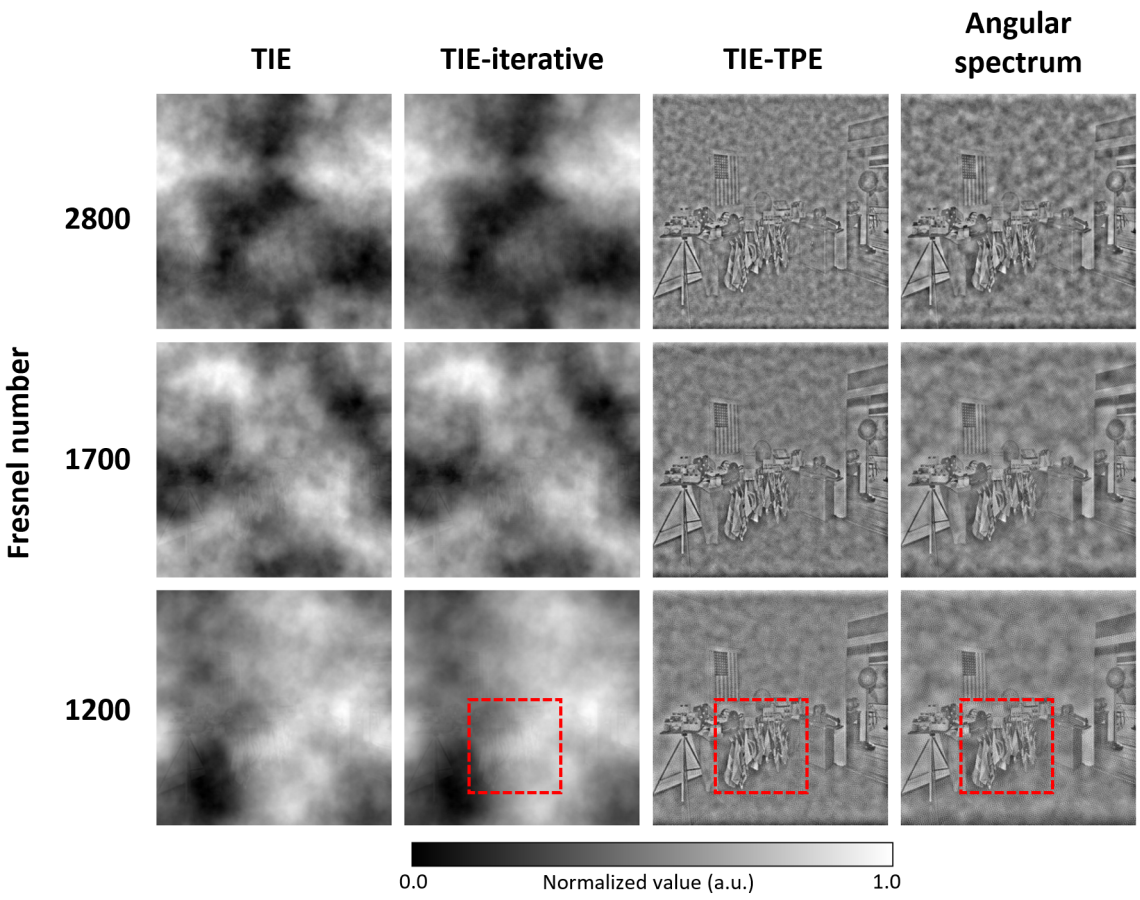

Figure 2-3: Estimated phases ($n_z = 3$) from different phase retrieval methods where the measurements are contaminated with shot noise. For the noise generation, we assume 1000 photon counts per pixel. For qualitative visualization, each phase estimation is normalized to $[0.0, 1.0]$. For quantitative comparison of estimations from different methods presented in this figure, see Tab. 2.2. It can be seen that estimations from TIE and TIE-iterative greatly suffer from the cloudy artifacts originating from numerical instabilities in classical TIE. The ground truth phase image is presented in Fig. (2-1).

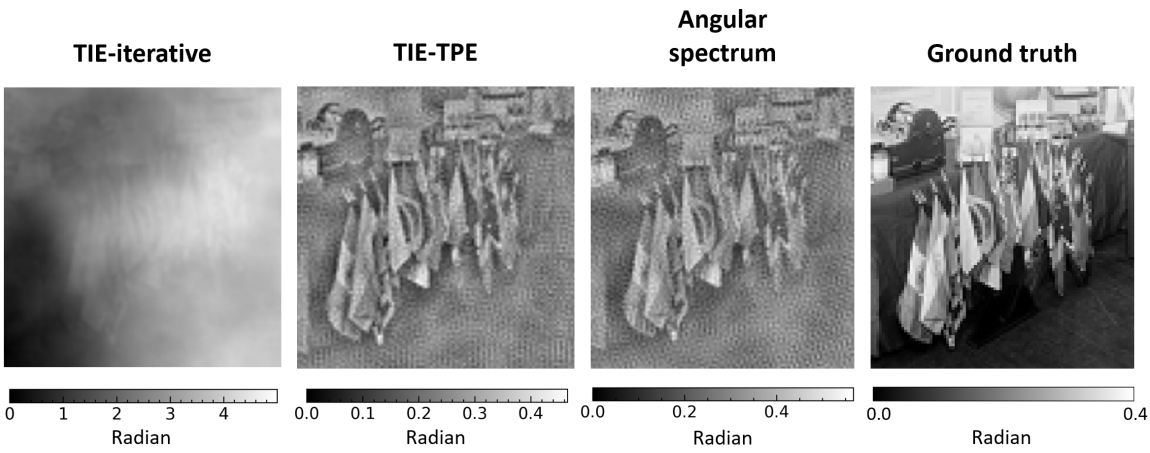

Figure 2-4: Enlarged views of the areas marked with red dotted boxes in Fig. (2-3) where $F = 1200$. The unif ot the phase is in radian. Note that TIE-iterative not only suffers from the cloudy artifacts, but the dynamic range of the estimated phase greatly deviates from that of the ground truth. Such behavior is also described quantitatively in Table 2.2.

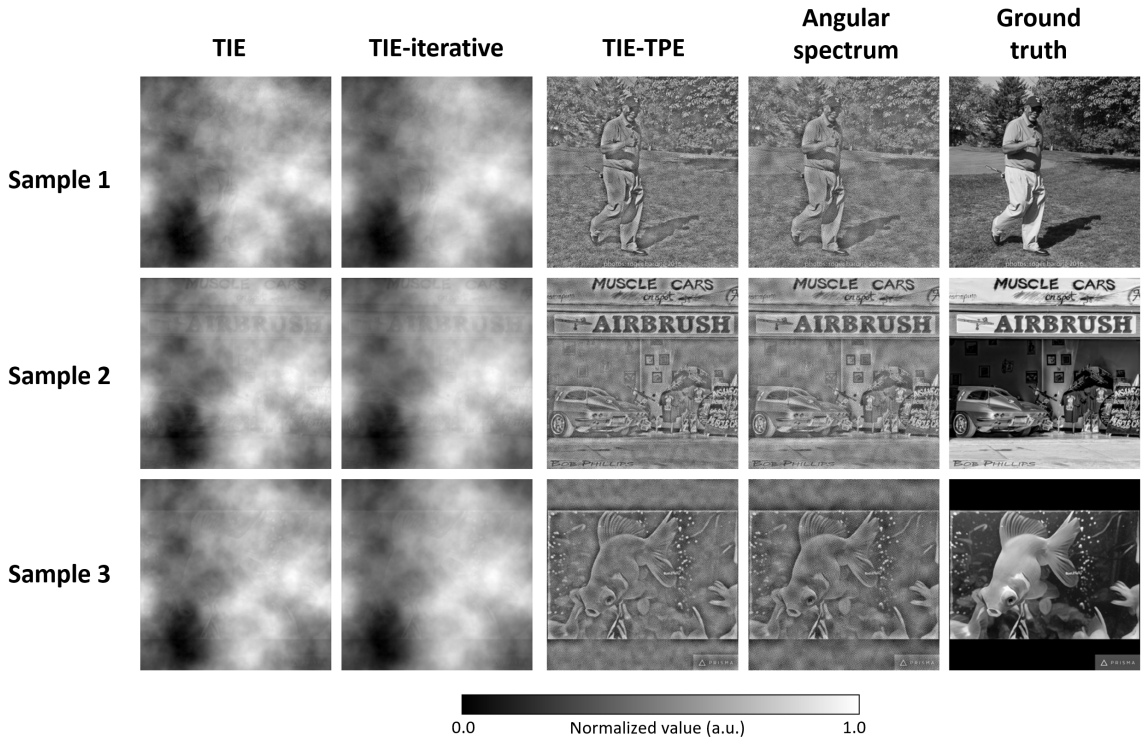

Figure 2-5: Estimated phases ($n_z = 3$, $F = 1200$) corresponding to various images sampled from the ImageNet dataset. The measurements are contaminated with shot noise. For the noise generation, we assume 1000 photon counts per pixel. For qualitative visualization, each phase estimation is normalized to $[0.0, 1.0]$. For quantitative comparison of estimations from different methods presented in this figure, see Tab. 2.2. It can be seen that estimations from TIE and TIE-iterative greatly suffer from the cloudy artifacts originating from numerical instabilities in classical TIE.

Chapter 3

Quantitative analysis on scalar optical scattering theory

This chapter contains contents from a journal publication: *Pang, S. & Barbastathis, G. Unified treatment of exact and approximate scalar electromagnetic wave scattering. Phys. Rev. E 106, 045301 (2022).* The original copyright is credited to American Physical Society.

3.1 Introduction

In Chapter 2, improvements on models regarding the phase retrieval problem are proposed, which enables the extraction of the phase information. Though this can greatly facilitate the use of information contained in the electromagnetic field, there are situations where additional procedures are required to decrypt such information into a useful form. An example situation is, as shown in Fig. (1-1), when the field faces complex scattering interaction with objects of interest. Here, just the recovery of the phase information is not sufficient, and the approximation on the scattering effect is indispensable.

In principle, the complex light-matter interactions leading to scattering are governed by Maxwell's equations or, under some assumptions, by the scalar Helmholtz equation, Eq. (3.1), that describes optical elastic scattering from objects that are large

compared to the wavelength. To simplify the process of modeling optical scattering and estimating object properties, there have been many studies on approximating solutions to the scalar Helmholtz equation. One of the most primitive is the projection approximation, where the scattered field is assumed to maintain the incident wavefront, e.g. a plane or spherical wave, while attenuation and phase delay accumulate proportional to the optical path length of rays through the object. When the incident wavefront is planar or spherical, this assumption leads to the Radon transform formulation, and is the basis of computed tomography. When it comes to relatively thin objects with non-negligible scattering, a more appropriate description is provided by the so-called single scattering approximations, including the first Born and Rytov methods [71]. As objects become dense and highly scattering, as expected, even single scattering methods start to fail, and models accounting for multiple scattering are required. Representative approaches are the Lippmann-Schwinger equation (LSE) [84, 15, 66], the multi-slice method [27, 41, 26, 59] and the beam propagation method (BPM) [88, 55, 56, 43], and the Born series [78, 95]. The multislice and beam propagation methods are very closely related, with the important distinction that the former was motivated by solving Schrödinger's equation, whereas the latter was for the Helmholtz equation.

Multiple scattering models can all be formulated starting from the scalar Helmholtz equation, but they rely on different approximations on the scattering process [60, 50, 37, 79, 25]. Subsequently, all three aforementioned methods may exhibit certain drawbacks compared to exact solutions of the scalar Helmholtz equation, and the discrepancies evidence themselves differently for each method. For example, the multislice method that preceded BPM historically often does include backscattering [20, 93] at the cost of added computational complexity. On the other hand, it has been reported that BPM cannot account for backscattering or reflection of fields and it would not be suitable for experimental conditions that significantly deviate from the paraxial approximation [64, 21]. Born series is numerically unstable, unless the optical potential is sufficiently weak. On the contrary, the LSE, by virtue of originating simply as an integral formulation of the scalar Helmholtz equation under the standard Rayleigh-

Sommerfeld radiation condition, requires no further assumptions. In principle, this can lead to high-precision solutions in numerically ideal cases [25, 50, 107]. However, solving the LSE may still be subject to numerical artifacts resulting from the inversion of the integral equation, and requires relatively intensive computational resources.

Hence, while the LSE promises the most reliable approximations of scattered fields and optical objects [84], BPM or Born series can also be considered if an error compared to LSE is bounded below a given acceptable threshold. In previous studies, conditions that can make such small error achievable are usually summarized qualitatively, e.g. laterally large objects, small illumination angles, and weak potential. This is because LSE, BPM, and Born series originate from different approximations and derivations. Subsequently, explicit and quantitative relationships between the different methods, especially between LSE and BPM, have not been addressed very clearly. For example, would there be a theoretical parameter that can be used for estimation deviations? What would be the relation between the successive application of phase delays [37] and the three-dimensional volume integral in LSE, Eq. (3.3)?

In fact, comparing the solution under the paraxial approximation to the original wave equation in the differential form, there have been multiple studies to provide quantitative measures. However, they often assume special conditions, or do not provide direct insight on the deviation from the integral formulation. One of the earliest studies to apply the paraxial approximation on the wave equation is [62], where the atmospheric propagation of electromagnetic fields in the troposphere is discussed. Subsequently, the very origin of the approximation is mathematically realized as a relationship between two differential operators that constitute the wave equation, e.g. $\hat{P}$ and $\hat{Q}$ in Chapter 3.4.2, in not only optics but related areas such as acoustics [96, 36, 61]. However, the impact of the operators on the validity of the paraxial approximation lacks of a quantitative measure on the error and its connection to the integral formulation. Studies like [98, 37] extend the idea and tries to provide an estimation on the paraxial error, but only under special conditions such as collimating illuminations on stratified media. There exist discussions that include the three-dimensional integration for a wave field in media [88, 10], but they are limited

to apply such integration only on the free-space propagation part in paraxial methods. It is also worth noting that studies that stem from quantum mechanics (e.g. electron optics) start from slightly different formulations and logical flows based on the first-principle [27, 41, 72]; however, their results can also be summarized by the operators above and the paraxial error is not further extended to the form in LSE. More recently, there have been studies that analyze the paraxial propagation operator in mathematically rigorous ways [30, 31, 77], which include meticulous bounds on the error in the paraxial propagation. However, the error discussed in these studies are, strictly speaking, corresponds to an accumulated error over the space in modeling a wavefront emanating from a source. While such wavefront can be viewed as a scattered wave from an object via the Huygens-Fresnel principle, successive scattering inside media, e.g. how a wave field from a source is refracted by an nearby object, is not directly described by the studies. In addition, they, and other relevant studies [111], do not develop the error bound toward the integral formulation, i.e. the very origin of the deviation in the process that non-paraxial methods are rewritten as paraxial methods and vice versa, if such rewriting is possible.

Note that the precision of a scattering model may not be the sole parameter to determine the quality of field or object estimations. This is because such estimations consist of complex optimization procedures, which would also depend on various mathematical conditions e.g. preconditioning and regularization. Nevertheless, a more concrete understanding of the relationships and relative strengths and weaknesses of each method would be beneficial for us to analyze estimation results, review numerical settings, and track origins of artifacts and errors by evaluating applicability of scattering models.

Therefore, in this section, a definitive and quantifiable relationship among LSE, Born series, and BPM is proposed. In addition, concrete conditions where the scattered fields estimated respectively from the three methods exhibit insignificant differences are introduced. Specifically, a simple and dimensionless parameter is suggested to test the validity of Born series solution. Furthermore, the BPM is directly derived from the LSE and its corresponding Born series. This leads to another dimensionless

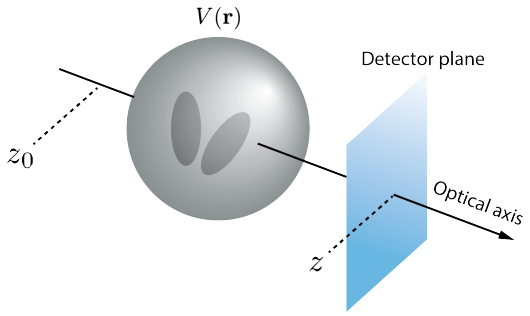

Figure 3-1: An example geometry for optical scattering from an optical potential V.

parameter based on explicit approximations adopted along the derivation.

3.2 Formulation of LSE

For mathematical convenience, the scalar Helmholtz equation Eq. (2.3) can be rewritten as

$$\left[\nabla^2 + (n_\mathrm{b}k_0)^2\right]\psi(\boldsymbol{r}) = -(n_\mathrm{b}k_0)^2\left[\left(\frac{n(\boldsymbol{r})}{n_\mathrm{b}}\right)^2 - 1\right]\psi(\boldsymbol{r}). \tag{3.1}$$

As a reminder, the phase velocities are obtained by dividing the vacuum light speed by the respective indices. Using the Green's function that satisfies the radiation condition [91],

$$G(\boldsymbol{r}-\boldsymbol{r}') = \frac{\exp\left(in_\mathrm{b}k_0\left\|\boldsymbol{r}-\boldsymbol{r}'\right\|\right)}{4\pi\left\|\boldsymbol{r}-\boldsymbol{r}'\right\|}, \tag{3.2}$$

an integral formulation identical to Eq. 3.1 may be derived, which is the LSE:

$$\psi(\boldsymbol{r}) = \psi_0(\boldsymbol{r}) + \int d\boldsymbol{r}'\; G(\boldsymbol{r}-\boldsymbol{r}')V(\boldsymbol{r}')\psi(\boldsymbol{r}'). \tag{3.3}$$

Here, $V(\boldsymbol{r}) = (n_\mathrm{b}k_0)^2\left[\left(\frac{n(\boldsymbol{r})}{n_\mathrm{b}}\right)^2 - 1\right]$ is the optical scattering potential and ψ_0 is the incident field.

The BPM describes the scattering process as a sequential application of 2D scattering layers, so it is not obvious how it can relate to the above LSE development. To develop the relationship later, it will be convenient to re-express the 3D Green's

function in terms of its Fourier spectrum. To this end, the Weyl expansion [7] is used,

$$\frac{\mathrm{e}^{in_\mathrm{b}k_0 r}}{r} = \frac{i}{2\pi}\int dk_x dk_y \, \frac{\mathrm{e}^{i(k_x x + k_y y + k_z|z|)}}{k_z}, \tag{3.4}$$

where $r = \|\boldsymbol{r}\|$, $k_z = \sqrt{(n_\mathrm{b}k_0)^2 - k_x^2 - k_y^2}$, and k_x and k_y are coordinates in the Fourier space. Setting z to be the optical axis, denote $\hat{\mathcal{F}}_{xy}$ as the 2D Fourier transform operator in the lateral dimensions. From the Weyl expansion, the original LSE can be rewritten as a composition of 2D Fourier transforms as

$$\psi(\boldsymbol{r}) - \psi_0(\boldsymbol{r}) \quad = \frac{i}{2}\int dz' \, \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{ik_z|z-z'|}}{k_z} \, \beta(k_x, k_y, z')\right], \tag{3.5}$$

where the subscript xy is to emphasize that the Fourier transform is applied on lateral dimensions, and

$$\beta(k_x, k_y, z) = \hat{\mathcal{F}}_{xy}\left[V(\boldsymbol{r})\psi(\boldsymbol{r})\right]. \tag{3.6}$$

The full derivation is in Appendix B.1. Without much loss of generality, it can be assumed that ψ_0 is incident from $z = -\infty$ and the optical detectors are located outside the support of V. In addition, set z_0 as an arbitrary point on the optical axis between the illumination source and the scattering potential V. Fig. 3-1 depicts the overall geometry. Consequently, it is obtained that

$$\begin{aligned}
&\psi(\boldsymbol{r}) - \psi_0(\boldsymbol{r}) \\
&\quad = \int d\boldsymbol{r}' \, G(\boldsymbol{r} - \boldsymbol{r}')V(\boldsymbol{r}')\psi(\boldsymbol{r}') \\
&\quad = \int_{z_0}^{z} dz' \int dx'dy' \, G(\boldsymbol{r} - \boldsymbol{r}')V(\boldsymbol{r}')\psi(\boldsymbol{r}') \\
&\quad = \frac{i}{2}\int_{z_0}^{z} dz' \, \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{ik_z(z-z')}}{k_z}\beta(k_x, k_y, z')\right],
\end{aligned} \tag{3.7}$$

i.e. the 3D convolution with the Green's function becomes a cascade of 2D convolutions at each z-slice.

3.3 From LSE to Born series

To derive a connection between LSE and BPM, we are required to express the original Born series in terms of the cascade of 2D convolutions in Eq. (3.7). For this, Eq. (3.7) is slightly modified first. Following the small-wavelength approximation underlying the scalar Helmholtz equation or noting that the wavefront envelope of ψ_0 would be much larger than objects in many imaging systems, it may be assumed that $\psi_0 = \exp(in_\mathrm{b}k_0 z)$, *i.e.* a pure plane wave. Dividing both sides of Eq. (3.7) by ψ_0 leads to

$$\varphi(\boldsymbol{r}) = 1 + \frac{i}{2}\int_{z_0}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z}\gamma(k_x,k_y,z')\right], \tag{3.8}$$

where $\varphi = \psi/\psi_0$, $\overline{k}_z = k_z - n_\mathrm{b}k_0$, and

$$\gamma(k_x,k_y,z) = \hat{\mathcal{F}}_{xy}\left[V(\boldsymbol{r})\varphi(\boldsymbol{r})\right]. \tag{3.9}$$

From Eqs. (3.7) and (3.8), define an LSE integral operator $\widehat{\mathrm{GV}}_\alpha$ as

$$\begin{aligned}\widehat{\mathrm{GV}}_\alpha : \varphi &\to \frac{1}{\psi_0}\int_{\alpha}^{z} dz' \int dx'dy'\ G(\boldsymbol{r}-\boldsymbol{r}')V(\boldsymbol{r}')\psi(\boldsymbol{r}') \\ &= \frac{i}{2}\int_{\alpha}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z}\gamma(k_x,k_y,z')\right],\end{aligned} \tag{3.10}$$

e.g. $\varphi = 1 + \widehat{\mathrm{GV}}_{z_0}\varphi$. In addition, using that $\mathrm{e}^{i\overline{k}_z(z-z')} = 1$ at the origin of the Fourier space and setting $z_0 = -\infty$, Eq. (3.8) is converted to a more generalized form as

$$\begin{aligned}\varphi(\boldsymbol{r}) &= 1 + \frac{i}{2}\int_{-\infty}^{z_1} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z}\gamma(k_x,k_y,z')\right] \\ &\quad + \frac{i}{2}\int_{z_1}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z}\gamma(k_x,k_y,z')\right] \\ &= \hat{\mathcal{F}}_{xy}^{\dagger}\mathrm{e}^{i\overline{k}_z(z-z_1)}\hat{\mathcal{F}}_{xy}\left[1 + \frac{i}{2}\int_{-\infty}^{z_1} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z_1-z')}}{k_z}\gamma(k_x,k_y,z')\right]\right] \\ &\quad + \frac{i}{2}\int_{z_1}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger}\left[\frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z}\gamma(k_x,k_y,z')\right]\end{aligned} \tag{3.11}$$

$$= \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\bar{k}_z(z-z_1)} \hat{\mathcal{F}}_{xy} \varphi(x,y,z_1) + \frac{i}{2} \int_{z_1}^{z} dz' \; \hat{\mathcal{F}}_{xy}^{\dagger} \left[\frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \gamma(k_x,k_y,z') \right],$$

where $z_1 \leq z$ is a point on the optical axis.

Assuming that the operator norm of $\widehat{\mathrm{GV}}_{z_0}$ is less than 1, the solution of the Fredholm integral equation of the second kind, Eq. (3.8), can be described as a convergent geometric series (Born series or Liouville-Neumann series) [52]:

$$\varphi(\boldsymbol{r}) = \sum_{j=0}^{\infty} \left(\frac{i}{2} \right)^j f_j(\boldsymbol{r}), \tag{3.12}$$

where

$$f_0(\boldsymbol{r}) = \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\bar{k}_z(z-z_0)} \hat{\mathcal{F}}_{xy} \varphi(x,y,z_0) \tag{3.13a}$$

$$\begin{aligned} f_j(\boldsymbol{r}) &= \int_{z_0}^{z} dz' \; \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} \left[V(\boldsymbol{r}') f_{j-1}(\boldsymbol{r}') \right] \\ &= \frac{2}{i} \widehat{\mathrm{GV}}_{z_0} f_{j-1}. \end{aligned} \tag{3.13b}$$

This may be shown by substituting Eq. (3.12) into Eq. (3.11). That f_j represents the j-th order scattering term becomes obvious if Eq. 3.12 is rewritten as

$$\varphi(\boldsymbol{r}) = f_0(\boldsymbol{r}) + \widehat{\mathrm{GV}}_{z_0} f_0(\boldsymbol{r}) + \left(\widehat{\mathrm{GV}}_{z_0} \right)^2 f_0(\boldsymbol{r}) + \cdots, \tag{3.14}$$

using Eq. (3.13). Eqs. (3.12) and (3.13) are the core connection between LSE and BPM that will be established in the next section.

