3D FINITE-DIFFERENCE FREQUENCY-DOMAIN METHOD FOR PLASMONICS AND NANOPHOTONICS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Wonseok Shin
August 2013

This dissertation is online at: http://purl.stanford.edu/jq442vm4326

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Shanhui Fan, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

David Miller, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jelena Vuckovic

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The finite-difference frequency-domain (FDFD) method is a conceptually simple method to solve time-dependent differential equations for steady-state solutions. In solving Maxwell's equations in three-dimensional (3D) space, however, the FDFD method has not been a popular method due to the slow convergence of iterative methods of solving a large system of linear equations $Ax = b$ constructed by the FDFD method. In this dissertation, we show that the convergence speed can be greatly accelerated for plasmonic and nanophotonic systems by carefully modifying the properties of A. First, we make the matrix A significantly better-conditioned by using the stretched-coordinate perfectly matched layer (SC-PML) rather than the more commonly used uniaxial PML (UPML) as an absorbing boundary. Second, we eliminate the high multiplicity of near-zero eigenvalues of A by utilizing the continuity equation. By combining these two techniques, we achieve 300-fold acceleration in the convergence of iterative methods for an example 3D plasmonic system. We also demonstrate successful application of the acceleration techniques to a real-world engineering problem of designing novel integrated optical circuit components, namely broadband sharp 90-degree bends and T-splitters, in plasmonic coaxial waveguides.

Topics

- Frequency-domain Maxwell's equations
- Finite-difference method to approximate differential equations
- Iterative methods to solve a system of linear equations $Ax = b$
- Preconditioners to accelerate iterative methods
- Eigenvalues and eigenvectors; singular values and singular vectors
- Plasmonics and nanophotonics

Denn ich hatte ja längst aus meinen Erfahrungen im Münchner Bürgerkrieg gelernt, daß man eine politische Richtung nie nach den Zielen beurteilen darf, die sie laut verkündet und vielleicht auch wirklich anstrebt, sondern nur nach den Mitteln, die sie zu ihrer Verwirklichung einsetzt. Schlechte Mittel beweisen ja, daß die Urheber an die Überzeugungskraft ihrer These selbst nicht mehr glauben.

Werner K. Heisenberg (1901–1976)
in *Der Teil und das Ganze*

For if I had learned one thing from my experiences during the civil war, it was that one must never judge a political movement by the aims it so loudly proclaims and perhaps genuinely strives to attain, but only by the means it uses to achieve them. The choice of bad means simply proves that those responsible have lost faith in the persuasive force of their original arguments.

Werner K. Heisenberg (1901–1976)
in *The Part and the Whole*

To my parents, parents-in-law
and to Kyuwon

Preface

For the last few decades around the dawn of the 21st century, photonics has emerged as a viable solution to ever increasing demands for faster communication of larger amounts of data. Because photonics uses electromagnetic waves as information carriers, efficient numerical solution of Maxwell's equations, which govern all electromagnetic phenomena, has become more and more important for designing integrated photonic devices and discovering novel optical phenomena to further advance photonics technology.

This dissertation is about one such numerical method to solve Maxwell's equations: the finite-difference frequency-domain (FDFD) method [1–3]. The FDFD method transforms the frequency-domain Maxwell's equations into numerically solvable forms using the finite-difference method. Compared with the other numerical methods to solve the frequency-domain equations such as the finite element method (FEM) [4] and method of moments (MoM) [5$^{\text{Ch. 2}}$], the FDFD method has an advantage in its conceptual simplicity, because the finite-difference method simply approximates derivatives (in the form of dy/dx) by ratios between two finite differences (in the form of $\Delta y/\Delta x$). The conceptual simplicity eventually leads to efficiency in parallel computing environment for large three-dimensional (3D) problems.

The FDFD method can also be compared with the finite-difference time-domain (FDTD) method [6]. Because one method solves the frequency-domain equations and the other solves the time-domain equations, the two methods have different domains of application. In terms of implementation, however, they are quite similar in that both use the finite-difference method. Despite this similarity, the FDFD method has been far less popular than the FDTD method in solving 3D problems. This dissertation addresses a few difficulties in implementing an efficient 3D FDFD solver of Maxwell's equations, and shows that once it is

implemented correctly it can be a practical method for solving 3D problems.

Especially, this dissertation aims to apply the FDFD method to plasmonic and nanophotonic systems. For these systems, the critical dimensions of objects are typically much smaller than a wavelength, and finite-difference grid cells smaller than 1/1000 of a wavelength are often required to represent the electromagnetic fields interacting with these objects accurately. Identifying the difficulties arising from this orders-of-magnitude difference and providing techniques to overcome them are the main contributions of this dissertation.