3.3.1 Convergence of the Born series

Before discussing the BPM, briefly take a pause to consider the validity of the Born series. Assuming that solutions of the LSE are continuous, the convergence of the Born series can be shown in a few different ways, e.g. using the Banach-Keissinger theorem [69], again given that the operator norm of $\widehat{\mathrm{GV}}_{z_0}$ is less than 1. Otherwise, the convergence of the series cannot be guaranteed and due to the divergent behavior

of $\left(\widehat{\mathrm{GV}}_{z_0}\right)^j$ as $n \gg n_\mathrm{b}$ and $j \to \infty$ it would be difficult to obtain the error bound between the series expansion and the true solution of the LSE. Hence, it is important to estimate the dependency of the operator norm on V. In other words, it is necessary to estimate conditions on V that make the operator norm of $\widehat{\mathrm{GV}}_{z_0}$ less than 1 in some domain. In numerical computations, the evaluation of $\varphi(\boldsymbol{r})$ is usually treated in a bounded subset $\mathcal{D}$ of $\mathbb{R}^3$, e.g. a box

$$\mathcal{D} = \left[-\frac{L_1}{2}, \frac{L_1}{2}\right] \times \left[-\frac{L_2}{2}, \frac{L_2}{2}\right] \times \left[-\frac{L_3}{2}, \frac{L_3}{2}\right], \tag{3.15}$$

which contains the support of V. We now evaluate the operator norm in $\mathcal{D}$.

From the definition of $\widehat{\mathrm{GV}}_{z_0}$, Eq. (3.10),

$$\left\|\widehat{\mathrm{GV}}_\alpha \varphi\right\| \leq \left\|\hat{\mathrm{G}}\right\| \|\varphi\| \sup_{\mathcal{D}}(V) \tag{3.16}$$

where $\left\|\hat{\mathrm{G}}\right\|$ is the operator norm of

$$\hat{\mathrm{G}}: \ \varphi \to \int_{\mathcal{D}} d\boldsymbol{r}' \ G(\boldsymbol{r} - \boldsymbol{r}')\varphi(\boldsymbol{r}'). \tag{3.17}$$

It is difficult to get an analytical expression for $\left\|\hat{\mathrm{G}}\right\|$, particularly due to the singularity of G at the origin. Instead, [76] suggests using a numerical method, which is a crude approximation on the true norm. To achieve a more analytical approach, it is possible to first try to remove the singularity using the discussion in [102]. It can be easily shown that

$$\widehat{\mathrm{GV}}_{z_0}\varphi = \frac{1}{\psi_0} \int_{\mathcal{D}} d\boldsymbol{r}' \ G(\boldsymbol{r} - \boldsymbol{r}') \operatorname{rect}\left(\frac{\|\boldsymbol{r} - \boldsymbol{r}'\|}{2L_M}\right) V(\boldsymbol{r}')\psi(\boldsymbol{r}'), \tag{3.18}$$

where L_M is the diagonal length of the smallest box containing the support of V, e.g. $\sqrt{L_1^2 + L_2^2 + L_3^2}$. Then $\left\|\hat{\mathrm{G}}\right\|$ becomes the norm of a convolution with a new kernel,

$$\bar{G}(\boldsymbol{r}) = G(\boldsymbol{r}) \operatorname{rect}\left(\frac{\|\boldsymbol{r}\|}{2L_M}\right), \tag{3.19}$$

whose Fourier transform is entire by virtue of the Paley-Wiener theorem:

$$\hat{\mathcal{F}}\bar{G}(\boldsymbol{r})(k) = \frac{1}{k}\frac{1}{(n_{\mathrm{b}}k_0 - k)(n_{\mathrm{b}}k_0 + k)}\left[\mathrm{e}^{in_{\mathrm{b}}k_0 L_M}(k\cos kL_M - in_{\mathrm{b}}k_0 \sin kL_M) - k\right]. \tag{3.20}$$

Since the Fourier transform is unitary, $\left\|\hat{\mathrm{G}}\right\|$ would be bound by the largest Fourier coefficient of $\bar{G}(\boldsymbol{r})$. Under the small wavelength approximation on which the scalar Helmholtz equation is based, $n_{\mathrm{b}}k_0 L_M \gg 1$ and subsequently, the absolute value of $\hat{\mathcal{F}}\bar{G}(\boldsymbol{r})(k)$ has two peaks at $k = n_{\mathrm{b}}k_0$ (from surface of momentum conservation) and $k = 0$ (from regularization of the singularity), which asymptotically approach $\frac{L_M}{n_{\mathrm{b}}k_0}$ and $\frac{L_M}{2n_{\mathrm{b}}k_0}$, respectively. Therefore,

$$\left\|\hat{\mathrm{G}}\right\| \leq \frac{L_M}{n_{\mathrm{b}}k_0}, \tag{3.21}$$

and subsequently,

$$\left\|\widehat{\mathrm{GV}}_{z_0}\right\| \leq \frac{L_M}{n_{\mathrm{b}}k_0}\sup_{\mathcal{D}}(V). \tag{3.22}$$

However, Eq. (3.22) would be too loose an estimate on the operator norm, *i.e.* the use of $\sup_{\mathcal{D}}(V)$ in Eq. (3.16). Hence it is suggested to alternatively use

$$\left\|\widehat{\mathrm{GV}}_{\alpha}\right\| \lesssim \frac{L_M}{n_{\mathrm{b}}k_0}\underset{\mathcal{D}}{\mathrm{mean}}(V) \tag{3.23}$$

as an approximation if the potential V is mostly smooth. Setting

$$V(\boldsymbol{r}) = (n_{\mathrm{b}}k_0)^2\left[\left(\frac{n(\boldsymbol{r})}{n_{\mathrm{b}}}\right)^2 - 1\right], \tag{3.24}$$

Eq. (3.23) can be rewritten as

$$\left\|\widehat{\mathrm{GV}}_{\alpha}\right\| \lesssim L_M n_{\mathrm{b}}k_0\left[\left(\frac{\underset{\mathcal{D}}{\mathrm{mean}}(n)}{n_{\mathrm{b}}}\right)^2 - 1\right]. \tag{3.25}$$

That is, roughly speaking, the validity of the Born series guarantee is inversely pro-

portional to the object scale with respect to the incident wavelength and the square of the refractive index. The estimation of the norm in Eq. (3.25) is tighter and simpler than previous reports e.g. [69, 57] as the size of optical objects becomes large. A detailed discussion is presented in Appendix B.2. The tightness of the bound also helps improve the truncation error estimate expressed as geometric series of the norm, e.g. [69],

$$\left\| \varphi - \sum_{j=0}^{N} \left(\widehat{\mathrm{GV}}_{z_0} \right)^j f_0 \right\| \leq \frac{\left\| \widehat{\mathrm{GV}}_{z_0} \right\|^{N+1}}{1 - \left\| \widehat{\mathrm{GV}}_{z_0} \right\|} \left\| f_0 \right\|. \tag{3.26}$$

3.4 From Born series to BPM

As discussed in the previous section, Eq. (3.13) plays a key role in connecting LSE and BPM. Further derivations on such connection begin with analyzing f_1, the first term in the Born series, representing a single scattering event:

$$f_1(\boldsymbol{r}) = \int_{z_0}^{z} dz' \; \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} \Big[V(\boldsymbol{r}') \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\bar{k}_z(z'-z_0)} \hat{\mathcal{F}}_{xy} \left[\varphi(x', y', z_0) \right] \Big]. \tag{3.27}$$

To derive the BPM, it is required that the two operators

$$\hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} \quad \text{and} \quad V(\boldsymbol{r})\times \tag{3.28}$$

commute. When it comes to their physical intuitions, these operators imply the propagation and change of momentum of photons, respectively. More specifically, note that the kernel $\mathrm{e}^{i\bar{k}_z(z-z')}/k_z$ originates from the Weyl expansion on the three-dimensional convolution with the Green's function. Hence, the commutation entails that original momentum of incoming photons does not change much, especially in terms of its direction.

Using the convolution theorem, it can be shown that

$$\hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} V(\boldsymbol{r}') = \frac{1}{(2\pi)^2} \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \left[\tilde{V}_{z'} \star \right] \hat{\mathcal{F}}_{xy}, \tag{3.29}$$

where $\tilde{V}_{z'\star}$ is a convolution operator:

$$\tilde{V}_{z'\star}: \ \varphi(\boldsymbol{k}) \to \int d\boldsymbol{k}' \hat{\mathcal{F}}_{xy}\left[V(x,y,z)\right](\boldsymbol{k}-\boldsymbol{k}')\varphi(\boldsymbol{k}'). \tag{3.30}$$

Here, assume that V is band-limited in each of its xy-slices. For brevity, first define the boxcar function in $\mathbb{R}^2$ as

$$\mathrm{rect}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \|\boldsymbol{x}\| > \frac{1}{2} \\ 1, & \text{otherwise,} \end{cases} \tag{3.31}$$

and approximate $\hat{\mathcal{F}}_{xy}\psi$ and $\tilde{V}_{z'}$ as

$$\hat{\mathcal{F}}_{xy}\varphi \approx C_\varphi \,\mathrm{rect}\left(\frac{\boldsymbol{k}}{2K_\varphi}\right) \tag{3.32a}$$

$$\tilde{V}_{z'} \approx C_V \,\mathrm{rect}\left(\frac{\boldsymbol{k}}{2K_V}\right), \tag{3.32b}$$

i.e. their support is confined to spheres of size K_φ and K_V, respectively, while C_φ and C_V are upper bounds on the approximate operator amplitudes. It follows that

$$\frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z}\left[\tilde{V}_{z'\star}\right]\hat{\mathcal{F}}_{xy}\varphi \approx C_\varphi C_V \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z}(\pi K_V^2)\,\mathrm{rect}\left(\frac{\boldsymbol{k}}{2(K_V+K_\varphi)}\right). \tag{3.33}$$

On the other hand,

$$\begin{aligned}&\left[\tilde{V}_{z'\star}\right]\frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z}\hat{\mathcal{F}}_{xy}\varphi \\ &\approx C_\varphi C_V \,\mathrm{rect}\left(\frac{\boldsymbol{k}}{2(K_V+K_\varphi)}\right)\left[\mathrm{e}^{-in_\mathrm{b}k_0(z-z')}\int_{B_{K_V}(\boldsymbol{k})} d\boldsymbol{k}' \,\frac{\mathrm{e}^{i(z-z')\sqrt{(n_\mathrm{b}k_0)^2-(k_x')^2-(k_y')^2}}}{\sqrt{(n_\mathrm{b}k_0)^2-(k_x')^2-(k_y')^2}}\right],\end{aligned} \tag{3.34}$$

where $B_{K_V}(\boldsymbol{k})$ is a ball of radius K_V centered at $\boldsymbol{k}$. Comparing Eqs. (3.33) and (3.34),

the two operators in Eq. (3.28) would commute if

$$\pi K_V^2 \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \approx \mathrm{e}^{-in_\mathrm{b}k_0(z-z')} \times \int_{B_{K_V}(\boldsymbol{k})} d\boldsymbol{k}' \, \frac{\mathrm{e}^{i(z-z')\sqrt{(n_\mathrm{b}k_0)^2-(k'_x)^2-(k'_y)^2}}}{\sqrt{(n_\mathrm{b}k_0)^2-(k'_x)^2-(k'_y)^2}}, \tag{3.35}$$

i.e. if the propagator (2D Fourier spectrum of the Green's function) is nearly constant in $B_{K_V}(\boldsymbol{k})$ for every $\boldsymbol{k}$ in $B_{K_\varphi+K_V}(\boldsymbol{0})$. This is consistent with the weak scattering approximation applied separately on each slice of the BPM. To satisfy condition (3.35), it is sufficient to require that

$$z - z' \text{ and } K_V \text{ are small.} \tag{3.36}$$

To further simplify the integrand in Eq. (3.35) toward obtaining an estimate of its validity bound, let us assume that $z-z'$ is sufficiently small so that the term $\mathrm{e}^{i\bar{k}_z(z-z')}$ can be considered locally constant in $B_{K_V}(\boldsymbol{k})$ and describe this term as a constant C_z. Then, at $\boldsymbol{k}=\boldsymbol{0}$,

$$\begin{aligned}
&\int_{B_{K_V}(\boldsymbol{0})} d\boldsymbol{k}' \, \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{\sqrt{(n_\mathrm{b}k_0)^2-(k'_x)^2-(k'_y)^2}} \\
&\qquad = \int_0^{2\pi} d\theta \int_0^{K_V} r dr \frac{C_z}{\sqrt{k^2-r^2}} \\
&\qquad = 2\pi C_z \left(n_\mathrm{b}k_0 - \sqrt{(n_\mathrm{b}k_0)^2 - K_V^2} \right),
\end{aligned} \tag{3.37}$$

and, subsequently,

$$\begin{aligned}
&\left| \pi K_V^2 \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} - \int_{B_{K_V}(\boldsymbol{0})} d\boldsymbol{k}' \, \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{\sqrt{(n_\mathrm{b}k_0)^2-(k'_x)^2-(k'_y)^2}} \right| \\
&\approx \pi C_z n_\mathrm{b} k_0 \left(2 - 2\sqrt{1-\mathcal{S}^2} - \mathcal{S}^2 \right),
\end{aligned} \tag{3.38}$$

where $\mathcal{S}$ is the dimensionless parameter

$$\mathcal{S} \equiv \frac{K_V}{n_\mathrm{b}k_0}. \tag{3.39}$$

The last term in Eq. (3.38) shall be referred to as

$$\delta_0 = 2 - 2\sqrt{1-\mathcal{S}^2} - \mathcal{S}^2 \approx \frac{\mathcal{S}^4}{2}. \tag{3.40}$$

The behavior of δ_0 *vs.* $\mathcal{S}$ is shown further down in Fig. 3-2 as part of a longer discussion on the BPM's validity. The approximation applies for $\mathcal{S} \ll 1$. As previously mentioned, $\mathcal{S}$ regulates the magnitude of the commutation error, illustrating the impact of the optical potential on the propagation of incoming photons. In other words, $\mathcal{S}$ refers to the scattering angle, which represents the degree of scattering from the original path. Note that K_V and k_0 are reciprocals of the size of objects and the wavelength, respectively; hence, $\mathcal{S}$ is closely connected to

$$\frac{\lambda}{L_{xy}}, \tag{3.41}$$

where L_{xy} stands for the size of objects in lateral dimensions. The equation refers to the angular resolution (diffraction limit). In this sense, $\mathcal{S}$ can be viewed as the degree of diffraction under an aperture, or an object in a general setting. The connection between $\mathcal{S}$, scattering angle, and the diffraction may become more clear in extremities. For example, $\mathcal{S} = 0$ in free space, leading to the angular spectrum method (Rayleigh-Sommerfeld diffraction) in BPM, i.e. Eq. (3.47). Incoming photons just propagate as they are without any perturbation. On the other hand, $\mathcal{S} = 1$ when the size of an object, or an aperture, matches the wavelength, which suffers from the largest amount of scattering that the scalar scattering theory can handle. In principle, $\mathcal{S}$ can be larger than 1 in Rayleigh scattering, but then scattering may no longer be described adequately by the scalar theory; instead, polarization and evanescent wave contributions must be taken into account.

From Eqs. (3.33) and (3.34), Eq. (3.38) corresponds to the error of the commutation at $\boldsymbol{k} = \mathbf{0}$ (more precisely, the error normalized by C_φ and C_V that are average amplitudes of φ and V in the Fourier space). When $\boldsymbol{k} \neq \mathbf{0}$ it is not straightforward to derive an analytical expression for the error, but it can be anticipated that it would

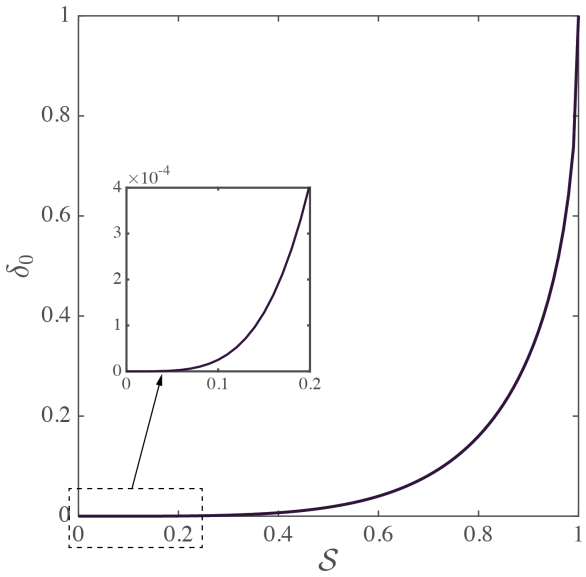

Figure 3-2: Dependence of δ_0 on $\mathcal{S}$. As $\mathcal{S}$ increases, δ_0 approaches its maximum value, 1. This implies that the Fourier transform of V has significant effects on the validity of the BPM.

be proportional to $\|\boldsymbol{k}\|$. This is because $1/\sqrt{(n_\mathrm{b}k_0)^2-(k'_x)^2-(k'_y)^2}$ in Eq. (3.34) changes rapidly as the domain of integral, $B_{K_V}(\boldsymbol{k})$, moves away from the origin in the Fourier space. Hence,

$$\begin{aligned}&\left|\frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z}\left[\tilde{V}_{z'}\star\right]\hat{\mathcal{F}}_{xy}\varphi-\left[\tilde{V}_{z'}\star\right]\frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z}\hat{\mathcal{F}}_{xy}\varphi\right|\\&\approx\underbrace{\pi C_\varphi C_z C_V n_\mathrm{b}k_0\delta_0}_{\varepsilon_0}+\varepsilon\left(K_V,K_\varphi\right),\end{aligned}\tag{3.42}$$

where ε represents the additional error originating from $\boldsymbol{k}\neq\boldsymbol{0}$ regions, which depends on the effective support of both V and φ in the Fourier space and increases more rapidly than ε_0.

From now on, assume that Eq. (3.36) is satisfied in our system. Then, Eq. (3.27) becomes

$$\begin{aligned}f_1(\boldsymbol{r})&=\int_{z_0}^{z}dz'\;V(x,y,z')\hat{\mathcal{F}}^\dagger_{xy}\frac{\mathrm{e}^{i\bar{k}_z(z-z_0)}}{k_z}\hat{\mathcal{F}}_{xy}\left[\varphi(x,y,z_0)\right]\\&=\left\{\int_{z_0}^{z}dz'\;V(x,y,z')\right\}\hat{\mathcal{F}}^\dagger_{xy}\frac{\mathrm{e}^{i\bar{k}_z(z-z_0)}}{k_z}\hat{\mathcal{F}}_{xy}\left[\varphi(x,y,z_0)\right].\end{aligned}\tag{3.43}$$

Subsequently, evaluating f_2 yields

$$\begin{aligned} f_2(\boldsymbol{r}) &= \int_{z_0}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\overline{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} \left[V(\boldsymbol{r}') \left\{ \int_{z_0}^{z'} dz''\ V(\boldsymbol{r}'') \right\} \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{ik_z(z'-z_0)}}{k_z} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right] \right] \\ &= \left\{ \int_{z_0}^{z} dz'\ V(x,y,z') \int_{z_0}^{z'} dz''\ V(x,y,z'') \right\} \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\overline{k}_z(z-z_0)}}{k_z^2} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right] \\ &= \frac{1}{2!} \left\{ \int_{z_0}^{z} dz'\ V(x,y,z') \right\}^2 \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\overline{k}_z(z-z_0)}}{k_z^2} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right], \end{aligned} \tag{3.44}$$

where the last equality is derived using integration-by-parts [52]. Repeating the same procedure, it can be deduced that

$$f_j(\boldsymbol{r}) = \frac{1}{j!} \left\{ \int_{z_0}^{z} dz'\ V(x,y,z') \right\}^j \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\overline{k}_z(z-z_0)}}{k_z^j} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right]. \tag{3.45}$$

From the analysis on the commutation error, BPM requires K_φ and K_V to be small. Hence, $|k_x|, |k_y| \ll n_\mathrm{b} k_0$ and $k_z \approx n_\mathrm{b} k_0$. Subsequently,

$$f_j(\boldsymbol{r}) \approx \frac{1}{j!} \left\{ \int_{z_0}^{z} dz'\ V(x,y,z') \right\}^j \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\overline{k}_z(z-z_0)}}{(n_\mathrm{b} k_0)^j} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right]. \tag{3.46}$$

Inserting Eq. (3.46) to Eq. (3.12) gives

$$\begin{aligned} \varphi(\boldsymbol{r}) &= \exp\left(\frac{i}{2n_\mathrm{b} k_0} \left\{ \int_{z_0}^{z} dz'\ V(x,y,z') \right\} \right) \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\overline{k}_z(z-z_0)} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right] \\ &= \exp\left(\frac{i n_\mathrm{b} k_0}{2} \left\{ \int_{z_0}^{z} dz'\ \left[\left(\frac{n(x,y,z')}{n_\mathrm{b}} \right)^2 - 1 \right] \right\} \right) \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\overline{k}_z(z-z_0)} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right] \\ &\approx \exp\left(\frac{i n_\mathrm{b} k_0}{\xi} (z-z_0) \left[\left(\frac{n(x,y,z)}{n_\mathrm{b}} \right)^{\xi} - 1 \right] \right) \hat{\mathcal{F}}_{xy}^{\dagger} \mathrm{e}^{i\overline{k}_z(z-z_0)} \hat{\mathcal{F}}_{xy} \left[\varphi(x,y,z_0) \right], \end{aligned} \tag{3.47}$$

where $\xi = 2$. Comparing Eqs. (3.13) and (3.46), it is implied that the j-th order scattering term in Born series corresponds to the j-th order polynomial in the Taylor expansion of the exponential modulation in the BPM. This successive application of the diffraction and phase modulation is also reminiscent of a similar result derived according to the multislice method [41, 59].

3.4.1 Difference between Born series and BPM

Though Born series and BPM both originate from the LSE and their mathematical structures are closely related, BPM imposes different assumptions on the scattering process. First, due to Eq. (3.36), it is required that $|z - z_0|$ be small. Hence, previous studies on BPM suggest slicing a thick V along the optical axis and applying BPM on each slice consecutively. However, this violates our assumption that z is outside of the support of V, as in Fig. 3-1. In other words, at each j^{th} slice inside V, BPM has a numerical discrepancy

$$\frac{i}{2}\int_{z_0}^{z} dz' \; \hat{\mathcal{F}}_{xy}^{\dagger} \frac{\mathrm{e}^{i\bar{k}_z(z-z')}}{k_z} \hat{\mathcal{F}}_{xy} \Big[V(\boldsymbol{r}') \big[\varphi_{j-1} - \varphi\big](\boldsymbol{r}') \Big], \tag{3.48}$$

where φ_{j-1} is a field at the $(j-1)^{\text{th}}$ slice in BPM and φ is that of LSE. The difference $\varphi_{j-1} - \varphi$ would approximately amount to backscattered fields from $V(x, y, z)$ where $z \geq z_j$ and z_j is the z-coordinate of the j^{th} slice. There have been studies on including backscattering effects in BPM [59], which may be an extension to the present work.

Despite Eq. (3.47) suggesting a close connection between Born series and BPM, they do exhibit different numerical convergence. Specifically, BPM is known to be numerically stable with high V, compared to the Born series. One may be able to speculate that such behavior can be attributed to the following conditions. First, in BPM, it is assumed that K_φ and K_V are small, which makes $1/k_z$ as small as possible in the expansion. In other words, all Fourier coefficients that are multiplied with large $1/k_z$ are effectively ignored, and that promotes convergence. Second, as in Eq. (3.48), BPM does not consider backscattered fields. This would decrease the norm of the LSE operator. We present numerical experiments on comparing the convergence behavior of Born series and BPM in Section 3.5.

3.4.2 On the appearance of a different value of ξ in BPM's wave modulation term

According to Eq. (3.47), BPM consists of two operations. First, an incident field is propagated with small distance $z-z_0$. Subsequently, the field undergoes a phase modulation. The modulation is proportional to $(n/n_\mathrm{b})^\xi/\xi$ where $\xi = 2$. This resembles BPM in previous studies except they suggest $\xi = 1$ [79, 37].

The difference in the assumed values of ξ originates from the respective assumptions. To track the differences, let us again start with the Helmholtz equation Eq. (3.1), rewritten here for convenience as

$$\left[\frac{\partial^2}{\partial^2 z} + \nabla_{xy}^2 + k_0^2 n^2\right] \psi = 0, \tag{3.49}$$

where ∇_{xy} refers to the gradient in the lateral dimensions. Setting $\hat{P}^2 = \frac{\partial}{\partial z}$ and $\hat{Q}^2 = \nabla_{xy}^2 + k_0^2 n^2$, the equation can be further simplified as

$$\left[(\hat{P} + i\hat{Q})(\hat{P} - i\hat{Q}) + i\left\langle P, Q\right\rangle\right] \psi = 0, \tag{3.50}$$

where $\langle , \rangle$ is the commutator. If the variation of n along the optical axis is negligible, then $\langle P, Q\rangle \to 0$ [37], which requires

$$\left[\hat{P} - i\hat{Q}\right] \psi = 0. \tag{3.51}$$

In fact, there is another set of solutions from $\left[\hat{P} + i\hat{Q}\right] \psi = 0$, but this represents fields propagating backwards [97]. Consequently, from Eq. (3.51), ψ can be expressed as

$$\psi(x, y, z) = \exp\left[i(z - z_0)\left(\nabla_{xy}^2 + k_0^2 n^2\right)^{1/2}\right] \psi(x, y, z_0). \tag{3.52}$$

Note that $n = 1$ leads to the propagation in free space, as shown in [97]. To derive the BPM, it is required to separate ∇_{xy}^2 from n^2 in the square root. A straightforward

way to separate them is to use the Taylor expansion:

$$\begin{aligned}\left(\nabla_{xy}^2 + k_0^2 n^2\right)^{1/2} &= k_0 \left(1 + \frac{1}{k_0^2}\nabla_{xy}^2 + (n^2 - 1)\right)^{1/2} \\ &\approx k_0 + \frac{1}{2k_0}\nabla_{xy}^2 + \frac{k_0}{2}(n^2 - 1). \end{aligned} \tag{3.53}$$

Eq. (3.53) would be satisfied if $\left\| \frac{1}{k_0^2}\nabla_{xy}^2 + (n^2 - 1) \right\|$ is small, i.e. both the scattering angle and the lateral variation of n are small [98]. Eq. (3.53) corresponds to the phase modulation with $\xi = 2$, which uses the same assumptions on fields leading to the derivation of Eq. (3.47). On the other hand, [36, 37] suggest that

$$\left(\nabla_{xy}^2 + k_0^2 n^2\right)^{1/2} \approx (\nabla_{xy}^2 + k_0^2)^{1/2} + k_0(n - 1), \tag{3.54}$$

which can be justified if the lateral variation of n is small. This corresponds to the phase modulation with $\xi = 1$.

Summarizing, Eqs. (3.53) for $\xi = 2$ and (3.54) for $\xi = 1$ require different assumptions. The former requires both $\nabla_{xy}^2\psi$ and $\nabla_{xy}^2 n$ to be small; whereas the latter does not need the small scattering angle condition. However, the small lateral variation of n indirectly implies that the scattering angle of ψ in the potentials also needs to be small. Hence, it is expected that the $\xi = 1$ modulation would not result in significant difference over the $\xi = 2$ modulation, especially when $\mathcal{S}$ is small. This was confirmed empirically by our numerical observations. Explicitly, the effect of ξ on spherical potentials is presented in Appendix B.3.

3.4.3 Validity of the BPM

Eqs. (3.36) and (3.42) imply that the BPM approaches the LSE as K_V, the upper bound of diffraction away from the optical axis, becomes smaller. Hence, the difference between BPM and LSE would also depend on K_V and $\mathcal{S}$. Since, again, the exact evaluation of such difference can be difficult, here we devise some simplifying

approximations that also lend some insight to the problem. From Eq. (3.32b),

$$\begin{aligned} V_z(\boldsymbol{x}) &\approx C_V K_V^2 \operatorname{sinc}\left(2K_V\|\boldsymbol{x}\|\right) \\ &\approx (k_0 n_\mathrm{b})^2 \left(\frac{n_z(\boldsymbol{x})}{n_\mathrm{b}}\right)^2 \end{aligned} \tag{3.55}$$

where the subscript z is used to represent a z-slice. In other words, V is a function whose amplitude is $(k_0 n_z)^2$ and effective support is K_V^{-1}. Assuming that the gradient of n_z in the xy plane is negligible, it may be derived that

$$\varepsilon_0 \approx C_\varphi C_z \mathcal{S}^{-2} \left(\frac{n_z}{n_\mathrm{b}}\right)^2 (n_\mathrm{b} k_0)\delta_0. \tag{3.56}$$

This is the commutation error at $\boldsymbol{k} = \boldsymbol{0}$ in Eq. (3.42). If $\mathcal{S}$ is sufficiently small, Eq. (3.40) gives

$$\varepsilon_0 \approx C_\varphi C_z \left(\frac{n_z}{n_\mathrm{b}}\right)^2 (n_\mathrm{b} k_0)\, \mathcal{S}^2. \tag{3.57}$$

Neglecting the diffraction effect between z and z_0, the commutation error in the first order scattering term, Eq. (3.27), becomes

$$\begin{aligned} \varepsilon_{z,z_0} &= \int_{z_0}^{z} dz'\ \hat{\mathcal{F}}_{xy}^{\dagger} \left[C_\varphi C_z \left(\frac{n_{z'}}{n_\mathrm{b}}\right)^2 (n_\mathrm{b} k_0)\mathcal{S}^2 + \varepsilon\right] \\ &\approx (z - z_0)\hat{\mathcal{F}}_{xy}^{\dagger} \left[C_\varphi C_z \left(\frac{n_{z_0}}{n_\mathrm{b}}\right)^2 (n_\mathrm{b} k_0)\mathcal{S}^2 + \varepsilon\right], \end{aligned} \tag{3.58}$$

where the subscripts in ε_{z,z_0} are used to emphasize that now we consider the total commutation error from a potential slice. If we approximate ε as a function whose amplitude is ε_0 and effective support is mostly governed by φ, then Eq. (3.58) finally becomes

$$\varepsilon_{z,z_0} \approx C(z - z_0) \left(\frac{n_{z_0}}{n_\mathrm{b}}\right)^2 (n_\mathrm{b} k_0)\, \mathcal{S}^2, \tag{3.59}$$

where C is a dimensionless number that is almost independent of the system configuration. In addition, since it is required that $\mathrm{e}^{i\bar{k}_z(z-z_0)}$ is nearly constant in the derivation of BPM, $n_\mathrm{b} k_0 (z - z_0)$ can be regarded as another dimensionless number

that is independent of the system configuration. Subsequently, one can further simplify ε_{z,z_0} as

$$\varepsilon_{z,z_0} \approx C \left(\frac{n_{z_0}}{n_\mathrm{b}} \right)^2 \mathcal{S}^2. \tag{3.60}$$

Using ε_{z,z_0}, the total commutation error, ε_t, in the first order scattering term from an entire potential can be expressed. Let us denote as $z_1, \cdots, z_N$ th locations of the z-slices along the optical axis. Then

$$\begin{aligned} \varepsilon_t &= \sum_{m=1}^{N} \varepsilon_{z_m, z_{m-1}} \\ &= C \sum_{m=1}^{N} \left(\frac{n_{z_{m-1}}}{n_\mathrm{b}} \right)^2 \mathcal{S}^2(z_{m-1}) \\ &\approx C(n_\mathrm{b} k_0) \int_{R_z/2}^{R_z/2} dz \ \left(\frac{n_z}{n_\mathrm{b}} \right)^2 \mathcal{S}^2(z) \end{aligned} \tag{3.61}$$

where the z dependency of $\mathcal{S}$ is due to K_V in $\mathcal{S}$, and that is approximately reciprocal to the size of the potential in the xy plane; whereas R_z is the size of the potential along the optical axis.

Eq. (3.61) implies that the error of BPM increases as the thickness of the potential increases and the lateral size of the potential decreases, which agrees with previous studies on optical scattering. What is important is that the effect of the lateral size is larger than that of the thickness. To be more specific, a case of Mie scattering can be considered where an incident planewave is scattered by a spherical potential of radius R_z with constant refractive index n. Then

$$K_V(z) \sim \frac{1}{\sqrt{R_z^2 - z^2}}, \quad z \in \left[-\frac{R_z}{2}, \frac{R_z}{2} \right], \tag{3.62}$$

which gives

$$\varepsilon_t \approx C \left(\frac{n}{n_\mathrm{b}} \right)^2 \frac{1}{n_\mathrm{b} k_0 R_z} \ln 3. \tag{3.63}$$

In other words, as the sphere becomes large with respect to the incident wavelength, the error decreases though the thickness of the potential grows. This is because the average error at each potential slice decreases more rapidly.

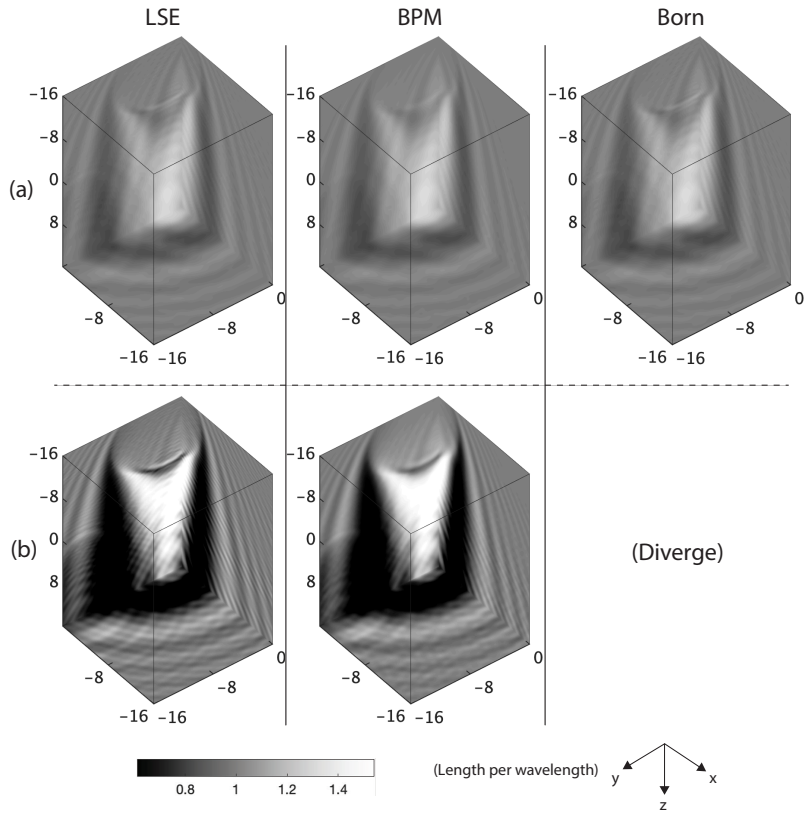

Figure 3-3: Comparison of scattered fields from LSE, BPM, and Born series. Two different dielectric spheres are considered where n is only changed to adjust the estimated norm of the LSE operator in Eq. (3.25). (a) The norm is 0.9. (b) The norm is 15.

Overall, Eq. (3.61) entails that BPM approximates the LSE if the magnitude of the refractive index n and the dimensionless parameter $\mathcal{S}$ are both small enough. Qualitatively, small $\mathcal{S}$ implies that the variation of n along the lateral direction should be small in the scale of the wavelength. In addition, Eq. (3.47) suggests that the variation of n should also be small along the optical axis. These ideas agree with previous studies [36, 37]. Due to the complex behavior of ε and the accumulation of commutation error in high order scattering terms in Eq. (3.46), the actual dependency of the difference between BPM and LSE may deviate from ε_t. Nevertheless, it can serve as a useful lower bound for the accuracy of the BPM.

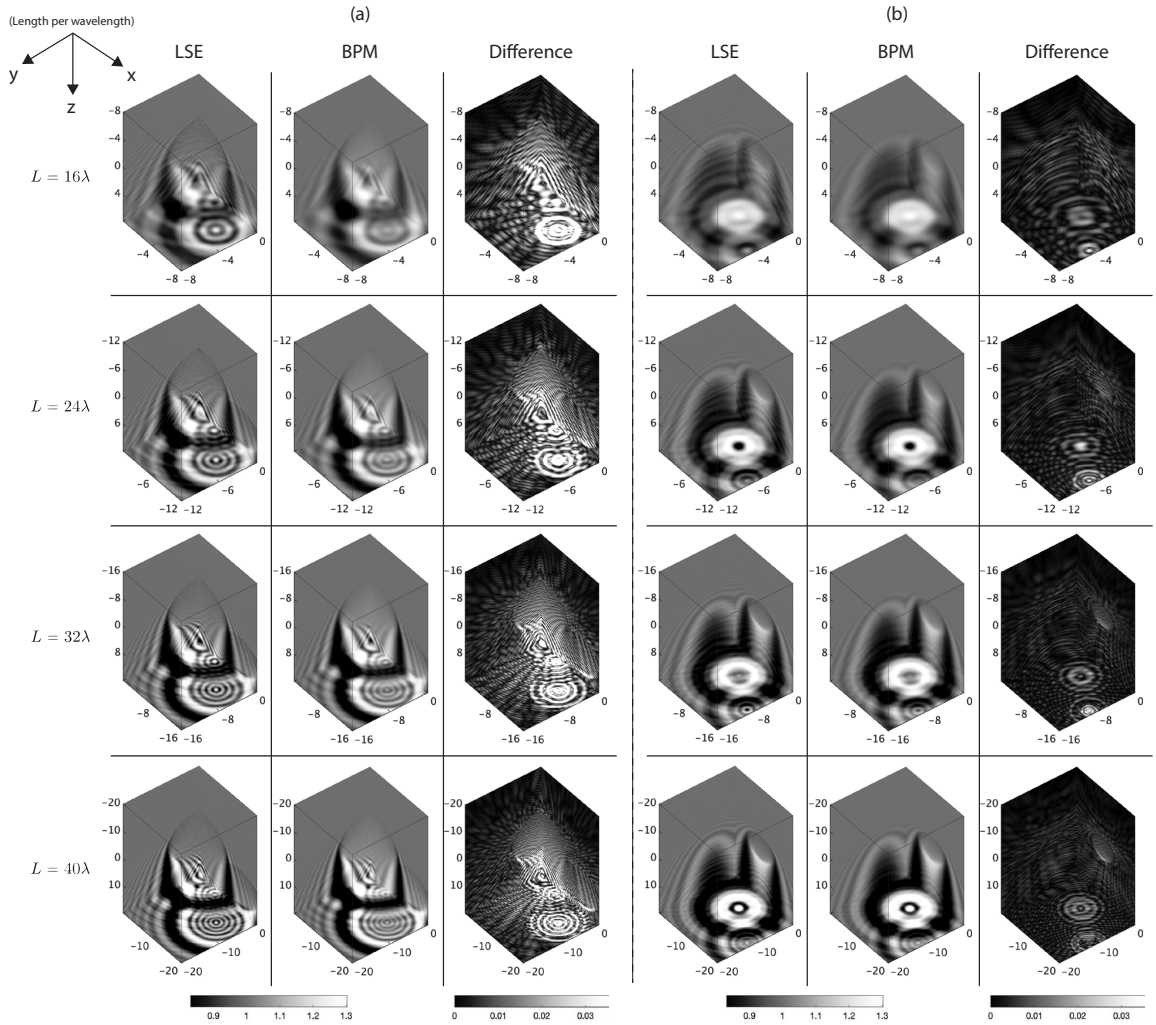

Figure 3-4: Scattered fields estimated from LSE and BPM when the size L of a cubic computational box changes. Here, two distinct potentials are considered, marked as (a) and (b), both consisting of dielectric spheres. The mean refractive index is 1.02. The difference refers to the elementwise absolute error divided by the maximum field amplitude.

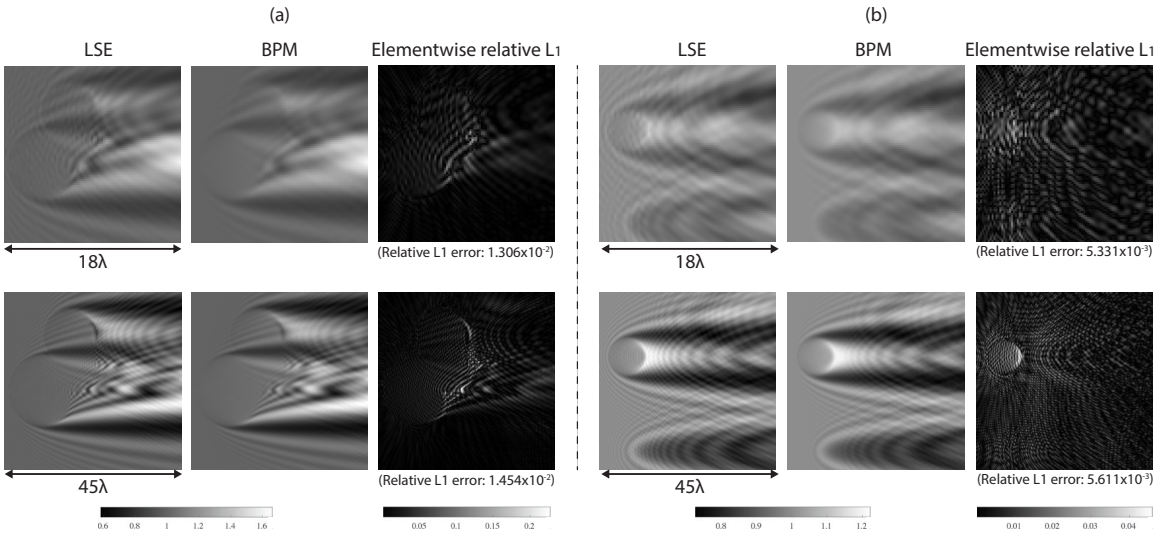

Figure 3-5: xz-view of scattered fields estimated from LSE and BPM for the objects as in Fig. 3-4, marked as (a) and (b).

3.5 Numerical discussion

In this section, numerical validations are conducted for previous discussions on LSE, Born series, and BPM. The Fourier transform in BPM is efficiently evaluated by using the fast Fourier transform (FFT). Similarly, the convolution integral with the Green's function in the Born series and the LSE is evaluated using the convolution theorem and FFT [102, 100]. We consider a uniform and cubic computational grid where 6 pixels per wavelength are used to discretize fields and refractive index functions. The LSE is solved using the QMRCGSTAB algorithm [19] until $\left\|\hat{A}\psi - \psi_0\right\|_2 / \|\psi_0\|_2$ reaches 10^{-5} where $\hat{A}\psi = \psi - \hat{G}(V\psi)$. Without much loss of generality, we set $n_b = 1$. Before proceeding further, we first demonstrate that LSE well approximates the finite-difference time-domain (FDTD) solutions in Appendix B.4.

In Section 3.4.1, we discuss the stronger convergence behavior of BPM compared to Born series. Mainly, this is because BPM neglects high $1/k_z$ portions in the field propagator, though both methods originate from the same polynomial series of f_j. Fig. 3-3 shows how scattered field estimations depend on the magnitude of n. As n increases, the upper bound of the operator norm of the LSE operator in Eq. (3.25) becomes high, which indicates the divergence of Born series. On the other hand,

BPM does not exhibit such divergence.

Subsequently, the difference between LSE and BPM is further investigated. Qualitatively speaking, it is controlled by the dimensionless parameter $\mathcal{S}$, which tells that large size and small refractive index induce small difference. In Fig. 3-4, it can seen that complex interference patterns near small objects are not well estimated in BPM. By measuring SSIM, PSNR, and the relative L_1 error (also referred to as MAE, mean absolute error), Table 3.1 additionally presents quantitative comparison between them. The quantitative metrics follow the same trend as the qualitative analysis, except the L_1 error in amplitude. This can be attributed to high frequency oscillations along the optical axis when ψ_0 is scattered by relatively large objects. For example, in Fig. 3-5, the good agreement between LSE and BPM is again presented as the size of potentials increases. At the same time, fine stripes of high relative L_1 errors appear, which originates from oscillatory patterns in amplitudes along the optical axis. Such patterns are numerically subtle to estimate accurately. On the other hand, Fig. 3-6 and Table 3.2 demonstrates strong reciprocity between the magnitude of the refractive index and the error between LSE and BPM, which agrees with our theoretical analysis.

Corroborating results in Fig. (3-4) and Table 3.1, conduct additional experiments are conducted regarding the size dependency of the error between LSE and BPM under a higher refractive index n condition. Specifically, $n = 1.08$ is considered. In Fig. (3-7), the expected tendency of BPM well approximates interference patterns of LSE as size increases, except at strong focal points. Table 3.3 lists corresponding quantitative results, which show decrease in SSIM and PSNR for the phase from large potentials. This may be attributed to the increased ill-conditionedness of the LSE operator [108] and fine oscillatory features, which reduces the numerical stability of the simulation.

In the aforementioned discussions, numerical results are mainly based on spherical objects. To further check the applicability of such discussions, tori are considered, which are topologically distinct to spheres while not containing edges that are sharp enough to destabilize numerical solvers [12]. In Fig. 3-8, it can be seen that high-

Table 3.1: Image quality metrics on fields from LSE and BPM when the size L of a cubic computational box changes. 15 different potentials are considered, which consist of dielectric spheres. The mean refractive index is 1.02. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$L = 16\lambda$, amplitude	0.948	37.996	6.683×10^{-3}
$L = 24\lambda$, amplitude	0.965	40.147	6.909×10^{-3}
$L = 32\lambda$, amplitude	0.974	41.732	7.127×10^{-3}
$L = 40\lambda$, amplitude	0.977	42.616	7.400×10^{-3}
$L = 16\lambda$, phase	0.991	38.067	4.101×10^{-2}
$L = 24\lambda$, phase	0.995	41.617	2.698×10^{-2}
$L = 32\lambda$, phase	0.997	44.126	2.010×10^{-2}
$L = 40\lambda$, phase	0.998	46.067	1.602×10^{-2}

frequency interference patterns start to appear in BPM as L increases. On the other hand, Fig. 3-10 depicts the difference between LSE and BPM as n changes. The agreement between LSE and BPM deteriorates in a high index setting and near the focal points, which implies that BPM may be inappropriate in this case. Table 3.4 and 3.5 provide quantitative analysis on the effects of L and n. Except for the relative L_1 error in amplitude seen in Fig. 3-9 and what is discussed already with reference to Fig. 3-5, the quantitative results are in a good agreement with the theoretical analysis.

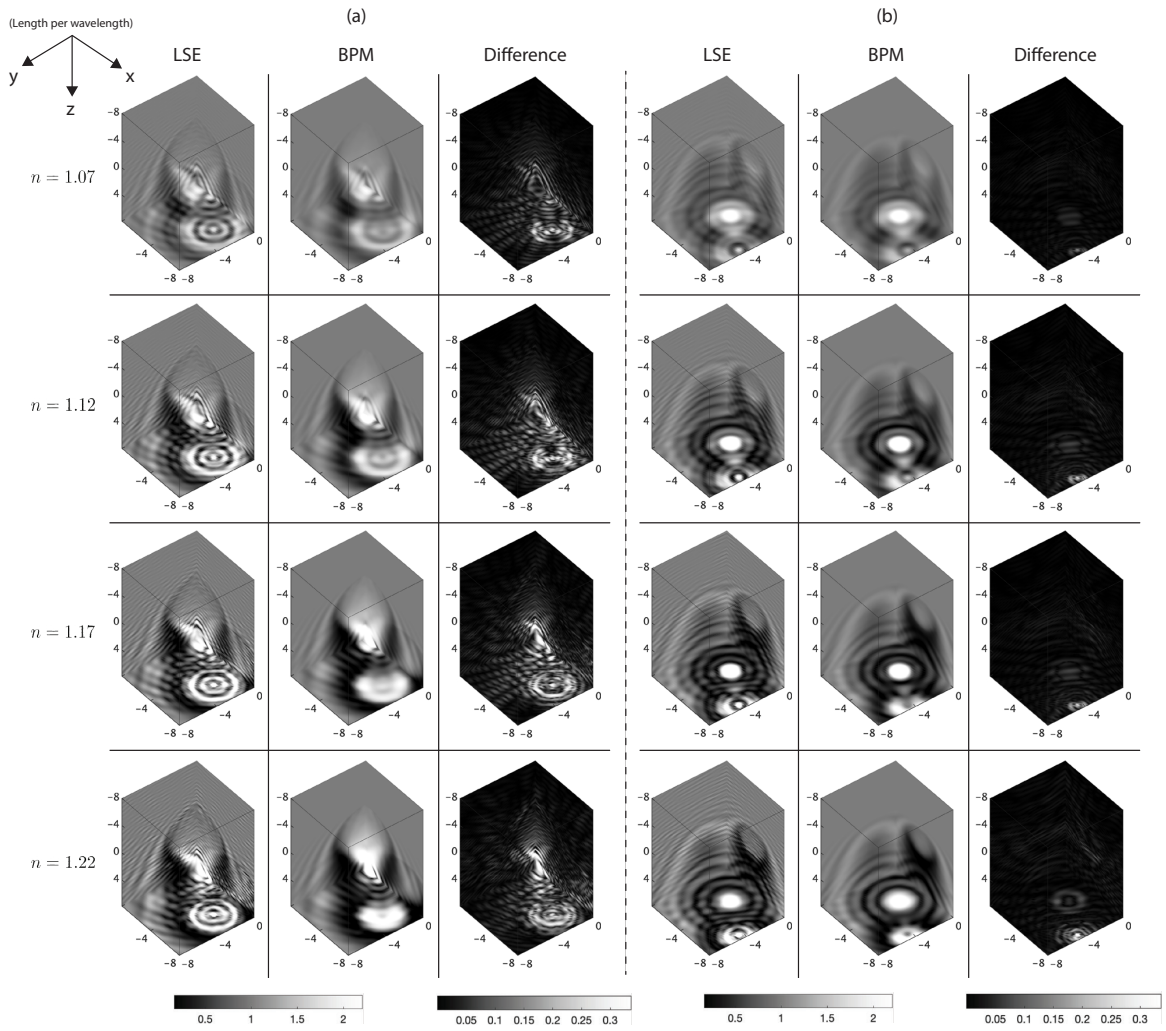

Figure 3-6: Scattered fields estimated from LSE and BPM when the mean refractive index n of spherical potentials changes. Here, potentials consist of spheres. The size of a cubic computational box is 16λ. We show two different objects, which are marked with (a) and (b). Difference refers to the elementwise absolute error divided by the maximum field amplitude.