This dissertation is organized as follows. Chapter 1 introduces the frequency-domain Maxwell's equations and carefully formulates a differential equation to solve. Then it discretizes the differential equation into a system of linear equations $Ax = b$ using the finite-difference method. It also reviews iterative methods of solving $Ax = b$, which are essential in solving large 3D problems. Three benchmark problems that have been solved repeatedly throughout the dissertation are also described here.

Chapter 2 addresses the first difficulty in implementing an efficient 3D FDFD solver: the ill-conditioned matrix A due to an inappropriate choice of the perfectly matched layer (PML). PML is an essential absorbing boundary that is widely used in simulation of electromagnetic wave propagation. There are mainly two kinds of PML, and between the two the more commonly used UPML is shown to ill-condition the matrix A and therefore to induce slow convergence of iterative methods. This slow convergence problem is solved by replacing UPML with the less popular SC-PML. A rigorous analysis to prove the superiority of SC-PML over UPML is provided. For cases where UPML is indispensable, SP-UPML, which is a combination of UPML and an effective diagonal preconditioning scheme, is developed.

Chapter 3 addresses the second difficulty in implementing an efficient 3D FDFD solver: the high multiplicity of near-zero eigenvalues of the matrix A. The matrix A for plasmonic and nanophotonic systems has a very high multiplicity of near-zero eigenvalues, which stagnates the convergence of iterative methods. This stagnation problem is solved by eliminating the near-zero eigenvalues using the continuity equation. An intuitive explanation for the impact of the near-zero eigenvalues and definiteness of A on the convergence behavior of iterative methods is also provided.

Chapter 4 demonstrates the usefulness of the iterative FDFD solver by using it in designing novel waveguide components for integrated optical circuits. The components are sharp 90° bends and T-splitters formed in plasmonic coaxial waveguides. These components bend and split optical waves almost perfectly with nearly no reflection and radiation loss over a broad range of wavelengths, including the telecommunication wavelength of 1.55 μm.

Finally, Chapter 5 concludes this dissertation with a few important remarks and outlooks. The difference in the use of iterative methods between this dissertation and the literature of general numerical linear algebra is highlighted.

This Ph.D. dissertation would have been impossible to finish without help and support of many people and organizations. First, I would like to thank my Ph.D. advisor, Prof. Shanhui Fan, for his guidance based on a thorough knowledge of this field of research as well as for his emotional support. I also would like to thank the members of my reading committee, Prof. David A. B. Miller and Prof. Jelena Vuckovic, for their inspirational lectures and pioneering research, which have been the source of my passion in nanophotonics. I am indebted to Dr. Georgios Veronis for the initial introduction to the FDFD method, and to Jesse Lu for the implementation of my FDFD solver on GPUs, which demonstrated the true capability of the FDFD method. All of the Fan group members have been a joy to work with, but I especially thank Peter, Xiaofang, Lieven, Aaswath, Ken, Linxiao, and Yasin for actually using my solver program and suggesting useful features, and Zheng, Zongfu, Sunil, Zhichao, Eden, Victor, and Sacha for useful and interesting discussions. Besides this group, I thank Prof. Wenshan Cai for the collaboration in plasmonic waveguide projects, and Paul Hansen for an opportunity to guest lecture about the FDFD method in his numerical electromagnetics class. I also acknowledge the generous support from Samsung Scholarship, the National Science Foundation (Grant No. DMS-0968809), the AFOSR MURI program (Grant No. FA9550-09-1-0704), and the Interconnect Focus Center, funded under the Focus Center Research Program, a Semiconductor Research Corporation entity.

On the domestic side, I thank my parents, sister, and in-laws in Korea for all their support and encouragement. Most important, I deeply thank my wife Kyuwon for always standing by me with endless love.