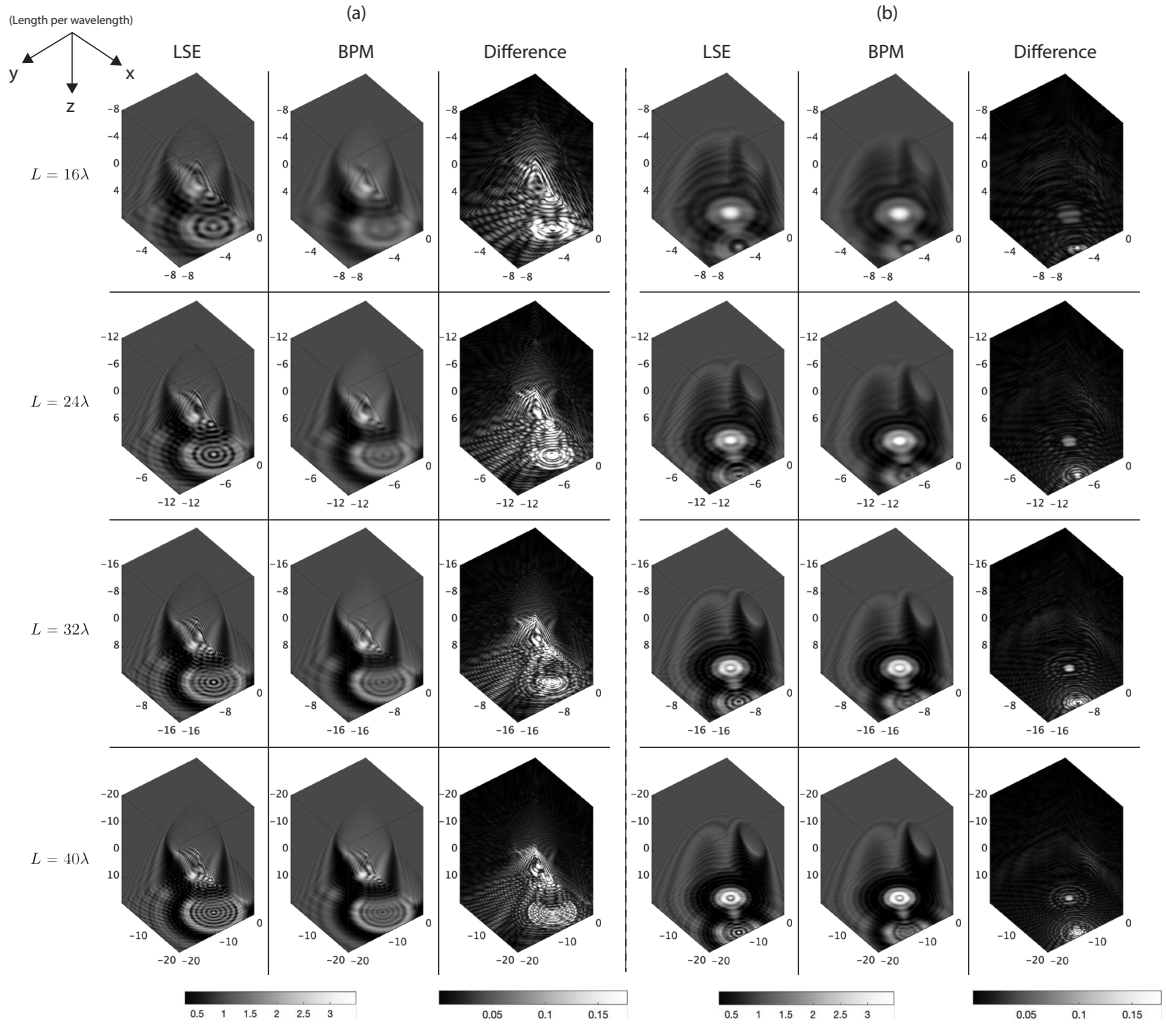

Figure 3-7: Scattered fields estimated from LSE and BPM when the size L of a cubic computational box changes. 15 different potentials are considered, which consist of spheres. The mean refractive index of spherical potentials is 1.08. We show two different objects, which are marked with (a) and (b). Difference refers to the elementwise absolute error divided by the maximum field amplitude.

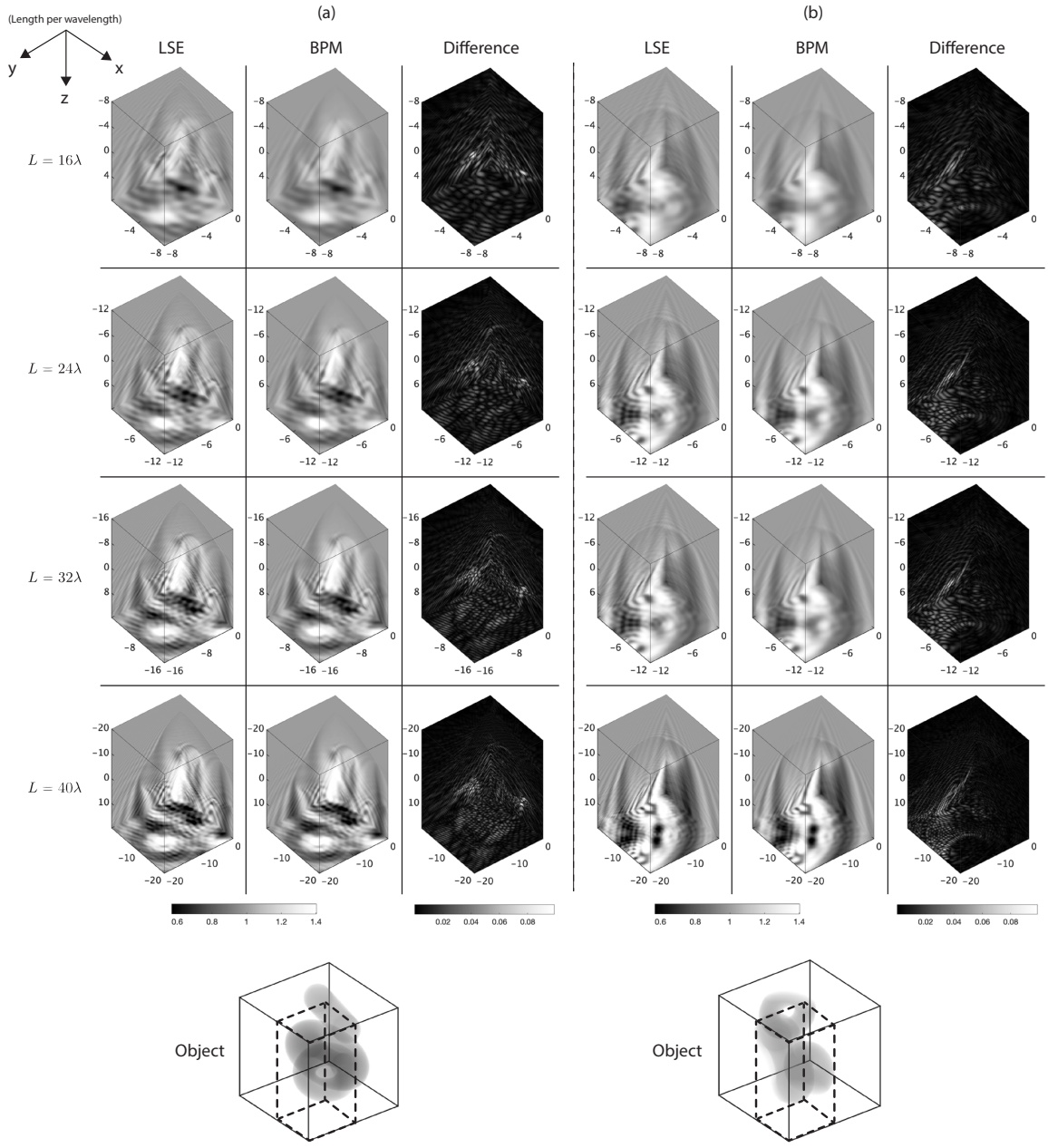

Figure 3-8: Scattered fields estimated from LSE and BPM when the size L of a cubic computational box changes. Two distinct potentials are considered, marked as (a) and (b), both consisting of dielectric tori. The mean refractive index is 1.02. The difference refers to the elementwise absolute error divided by the maximum field amplitude. For more visibility, we show the shape of objects where the scattered field boxes correspond to the regions enclosed with dotted black lines.

Table 3.2: Image quality metrics on fields from LSE and BPM when the mean refractive index n of spherical potentials changes. 15 different potentials are considered, which consist of dielectric spheres. The size of the cubic computational box is 16λ. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$n = 1.07$, amplitude	0.931	36.722	2.790×10^{-2}
$n = 1.12$, amplitude	0.888	34.429	6.291×10^{-2}
$n = 1.17$, amplitude	0.838	32.394	9.715×10^{-2}
$n = 1.22$, amplitude	0.812	31.137	12.076×10^{-2}
$n = 1.07$, phase	0.990	39.126	4.114×10^{-2}
$n = 1.12$, phase	0.971	36.198	4.339×10^{-2}
$n = 1.17$, phase	0.933	31.826	5.272×10^{-2}
$n = 1.22$, phase	0.910	29.105	6.042×10^{-2}

Table 3.3: Image quality metrics on fields from LSE and BPM when the size L of a cubic computational box changes. 15 different potentials are considered, which consist of spheres. The mean refractive index of spherical potentials is 1.08. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$L = 16\lambda$, amplitude	0.923	36.243	3.392×10^{-2}
$L = 24\lambda$, amplitude	0.930	37.200	4.170×10^{-2}
$L = 32\lambda$, amplitude	0.932	37.691	4.932×10^{-2}
$L = 40\lambda$, amplitude	0.937	38.330	5.528×10^{-2}
$L = 16\lambda$, phase	0.989	38.068	4.121×10^{-2}
$L = 24\lambda$, phase	0.990	40.656	2.769×10^{-2}
$L = 32\lambda$, phase	0.986	41.105	2.220×10^{-2}
$L = 40\lambda$, phase	0.983	40.212	1.938×10^{-2}

Table 3.4: Image quality metrics on fields from LSE and BPM when the size L of a cubic computational box changes. This table considers 15 different potentials which consist of dielectric tori. The mean refractive index is 1.02. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$L = 16\lambda$, amplitude	0.947	39.372	7.497×10^{-3}
$L = 24\lambda$, amplitude	0.965	41.841	7.795×10^{-3}
$L = 32\lambda$, amplitude	0.972	42.954	8.147×10^{-3}
$L = 40\lambda$, amplitude	0.975	43.626	8.514×10^{-3}
$L = 16\lambda$, phase	0.991	38.066	4.102×10^{-2}
$L = 24\lambda$, phase	0.995	41.622	2.699×10^{-2}
$L = 32\lambda$, phase	0.997	44.137	2.011×10^{-2}
$L = 40\lambda$, phase	0.998	46.081	1.603×10^{-2}

Table 3.5: Image quality metrics on fields from LSE and BPM when the mean refractive index n of spherical potentials changes. This table considers 15 different potentials which consist of dielectric tori. The size of the cubic computational box is 16λ. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$n = 1.07$, amplitude	0.915	36.019	3.534×10^{-2}
$n = 1.12$, amplitude	0.835	32.101	8.188×10^{-2}
$n = 1.17$, amplitude	0.740	29.499	13.925×10^{-2}
$n = 1.22$, amplitude	0.657	27.537	17.640×10^{-2}
$n = 1.07$, phase	0.986	37.707	4.143×10^{-2}
$n = 1.12$, phase	0.951	33.572	4.681×10^{-2}
$n = 1.17$, phase	0.899	28.539	5.971×10^{-2}
$n = 1.22$, phase	0.852	25.029	8.073×10^{-2}

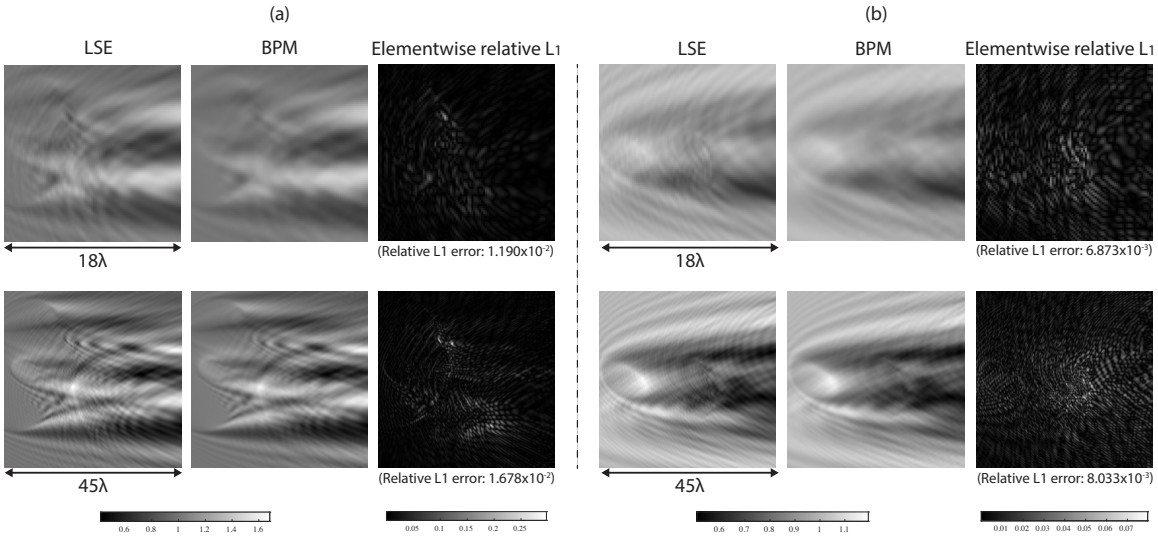

Figure 3-9: xz-view of scattered fields estimated from LSE and BPM for the objects as in Fig. 3-8, marked as (a) and (b).

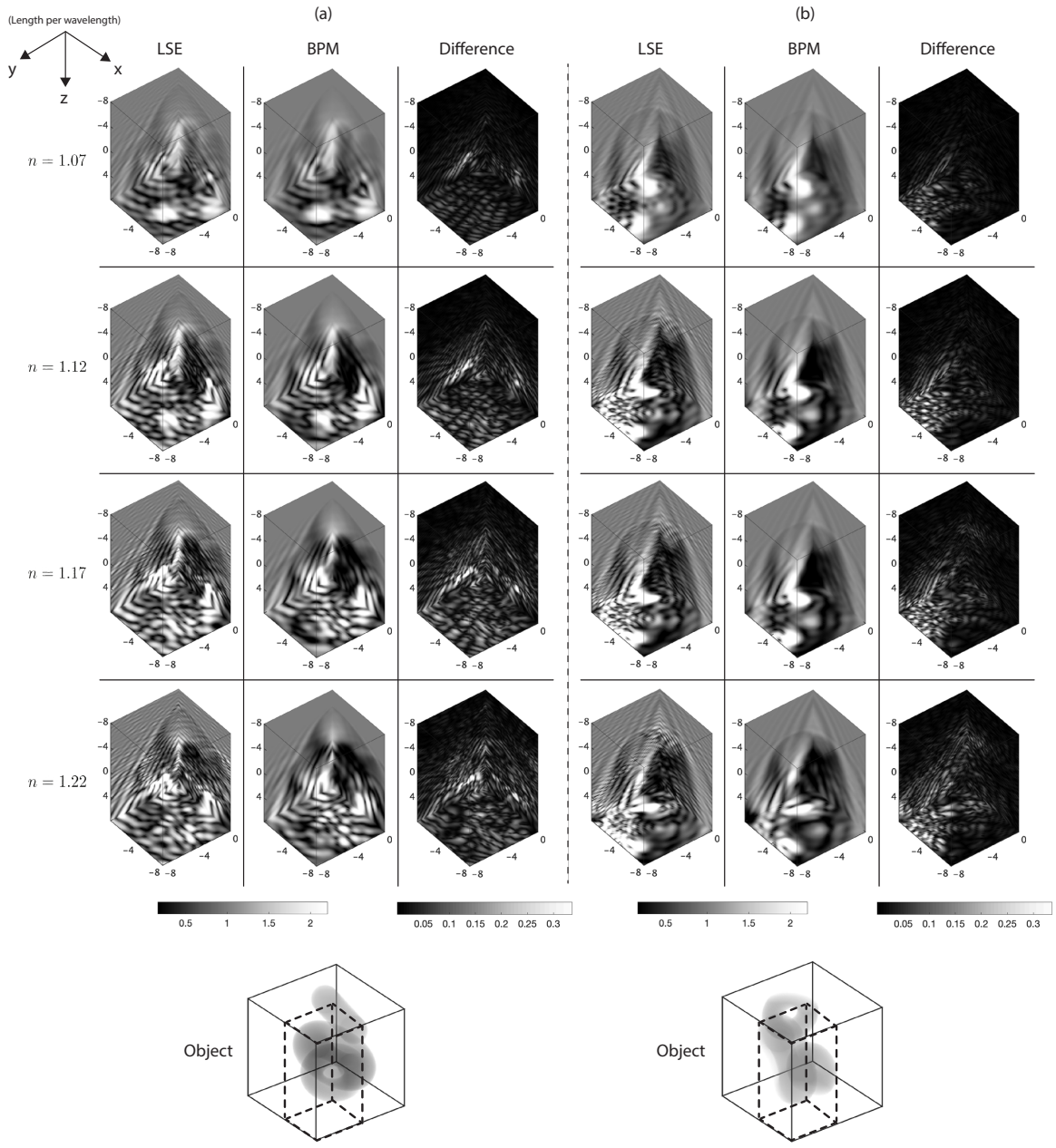

Figure 3-10: Scattered fields estimated from LSE and BPM when the mean refractive index n of a cubic computational box changes. In this figure, potentials consist of tori. The size of a cubic computational box is 16λ. This figure considers two different objects, which are marked with (a) and (b). Difference refers to the elementwise absolute error divided by the maximum field amplitude. For more visibility, we show the shape of objects where the scattered field boxes correspond to the regions enclosed with dotted black lines.

Chapter 4

Neural regularization on LSE for fast and differentiable forward scattering

4.1 Introduction

In the approximation of optical scattering under the scalar wave approximation, LSE has been studied as one of the most accurate methods. Furthermore, in Chapter 3, the theoretical analysis on such accuracy is presented. In fact, due to its equivalency to the scalar Helmholtz equation, it has been used not only in optical scattering but many branches of physics such as acoustics [25, 73], seismic imaging [68], microscopy [14], and quantum scattering [3, 38].

Despite of such importance, however, alternative scattering models are often adopted instead of the LSE. As illustrated previously, this is mainly due to computational complexities in the LSE; it stands out as a complex integral equation when compared to simpler models, lacking a general analytical solution. Subsequently, the resolution of the LSE necessitates the application of iterative methods for linear systems, which may result in very slow convergence.

Consider $\mathcal{D}$ in Eq. (3.15) as a domain of scattered waves in LSE. For computational

purposes, the domain $\mathcal{D}$ is discretized into a lattice, defined by

$$D_N = \left\{ \left(\frac{L_1}{N_1} j_1, \frac{L_2}{N_2} j_2, \frac{L_3}{N_3} j_3 \right) \,:\, \mathbf{j} \in \mathcal{I}_{\mathbf{N}} \right\}, \tag{4.1}$$

where L_d and N_d represent the size and the number of grid points in the d-th dimension, and

$$\mathcal{I}_N = \left\{ (j_1, j_2, j_3) \,:\, -\frac{N_d}{2} \leq j_d < \frac{N_d}{2},\ j_d \in \mathbb{Z},\ d \in \{1,2,3\} \right\}, \tag{4.2}$$

i.e. $\mathcal{D}_N$ represents a cubic grid with N_i divisions along each dimension of length L_i. In this condition, a reformulation of the Lippmann-Schwinger equation is considered to treat it as a linear system:

$$\left[I - \int d\mathbf{r}'\ G(\mathbf{r} - \mathbf{r}') V(\mathbf{r}') \cdot \right] \psi(\mathbf{r}) = \psi_0(\mathbf{r}). \tag{4.3}$$

where I symbolizes the identity operator. This equation has an equivalent form to:

$$Ax = b, \tag{4.4}$$

where b corresponds to ψ_0 and x to ψ, with a linear operator A. Accordingly, the LSE can be iteratively solved with iterative solvers, including Krylov subspace techniques like the conjugate gradient (CG) or the generalized minimal residual method (GMRES), as reviewed in literature [32, 42].

However, the convergence of Eq. 4.3, also referred to as Eq. 4.4, is highly dependent on the numerical conditions present within the system. In short, the condition number of A has a proportionality to the scale of the system and the strength of the scattering potential V [107, 108], which can also be expected from a scenario where the eigenvalues of A correlate directly with the strength of the convolution kernel [33]. This is exemplified by the quadratic relationship observed in the convergence of the CG method, as illustrated in Fig. (4-1). Beyond the number of iterations, the operational time required for applying A to a field becomes a critical consideration,

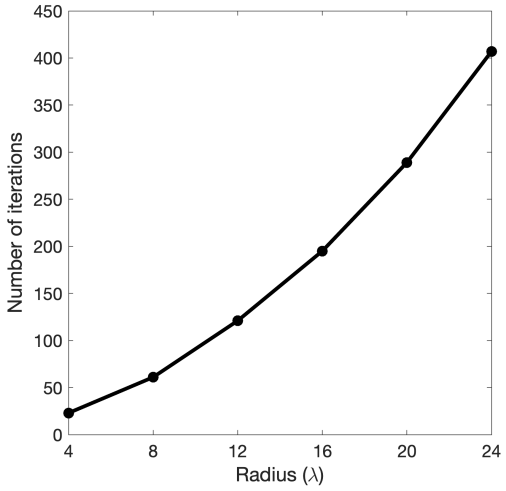

Figure 4-1: The number of CG iterations required to solve the LSE. Spherical objects with various radii are considered. Their refractive index contrast is 1.03. 3 pixels per wavelength are used to sample ψ. The iteration is stopped when the relative L_2 norm error, $\|A\psi - \psi_0\|_2 \,/\, \|\psi_0\|_2$, reaches 10^{-6}. Retrieved from [81].

with the computational effort approximately increasing in proportion to L^3 with each iteration.

Given these considerations, despite the LSE's capability to accurately model light-matter interactions, the computational demands for analyzing relatively large objects are prohibitive. This challenge underscores the necessity for the development of a more efficient LSE solver. In fact, the very origin of the slow convergence under ill-conditioned A is that it is difficult to express the solution x in terms of basis vectors in the Krylov subspace. In other words, mathematical characteristics of the optical scattering phenomenon requires many basis vectors. Hence, one may think of ways to promote mathematical properties that scattered waves may share during the optimization. Such promotion, which *regularizes* the prospective form of the solution x, constitutes the primary aim of this section. As a result, it is further presented that the regularization on the solution can lead to better convergence.

4.2 Motivation on using regularizers in solutions to LSE

In the simplest condition, one can think of an optimization problem to solve Eq. (4.4), as

$$x^\star = \arg\min_x L(Ax, b), \tag{4.5}$$

where L is a convex loss function. If A is Hermitian and positive-definite, one of the prevalent examples of L is

$$L(Ax, b) = \frac{1}{2} \langle x, Ax \rangle - \langle x, b \rangle . \tag{4.6}$$

It is well-known that CG is effectively minimizing this form of L where the residual at each iteration becomes the direction of the gradient descent. On the other hand, if A is not Hermitian or positive-definite, A may be multiplied by its adjoint to form a new Hermitian operator $A^\dagger A$ and CG can be applied, resulting in a formulation that can be translated to a minimization of the L_2 loss:

$$L(Ax, b) = \frac{1}{2} \|Ax - b\|^2 . \tag{4.7}$$

Hence, roughly speaking, simple optimization problems like Eq. (4.5) may be compared to the application of CG without additional considerations, such as preconditioners. Accordingly, it can be expected that the convergence rate of gradient-based methods on Eq. (4.5) would be comparable to that of CG, which can be a problem for the ill-conditioned LSE.

As discussed in the previous section, problems originating from the ill-conditionedness of A can be represented by the difficulty in expressing the optical scattering in mathematical ways, e.g. via basis vectors in Krylov subspaces or additional information per each gradient descent step. To guide the solution of Eq. (4.5) toward a space of

scattered waves, one can introduce an additional function:

$$x^\star = \arg\min_x L(Ax, b) + R(x), \tag{4.8}$$

where R is often referred to as the regularizer. In recent studies, one of the well-known regularizers in the sense of the convex optimization and the inverse problem is the sparsifying regularizer, in combination with the compressive sensing (CS) in the signal processing. CS leverages the concept that signals, which are sparse or can be made sparse through transformation with a particular basis set, can be reconstructed with fewer samples than traditionally dictated by the Nyquist-Shannon criterion [17]. It should be emphasized that CS targets ill-posed situations, e.g. non-negligible noises and A with very small number of rows. This innovative approach has had a profound impact not only on the field of signal processing but also across a broad spectrum of disciplines that deal with the challenge of solving ill-posed inverse problems [17, 94]. In particular, the use of wavelets for image processing, predicated on the sparsity of natural images in the wavelet domain, exemplifies an application of this principle [67]. Moreover, the emergence of dictionary learning methods, which derive optimal sparse representations from training data, represents another significant stride in leveraging sparsity for signal processing [2, 34].

Without promoting the sparsity, the minimization of the original loss function would be stuck at several degenerate solutions. In this sense, the regularizer, roughly speaking, *promotes* desirable mathematical properties during the solution process, consequently leading to a good solution. Similarly, if there exist some mathematical properties that solutions to the LSE share, such properties can also be promoted by a regularizer. Furthermore, if such properties are related to degeneracies that can appear during the solution process on Eq. (4.5), the inclusion of R may be able to help the minimization. The core idea of this work is to adopt R as in Eq. (4.8), and to find R suited for the LSE.

4.3 Learning regularization with neural networks

As illustrated in the previous section, the introduction of R can be helpful in solving the LSE by promoting certain mathematical properties that may be beneficial to reach the solution. However, it is not straightforward to apply R on the LSE, because the form of an appropriate regularizer R is not known *a priori* in many cases, including the LSE. Furthermore, the appropriate form of R can be very complex, which would make the choice of R infeasible. Hence, studies often assume R in an *ad hoc* manner; one of the well-known examples is the total variation regularization under a prior that objects are piecewise continuous and have sharp edges. When it comes to the LSE, R is introduced to represent a physical prior on different entities, scattered fields, but such prior is still connected to the information from the space of objects that scatter incident fields. In this sense, R in LSE also implies a certain physical prior on objects of interest as in previous studies, while its definition becomes even more involved. Specifically, it is difficult to design a prior that can be helpful in evaluating the validity of a scattered field, where such prior is influenced by some physical properties of objects that are not well-known. Thus, instead of knowing R a priori, R may be learned from examples by approximating it as a sufficiently complex function. A neural network can a candidate for this purpose, because of its good ability to approximate arbitrarily complex functions.