Wonseok Shin

Stanford, California

Contents

Abstract **v**

Preface **ix**

1 Basic formulation of the FDFD method for Maxwell's equations **1**

1.1 Equation formulation 3

1.2 Finite-difference approximation 5

1.3 Iterative methods to solve $Ax = b$ 8

1.4 Benchmark problems 11

2 Accelerated solution by the correct choice of PML **15**

2.1 Review of SC-PML and UPML for the frequency-domain Maxwell's equations 16

2.2 Convergence speed of iterative methods to solve the UPML and SC-PML equations 19

2.3 Condition numbers of the UPML and SC-PML matrices 21

2.3.1 Mathematical background 22

2.3.2 Maximum singular values of homogeneous media 25

2.3.3 Minimum singular values of homogeneous media 30

2.3.4 Minimum singular values of homogeneous media with $\varepsilon > 0$ in a bounded domain 31

2.3.5 Variational method to estimate extreme singular values of inhomogeneous EM systems 38

2.3.6 Numerical validation 44

2.4 Diagonal preconditioning scheme for UPML 53

2.4.1 Relation between UPML and SC-PML 54

2.4.2 Scale-factor-preconditioned UPML 55
2.5 Summary and remarks 57

3 Accelerated solution by engineering the eigenvalue distribution 59
3.1 Eigenvalue distribution of the operator for a homogeneous system 62
3.2 Impact of the eigenvalue distribution on the convergence behavior of GMRES 66
3.3 Convergence behavior of QMR for 3D inhomogeneous systems 76
3.4 Summary and remarks 78

4 Design of plasmonic coaxial waveguide bends and splitters by the FDFD method 81
4.1 Properties of plasmonic coaxial waveguides 83
4.2 Performance of sharp 90° bends 85
4.3 Performance of T-splitters 88
4.4 Summary and remarks 92

5 Conclusion and final remarks 93

A First-order perturbation method for nondegenerate singular values of symmetric matrices 97

B Maximum singular values of homogeneous media accounting for finite-difference approximation 101

C Lengthy derivations of various formulae in Sec. 2.3.4 105
C.1 k_x's minimizing $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for a given k_y 105
C.2 Estimates of $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_{x0}}$ and $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_{x0}}$ for a given $k_y > \omega/c$ 109
C.3 Lowest-order approximation of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})/\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ around $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ . . 113

D Eigenvalues and eigenfunctions of $\nabla\times(\nabla\times\quad)$ and $\nabla(\nabla\cdot\quad)$ 115

E Effect of the smallest root of $\tilde{p}_m \in \mathcal{P}_m$ on the slopes at the roots 117

F Trend in the slopes of a polynomial at the roots 119

Bibliography 121

Index 135

List of Figures

1.1 Comparison of popularity of the FDTD versus FDFD method in the academia 2
1.2 Yee's finite-difference grid 6
1.3 A typical convergence plot of QMR 10
1.4 Benchmark problem "Slot": a plasmonic slot waveguide bend 11
1.5 Benchmark problems "Diel" and "Array": a dielectric waveguide and metallic pillar array 12

2.1 An example of an EM system surrounded by PML 17
2.2 Convergence of QMR for the three benchmark problems surrounded by UPML and SC-PML 20
2.3 The 3D plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$ as functions of $\mathbf{k}$ 32
2.4 The 2D contour plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ as a function of $\mathbf{k}$ 34
2.5 An example of an inhomogeneous EM system 38
2.6 Two inhomogeneous EM systems whose extreme singular values and condition numbers are numerically calculated: a vacuum and an metal-dielectric-metal waveguide bend 45
2.7 Maximum right singular vectors of vacua surrounded by UPML and SC-PML 46
2.8 Minimum right singular vectors of vacua surrounded by UPML and SC-PML 48
2.9 Maximum right singular vectors of metal-dielectric-metal waveguide bends surrounded by UPML and SC-PML 51
2.10 Minimum right singular vectors of metal-dielectric-metal waveguide bends surrounded by UPML and SC-PML 52

2.11 Convergence of QMR for the UPML equation, SC-PML equation, SP-UPML equation, and the UPML equation preconditioned by the Jacobi preconditioner 56

3.1 A vacuum-filled 2D square domain for which the eigenvalue distribution is calculated numerically for $s = 0, \pm 1$ 64
3.2 The eigenvalue distribution for $s = 0, \pm 1$ for a vacuum-filled 2D domain . . . 65
3.3 Convergence behavior of GMRES for a vacuum-filled 2D domain 67
3.4 Initial evolution of $r_m/\|b\|$ for $s = -1$ for a vacuum-filled 2D domain 69
3.5 Initial evolution of $r_m/\|b\|$ for $s = 0$ for a vacuum-filled 2D domain 71
3.6 Impact of the magnitude of the smallest root of a polynomial $\tilde{p}_m \in \mathcal{P}_m$ on the oscillation amplitudes of $\tilde{p}_m$ 72
3.7 Evolution of $r_m/\|b\|$ for $s = +1$ for a vacuum-filled 2D domain 73
3.8 Candidates for the residual polynomials for a nearly positive-definite matrix and strongly indefinite matrix 74
3.9 Convergence behavior of QMR for the benchmark problems with the modified Maxwell's equations 77