However, substituting R as a neural network has its own problems. For example,

$$x^\star = \arg\min_x L(Ax, b) + \mathcal{C}(x), \tag{4.9}$$

where R is substituted by a neural network $\mathcal{C}$ in Eq. (4.8). In this scheme, one has to face two optimization problems at the same time, the original problem and the optimization of weights of $\mathcal{C}$, which can be extremely difficult. Hence, it would be helpful if the equation above can be approximated or expressed as simple steps. For such purpose, in the following sections, a method called the proximal gradient method (PGM) is introduced and a corresponding architecture for $\mathcal{C}$ is proposed.

4.3.1 Proximal gradient descent

Originally, PGM has been introduced for convex optimization problems where the regularization R is non-differentiable or the direct evaluation of the gradient of $L+R$ in Eq. (4.8) is not trivial. Assume that L is convex and everywhere differentiable, and R is convex but not necessarily differentiable. Denoting $x^{(i)}$ as a current guess on $x^\star$, solving the following problem

$$x^{(i+1)} = \arg\min_x L(Ax, b) + R(x) \quad \text{s.t. } x \text{ is close to } x^{(i)}, \tag{4.10}$$

may improve $x^{(i)}$, as both L and R are convex. A simple intuition behind PGM is to approximate L as a quadratic function near $x^{(i)}$:

$$L(x) \approx L(x^{(i)}) + \left\langle \nabla L(x^{(i)}), (x - x^{(i)}) \right\rangle + \frac{1}{2t} \left\| x - x^{(i)} \right\|_2^2 + R(x). \tag{4.11}$$

Accordingly, Eq. (4.10) becomes

$$x^{(i+1)} = \arg\min_x \frac{1}{2t} \left\| x - \left(x^{(i)} - t\nabla L(x^{(i)}) \right) \right\|_2^2 + R(x), \tag{4.12}$$

which implies that the next guess is a point that minimizes R while being close to the gradient update with respect to L. In what follows, the proximal operator is introduced [83]:

$$\text{prox}_R[v] \equiv \arg\min_x R(x) + \|x - v\|_2^2. \tag{4.13}$$

Subsequently, the expression for $x^{(i+1)}$ can be rewritten as

$$x^{(i+1)} = \text{prox}_{2tR} \left[x^{(i)} - t\nabla L(x^{(i)}) \right], \tag{4.14}$$

which is referred to as the PGM. Returning to Eq. (4.8), Eq. (4.14) can be applied to find a regularized solution to LSE. Substituting the L_2 loss in Eq. (4.7) to Eq. (4.14)

gives

$$\begin{aligned} x^{(i+1)} &= \mathrm{prox}_{2tR}\left[tA^\dagger b + (I - tA^\dagger A)x^{(i)}.\right] \\ &= \mathrm{prox}_{2tR}\left[A^\dagger\left(t(b - Ax^{(i)})\right) + x^{(i)}\right]. \end{aligned} \tag{4.15}$$

In some cases, the analytical form of $\mathrm{prox}_{2tR}[v]$ is already known, e.g. the soft-thresholding operator when R is the L_1 norm, which enables the fast computation of Eq. (4.15) regardless of the non-differentiability of R. Under a choice of t, if the iteration in Eq. (4.15) leads to a stationary point, such point corresponds to the minimizer of Eq. (4.8). The convergence behavior of Eq. (4.15) has been discussed in many studies, e.g. [4, 29, 83].

4.3.2 Recurrent networks for proximal operator

By the virtue of the PGM Eq. (4.15), one may be able to solve Eq. (4.8) using

$$\begin{aligned} x^{(i+1)} &= \mathcal{C}\left[tA^\dagger b + (I - tA^\dagger A)x^{(i)}\right] \\ &= \mathcal{C}\left[A^\dagger\left(t(b - Ax^{(i)})\right) + x^{(i)}\right], \end{aligned} \tag{4.16}$$

where the proximal operator is substituted by a neural network $\mathcal{C}$. By setting the proximal operator learnable, Eq. (4.16) has a more convenient form than Eq. (4.9), as the optimization with respect to x is changed to the iteration of a simple gradient step and an application of $\mathcal{C}$.

When it comes to the inverse problem with b (measurements) contaminated with external perturbations such as noises, there have been several studies have discussed possible ways to learn R from available data in a similar manner to the previous sections (and solving LSE should be referred to as a forward problem). However, they are not directly applicable to solve the LSE. For example, it has been suggested to assume specific forms on R, e.g. using basis functions or parameterizing a part of a proximal operator corresponding to a known regularizer [46, 54, 53, 110]; however, such R would be too simple to be used for the regularization in the LSE. Learning dictionaries is another popular approach, but is also of limited use because it assumes a specific representation; the sparse basis itself is learned from the data by imposing

a sparsity criterion on the representation [2, 34]. On the other hand, there have been studies that use neural networks, not basis functions, to extend the scope of functions that can be learned [70, 106, 44, 75, 22]. However, they typically leverage large neural networks, which require significant amount of computation compared to iterative linear solvers. Furthermore, they often presume an equality constraint $Ax = b$ under ill-posed conditions, which is not perfectly suitable for systems without ill-posed measurements.

Accordingly, it is not computationally efficient if $\mathcal{C}$ is represented by a conventional neural network with large computational requirements. Furthermore, successive application of neural networks at each iteration of the proximal gradient process leads to the destabilization of the network training, i.e. numerical issues regarding naive recurrent networks [48]. Consequently, it is recognized that the recurrence can achieve numerical stability when the neural network incorporates a specific configuration: a long short-term memory (LSTM) where the ability to selectively forget over time is crucial [49]. Furthermore, the LSTM's capability to maintain historical information across iterations allows it to respond variably at different iteration stages, which could be advantageous during the proximal gradient process. Based on the structure of the LSTM, there have been proposed multiple closely relevant architectures for the recurrent process such as GRU [24], and others [112, 63, 89]. Subsequently, $\mathcal{C}$ in Eq. (4.16) is substituted by a recurrent neural network with considerations on the numerical stability and sequential information as in the LSTM. To summarize, the expected advantages of adopting the recurrence-stable architecture can be listed as follows:

- The amount of computation required in each step of the architecture is usually significantly lightweight compared to conventional neural networks.

- The architecture can alleviate the gradient problem regarding conventional neural networks with naive recurrence, and it is aware of iteration stages. In addition, it has been shown that the architectural design is suitable for handling sequential processing of information.

- The physical information is directly provided through the operator A, which means that it would not be required for the neural network to learn the physics of a system as in simple end-to-end networks.

In particular, the fact that the physical information is fused into the training process can greatly benefit the generalizabiilty of $\mathcal{C}$. Put differently, with proper training, $\mathcal{C}$ can act as an optimal regularizer for a specific system A when dealing with a carefully chosen category of objects x. This category should be wide enough to maintain relevance yet narrow enough to preserve the strength of the prior. Under these conditions, Eq. (4.16) is capable of delivering satisfactory outcomes and exhibiting good generalization capabilities. This implies that the system can accurately process inputs that were not part of the training data, even if these inputs slightly deviate from the exact category of the training set. The aspect of network generalizability is explored in more depth in subsequent sections.

4.3.3 Mathematical inspiration from preconditioned CG

Algorithm 1 CG with a preconditioner M^{-1} and a tolerance ε

$r^{(0)} \leftarrow b - Ax^{(0)}$
$z^{(0)} \leftarrow M^{-1}r^{(0)}$
$p^{(0)} \leftarrow z^{(0)}$
$i \leftarrow 0$
while $\frac{\|r^{(i)}\|}{\|b\|} \leq \varepsilon$ **do**
    $\alpha^{(i)} \leftarrow \frac{\langle r^{(i)}, z^{(i)} \rangle}{\langle p^{(i)}, Ap^{(i)} \rangle}$
    $x^{(i+1)} \leftarrow x^{(i)} + \alpha^{(i)} p^{(i)}$
    $r^{(i+1)} \leftarrow r^{(i)} - \alpha^{(i)} Ap^{(i)}$
    $z^{(i+1)} \leftarrow M^{-1} r^{(i+1)}$
    $\beta^{(i)} \leftarrow \frac{\langle r^{(i+1)}, z^{(i+1)} \rangle}{\langle r^{(i)}, z^{(i)} \rangle}$
    $p^{(i+1)} \leftarrow z^{(i+1)} + \beta^{(i)} p^{(i)}$
    $i \leftarrow i + 1$
end while

As illustrated previously, Krylov subspace methods are well-known methods to solve the LSE in many previous studies. Hence, the mathematical intuition behind $\mathcal{C}$

in Eq. (4.16) in terms of such methods would be beneficial for us to better understand the role of $\mathcal{C}$. Moreover, one may be able to obtain useful inspiration from such methods to improve Eq. (4.16) further.

In this section, the CG with a preconditioner is considered, because it has one one of the simplest structures and it has a direct relationship to the gradient descent. Algorithm 1 depicts the process in the CG with a preconditioner M^{-1}, i.e.

$$M^{-1}Ax = M^{-1}b. \tag{4.17}$$

Note that CG requires A to be positive-definite and Hermitian. Otherwise, one can substitute $A \to A^\dagger A$ and $b \to A^\dagger b$, converting the line 7 and 8 in Algorithm 1 as

$$\begin{aligned} x^{(i+1)} &= x^{(i)} + \alpha^{(i)} p^{(i)} \\ &= x^{(i)} + \alpha^{(i)} M^{-1} r^{(i)} - \underbrace{\alpha^{(i)} \frac{\beta^{(i-1)}}{\alpha^{(i-1)}} (x^{(i)} - x^{(i-1)})}_{\text{Heavy-ball method [85]}}, \end{aligned} \tag{4.18}$$

and

$$r^{(i)} = A^\dagger b - A^\dagger A x^{(i)}. \tag{4.19}$$

These equations can be closely related to Eq. (4.16), if one assumes that

$$\mathcal{C}: \quad t r^{(i)} + x^{(i)} \to t M^{-1} r^{(i)} + x^{(i)}, \tag{4.20}$$

and $t = \alpha^{(i)}$. Specifically, Eq. (4.16) can be roughly viewed as a preconditioned CG with a nonlinear preconditioner[1] and without the Heavy-ball acceleration.

Under this view, Eq. (4.16) can be interpreted as learning a nonlinear preconditioner with respect to a matrix $A^\dagger A$. However, taking $A^\dagger A$ squares the condition number of the original matrix A, which is mathematically unfavorable. Furthermore, applying $A^\dagger A$ doubles the amount of the matrix-vector multiplication at each iteration. Hence, inspired by the CG, it is proposed that $\mathcal{C}$ in Eq. (4.16) has the following

[1]Though it is not mathematically robust to call this as a *preconditioner* anymore.

form:

$$\mathcal{C}: \quad A^\dagger\left(t(b - Ax^{(i)})\right) + x^{(i)} \to t\mathrm{NN}(b - Ax^{(i)}) + x^{(i)}, \tag{4.21}$$

where NN represents a recurrent neural network part and the application of $A^\dagger$ is implicitly fused into NN. The form in Eq. (4.21) may also be interpreted as the residual learning [47], which has been shown to improve the gradient flow in deep neural networks.

4.4 Numerical experiments

4.4.1 Architecture of neural regularizer

In this study, the minimal gated unit (MGU) [112] is chosen as the architecture of $\mathcal{C}$. The cell consists of the following key equations:

$$\begin{aligned} f^{(i+1)} &= \sigma\left(W_{yf,j} \star y^{(i)} + W_{hf,j} \star h^{(i)}\right) \\ o^{(i+1)} &= \tanh\left(W_{yo,j} \star y^{(i)} + W_{ho,j} \star h^{(i)}\right) \\ h^{(i+1)} &= (1 - f^{(i+1)}) \circ h^{(i)} + f^{(i+1)} \circ o^{(i+1)} \end{aligned} \tag{4.22}$$

where W, σ, and $\circ$ represent the weight, the sigmoid function, and the element-wise product. The subscript j in weights is to distinguish different cells if multiple MGU cells are used. Here, $\star$ is represented by the Fourier convolution [86]:

$$W\star: \ y \ \to \ \hat{\mathcal{F}}^\dagger(W\circ)\hat{\mathcal{F}}y. \tag{4.23}$$

Note that the bias terms in the original MGU are omitted. At each iteration, $\mathcal{C}$ receives two inputs: $y^{(i)}$ and $h^{(i)}$. $y^{(i)}$ represents input information provided to the network at each iteration. In this work, information on the optical potential V is delivered; however, as a potential and a scattered field have mathematically distinct properties, naively passing V to the network can induce extra difficulty in estimating

fields. Hence, $y^{(i)}$ is set as the first-Born approximation with V:

$$y^{(i)} = b + \int d\mathbf{r}' \; G(\mathbf{r} - \mathbf{r}')V(\mathbf{r}')b(\mathbf{r}'). \tag{4.24}$$

On the other hand, $h^{(i)}$ is referred to as the hidden state, where the cell extracts useful information based on $y^{(i)}$. Based on Eq. (4.21), the input hidden vector is set as $b - Ax^{(i)}$, i.e.

$$h^{(i)} = b - Ax^{(i)}. \tag{4.25}$$

The final output from $\mathcal{C}$ is then expressed by using the hidden state $h^{(i+1)}$:

$$x^{(i+1)} = h^{(i+1)} + x^{(i)}, \tag{4.26}$$

i.e. $t = 1$ in Eq. (4.21).

In every step, the most expensive operation in $\mathcal{C}$ is Eq. (4.23). Thus, denoting N as the number of voxels in the uniform discretization, Eq. (4.1), the amount of computation N_c in every step is approximately proportional to $\approx 4N \log N + 8N \log 8N$:

- $4N \log N$ originates from four operations of $W\star$ in $\mathcal{C}$.
- $8N \log 8N$ stems from A, which is evaluated by the Fourier convolution with zero-padding [102, 84].

The network is deliberately designed to consist of relatively cheap operations, compared to classical linear solvers. Specifically, Table 4.1 presents approximated values on N_c in well-known Krylov subspace methods, showing the computational efficiency of the proposed method.

4.4.2 Training procedure

For training of $\mathcal{C}$, scattering of a planewave, $\psi_0(\mathbf{x}) = e^{ikz}$, is considered. 9600 examples of V are generated and split into 9000 training, 300 validation, and 300 test examples. For each example, the refractive index n consists of 1 to 3 spherical objects whose magnitude ranges from 1.01 to 1.13 and diameter from 11λ to 21λ. The index and

Table 4.1: The amount of computation N_c required per each iteration in different methods. Here, two Krylov subspace methods are compared with the proposed method: BiCGSTAB-2 and QMRCGSTAB. N_c is approximated by counting the number of matrix-vector multiplications and considering their computational complexity. For instance, $N \log N$ and $8N \log 8N$ originate from the Fourier convolution without and with zero-padding, respectively.

	BiCGSTAB-2	QMRCGSTAB	Proposed
N_c	$4 \times (8N \log 8N)$	$2 \times (8N \log 8N)$	$4 \times (N \log N) + (8N \log 8N)$

diameter ranges are chosen based on previous studies regarding the LSE [84, 55, 56, 21], where experimental conditions in such studies are well covered (and even worse in this study). The background index n_b is set 1. The computational domain D_N in Eq. (4.1) is set as a cubic box with length 64λ in each dimension.

Starting from $x^{(0)} = b = \psi_0$, Eq. (4.16) is repeated eight times with four different MGU cells (two iterations per cell), i.e. until $x^{(8)}$ is obtained. In other words, the total amount of computation during the inference is

$$N_c = 32N \log N + 64N \log N. \tag{4.27}$$

Subsequently, the difference between $Ax^{(8)}$ and b is measured based on the training loss function [65], which is the negative Pearson correlation coefficient (NPCC) [43]. In this way, the preparation of ground truth fields corresponding to training potentials can be skipped, which is computationally expensive in practical situations. For the update of weights in the network based on the difference, the Adam optimizer [58] with an initial learning rate 10^{-3} is utilized. For the efficient evaluation of the operator A in the LSE during the training, additional numerical techniques have been applied, as illustrated in Section C.1 and Section C.2.

4.4.3 Results

The training result is summarized in Table 4.2 where the proposed model is checked with test examples consisting of spherical objects. Since ground truth scattered fields are not prepared during the training (and such preparation may not be feasible in

practice), all metrics listed in the table are evaluated between $Ax^{(8)}$ and b. In addition, for the comparison to the classical linear solvers, BiCGSTAB-(l) [101, 92] with $l = 2$ and QMRCGSTAB [45] are chosen. When it comes to BiCGSTAB-2, there are four matrix-vector multiplications per iteration, which approximately leads to $N_c \approx 32 \log 8N$ per iteration. Table 4.1 summarizes estimations of N_c in different methods, where similar calculation is done on QMRCGSTAB. N_c in Table 4.2 is the gross amount of computation over all iterations required to achieve the same tolerance to the proposed method. While the relative L_2 error may not be extremely low, the convergence to feasible solutions implies that $\mathcal{C}$ is able to learn a regularization that promotes a space of scattered fields. As a result, it is worth noting that in the proposed model, N_c is approximately 7-10 times smaller than BiCGSTAB-2 and 2-3 times smaller than QMRCGSTAB, significantly reducing the computational resource.

In Fig. (4-2), the convergence behavior of the proposed model is depicted in more detail. As the iteration in Eq. (4.16) proceeds, the relative L_2 error decreases exponentially. However, it has been also observed that increasing the number of iterations in Eq. (4.16) does not further decrease the error; rather, the error shows an oscillating behavior. Hence, it is speculated that the behavior of $\mathcal{C}$ is subject to the number of iterations chosen during the training, e.g. $\mathcal{C}$ tries to achieve the training objective in 8 iterations and information on the iteration via $h^{(i)}$ may be valid only up to the 8^{th} iteration.

In addition to the faster convergence by specifically setting the role of $\mathcal{C}$ to learn a regularization, the inclusion of the physical information A during the training may have other advantages, compared to conventional neural networks. Note that the network is initially trained using objects composed of several spheres. From a supervised learning perspective, one might anticipate inaccuracies in the scattering fields generated by the network when it encounters objects devoid of spherical components. Surprisingly, the network demonstrates a robust ability to closely estimate scattering fields for test cases involving shapes vastly different from those seen during training, such as polyhedra.

Fig. (4-3) depicts the network's output when it processes combinations of polyhe-

dral objects. For a comparison, an estimation on the scattered field from BiCGSTAB-2 is also depicted whose relative L_2 error is 10^{-3}. The resulting fields exhibit certain imperfections near sharp edges, possibly due to the network's unfamiliarity with sharp edges and vertices of polyhedral shapes during its training phase, which could degrade the quality of its estimations. Nevertheless, the qualitative performance on unseen objects does not exhibit a serious deviation from the solution with lower tolerance.

Furthermore, both Table 4.2 and Fig. (4-2) confirm that the network's estimated fields align numerically with the spherical case. This level of adaptability suggests that the network efficiently incorporates physical principles from A throughout its training, with the MGU layer acting as a suitable proximal operator for the problem at hand. This is in contrast to a scenario where the network's parameters are simply too closely fitted to spherical training samples. In such a case, ψ from the physics layer would be inaccurately regularized, leading to suboptimal convergence outcomes.

In overall, the proposed method Eq. (4.16) is inspired by the proximal gradient process where the role of $\mathcal{C}$ is implicitly set as a regularizer for the LSE. Furthermore, unlike conventional neural networks, $\mathcal{C}$ can leverage information from the physical part $b - Ax$ in the architecture, potentially alleviating the need for networks to approximate A (or relevant operations such as the adjoint and the inverse). Though the current performance may require further improvement in terms of the residual error, it exhibits promising generalizability with tolerable performance in the qualitative manner. It may be emphasized that $\mathcal{C}$ is fast and easily differentiable with respect to V and ψ. Hence, the application of Eq. (4.16) in the inverse scattering is straightforward and may be able to improve the performance of the object retrieval. Lastly, there are many systems in physics that can be expressed as a linear problem, Eq. (4.4). It should be noted that the proposed technique in this work can be directly applied on such systems. The technique exhibits a stable convergence via a learned regularization, while linear solvers such as BiCGSTAB and QMRCGSTAB discussed in this work may have erratic convergence behaviors depending on a system [39]. Hence, it is expected that this work may be able to facilitate other studies in relevant fields of physics.

Table 4.2: Image quality metrics based on fields estimated by the proposed method. Each potential consists of up to 3 spherical or polyhedral objects. The range of the refractive index is $[1.01, 1.1]$. For each spherical or polyhedral object, the metrics are evaluated with 30 random potentials. In this table, N_c is used to denote the amount of computation (see, Section 4.4.1) required to achieve the same residual tolerance to the proposed method. In the proposed method, $N_c = 32N \log N + 64N \log 8N$. The relative L_2 error corresponds to $\frac{\left\|Ax^{(8)} - b\right\|}{\|b\|}$.

	Relative L_2	N_c via BiCGSTAB-2	N_c via QMRCGSTAB
Spherical	0.017	$1032N \log 8N$	$336N \log 8N$
Polyhedral	0.016	$712N \log 8N$	$211N \log 8N$

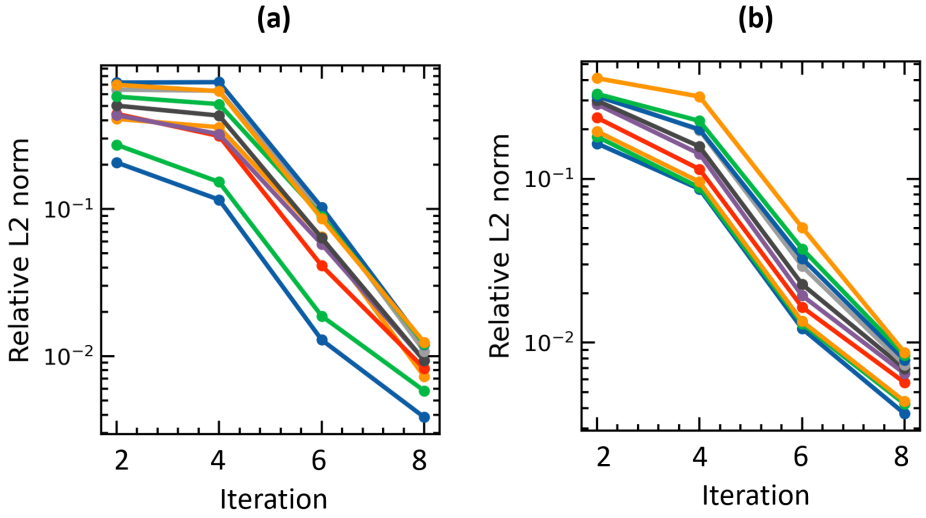

Figure 4-2: Convergence behavior of the proposed method on different objects, (a) spherical objects and (b) polyhedral objects, depending on the iteration number of $\mathcal{C}$. This figure shares the same experimental condition to Table 4.2, where 10 example potentials are randomly selected for the visualization.

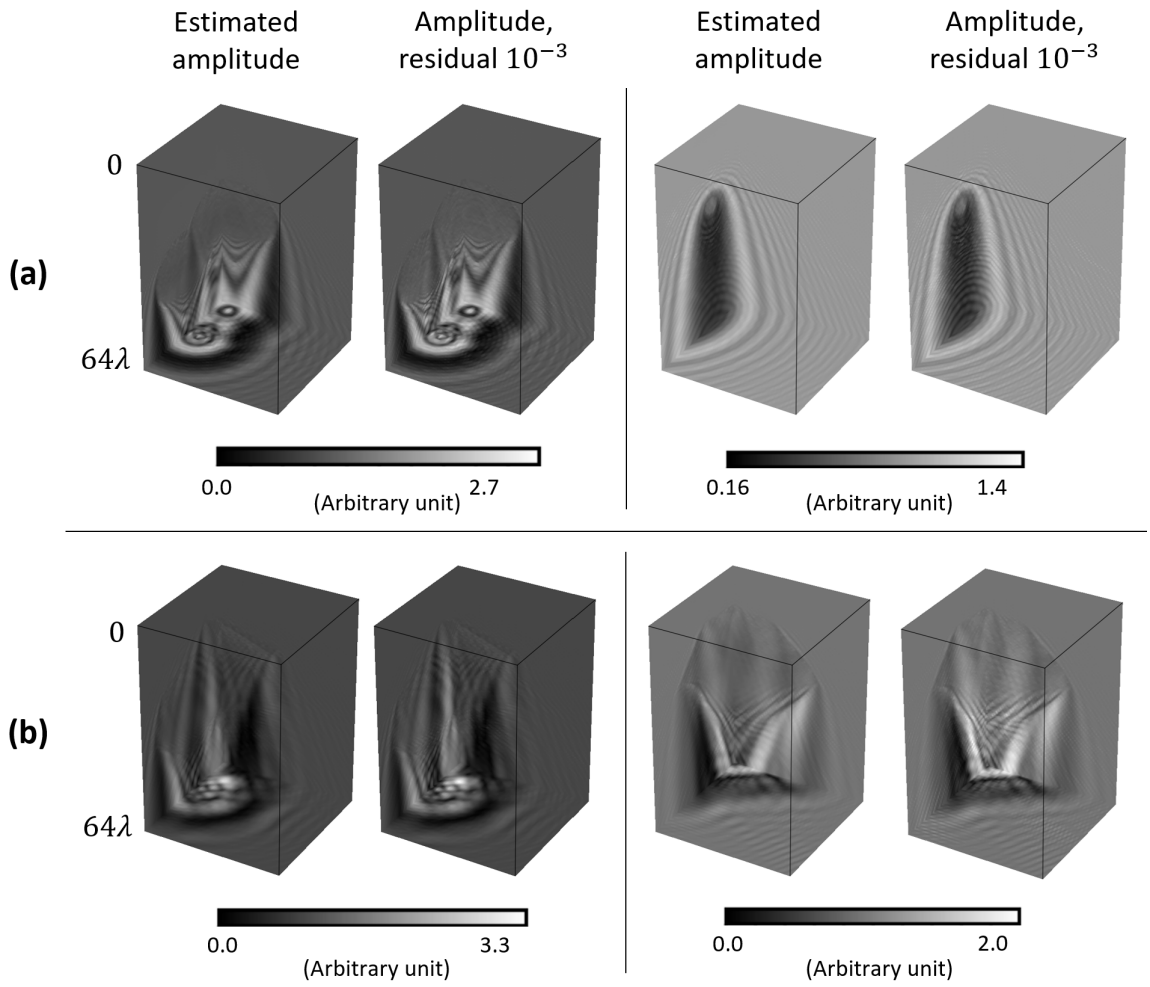

Figure 4-3: Scattered intensities estimated from the proposed method and the BiCGSTAB-2 with the relative L_2 error 10^{-3}. The incident wave propagates downward in the illustration. The figure considers two types of objects: (a) objects consisting of dielectric spheres and (b) objects consisting of dielectric polyhedra.