4.1 Structures of a sharp 90° bend and T-splitter in plasmonic coaxial waveguides 82
4.2 Properties of the reference coaxial waveguide 84
4.3 Performance of sharp 90° bends in plasmonic coaxial waveguides 86
4.4 Performance of sharp 90° bends in PEC coaxial waveguides with various cross-sectional dimensions 87
4.5 Performance of T-splitters in plasmonic coaxial waveguides 88
4.6 Optimization of T-splitters in plasmonic coaxial waveguides 90

List of Tables

1.1 Specification of the finite-difference grids used for three benchmark problems 13

2.1 Maximum singular values of vacua surrounded by UPML and SC-PML . . . 47
2.2 Minimum singular values of vacua surrounded by UPML and SC-PML . . . 49
2.3 Extreme singular values of metal-dielectric-metal waveguide bends surrounded by UPML and SC-PML . . . 50

3.1 Properties of the eigenvalue distributions for $s = 0$, $s < 0$, $s > 0$ for a vacuum-filled 2D domain . . . 63
3.2 Benchmark problems' parameters to test the condition for the low-frequency regime . . . 76

Chapter 1

Basic formulation of the FDFD method for Maxwell's equations

We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil.

DONALD E. KNUTH (1938–present)

MAXWELL'S EQUATIONS are partial differential equations that govern optical and electromagnetic (EM) phenomena. In the frequency domain, they are

$$\nabla \times \mathbf{E}(\mathbf{r}, \omega) = -i\omega\mu(\mathbf{r}, \omega)\mathbf{H}(\mathbf{r}, \omega) - \mathbf{M}(\mathbf{r}, \omega), \tag{1.1a}$$

$$\nabla \times \mathbf{H}(\mathbf{r}, \omega) = i\omega\varepsilon(\mathbf{r}, \omega)\mathbf{E}(\mathbf{r}, \omega) + \mathbf{J}(\mathbf{r}, \omega), \tag{1.1b}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic fields (or alternatively called the E- and H-fields); $\mathbf{J}$ and $\mathbf{M}$ are the electric and magnetic current source densities; ε and μ are the electric permittivity and magnetic permeability. All these quantities are functions of position $\mathbf{r}$ and angular frequency ω.

The above frequency-domain Maxwell's equations are derived from the time-domain

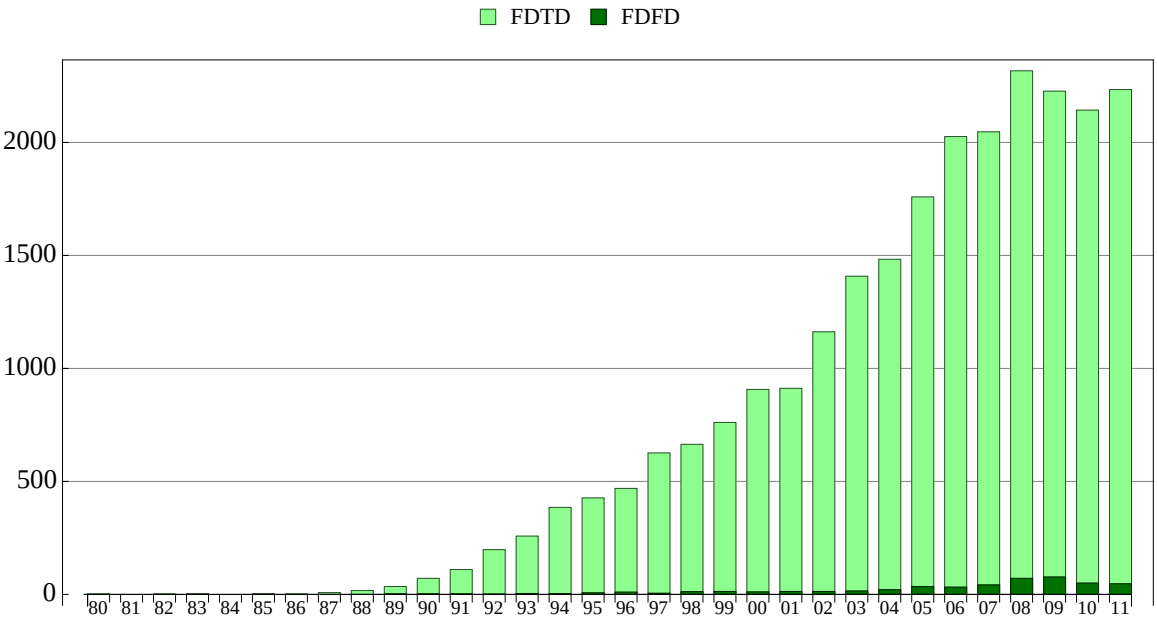


Figure 1.1: Comparison of popularity of the FDTD versus the FDFD method in the academia. In each year from 1980 to 2011, the numbers of academic papers published with "FDTD" and "FDFD" as keywords are counted and shown as bright and dark columns, respectively. The term "FDTD" was first coined in 1980 by Taflove [7]. The term "FDFD" was first used in 1989 by Ling [8]. In terms of popularity, the FDTD method has been much more successful than the FDFD method.

Maxwell's equations

$$\nabla \times \mathcal{E}(\mathbf{r}, t) = -\partial_t \mathcal{B}(\mathbf{r}, t) - \mathcal{M}(\mathbf{r}, t), \tag{1.2a}$$

$$\nabla \times \mathcal{H}(\mathbf{r}, t) = \partial_t \mathcal{D}(\mathbf{r}, t) + \mathcal{J}(\mathbf{r}, t) \tag{1.2b}$$

for a given ω by assuming a time dependence of $e^{+i\omega t}$ in each time-dependent quantity $\mathcal{F}(\mathbf{r}, t)$ to have $\mathcal{F}(\mathbf{r}, t) = \mathbf{F}(\mathbf{r}, \omega) e^{i\omega t}$ and then using the constitutive equations $\mathbf{D} = \varepsilon \mathbf{E}$ and $\mathbf{B} = \mu \mathbf{H}$. Alternatively, we can also obtain Eq. (1.1) by Fourier-transforming Eq. (1.2) in time.

The finite-difference method is a method to construct numerically solvable difference equations out of differential equations, by approximating derivatives by ratios between finite differences. Therefore, the method can be applied to both the time- and frequency-domain

Maxwell's equations. Nevertheless, the finite-difference method has been much more popular in solving the time-domain Maxwell's equations than the frequency-domain equations. Figure 1.1 directly shows how successful the finite-difference time-domain (FDTD) method has been compared to the finite-difference frequency-domain (FDFD) method in the academic community.

The time- and frequency-domain Maxwell's equations, however, are complementary tools for understanding optical and EM phenomena. The time-domain equations are indispensable to investigating transient states and dynamics, but the frequency-domain equations are also crucial to studying steady states and treating dispersive materials accurately. Therefore, providing techniques to implement an efficient FDFD solver should benefit the research community in a way that has not been possible with the FDTD method.

In this chapter, we review the basics of the FDFD method that form the foundation of this dissertation. In Sec. 1.1 we formulate a frequency-domain differential equation to solve. In Sec. 1.2 we discretize the differential equation by the finite-difference method. In Sec. 1.3 we briefly review iterative methods of solving the discretized equation. Lastly, in Sec. 1.4 we introduce benchmark problems against which we will test the efficiency of our techniques developed in the rest of the dissertation.

1.1 Equation formulation

The FDFD method solves the frequency-domain Maxwell's equations (1.1) for the E- and H-fields for given current source densities $\mathbf{J}$ and $\mathbf{M}$. There are a few different choices of actual equations to solve, though. First, we can choose to solve Eq. (1.1) directly for the E- and H-fields at the same time. This choice is equivalent to solving

$$\begin{bmatrix} -i\omega\varepsilon & \nabla\times \\ \nabla\times & i\omega\mu \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} \mathbf{J} \\ -\mathbf{M} \end{bmatrix}. \tag{1.3}$$

However, when described in the SI unit, Eq. (1.3) becomes an ill-conditioned equation, which is not favorable to numerical solvers. The concept of ill- and well-conditioned equations will be discussed in detail in Ch. 2, but we can easily see that Eq. (1.3) is ill-conditioned

from the formula $|\mathbf{E}|/|\mathbf{H}| = |\eta|$, which holds for a plane wave propagating in a medium with the impedance $\eta = \sqrt{\mu/\varepsilon}$. In vacuum the impedance is $\eta_0 = 377\,\Omega$, which means that the solution E-field is typically two orders-of magnitude stronger than the H-field. Such a huge contrast in magnitude between the elements of a solution is a sign of an ill-conditioned equation. To obtain a better-conditioned equation, we should introduce a normalized H-field variable $\widetilde{\mathbf{H}} = \eta\mathbf{H}$, but then it becomes harder to enforce the continuity of the normal and tangential components of the H-field at the interfaces between different materials. Using η_0 as a normalization factor is another possibility [9[Sec. 3.4.1], 10], but then the normalization is no longer optimal.