Chapter 5

Conclusion

In computational imaging, information on various physical systems are carried by photons and there have been many studies on how to retrieve such information from photons in a numerical sense. However, due to the difficulty in capturing the phase part of photons, loss of some information is inevitable in many scenarios. Moreover, the relationship between the wave function of photons and systems becomes complex, the characterization of systems often requires sophisticated *decoding* of information in photons. These issues lead to two important problems in computational imaging: the phase retrieval problem and the approximation on the optical scattering. In this thesis, phenomenal models on each problem are considered and several improvements are suggested to mitigate numerical obstacles behind them.

Regarding the phase retrieval, TIE is one of the famous models, due to its simplicity and guaranteed convergence. TIE is originated from the paraxial wave equation and relates the intensity derivative to the phase, which enables the estimation of the phase from multiple intensity measurements. However, because of its sensitivity to the choice of defocus distances Δz and the singularity at the origin of the Fourier space in its kernel, it has been studied that TIE is vulnerable to experimental conditions and noises. The proposed model, TIE-TPE, can mitigate such problems by interpreting TIE as an ODE, theoretically leading to reduced Δz-dependency and better suppression of artifacts from noises. In numerical experiments, it is shown that classical algorithms on TIE only work in very ideal situations, e.g. very small

Δz and noises. Otherwise, TIE-TPE greatly stabilizes the phase estimation process, showing good performance regardless of experimental conditions. As a non-paraxial extension from TIE-TPE, a method based on the angular spectrum propagation is also discussed. It is expected that TIE-TPE and the angular spectrum-based method may be connected to famous iterative methods in the phase retrieval problem; accordingly, a more unified treatment on various algorithms on the problem may be possible in the future.

As illustrated above, optical scattering models should be considered to further decrypt information in photon wave functions, as their intensities and phases undergo serious interaction with systems in intangible ways. LSE, BPM, and Born series are three scattering models that have been widely used in previous studies. However, detailed relationships between them have remained obscure, though they are all originated from the scalar Helmholtz equation. This is because they have different assumptions, leading to a deviation in their forms. For more accurate and quantitative comparison between different models, estimating discrepancies in a robust way, analytical relationships between LSE, BPM, and Born series are discussed. It is shown that BPM and Born series both can originate from the series expansion of LSE. However, they exhibit different convergence behavior. Analyzing this behavior, a simple and dimensionless condition is suggested to guarantee the convergence of Born series that is tighter than previous studies. Furthermore, assumptions behind BPM that field propagation and modulation from optical potentials commute can effectively reduce the operator norm of the LSE operator, leading to a stronger convergence than Born series. The errors resulting from such commutation assumption can be estimated by a dimensionless parameter $\mathcal{S}$. Subsequently, numerical experiments are conducted, which corroborate the feasibility of our theoretical analysis. It is expected that approaches in this section can be extended to other models e.g. [11, 21] that are not discussed here but closely relate to LSE, Born series, and BPM. Lastly, this study may be able to help analysis not only of scattering models but also field and object estimations in the inverse scattering.

Consequently, LSE can be considered as one of the most accurate models for es-

timating of the optical scattering. However, as outlined in previous studies, it has a serious drawback, which is its large computational requirement. Such drawback originates from the ill-conditionedness and the large 3D integral regarding the forward operator in the LSE. Numerous studies have already ventured into solving relevant problems through the utilization of neural networks with large size and computational cost. Nonetheless, this strategy might not yield high-quality predictions due to the challenge of adequately capturing the physics of the systems. In this thesis, a neural network with an alternative architecture is proposed for the LSE, which is inspired from the proximal gradient descent and partially from the classical linear solvers. Within this inspiration, a neural network is implicitly regarded as an efficient regularizer for the LSE and knowledge on the physics of the LSE is conveyed via separate physics layers. In turn, this architecture exhibits significant improvement in terms of the computational cost compared to one of the well-known iterative linear solvers. Moreover, having been trained on entities like dielectric spheres, the proposed network demonstrates a generalizability in predicting scattered fields from objects that are unseen during the training, both from a visual and numerical standpoint. This superiority is attributed to the integration of the physical operator A during the training phase, a decision rooted in solid mathematical principles. Furthermore, our approach deliberately designs tunable parameters to serve as a regularizing factor, enhancing the model's ability to generalize. Though the absolute performance of the proposed model may require further optimization, we underscore the importance of incorporating physical principles into neural networks, not just for the LSE but for tackling general physical inverse problems.

Appendix A

Adjoint method in equality-constrained optimizations

A.1 Adjoint equation from method of Lagrange multipliers

In various fields of mathematics and physics, we are often interested in the following optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{u}} \quad & f(\mathbf{u}) \\ \text{subject to} \quad & g(\mathbf{u}) = \mathbf{0} \end{aligned} \tag{A.1}$$

where $f : \mathbf{u} \in \mathbb{H}_f \to \mathbb{R}$, $g : \mathbf{u} \in \mathbb{H} \to \mathbb{H}_g$, and $\mathbb{H}_f$ and $\mathbb{H}_g$ are Hilbert spaces. For simplicity, we assume that f is convex. We often call f and g as the objective and the equality constraint, respectively. Intuitively, in physical applications, f can be considered as the deviation between our estimation and observations. g is a physical system that governs observations or an ad-hoc model such as neural networks.

The minimization problem Eq. (A.1) can be regarded as an unconstrained optimization problem using the Lagrange formulation:

$$\mathcal{L}(\mathbf{u}, \mathbf{h}) \equiv f(\mathbf{u}) + \langle \mathbf{h}, g(\mathbf{u}) \rangle \tag{A.2}$$

where $\mathbf{h}$ is the Lagrange multiplier[1]. It is well studied that finding the saddle points of Eq. (A.2), i.e. points that are minima with respect to u and at the same time, are maxima with respect to $\mathbf{h}$, is the necessary condition to finding an optimal solution to Eq. (A.1). Roughly, if f and g are differentiable, the derivative of $\mathcal{L}$ will be zero with respect to $\mathbf{u}$ and $\mathbf{h}$, if we are at an optimal point. This is quantitatively expressed as follows.

Definition 1 (Frechet derivative)**.** *Let $f : \mathbb{X} \to \mathbb{Y}$ where $\mathbb{X}$ and $\mathbb{Y}$ are Hilbert spaces. If there exists a linear operator $f'(\mathbf{u})$ that satisfies*

$$\lim_{\|\boldsymbol{\eta}\|_X \to 0} \frac{\|f(\mathbf{u}+\boldsymbol{\eta}) - f(\mathbf{u}) - f'(\mathbf{u})(\boldsymbol{\eta})\|_X^2}{\|\boldsymbol{\eta}\|_{\mathbb{X}}} = 0 \qquad \boldsymbol{\eta} \in \mathbb{X}, \tag{A.3}$$

the operator is called the Frechet derivative of f at $\mathbf{u}$.

Definition 2 (Gradient)**.** *For a function $f : \mathbb{X} \to \mathbb{R}$ where $\mathbb{X}$ is a Hilbert space, assume that f has the Frechet derivative at $\mathbf{u} \in \mathbb{X}$. If there exists a vector $\nabla f(\mathbf{u}) \in \mathbb{X}$ that satisfies*

$$f'(\mathbf{u})(\boldsymbol{\eta}) = \langle \nabla f(\mathbf{u}), \boldsymbol{\eta} \rangle \qquad \forall \boldsymbol{\eta} \in \mathbb{X} \tag{A.4}$$

then it is called the gradient of f at $\mathbf{u}$.

Definition 3 (Jacobian)**.** *Let f be a function between $\mathbb{R}^n$ and $\mathbb{R}^m$. If f is Frechet differentiable at $\mathbf{x}$, there exists a matrix F such that*

$$f'(\mathbf{u})(\boldsymbol{\eta}) = F\boldsymbol{\eta} \qquad \forall \boldsymbol{\eta} \in \mathbb{R}^n. \tag{A.5}$$

A is a n-by-m matrix and it can be shown that its elements are $F_{ij} = \partial f_i(\mathbf{u})/\partial u_j$ where f_i is the i^{th} element of the column-vector representation of f. F is called the Jacobian of f. If $m = 1$, we can show that $F = \nabla f$.

Theorem 1. *Let $\mathbb{X}$ and $\mathbb{Y}$ be Hilbert spaces. Assume that $f : \mathbb{X} \to \mathbb{R}$ is convex and continuously Frechet differentiable. In addition, assume that f has an extremum*

[1]In a more general sense, $\mathbf{h}$ is the linear functional on $g(\mathbf{u})$, but by the virtue of the Riesz representation theorem in Hilbert spaces, we use the notion of the inner product to interpret $\mathbf{h}$ as a vector.

at $\mathbf{u}^\star$ *subject to a constraint* $g(\mathbf{u}) = \mathbf{0}$ *where* $g : \mathbb{X} \to \mathbb{Y}$ *is continuously Frechet differentiable. If* $f'(\mathbf{u})$ *maps* $\mathbb{X}$ *onto* $\mathbb{Y}$*, there exists* $\mathbf{h}$ *such that the Lagrangian*

$$\mathcal{L}(\mathbf{u}; \mathbf{h}) = f(\mathbf{u}) + \langle \mathbf{h}, g(\mathbf{u}) \rangle \tag{A.6}$$

is stationary at $\mathbf{u}^\star$ *[5, 13], i.e.*

$$\mathcal{L}'(\mathbf{u}^\star)(\boldsymbol{\eta}) = f'(\mathbf{u}^\star)(\boldsymbol{\eta}) + \langle \mathbf{h}, g'(\mathbf{u}^\star)(\boldsymbol{\eta}) \rangle = \mathbf{0} \qquad \forall \boldsymbol{\eta} \in \mathbb{X}. \tag{A.7}$$

Note that the stationarity of $\mathcal{L}$ *is automatically implied in the constraint* $g(\mathbf{u}) = \mathbf{0}$*.*

In practice, the optimizable vector $\mathbf{u}$ may be decomposed into two parts, $\boldsymbol{\varphi}$ and $\mathbf{p}$, each of which represents a state variable and an adjustable parameter, respectively. They are generally not independent to each other, as they are coupled by the system $g(\boldsymbol{\varphi}, \mathbf{p}) = \mathbf{0}$. Considering the Cartesian space for simplicity, we can derive another version of Theorem 1 for such decomposition:

Corollary 1. *From Theorem 1, let* $\mathbb{X}$ *and* $\mathbb{Y}$ *be* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively. Let* $\mathbf{u}$ *be decomposable into* $\boldsymbol{\varphi}$ *and* $\mathbf{p}$*, i.e.* $\mathbf{u} = \boldsymbol{\varphi} \oplus \mathbf{p}$ *and* $f(\mathbf{u}) = f(\boldsymbol{\varphi}, \mathbf{p})$*. Then at* $\mathbf{u}^\star = \boldsymbol{\varphi}^\star \oplus \mathbf{p}^\star$*, we have the stationarity condition expressed as [13]*

$$\begin{aligned} &\textit{Forward equation} && g(\boldsymbol{\varphi}, \mathbf{p}) = \mathbf{0} \\ &\textit{Adjoint equation} && \nabla_{\boldsymbol{\varphi}} f(\boldsymbol{\varphi}^\star, \mathbf{p}^\star) + G_{\boldsymbol{\varphi}}^\dagger \mathbf{h} = \mathbf{0} \\ &\textit{Control equation} && \nabla_{\mathbf{p}} f(\boldsymbol{\varphi}^\star, \mathbf{p}^\star) + G_{\mathbf{p}}^\dagger \mathbf{h} = \mathbf{0}, \end{aligned} \tag{A.8}$$

where $G_{\boldsymbol{\varphi}}^\dagger$ *is the Jacobian of* g *with respect to* $\boldsymbol{\varphi}$ *at* $\boldsymbol{\varphi}^\star \oplus \mathbf{p}^\star$*.* $\dagger$ *denotes the matrix adjoint.*

Similar versions with $\mathbb{X}$ and $\mathbb{Y}$ being non-Cartesian spaces exist. All such versions decompose the original stationary condition into two parts: the adjoint equation and the control equation. In practice, we want to optimize for $\mathbf{p}$ where $\boldsymbol{\varphi}$ is a variable with an indirect dependency via g. Hence, $\nabla_{\mathbf{p}} f$ in the control equation usually becomes the entity we want to compute; the equation *controls* our optimization process by giving

feedback on our guess on $\mathbf{p}$. Subsequently, we may adjust $\mathbf{p}$ to reach an optimal point via the gradient descent as follows.

1. From our current estimation on $\mathbf{p}$, solve for $\boldsymbol{\varphi}$ using the forward equation.

2. To satisfy the stationary condition, we first solve the adjoint equation from the current $\boldsymbol{\varphi}$ and $\mathbf{p}$ to get $\mathbf{h}$.

3. Evaluate $\nabla_{\mathbf{p}} f$ using $\mathbf{h}$ and the control equation.

4. Adjust the current estimation on $\mathbf{p}$ to reach a point $\mathbf{p}^\star$ that satisfies the stationary condition and *can* be an optimal solution.

As noted above, because Theorem 1 only tells us about the necessary condition for optimal solutions, satisfying Eq. (A.8) does not necessarily mean that we are at an optimal point. However, in practice, we will approach to a point $\mathbf{p}$ where the objective becomes invariant near $\mathbf{p}$ (as $\nabla_{\mathbf{p}} f \to \mathbf{0}$) and the stationary condition is satisfied, which would lead to the good estimation on the parameter that we can obtain from observations in most cases.

A.2 Adjoint equation for ordinary differential equation

Consider an ordinary differential equation:

$$\frac{d\varphi}{dz} = y(\varphi, p, z), \quad \varphi(z_0) = \varphi_0, \tag{A.9}$$

where φ and y are arbitrary functions and p is a parameter of interest. Note that this equation is the same form to the coupled TIE-TPE (the former can easily be extended for multi-dimensional functions). Suppose that we want to minimize an objective

$$f(\varphi, p) = \int_{z_0}^{Z} dz\, l(\varphi, p), \tag{A.10}$$

where l stands for a loss per each z. Then, it is straightforward to derive the corresponding stationary condition:

$$\begin{aligned} &\text{Forward equation} && \frac{d\varphi}{dz} - y(\varphi, p, z) = 0,\ \varphi(z_0) = \varphi_0, z_0 \le z \le Z \\ &\text{Adjoint equation} && \int_{z_0}^{Z} dz\, \frac{\partial l(\varphi, p)}{\partial \varphi} + \left[\frac{d}{dz} - \frac{\partial y}{\partial \varphi}\right]^{\dagger} h = 0 \\ &\text{Control equation} && \nabla_p f = \int_{z_0}^{Z} dz\, \frac{\partial y}{\partial p}^{\dagger} h \end{aligned} \tag{A.11}$$

One caveat in the equations above is that we should evaluate the adjoint of the derivative operators. Roughly speaking, using the definition of the adjoint in the Hilbert space, one can show that

$$\frac{d}{dz}^{\dagger} \to -\frac{d}{dz} \quad \text{s.t. additional boundary conditions,} \tag{A.12}$$

which, in fact, adds a boundary condition $h(Z) = 0$ in the adjoint equation above.

Moving one step forward from $\nabla_p f$, one can evaluate the total differential of f with respect to p. Note that the adjoint equation above implies that the Lagrangian multiplier h satisfies

$$\frac{dh}{dz} = -\frac{\partial y}{\partial \varphi}^{\dagger} h + \frac{\partial l(\varphi, p)}{\partial \varphi}, \quad h(Z) = 0. \tag{A.13}$$

With the forward equation, $\frac{df}{dp}$ is equivalent to $\frac{d\mathcal{L}}{dp}$. Subsequently, by removing terms in the adjoint equation from $\frac{d\mathcal{L}}{dp}$, one can derive that [87]

$$\frac{df}{dp} = -h^{\dagger}(z_0)\frac{d\varphi}{dp}(z_0) + \int_{z_0}^{Z} dz\, \frac{\partial l}{\partial p} - h^{\dagger}\frac{\partial y}{\partial p}. \tag{A.14}$$

Note that in reality, l is usually discrete in z; i.e. we only have discrete observations:

$$l(\varphi, p) = \sum_i l_i(\varphi, p, o_i)\delta(z - z_i) \tag{A.15}$$

Hence, $\frac{\partial l}{\partial \varphi}$ in Eq. (A.13) should be applied as spikes. The integral on the right hand

side of Eq. (A.14) is equivalent to solve an ODE such that

$$\frac{da}{dz} = \frac{\partial l}{\partial p} - h^{\dagger}\frac{\partial y}{\partial p}, \tag{A.16}$$

where $a(Z)$ becomes the integration value.

Appendix B

Supplemental materials on comparison between optical scattering models

B.1 LSE as a composition of 2D Fourier transforms

In this section, we derive Eq. (3.5). Fourier transforming $\psi - \psi_0$ yields

$$\begin{aligned}
&\hat{\mathcal{F}}_{xy}\left[\psi(\boldsymbol{r}) - \psi_0(\boldsymbol{r})\right](k_x, k_y, z) \\
&= \int dx\, dy\, \mathrm{e}^{-ik_x x - ik_y y}\left[\psi(\boldsymbol{r}) - \psi_0(\boldsymbol{r})\right] \\
&= \int dx\, dy\, \mathrm{e}^{-ik_x x - ik_y y} \int d\boldsymbol{r}'\; G(\boldsymbol{r} - \boldsymbol{r}')V(\boldsymbol{r}')\psi(\boldsymbol{r}'). \qquad \text{(B.1)}
\end{aligned}$$

Using the Weyl expansion, Eq. (3.4), the Green's function can also be expressed as a 2D Fourier transform. Then we obtain

$$\begin{aligned}
&\hat{\mathcal{F}}_{xy}\left[\psi(\boldsymbol{r}) - \psi_0(\boldsymbol{r})\right](k_x, k_y, z) \\
&= \frac{i}{8\pi^2} \int dx\, dy \int d\boldsymbol{r}' \int dk_x' dk_y'\; \mathrm{e}^{-ik_x x - ik_y y}
\end{aligned}$$

$$
\begin{aligned}
&\times \frac{\mathrm{e}^{i\left(k'_x(x-x')+k'_y(y-y')+k'_z|z-z'|\right)}}{k'_z} V(\boldsymbol{r}')\psi(\boldsymbol{r}') \\
&= \frac{i}{8\pi^2}\int d\boldsymbol{r}' V(\boldsymbol{r}')\psi(\boldsymbol{r}') \int dk'_x dk'_y \\
&\times \frac{\mathrm{e}^{-i\left(k'_x x'+k'_y y'-k'_z|z-z'|\right)}}{k'_z} \underbrace{\int dx\, dy\ \mathrm{e}^{i\left(x(k'_x-k_x)+y(k'_y-k_y)\right)}}_{(2\pi)^2\delta(k_x-k'_x)\delta(k_y-k'_y)} \\
&= \frac{i}{2}\int d\boldsymbol{r}' V(\boldsymbol{r}')\psi(\boldsymbol{r}')\ \frac{\mathrm{e}^{-i(k_x x'+k_y y'-k_z|z-z'|)}}{k_z} \\
&= \frac{i}{2}\int dz'\ \frac{\mathrm{e}^{ik_z|z-z'|}}{k_z} \int dx'dy' V(\boldsymbol{r}')\psi(\boldsymbol{r}')\ \mathrm{e}^{-i(k_x x'+k_y y')} \\
&= \frac{i}{2}\int dz'\ \frac{\mathrm{e}^{ik_z|z-z'|}}{k_z} \hat{\mathcal{F}}_{xy}\left[V(\boldsymbol{r})\psi(\boldsymbol{r})\right](k_x, k_y, z'). \qquad \text{(B.2)}
\end{aligned}
$$

Taking the inverse Fourier transform in Eq. (B.2) finalizes the derivation leading to Eq. (3.5).

B.2 Potential bound for convergence of the Born series

Previous studies discuss how to estimate the operator norm of the LSE integral operator and thus guarantee the convergence of the Born series. For example, [69] requires

$$
2\int \max_{\theta,\phi} |V(r,\theta,\phi)|\, r dr < 1, \qquad \text{(B.3)}
$$

where r, θ, and ϕ are radial distance, polar angle, and azimuthal angle in the spherical coordinate system. Considering the simplest case, let us assume a Mie scattering condition in which a sphere of radius R scatters a plane wave. Then Eq. (B.3) becomes

$$
\left(\frac{n}{n_\mathrm{b}}\right)^2 < 1 + \frac{1}{(n_\mathrm{b}k_0R)^2}. \qquad \text{(B.4)}
$$

Similarly, [57] suggests

$$
\left(\frac{n}{n_\mathrm{b}}\right)^2 < 1 + \frac{1}{17/2(n_\mathrm{b}k_0R)^2 + 2\sqrt{74}(n_\mathrm{b}k_0R) + 105}. \qquad \text{(B.5)}
$$

By comparison, our discussion in Sec. 3.3.1 concludes that it is sufficient to satisfy

$$\left(\frac{n}{n_{\mathrm{b}}}\right)^2 < 1 + \frac{1}{2\sqrt{3}(n_{\mathrm{b}}k_0R)} \tag{B.6}$$

to make the Born series convergent. The scalar wave approximation already requires $n_{\mathrm{b}}k_0R \gg 1$, which means that $(n_{\mathrm{b}}k_0R)^2$ terms in Eqs. (B.3)-(B.5) increase quickly. This makes the estimation on the upper bound of n too close to 1. On the contrary, Eq. (B.6) shows the first-order dependency on $n_{\mathrm{b}}k_0R$, which relaxes the requirement on n.

B.3 Numerical comparison on different ξ in BPM's wave modulation

Based on the discussion in Chapter 3.4.2, we compare field estimations from different ξ in BPM. In Fig. B-1, it is shown that there is no significant difference in scattered amplitudes and the elementwise difference is less than one percent of maximum amplitude value. This can be quantitatively validated in Table B.1 where SSIM and PSNR exhibit very high values. Hence, we may conclude that $\xi = 1$ and $\xi = 2$ in the phase modulation term would not significantly influence the field estimation, except some unusual cases.

B.4 Comparison between FDTD and LSE

To test the estimation quality of LSE, we compare it with FDTD solutions from the Lumerical [1] 3D Electromagnetic Simulator. In Fig. (B-2), it can be observed that the high frequency interference patterns are approximated well by the LSE. The numerical difference in each voxel is less than one percent of the maximum amplitude value. In Table B.2, we list quantitative results considering six different potentials. These results further corroborate the validity of the LSE.

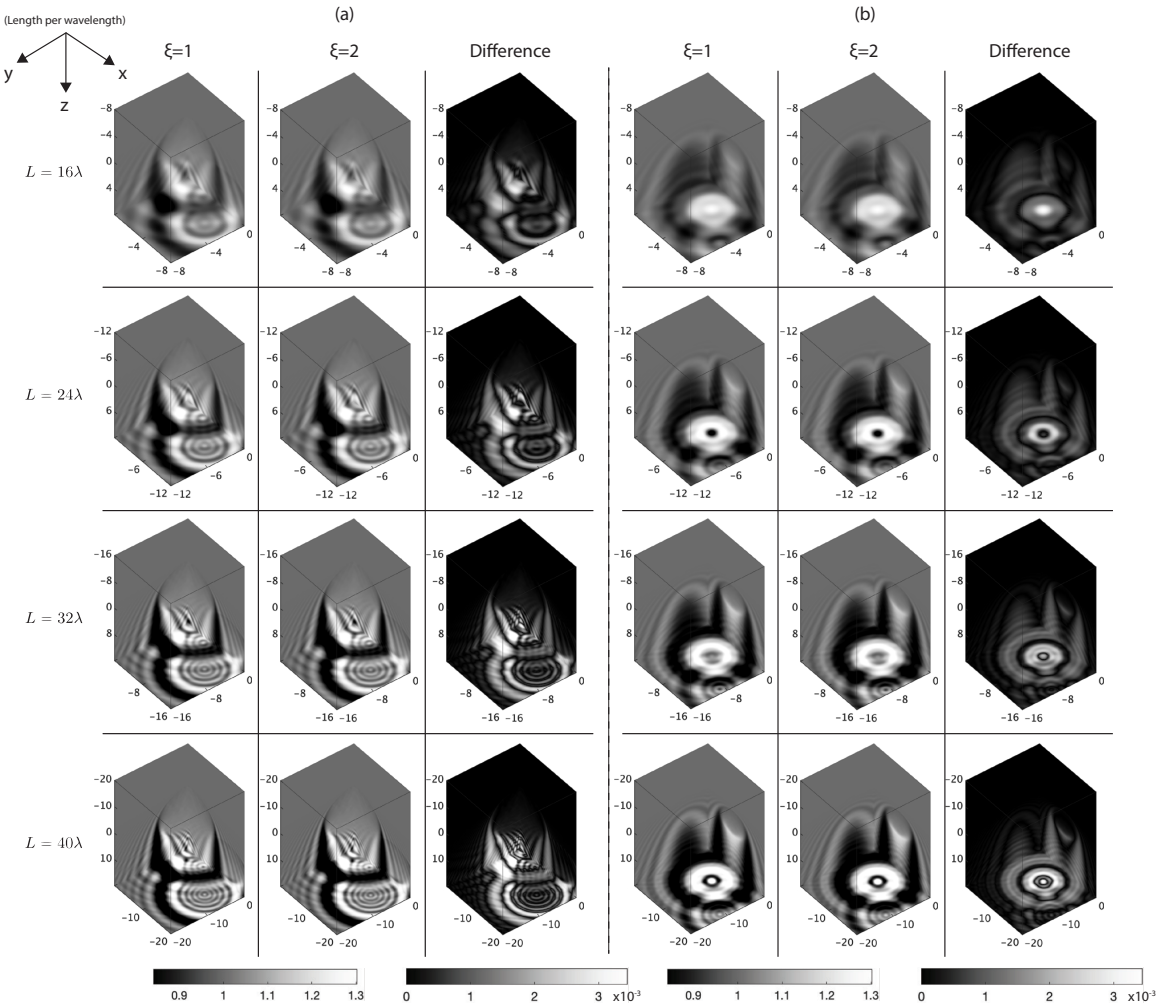

Figure B-1: Scattered fields estimated from BPM with different ξ choices: $\xi = 1$ and $\xi = 2$. We consider potentials which consist of spheres. The size L of a cubic computational box is changed from 16λ to 40λ. The mean refractive index of spherical potentials is 1.02. We show two different objects, which are marked with (a) and (b). Difference refers to the elementwise absolute error divided by the maximum field amplitude.