A better alternative is to eliminate either the E- or H-field from Eq. (1.1) to obtain

$$\nabla \times \mu^{-1}\nabla \times \mathbf{E} - \omega^2\varepsilon\mathbf{E} = -i\omega\mathbf{J} - \nabla \times \mu^{-1}\mathbf{M} \tag{1.4}$$

or

$$\nabla \times \varepsilon^{-1}\nabla \times \mathbf{H} - \omega^2\mu\mathbf{H} = -i\omega\mathbf{M} + \nabla \times \varepsilon^{-1}\mathbf{J}. \tag{1.5}$$

Both equations have only one of the E- and H-fields as an unknown, so they are no longer ill-conditioned due to the huge contrast in magnitude between the two fields. An additional benefit of this formulation is that the size of the solution vector is halved from that of Eq. (1.3). Once either Eq. (1.4) or Eq. (1.5) is solved for one field, the other field can be easily recovered by simple substitution of the solved field into Eq. (1.1).

Between the above two equations, we choose to solve Eq. (1.4) because it is a better formulation for nanophotonic systems by the following reason, which will be elaborated in Ch 3. In nanophotonics, the second term of the left-hand side of Eq. (1.4) is usually much smaller than the first term, i.e., $|\omega^2\varepsilon\mathbf{E}| \ll |\nabla \times \mu^{-1}\nabla \times \mathbf{E}|$, because nanophotonic objects are much smaller than a wavelength. Therefore the operator of Eq. (1.4) can be well-approximated by $\nabla \times \mu^{-1}\nabla\times$. Because $\mu = \mu_0$ for most materials used in nanophotonic devices, this operator is Hermitian positive-semidefinite. In other words, the operator of Eq. (1.4) is close to a Hermitian positive-semidefinite operator, which is very favorable to numerical solvers. The second term of the left-hand side of Eq. (1.5) is also ignorable for the same reason, but the operator of the first term is neither Hermitian nor positive-semidefinite, because ε is complex and has a negative real part for metals.

In solving Eq. (1.4) we can ignore $\mathbf{M}$ without loss of generality, because for given $\mathbf{J}$ and $\mathbf{M}$ a new current source densities $\mathbf{J}' = \mathbf{J} + \frac{1}{i\omega}\nabla \times \mu^{-1}\mathbf{M}$ and $\mathbf{M}' = 0$ generate the same right-hand side of Eq. (1.4). Therefore, the equation we focus on solving in the rest of this dissertation is

$$\nabla \times \mu^{-1}\nabla \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E} = -i\omega \mathbf{J}. \tag{1.6}$$

1.2 Finite-difference approximation

In the Cartesian coordinate system, the frequency-domain Maxwell's equations (1.1a) and (1.1b) are written as

$$\partial_y E_z - \partial_z E_y = -i\omega\mu H_x - M_x, \tag{1.7a}$$
$$\partial_z E_x - \partial_x E_z = -i\omega\mu H_y - M_y, \tag{1.7b}$$
$$\partial_x E_y - \partial_y E_x = -i\omega\mu H_z - M_z, \tag{1.7c}$$

and

$$\partial_y H_z - \partial_z H_y = i\omega\varepsilon E_x + J_x, \tag{1.8a}$$
$$\partial_z H_x - \partial_x H_z = i\omega\varepsilon E_y + J_y, \tag{1.8b}$$
$$\partial_x H_y - \partial_y H_x = i\omega\varepsilon E_z + J_z. \tag{1.8c}$$

To solve Eqs. (1.7) and (1.8) numerically by the finite-difference method, we approximate each derivative by a ratio between finite differences as

$$\frac{E_z^{i,j+1,k} - E_z^{i,j,k}}{\Delta_y^j} - \frac{E_y^{i,j,k+1} - E_y^{i,j,k}}{\Delta_z^k} = -i\omega\mu_x^{i,j,k} H_x^{i,j,k} - M_x^{i,j,k}, \tag{1.9a}$$
$$\frac{E_x^{i,j,k+1} - E_x^{i,j,k}}{\Delta_z^k} - \frac{E_z^{i+1,j,k} - E_z^{i,j,k}}{\Delta_x^i} = -i\omega\mu_y^{i,j,k} H_y^{i,j,k} - M_y^{i,j,k}, \tag{1.9b}$$
$$\frac{E_y^{i+1,j,k} - E_y^{i,j,k}}{\Delta_x^i} - \frac{E_x^{i,j+1,k} - E_x^{i,j,k}}{\Delta_y^j} = -i\omega\mu_z^{i,j,k} H_z^{i,j,k} - M_z^{i,j,k}, \tag{1.9c}$$

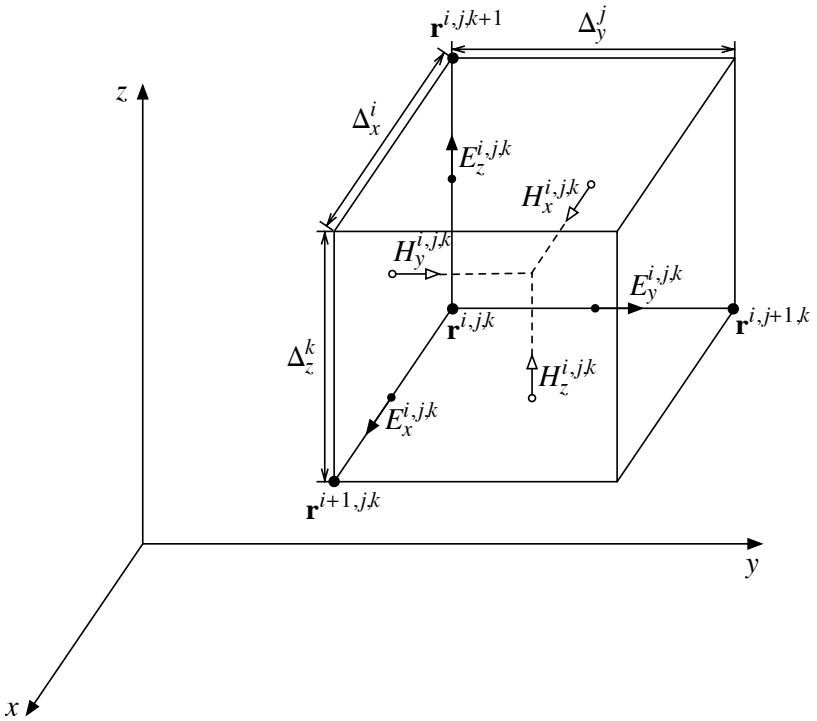


Figure 1.2: Yee's finite-difference grid. The box depicts the (i, j, k)th cell of Yee's grid whose corner with the smallest x-, y-, z-coordinates is at $\mathbf{r}^{i,j,k}$. Repetition of the box in the x-, y-, z-directions generates the entire grid. The x-, y-, z-components of the E-field (solid-headed arrows) are defined at the centers of the edges of the box. The x-, y-, z-components of H-field (open-headed arrows) are defined at the centers of the faces of the box.

and

$$\frac{H_z^{i,j,k} - H_z^{i,j-1,k}}{\widetilde{\Delta}_y^j} - \frac{H_y^{i,j,k} - H_y^{i,j,k-1}}{\widetilde{\Delta}_z^k} = i\omega\varepsilon_x^{i,j,k} E_x^{i,j,k} + J_x^{i,j,k}, \tag{1.10a}$$

$$\frac{H_x^{i,j,k} - H_x^{i,j,k-1}}{\widetilde{\Delta}_z^k} - \frac{H_z^{i,j,k} - H_z^{i-1,j,k}}{\widetilde{\Delta}_x^i} = i\omega\varepsilon_y^{i,j,k} E_y^{i,j,k} + J_y^{i,j,k}, \tag{1.10b}$$

$$\frac{H_y^{i,j,k} - H_y^{i-1,j,k}}{\widetilde{\Delta}_x^i} - \frac{H_x^{i,j,k} - H_x^{i,j-1,k}}{\widetilde{\Delta}_y^j} = i\omega\varepsilon_z^{i,j,k} E_z^{i,j,k} + J_z^{i,j,k}, \tag{1.10c}$$

where $\varepsilon_w^{i,j,k}$ and $\mu_w^{i,j,k}$ are the electric permittivity and magnetic permeability evaluated at the locations where $E_w^{i,j,k}$ and $H_w^{i,j,k}$ are defined, respectively, and $\widetilde{\Delta}_w^l = (\Delta_w^l + \Delta_w^{l-1})/2$.