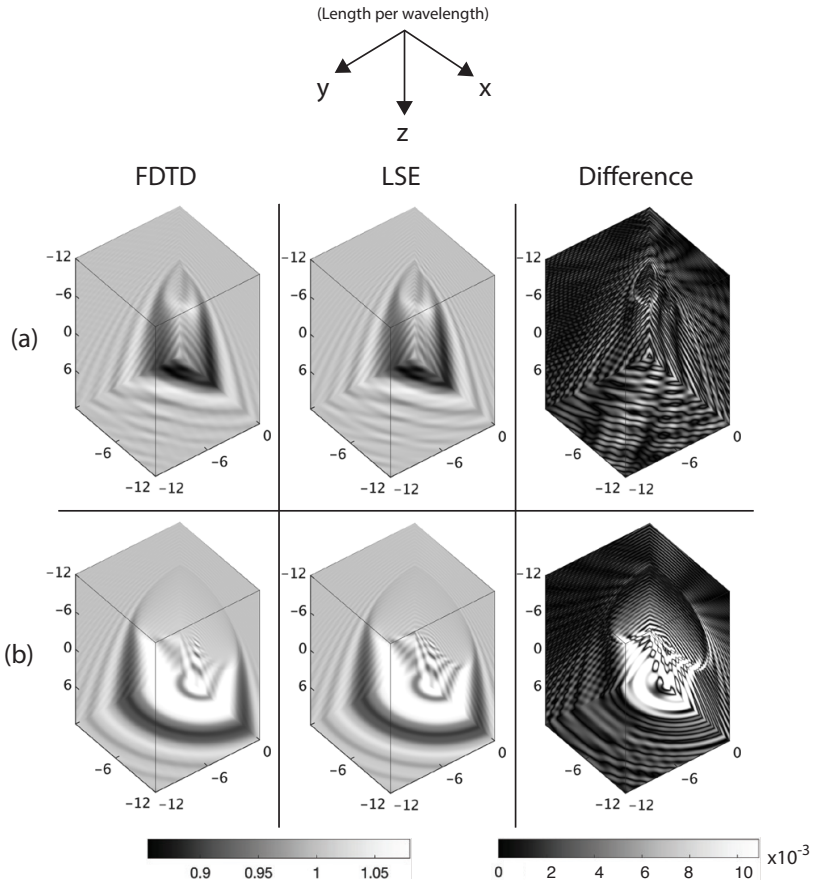

Figure B-2: Comparison of scattered fields from FDTD and LSE. Two different potentials are considered where the mean refractive index is 1.02. These potentials are marked with (a) and (b). Difference refers to the elementwise absolute error divided by the maximum field amplitude.

Table B.1: Image quality metrics on fields from BPM with $\xi = 1$ and $\xi = 2$ when the size L of a cubic computational box changes. We consider 15 different potentials which consist of spheres. The mean refractive index of spherical potentials is 1.02. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
$L = 16\lambda$, amplitude	1.000	64.461	2.911×10^{-4}
$L = 24\lambda$, amplitude	1.000	62.679	3.636×10^{-4}
$L = 32\lambda$, amplitude	1.000	62.884	4.319×10^{-4}
$L = 40\lambda$, amplitude	1.000	62.849	4.999×10^{-4}
$L = 16\lambda$, phase	1.000	91.573	1.944×10^{-5}
$L = 24\lambda$, phase	1.000	91.587	1.882×10^{-5}
$L = 32\lambda$, phase	1.000	91.601	1.841×10^{-5}
$L = 40\lambda$, phase	1.000	91.416	1.813×10^{-5}

Table B.2: Image quality metrics on fields from LSE and FDTD. We consider 6 different potentials which consist of spheres. The mean refractive index of spherical potentials is 1.02. The size of a cubic computational box is 24λ. The phase is unwrapped along the optical axis. The full width at half maximum of the Gaussian window in SSIM is $\lambda/2$.

	SSIM	PSNR	Relative L_1
Amplitude	0.982	42.162	3.592×10^{-3}
Phase	0.999	42.465	2.486×10^{-2}

Appendix C

Supplemental materials for neural regularization of LSE

C.1 Convolution integral without explicit zero-padding

where L_d and N_d represent the size and the number of grid points in the d-th dimension, and

$$\mathcal{I}_{\mathbf{N}} = \left\{ (j_1, j_2, j_3) \,:\, -\frac{N_d}{2} \leq j_d < \frac{N_d}{2},\ j_d \in \mathbb{Z},\ d \in \{1,2,3\} \right\}. \tag{C.1}$$

It is assumed that N_d is a even number. As in Eq. 3.3, vectors are denoted with boldfaced letters, e.g. $\mathbf{j} = (j_1, j_2, j_3)$ and $\mathbf{N} = (N_1, N_2, N_3)$.

Using the convolution theorem, the integral in Eq. 3.3 can be rewritten as

$$\psi(\mathbf{x}) - \psi_0(\mathbf{x}) = \xi(\mathbf{x}) = \mathcal{F}^{-1}\left(\tilde{G}(\boldsymbol{\nu})\tilde{\psi}(\boldsymbol{\nu}) \right) = \int_{\mathbb{R}^3} d\boldsymbol{\nu}\ e^{2\pi i \langle \boldsymbol{\nu}, \mathbf{x} \rangle} \tilde{G}(\boldsymbol{\nu})\tilde{\psi}(\boldsymbol{\nu}), \tag{C.2}$$

where $\tilde{\psi}$ refers to the Fourier transform of ψ and $\langle \cdot, \cdot \rangle$ represents the inner product. The direct discretization of Eq. C.2 is to assume that $\boldsymbol{\nu} = \mathrm{diag}\,(\mathbf{L})^{-1}\,\mathbf{m}$ where $\mathbf{m} \in \mathcal{I}_{\mathbf{M}}$ and $\mathrm{diag}\,(\mathbf{L})$ is a diagonal matrix whose diagonal elements are $\mathbf{L}$. Note that each entry of $\mathbf{M}$, which is the number of grid points in the Fourier space, should be larger than that of $\mathbf{N}$. This is because $\tilde{G}(\boldsymbol{\nu})$ is oscillatory in the Fourier space and one has to

prevent the periodic folding of ψ, which requires $\mathbf{M} = 4\mathbf{N}$, i.e. the zero-padding of a factor 4 in the spatial grid. In Chapter C.2, precomputation steps are adopted, which are suggested in [102, 84] to reduce the zero-padding factor from 4 to 2 and to mitigate singularities in the analytical expression of $\tilde{G}(\boldsymbol{\nu})$. Hence, at this moment, it is assumed that $\mathbf{M} = 2\mathbf{N}$ where N_d zeros are added to ψ in the d^{th} dimension.

Explicitly, Eq. C.2 is discretized as

$$\xi_p(\mathbf{j}) = \frac{1}{M_1 M_2 M_3} \sum_{\mathbf{m} \in \mathcal{I}_{\mathbf{M}}} e^{2\pi i \left\langle \operatorname{diag}(\mathbf{M})^{-1} \mathbf{m}, \mathbf{j} \right\rangle} \tilde{G}(\mathbf{m})\, \hat{\psi}_p(\mathbf{m}), \tag{C.3}$$

where the subscript p is used to denote the zero-padding and $\hat{\psi}_p$ is the discrete Fourier transform of ψ_p, which is a periodic extension of the zero-padded version of ψ:

$$\psi_p(\mathbf{j}) = \sum_{\mathbf{q} \in \mathbb{Z}^3} \psi_p(\mathbf{j} + \operatorname{diag}(\mathbf{M})\, \mathbf{q}), \tag{C.4}$$

where

$$\psi_p(\mathbf{j}) = \begin{cases} \psi(\mathbf{j}) & \mathbf{j} \in \mathcal{I}_{\mathbf{N}}, \\ 0 & \text{otherwise}, \end{cases} \tag{C.5}$$

for $\mathbf{j} \in D_{\mathbf{M}}$. Here, with a slight abuse of notation, the arguments of functions are set as discrete indices in the spatial and Fourier grids, since the spatial sampling rate L_d/N_d in each d^{th} dimension is considered fixed. However, Eq. C.5 indicates that the evaluation of Eq. C.3 requires memory 16 times larger than that to store $\psi(\mathbf{j})$ (8 from ξ_p and 8 from ψ_p). Furthermore, additional array slicing is required to make ψ_p from ψ and ξ from ξ_p, which may not be handled in a straight-forward manner especially with graphic processing units (GPU).

[9] suggests a way to calculate ξ in $D_{\mathbf{N}}$ without any explicit zero-padding. The basic idea is to decompose the summation over $\mathbf{M}$ to multiple summations over $\mathbf{N}$ in the discrete Fourier transform. All implementation examples discussed in [9] are described in terms of one-dimensional operations, but calling small array operations many times is not favorable compared to calling large operations small times in massively parallel setups, e.g. GPU. Moreover, in many numerical simulations, it is often

more convenient to store functions $\psi(\mathbf{j} \in \mathcal{I}_{\mathbf{N}})$ indexed as $\mathbf{j}+\mathbf{N}/2$, contiguously in numerical arrays. With these considerations, one can extend the idea as follows. Using the definition of the discrete Fourier transform,

$$\begin{aligned}
\hat{\psi}_p(\mathbf{m}-\mathbf{N}) &= \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{M}}} e^{-2\pi i\left\langle \mathrm{diag}(\mathbf{M})^{-1}(\mathbf{m}-\mathbf{N}),\mathbf{j}\right\rangle}\psi_p(\mathbf{j}) \\
&= \sum_{\mathbf{j}\in\mathcal{I}'_{\mathbf{N}}} e^{-2\pi i\left\langle \mathrm{diag}(\mathbf{M})^{-1}(\mathbf{m}-\mathbf{N}),\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right)\right\rangle}\psi_p\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right) \\
&= e^{\frac{\pi i}{2}(m_1+m_2+m_3)}e^{-\frac{\pi i}{2}(N_1+N_2+N_3)} \\
&\quad\times \sum_{\mathbf{j}\in\mathcal{I}'_{\mathbf{N}}} e^{-2\pi i\left\langle \mathrm{diag}(\mathbf{M})^{-1}\mathbf{m},\mathbf{j}\right\rangle}e^{\pi i(j_1+j_2+j_3)}\psi_p\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right)
\end{aligned} \tag{C.6}$$

where $\mathcal{I}'_{\mathbf{N}}$ is $\mathcal{I}_{\mathbf{N}}$ shifted by $\mathbf{N}/2$, i.e.

$$\mathcal{I}'_{\mathbf{N}} = \Big\{(j_1,j_2,j_3)\,:\,0\le j_d<N_d,\ j_d\in\mathbb{Z},\ d\in\{1,2,3\}\Big\}, \tag{C.7}$$

and $\mathbf{m}\in\mathcal{I}'_{\mathbf{M}}$. Similarly, Eq. C.3 can be rewritten as

$$\begin{aligned}
\xi_p\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right) &= \frac{1}{M_1M_2M_3}\sum_{\mathbf{m}\in\mathcal{I}'_{\mathbf{M}}} e^{2\pi i\left\langle \mathrm{diag}(\mathbf{M})^{-1}(\mathbf{m}-\mathbf{N}),\mathbf{j}-\frac{\mathbf{N}}{2}\right\rangle}\tilde{G}(\mathbf{m}-\mathbf{N})\,\hat{\psi}_p(\mathbf{m}-\mathbf{N}) \\
&= \frac{1}{M_1M_2M_3}e^{\frac{\pi i}{2}(N_1+N_2+N_3)}e^{-\pi i(j_1+j_2+j_3)} \\
&\quad\times \sum_{\mathbf{m}\in\mathcal{I}'_{\mathbf{M}}} e^{2\pi i\left\langle \mathrm{diag}(\mathbf{M})^{-1}\mathbf{m},\mathbf{j}\right\rangle}e^{-\frac{\pi i}{2}(m_1+m_2+m_3)}\tilde{G}(\mathbf{m}-\mathbf{N})\,\hat{\psi}_p(\mathbf{m}-\mathbf{N}),
\end{aligned} \tag{C.8}$$

where $\mathbf{j}\in\mathcal{I}'_{\mathbf{N}}$. To evaluate Eqs. C.6 and C.8 using FFT, one can decompose $\mathbf{m}$ as $2\mathbf{t}+\mathbf{s}$ where $\mathbf{t}\in\mathcal{I}'_{\mathbf{N}}$ and $\mathbf{s}\in\{0,1\}^3$, leading to

$$\begin{aligned}
\hat{\psi}_p\left(2\mathbf{t}+\mathbf{s}-\frac{\mathbf{N}}{2}\right) &= e^{\pi i(t_1+t_2+t_3)}e^{\frac{\pi i}{2}(s_1+s_2+s_3)}e^{-\frac{\pi i}{2}(N_1+N_2+N_3)} \\
&\quad\times \sum_{\mathbf{j}\in\mathcal{I}'_{\mathbf{N}}} e^{-2\pi i\left\langle \mathrm{diag}(\mathbf{N})^{-1}\mathbf{t},\mathbf{j}\right\rangle}e^{\pi i(j_1+j_2+j_3)}e^{-\pi i\left\langle \mathrm{diag}(\mathbf{N})^{-1}\mathbf{s},\mathbf{j}\right\rangle}\psi_p\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right),
\end{aligned} \tag{C.9}$$

and

$$\xi_p\left(\mathbf{j}-\frac{\mathbf{N}}{2}\right) = \frac{1}{M_1M_2M_3}e^{\frac{\pi i}{2}(N_1+N_2+N_3)}e^{-\pi i(j_1+j_2+j_3)} \times \sum_{\mathbf{s}\in\{0,1\}^3} e^{-\frac{\pi i}{2}(s_1+s_2+s_3)}e^{\pi i\left\langle \mathrm{diag}(\mathbf{N})^{-1}\mathbf{s},\mathbf{j}\right\rangle} \times \sum_{\mathbf{t}\in\mathcal{I}'_{\mathbf{N}}} e^{2\pi i\left\langle \mathrm{diag}(\mathbf{N})^{-1}\mathbf{t},\mathbf{j}\right\rangle}e^{-\pi i(t_1+t_2+t_3)}\tilde{G}(2\mathbf{t}+\mathbf{s}-\mathbf{N})\,\hat{\psi}_p(2\mathbf{t}+\mathbf{s}-\mathbf{N}). \tag{C.10}$$

Note that $\xi_p\left(\mathbf{j}-\mathbf{N}/2\right)$ and $\psi_p\left(\mathbf{j}-\mathbf{N}/2\right)$ corresponds to ξ and ψ in $D_{\mathbf{N}}$ that are contiguously indexed in numerical arrays, respectively. Eqs. C.9 and C.10 imply that the original padded convolution can be expressed as 2^3 convolutions without explicit padding. Terms such as $e^{\pm\pi i(t_1+t_2+t_3)}$, $e^{\pm\pi i(s_1+s_2+s_3)/2}$, and $e^{\pm\pi i(N_1+N_2+N_3)/2}$ are canceled each other, hence they can be omitted in the actual implementation. The other additional terms, e.g. $e^{\pm 2\pi i\left\langle \mathrm{diag}(\mathbf{N})^{-1}\mathbf{t},\mathbf{j}\right\rangle}$, only require small amount of memory, since they are separable in each dimension. Application to dimensions other than 3 is straightforward.

C.1.1 Application to vectorial optical scattering

In optics, Eq. 3.3 originates from the scalar Helmholtz equation where the birefringence is negligible. On the contrary, this section considers objects whose refractive index depends on the polarization. Under time-independent and non-magnetic systems, an eigenproblem for the electric field $\mathbf{E}$ can be derived from the Maxwell's equations as

$$\left[\nabla^2+(n_bk_0)^2+\nabla\nabla\cdot\right]\mathbf{E}(\mathbf{x}) = -(n_bk_0)^2\left[\left(\frac{\overline{\overline{n}}(\mathbf{x})}{n_b}\right)^2-1\right]\mathbf{E}(\mathbf{x}), \tag{C.11}$$

where $k_0=\|\mathbf{k}_0\|$ and $\overline{\overline{n}}$ is the refractive index tensor. The dyadic Green's function for Equation C.11 is

$$\overline{\overline{G}}(\mathbf{x}) = \mathrm{diag}\left(G(\mathbf{x})\right)+\frac{1}{(n_bk_0)^2}\overline{\overline{H}}G(\mathbf{x}), \tag{C.12}$$

where $\overline{\overline{H}}G$ is the Hessian matrix for G [23]. Here, two-rank tensors are described as matrices for notational simplicity. Subsequently, the response function in Eq. C.12 leads to an integral form of the eigenproblem,

$$\mathbf{E}(\mathbf{x}) = \mathbf{E}_0(\mathbf{x}) + \int \mathrm{d}\mathbf{x}' \, \overline{\overline{G}}\left(\mathbf{x} - \mathbf{x}'\right) \overline{\overline{V}}\left(\mathbf{x}'\right) \mathbf{E}\left(\mathbf{x}'\right), \tag{C.13}$$

where $\overline{\overline{V}}\left(\mathbf{x}\right) = \left(n_b k_0\right)^2 \left[\left(\frac{\overline{\overline{n}}(\mathbf{x})}{n_b}\right)^2 - 1\right]$ and $\mathbf{E}_0$ is the incident electric field.

Obviously, Eq. C.13 has a similar form to Eq. 3.3, which can be referred to as the vectorial LSE. Hence, a demodulation technique $\mathbf{E}(\mathbf{x}) = e^{i\langle n_b \mathbf{k}_0, \mathbf{x}\rangle} \boldsymbol{\mathcal{E}}(\mathbf{x})$ may also be applied under monochromatic illumination and $\boldsymbol{\mathcal{E}}$ would require a small number of Fourier components. However, unlike the scalar LSE, the computation of Eq. C.13 now suffers from additional singularities from the Hessian of G in $\overline{\overline{G}}$. In fact, G itself has singularities both in the spatial and Fourier domain, but they can be mitigated by the virtue of the Paley-Wiener theorem, as reviewed in Chapter C.2. Application of the same strategy on $\overline{\overline{G}}$ gives

$$\mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right) \overline{\overline{G}}(\mathbf{x}) = \mathrm{diag}\left(\mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right) G(\mathbf{x})\right) + \frac{1}{\left(n_b k_0\right)^2} \mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right) \overline{\overline{H}} G(\mathbf{x}), \tag{C.14}$$

where L_d is a constant larger than the diagonal length of the box $D_{\mathbf{N}}$. The first term on the right side of Eq. C.14 corresponds to a simple multi-dimensional extension of Chapter C.2, while it is not straightforward to derive the Fourier transform of the second term. Instead of numerically evaluating the rect function and the partial derivatives in the spatial grid, one can consider a practical trick. Due to the property of the rect function,

$$\mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right) \overline{\overline{H}} G(\mathbf{x}) \approx \overline{\overline{G}}_H \equiv \overline{\overline{H}} \left[\mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right) G(\mathbf{x})\right], \tag{C.15}$$

i.e. the rect function may be interchangeable with the partial derivatives except at the boundary of support of the rect function, which is barely touched by grid points in $D_{\mathbf{N}}$. Compared to the original Hessian, it is now straightforward to evaluate the Fourier transform of $\overline{\overline{G}}_H$.

C.1.2 Application to quantum scattering

In the Hartree atomic units, the time-independent and non-relativistic Schrödinger equation is written as

$$\left[\nabla^2 - \frac{2mP(\mathbf{x})}{m_e}\right]\psi(\mathbf{x}) = -k_0^2\psi(\mathbf{x}), \tag{C.16}$$

where m and m_e are the mass of a particle of interest and an electron, respectively. Specifically, Eq. C.16 considers an elastic scattering of a particle with an incident kinetic energy $\frac{\hbar^2 k_0^2}{2m}$ by a potential P where $\hbar$ is the reduced Planck constant. Since Eq. C.16 has the same form to the scalar Helmholtz equation in optics, the corresponding LSE can be derived under the radiation condition, i.e. by converting $V = 2mP/(m_e k_0^2)$. Similar interpretation on a scattering event is also possible, using the matter wave formulation in quantum mechanics.

However, compared to optical scattering, some additional issues should be addressed in quantum scattering. First, it is general that quantum mechanical potentials are not compactly supported in a strict sense. Hence, the naive application of the Paley-Wiener theorem is not valid, which requires different approaches to mitigate singularities in G, e.g. [6]. However, if P decays fast or can be approximated as zero at boundaries of D, it would not cause significant errors to apply window functions on P and to use methods in Chapter C.2. For example, one may consider a large D or a Coulomb potential with non-negligible screening effects [40]. Another issue regarding quantum scattering is to estimate the degree of refraction, which would require more sophisticated analysis than optical scattering. For example, one of the quantitative ways to deal with the refraction strength in quantum mechanics is to consider the angular dependency of the scattering cross section. On the other hand, one expects that empirical knowledge on quantum scattering might also be used, which would not cause critical errors in similar experimental setups to optical cases.

C.2 Precomputation steps for scalar Green's function

To solve the LSE using numerical techniques in Chapter C.1, the analytical expression on $\tilde{G}$ is required. This is not straightforward, since both G and $\tilde{G}$ contain singularities. In addition, the aperiodic convolution should be evaluated on $D_{4\mathbf{N}}$ to avoid the aliasing error, which consumes computational memory significantly. This section briefly reviews previous studies [102, 84] to mitigate these issues, following discussions in [81].

First, the singularities of G are mitigated in the following way. As V is a compactly supported function in D, Equation 3.3 can be re-expressed as

$$\psi(\mathbf{x}) = \psi_0(\mathbf{x}) + \int \mathrm{d}\mathbf{x}' \; G\left(\mathbf{x}-\mathbf{x}'\right) \mathrm{rect}\left(\frac{\mathbf{x}-\mathbf{x}'}{2L_d}\right) V\left(\mathbf{x}'\right) \psi\left(\mathbf{x}'\right), \tag{C.17}$$

where L_d is a constant larger than the maximum distance between any two points in D. For example, for a rectangular domain, L_d should be larger than the diagonal length of the three-dimensional box. Eq. C.17 implies a new kernel,

$$G\left(\mathbf{x}\right) \mathrm{rect}\left(\frac{\mathbf{x}}{2L_d}\right), \tag{C.18}$$

which is compactly supported. According to the Paley-Wiener theorem, the Fourier transform of this new kernel is entire. For the analytic form of the transform, refer to [81].

The zero-padding of factor 4 can be reduced to 2 via some precomputation steps. Such precomputation leverages a similar technique to Chapter C.1, which decompose a large Fourier grid into small grids. For simplicity in notation, consider the computation of $\xi(\mathbf{j})$ in a one-dimensional space. Extension to higher dimensions is straightforward. Then,

$$\xi(j) = \frac{1}{4N} \sum_{k \in [0,4N)_{\mathbb{Z}}} \tilde{G}[k] \hat{\psi}_p[k] \exp\left(\frac{2\pi i}{4N} jk\right)$$

$$
\begin{aligned}
&= \frac{1}{4N} \sum_{k\in[0,4N)_{\mathbb{Z}}} \tilde{G}[k] \sum_{q\in[0,4N)_{\mathbb{Z}}} \psi_p[q] \exp\left(-\frac{2\pi i}{4N} qk\right) \exp\left(\frac{2\pi i}{4N} jk\right) \\
&= \frac{1}{4N} \sum_{k\in[0,4N)_{\mathbb{Z}}} \tilde{G}[k] \sum_{q\in[0,\frac{N}{2})_{\mathbb{Z}} \bigcup [4N-\frac{N}{2},4N)_{\mathbb{Z}}} \psi_p[q] \exp\left(\frac{2\pi i}{4N} k(j-q)\right) \\
&= \frac{1}{4N} \sum_{q\in[0,\frac{N}{2})_{\mathbb{Z}} \bigcup [4N-\frac{N}{2},4N)_{\mathbb{Z}}} \psi_p[q] \sum_{k\in[0,4N)_{\mathbb{Z}}} \tilde{G}[k] \exp\left(\frac{2\pi i}{4N} k(j-q)\right) \\
&= \frac{1}{4N} \sum_{q\in[0,\frac{N}{2})_{\mathbb{Z}} \bigcup [4N-\frac{N}{2},4N)_{\mathbb{Z}}} \psi_p[q] \sum_{k\in[0,2N)_{\mathbb{Z}}} \sum_{s\in[0,1]_{\mathbb{Z}}} \tilde{G}[2k-s] \exp\left(\frac{2\pi i}{4N} (2k-s)(j-q)\right) \\
&= \frac{1}{4N} \sum_{q\in[0,\frac{N}{2})_{\mathbb{Z}} \bigcup [4N-\frac{N}{2},4N)_{\mathbb{Z}}} \psi_p[q] \\
&\quad \times \sum_{k\in[0,2N)_{\mathbb{Z}}} \sum_{s\in[0,1]_{\mathbb{Z}}} \tilde{G}[2k-s] \exp\left(\frac{2\pi i}{4N} 2k(j-q)\right) \exp\left(-\frac{2\pi i}{4N} s(j-q)\right),
\end{aligned}
\tag{C.19}
$$

where a subscript $\mathbb{Z}$ is used to denote that only integers are considered in an interval. Let us further denote that

$$
G^{(s)}[q] \equiv \frac{1}{2N} \exp\left(-\frac{2\pi i}{4N} sq\right) \sum_{k\in[0,2N)_{\mathbb{Z}}} \tilde{G}[2k-s] \exp\left(\frac{2\pi i}{2N} kq\right), \tag{C.20}
$$

which is just a discrete inverse Fourier transform of $\tilde{G}[2k-s]$ indexed at $k \in [0, 2N)_{\mathbb{Z}}$, modulated by a factor of $\exp\left(-\frac{2\pi i}{4N} sq\right)$ at each index q. Subsequently, Eq. C.19 can be formulated as

$$
\xi(j) = \frac{1}{2} \sum_{s\in[0,1]_{\mathbb{Z}}} \sum_{q\in[0,N)_{\mathbb{Z}} \bigcup [3N,4N)_{\mathbb{Z}}} \psi_p[q] \hat{\tilde{G}}^{(s)}[j-q], \tag{C.21}
$$

because $G^{(s)}$ has a period of $2N$. Eq. C.21 is simply a aperiodic convolution, effectively in a domain $D_{\mathbf{2N}}$. Hence, with a preconditioning step as in Eq. C.20, one can reduce the number of zeros from $3N_d$ to N_d in each d^{th} dimension.

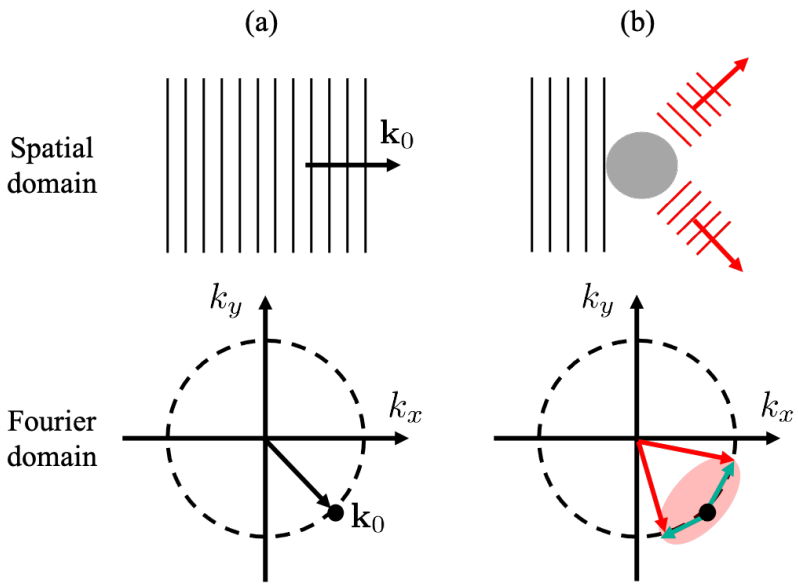

Figure C-1: Illustration of optical scattering in the spatial and Fourier domain. Dotted circles represent the Ewald sphere. (a) A monochromatic illumination with the wavevector $\mathbf{k}_0$ can be approximately represented by low-frequency wavefront in the spatial domain while it would appear as a delta-like peak in the Fourier domain. (b) As the illumination approaches an object (gray circle), it is refracted from the original illumination direction, denoted by green arrows. These green arrows indicate the degree of refraction. The scattered field has new wavevectors represented by red arrows and its Fourier transform is mostly supported on the area covered by the red circle.

C.2.1 LSE with demodulated fields

Though the memory requirement of the LSE can be reduced by the decomposition of the padded FFT, it still significantly depends on the size scale of scattering objects with respect to the incident field wavelength. In other words, the number of grid points becomes large as we describe high-frequency scattered fields in large objects.

Under monochromatic illumination, scattering of the incident field is described by the refraction from the original wavevector direction as illustrated in Fig. C-1. Accordingly, the scattered field ψ can be demodulated as illumination and refraction parts,

$$\psi(\mathbf{x}) = e^{i\langle n_b \mathbf{k}_0, \mathbf{x}\rangle}\varphi(\mathbf{x}), \tag{C.22}$$

where $\mathbf{k}_0$ and n_b represent the vacuum wavevector of illumination and the background

refractive index. Similarly, $\psi_0(\mathbf{x}) = e^{i\langle n_b \mathbf{k}_0, \mathbf{x}\rangle}\varphi_0(\mathbf{x})$ where $\varphi_0(\mathbf{x})$ is a low-frequency wavefront of illumination. Compared to the original field ψ, it is expected that the Fourier transform of φ is centered at the origin and decays fast as the size scale of V becomes large with respect to the wavelength [82]. That the refraction would not be significant is based on physics, which is naively validated with the lens equation. These considerations imply that it is more convenient to use φ to describe scattering. In fact, one can derive the LSE with respect to φ by dividing the both sides of Eq. 3.3 with the high-frequency carrier field $e^{i\langle n_b \mathbf{k}_0, \mathbf{x}\rangle}$,

$$\varphi(\mathbf{x}) = \varphi_0(\mathbf{x}) + \int \mathrm{d}\mathbf{x}' \; \mathcal{G}\left(\mathbf{x} - \mathbf{x}'\right) V\left(\mathbf{x}'\right) \varphi\left(\mathbf{x}'\right), \tag{C.23}$$

where

$$\mathcal{G}(\mathbf{x}) = G\left(\mathbf{x}\right) e^{-i\langle n_b \mathbf{k}_0, \mathbf{x}\rangle}. \tag{C.24}$$

Eq. C.23 is a low-frequency version of the original LSE and the sampling rate to prevent significant aliasing error depends on not $n_b\|\mathbf{k}_0\|$ but the degree of refraction from objects V. It is straightforward to apply Eqs. C.9 and C.10 to this low-frequency LSE. ψ is substituted by φ and we consider the analytical Fourier transform of the new kernel $\mathcal{G}$ instead of G.

An exact description on the degree of refraction (green arrows in Fig. C-1) may not be simple. Hence, determination of $\mathbf{N}$ given $\mathbf{L}$ for φ would not always be completely concrete, but we expect that the degree of refraction can be deduced from empirical knowledge without critical errors. If Eq. C.23 is applied on the inverse scattering, information on the Fourier components of φ is retrieved from intensity/field at detectors. However, the effective maximum frequency of φ at optical detectors is further limited by the optical bandwidth of systems, which depends on aperture setups and detector pixel sizes, etc. Hence, it can be anticipated that moderate estimation on the degree of refraction would be sufficient in real experiments.

Bibliography

[1] Lumerical Inc. `https://www.lumerical.com/`.

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[3] Jens Hjorleifur Bardarson, Ingibjorg Magnusdottir, Gudny Gudmundsdottir, Chi-Shung Tang, Andrei Manolescu, and Vidar Gudmundsson. Coherent electronic transport in a multimode quantum channel with Gaussian-type scatterers. *Physical Review B*, 70(24):245308, 2004.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] Dimitri P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 2014.

[6] Gregory Beylkin, Christopher Kurcz, and Lucas Monzón. Fast convolution with the free space Helmholtz Green's function. *Journal of Computational Physics*, 228(8):2770–2791, 2009.

[7] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 6 edition, 2013.

[8] Emrah Bostan, Emmanuel Froustey, Masih Nilchian, Daniel Sage, and Michael Unser. Variational phase imaging using the transport-of-intensity equation. *IEEE Transactions on Image Processing*, 25(2):807–817, 2015.

[9] John C. Bowman and Malcolm Roberts. Efficient dealiased convolutions without padding. *SIAM Journal on Scientific Computing*, 33(1):386–406, 2011.

[10] H. Bremmer. On the asymptotic evaluation of diffraction integrals with a special view to the theory of defocusing and optical contrast. *Physica*, 18(6-7):469–485, 1952.

[11] K.-H. Brenner and W. Singer. Light propagation through microlenses: a new simulation method. *Applied Optics*, 32(26):4984–4988, 1993.

[12] Oscar P. Bruno and E. McKay Hyde. Higher-order Fourier approximation in scattering by two-dimensional, inhomogeneous media. *SIAM Journal on Numerical Analysis*, 42(6):2298–2319, 2005.

[13] Tan Bui-Thanh. Adjoint and its roles in sciences, engineering, and mathematics: A tutorial. *arXiv preprint arXiv:2306.09917*, 2023.

[14] John Buker and George Kirczenow. Two-probe theory of scanning tunneling microscopy of single molecules: Zn (II)-etioporphyrin on alumina. *Physical Review B*, 72(20):205338, 2005.

[15] Florian Bürgel, Kamil S Kazimierski, and Armin Lechleiter. A sparsity regularization and total variation based computational framework for the inverse medium problem in scattering. *Journal of Computational Physics*, 339:1–30, 2017.

[16] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[17] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[18] Yang Cao, Shengtai Li, Linda Petzold, and Radu Serban. Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution. *SIAM Journal on Scientific Computing*, 24(3):1076–1089, 2003.

[19] Tony F. Chan, Efstratios Gallopoulos, Valeria Simoncini, Tedd Szeto, and Charles H. Tong. A quasi-minimal residual variant of the Bi-CGSTAB algorithm for nonsymmetric systems. *SIAM Journal on Scientific Computing*, 15(2):338–347, 1994.

[20] J. H. Chen and D. Van Dyck. Accurate multislice theory for elastic electron scattering in transmission electron microscopy. *Ultramicroscopy*, 70(1-2):29–44, 1997.

[21] Michael Chen, David Ren, Hsiou-Yuan Liu, Shwetadwip Chowdhury, and Laura Waller. Multi-layer Born multiple-scattering model for 3D phase microscopy. *Optica*, 7(5):394–403, 2020.

[22] Jing Cheng, Haifeng Wang, Leslie Ying, and Dong Liang. Model learning: Primal dual networks for fast MR imaging. In *Medical Image Computing and Computer Assisted Intervention*, pages 21–29. Springer, 2019.

[23] Weng Cho Chew. *Waves and Fields in Inhomogenous Media*, volume 16 of *IEEE Press Series on Electromagnetic Wave Theory*. John Wiley & Sons, 1999.

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[25] David Colton and Rainer Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93. Springer Nature, 2019.

[26] J. M. Cowley. *Diffraction Physics*. North-Holland, 3 edition, 1995.

[27] John M Cowley and A. F. Moodie. The scattering of electrons by atoms and crystals. I. A new theoretical approach. *Acta Crystallographica*, 10(10):609–619, 1957.

[28] Kevin Cowtan. Phase problem in X-ray crystallography, and its solution. *eLS*, 2001.

[29] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[30] Maarten V. de Hoop, Josselin Garnier, Sean F. Holman, and Knut Sølna. Scattering enabled retrieval of Green's functions from remotely incident wave packets using cross correlations. *Comptes Rendus Geoscience*, 343(8-9):526–532, 2011.

[31] Maarten V. De Hoop, Sean F. Holman, Hart F. Smith, and Gunther Uhlmann. Regularity and multi-scale discretization of the solution construction of hyperbolic evolution equations with limited smoothness. *Applied and Computational Harmonic Analysis*, 33(3):330–353, 2012.

[32] Ernest D. Eason. A review of least-squares methods for solving partial differential equations. *International journal for numerical methods in engineering*, 10(5):1021–1046, 1976.

[33] M. Ekstrom. A spectral characterization of the ill-conditioning in numerical deconvolution. *IEEE Transactions on Audio and Electroacoustics*, 21(4):344–348, 1973.

[34] Michael Elad and Michal Aharon. Image denoising via learned dictionaries and sparse representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 895–900. IEEE, 2006.

[35] Michael W. Farn and Joseph W. Goodman. Comparison of Rayleigh–Sommerfeld and Fresnel solutions for axial points. *Journal of the Optical Society of America A*, 7(5):948–950, 1990.

[36] M. D. Feit and J. A. Fleck. Light propagation in graded-index optical fibers. *Applied Optics*, 17(24):3990–3998, 1978.

[37] M. D. Feit and J. A. Fleck. Beam nonparaxiality, filament formation, and beam breakup in the self-focusing of optical beams. *Journal of Optical Society of America B*, 5(3):633–640, 1988.

[38] Aires Ferreira, J. Viana-Gomes, Johan Nilsson, Eduardo R. Mucciolo, Nuno M. R. Peres, and A. H. Castro Neto. Unified description of the dc conductivity of monolayer and bilayer graphene at finite densities based on resonant scatterers. *Physical Review B*, 83(16):165402, 2011.

[39] Aditi Ghai, Cao Lu, and Xiangmin Jiao. A comparison of preconditioned Krylov subspace methods for large-scale nonsymmetric linear systems. *Numerical Linear Algebra with Applications*, 26(1):e2215, 2019.

[40] Cristina E. González-Espinoza, Paul W. Ayers, Jacek Karwowski, and Andreas Savin. Smooth models for the Coulomb potential. *Theoretical Chemistry Accounts*, 135(12):256, 2016.

[41] P. Goodman and A. F. Moodie. Numerical evaluations of N-beam wave functions in electron scattering by the multi-slice method. *Acta Crystallographica Section A*, 30(2):280–290, 1974.

[42] Nicholas I. M. Gould, Jennifer A. Scott, and Yifan Hu. A numerical evaluation of sparse direct solvers for the solution of large sparse symmetric linear systems of equations. *ACM Transactions on Mathematical Software*, 33(2):10–es, 2007.

[43] Alexandre Goy, Girish Rughoobur, Shuai Li, Kwabena Arthur, Akintunde I Akinwande, and George Barbastathis. High-resolution limited-angle phase tomography of dense layered objects using deep neural networks. *Proceedings of the National Academy of Sciences*, 116(40):19848–19856, 2019.

[44] Harshit Gupta, Kyong Hwan Jin, Ha Q. Nguyen, Michael T. McCann, and Michael Unser. CNN-based projected gradient descent for consistent CT image reconstruction. *IEEE transactions on medical imaging*, 37(6):1440–1453, 2018.

[45] Masoud Hajarian. The generalized QMRCGSTAB algorithm for solving Sylvester-transpose matrix equations. *Applied Mathematics Letters*, 26(10):1013–1017, 2013.

[46] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[48] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991.

[49] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[50] Thorsten Hohage. Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem. *Journal of Computational Physics*, 214(1):224–238, 2006.

[51] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, 2010.

[52] Kazuo Ishizuka and Natsu Uyeda. A new theoretical and practical approach to the multislice method. *Acta Crystallographica Section A*, 33(5):740–749, 1977.

[53] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[54] Ulugbek S. Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.

[55] Ulugbek S. Kamilov, Ioannis N. Papadopoulos, Morteza H. Shoreh, Alexandre Goy, Cedric Vonesch, Michael Unser, and Demetri Psaltis. Learning approach to optical tomography. *Optica*, 2(6):517–522, 2015.

[56] Ulugbek S. Kamilov, Ioannis N. Papadopoulos, Morteza H. Shoreh, Alexandre Goy, Cedric Vonesch, Michael Unser, and Demetri Psaltis. Optical tomographic image reconstruction based on beam propagation and sparse regularization. *IEEE Transactions on Computational Imaging*, 2(1):59–70, 2016.

[57] Kimberly Kilgore, Shari Moskow, and John C. Schotland. Convergence of the Born and inverse Born series for electromagnetic scattering. *Applicable Analysis*, 96(10):1737–1748, 2017.

[58] Diederik P. Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[59] Earl J. Kirkland. *Advanced computing in electron microscopy.* Springer, 2 edition, 2010.

[60] Benjamin Krüger, Thomas Brenner, and Alwin Kienle. Solution of the inhomogeneous Maxwell's equations using a Born series. *Optics Express*, 25(21):25165–25182, 2017.

[61] James R. Kuttler and G. Daniel Dockery. Theoretical description of the parabolic approximation/Fourier split-step method of representing electromagnetic propagation in the troposphere. *Radio Science*, 26(2):381–393, 1991.

[62] Mikhail Aleksandrovich Leontovich and Vladimir Aleksandrovich Fock. Solution of the problem of propagation of electromagnetic waves along the earth's surface by the method of parabolic equation. *Journal of Physics – USSR*, 10(1):13–23, 1946.

[63] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466, 2018.

[64] Joowon Lim, Ahmed B. Ayoub, Elizabeth E. Antoine, and Demetri Psaltis. High-fidelity optical diffraction tomography of multiple scattering samples. *Light: Science & Applications*, 8(1):1–12, 2019.

[65] Joowon Lim and Demetri Psaltis. MaxwellNet: Physics-driven deep neural network training based on Maxwell's equations. *APL Photonics*, 7(1), 2022.

[66] Hsiou-Yuan Liu, Dehong Liu, Hassan Mansour, Petros T. Boufounos, Laura Waller, and Ulugbek S. Kamilov. SEAGLE: Sparsity-driven image reconstruction under multiple scattering. *IEEE Transactions on Computational Imaging*, 4(1):73–86, 2017.

[67] Stéphane Mallat. *A wavelet tour of signal processing.* Elsevier, 1999.

[68] Mikail Malovichko, Nikolay Khokhlov, Nikolay Yavich, and Mikhail Zhdanov. Approximate solutions of acoustic 3D integral equation and their application to seismic modeling and full-waveform inversion. *Journal of Computational Physics*, 346:318–339, 2017.

[69] Irwin Manning. Error and convergence bounds for the Born expansion. *Physical Review*, 139(2B):B495, 1965.

[70] Morteza Mardani, Hatef Monajemi, Vardan Papyan, Shreyas Vasanawala, David Donoho, and John Pauly. Recurrent generative adversarial networks for proximal learning and automated compressive image recovery. *arXiv preprint arXiv:1711.10046*, 2017.

[71] Daniel L. Marks. A family of approximations spanning the Born and Rytov scattering series. *Optics Express*, 14(19):8837–8848, 2006.

[72] Monika A. M. Marte and Stig Stenholm. Paraxial light and atom optics: the optical schrödinger equation and beyond. *Physical Review A*, 56(4):2940, 1997.

[73] Paul A. Martin. Acoustic scattering by inhomogeneous obstacles. *SIAM Journal on Applied Mathematics*, 64(1):297–308, 2003.

[74] Soheil Mehrabkhani and Thomas Schneider. Is the Rayleigh-Sommerfeld diffraction always an exact reference for high speed diffraction algorithms? *Optics Express*, 25(24):30229–30240, 2017.

[75] Tim Meinhardt, Michael Moller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1781–1790, 2017.

[76] F. Natterer. An error bound for the Born approximation. *Inverse problems*, 20(2):447–452, 2004.

[77] T. J. P. M. Op't Root and C. C. Stolk. One-way wave propagation with amplitude based on pseudo-differential operators. *Wave Motion*, 47(2):67–84, 2010.

[78] Gerwin Osnabrugge, Saroch Leedumrongwatthanakun, and Ivo M. Vellekoop. A convergent Born series for solving the inhomogeneous Helmholtz equation in arbitrarily large media. *Journal of Computational Physics*, 322:113–124, 2016.

[79] D. M. Paganin. *Coherent X-ray optics*. Number 6. Oxford University Press, 2006.

[80] David Paganin and Keith A. Nugent. Noninterferometric phase imaging with partially coherent light. *Physical Review Letters*, 80(12):2586, 1998.

[81] Subeen Pang. Machine learning regularized solution of the lippmann-schwinger equation. Master's thesis, Massachusetts Institute of Technology, 2021.

[82] Subeen Pang and George Barbastathis. Unified treatment of exact and approximate scalar electromagnetic wave scattering. *Physical Review E*, 106(4):045301, 2022.

[83] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[84] Thanh-An Pham, Emmanuel Soubies, Ahmed Ayoub, Joowon Lim, Demetri Psaltis, and Michael Unser. Three-dimensional optical diffraction tomography with Lippmann-Schwinger model. *IEEE Transactions on Computational Imaging*, 6:727–738, 2020.

[85] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[86] Harry Pratt, Bryan Williams, Frans Coenen, and Yalin Zheng. Fcnn: Fourier convolutional neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 786–798. Springer, 2017.

[87] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

[88] J. Van Roey, J. van der Donk, and P. E. Lagasse. Beam-propagation method: analysis and assessment. *Journal of the Optical Society of America*, 71(7):803–810, 1981.

[89] T. Konstantin Rusch and Siddhartha Mishra. UnICORNN: A recurrent model for learning very long time dependencies. In *International Conference on Machine Learning*, pages 9168–9178. PMLR, 2021.

[90] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[91] Jelena A. Schmalz, Gerd Schmalz, Timur E. Gureyev, and Konstantin M. Pavlov. On the derivation of the Green's function for the Helmholtz equation using generalized functions. *American Journal of Physics*, 78(2):181–186, 2010.

[92] Gerard L. G. Sleijpen and Diederik R. Fokkema. BiCGstab (l) for linear equations involving unsymmetric matrices with complex spectrum. *Electronic Transactions on Numerical Analysis*, 1(11):2000, 1993.

[93] Jakob Spiegelberg and Ján Rusz. A multislice theory of electron scattering in crystals including backscattering and inelastic effects. *Ultramicroscopy*, 159:11–18, 2015.

[94] J.-L. Starck and Mohamed-Jalal Fadili. An overview of inverse problem regularization using sparsity. In *IEEE International Conference on Image Processing*, pages 1453–1456. IEEE, 2009.

[95] Waleed Tahir, Ulugbek S. Kamilov, and Lei Tian. Holographic particle localization under multiple scattering. *Advanced Photonics*, 1(3):036003, 2019.

[96] Fred D. Tappert. The parabolic approximation method. *Wave Propagation and Underwater Acoustics*, 70:224, 1977.

[97] Michael Reed Teague. Deterministic phase retrieval: a Green's function solution. *Journal of Optical Society of America*, 73(11):1434–1441, 1983.

[98] David J. Thomson and N. R. Chapman. A wide-angle split-step algorithm for the parabolic equation. *The Journal of the Acoustical Society of America*, 74(6):1848–1854, 1983.

[99] Ch. Tsitouras. Runge–Kutta pairs of order 5 (4) satisfying only the first column simplifying assumption. *Computers & Mathematics with Applications*, 62(2):770–775, 2011.

[100] Gennadi Vainikko. Fast solvers of the Lippmann-Schwinger equation. In *Direct and inverse problems of mathematical physics*, pages 423–440. Springer, 2000.

[101] Henk A. Van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992.

[102] Felipe Vico, Leslie Greengard, and Miguel Ferrando. Fast convolution with free-space Green's functions. *Journal of Computational Physics*, 323:191–203, 2016.

[103] Laura Waller, Lei Tian, and George Barbastathis. Transport of intensity phase-amplitude imaging with higher order intensity derivatives. *Optics express*, 18(12):12552–12561, 2010.

[104] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[105] Emil Wolf and E. W. Marchand. Comparison of the Kirchhoff and the Rayleigh–Sommerfeld theories of diffraction at an aperture. *Journal of the Optical Society of America A*, 54(5):587–594, 1964.

[106] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. ADMM-CSNet: A deep learning approach for image compressive sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):521–538, 2018.

[107] Lexing Ying. Sparsifying preconditioner for the Lippmann–Schwinger equation. *Multiscale Modeling & Simulation*, 13(2):644–660, 2015.

[108] Leonardo Zepeda-Núnez and Hongkai Zhao. Fast alternating bidirectional preconditioner for the 2D high-frequency Lippmann–Schwinger equation. *SIAM Journal on Scientific Computing*, 38(5):B866–B888, 2016.

[109] Jialin Zhang, Qian Chen, Jiasong Sun, Long Tian, and Chao Zuo. On a universal solution to the transport-of-intensity equation. *Optics Letters*, 45(13):3649–3652, 2020.

[110] Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018.

[111] Pei Zhang, Lu Bai, Zhensen Wu, and Lixin Guo. Applying the parabolic equation to tropospheric groundwave propagation: A review of recent achievements and significant milestones. *IEEE Antennas and Propagation Magazine*, 58(3):31–44, 2016.

[112] Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou. Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, 13(3):226–234, 2016.

[113] Haowen Zhou, Haiyun Guo, and Partha P. Banerjee. Non-recursive transport of intensity phase retrieval with the transport of phase. *Applied Optics*, 61(5):B190–B199, 2022.

[114] Chao Zuo, Jiaji Li, Jiasong Sun, Yao Fan, Jialin Zhang, Linpeng Lu, Runnan Zhang, Bowen Wang, Lei Huang, and Qian Chen. Transport of intensity equation: a tutorial. *Optics and Lasers in Engineering*, 135:106187, 2020.