The locations of $E_w^{i,j,k}$ and $H_w^{i,j,k}$ are indicated in Yee's finite-difference grid cell in Fig. 1.2.

In his seminal paper [11], Yee cleverly interlaced the E- and H-field grids as shown in the figure. Such interlaced placement makes the finite differences in Eqs. (1.9) and (1.10) central differences; the use of the central difference is crucial, because its error is $O(\Delta_w^2)$, which decreases much faster as Δ_w decreases than $O(\Delta_w)$ of the forward and backward differences [3[Sec. 3.3], 6[Secs. 2.4, 3.6]]. In addition, the interlaced grid ensures that the vector calculus identities $\nabla \times \nabla \varphi = 0$ and $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ hold for arbitrary scalar and vector fields φ and $\mathbf{F}$ even after the derivatives are approximated with finite differences [12].

The three difference equations (1.9) are obtained for each grid cell. We collect them from all grid cells to construct

$$C_e e = -i\omega D_\mu h - m, \tag{1.11}$$

where

$$e = \begin{bmatrix} \vdots \\ E_x^{i,j,k} \\ E_y^{i,j,k} \\ E_z^{i,j,k} \\ \vdots \end{bmatrix}, \; h = \begin{bmatrix} \vdots \\ H_x^{i,j,k} \\ H_y^{i,j,k} \\ H_z^{i,j,k} \\ \vdots \end{bmatrix}, \; m = \begin{bmatrix} \vdots \\ M_x^{i,j,k} \\ M_y^{i,j,k} \\ M_z^{i,j,k} \\ \vdots \end{bmatrix} \tag{1.12}$$

are column vectors representing the relevant fields; C_e, whose nonzero elements are $\pm 1/\Delta_w^l$, is a matrix for the curl operator on the E-field; $D_\mu = \text{diag}(\ldots, \mu_x^{i,j,k}, \mu_y^{i,j,k}, \mu_z^{i,j,k}, \ldots)$ is a diagonal matrix for the magnetic permeability. Similarly, collecting Eq. (1.10) from all grid cells produces

$$C_h h = i\omega D_\varepsilon e + j, \tag{1.13}$$

where C_h, D_ε, j are defined similarly to C_e, D_μ, m.

Note that the index of each element of the vectors in Eq. (1.12) is completely determined by the directional index w and positional indices i, j, k. In other words, for the nth element of a column vector, there exists a unique combination (w, i, j, k) that corresponds to the sequential index n. Hence, (w, i, j, k) can be used to index an element of a column vector. We can similarly index an element of a matrix with $(w_1, i_1, j_1, k_1; w_2, i_2, j_2, k_2)$, where the first set (w_1, i_1, j_1, k_1) specifies a row and the second set (w_2, i_2, j_2, k_2) specifies a column. The described indexing scheme is useful in mapping difference equations to the corresponding matrix-vector representation. For example, from the right-hand side of Eq. (1.9a), we can

see that the equation corresponds to the (x, i, j, k)th row of Eq. (1.11). However, the left-hand side of Eq. (1.9a) does not have $E_x^{i,j,k}$, the (x, i, j, k)th element of e. This means that the $(x, i, j, k; x, i, j, k)$th element of C_e, which is on the diagonal of the matrix, is zero. We can argue the same way to show that Eqs. (1.9b) and (1.9c) do not provide diagonal elements to C_e, and therefore that the diagonal of C_e is completely filled with zeros.

Now, by eliminating h from Eqs. (1.11) and (1.13) and ignoring m we can easily formulate the finite-difference approximation of Eq. (1.6):

$$\left(C_h D_\mu^{-1} C_e - \omega^2 D_\varepsilon\right) e = -i\omega j, \tag{1.14}$$

which is simply a system of linear equations

$$Ax = b, \tag{1.15}$$

where A represents the operator, x represents the E-field we solve for, and b is a column vector determined by a given electric current source density.

1.3 Iterative methods to solve $Ax = b$

There are two categories of methods to solve the system of linear equations (1.15): direct methods and iterative methods [13Ch. 2]. Direct methods factorize A into a few (typically two or three) factors with which $A^{-1}b$ can be calculated efficiently, and they produce a solution in a fixed number of steps. Depending on the structures and properties of A, different factorization methods such as the Cholesky, LU, $LDM^\top$, $LDL^\top$ factorizations are used.

The other category of methods, i.e., iterative methods, produce an approximate solution at each iteration step until the solution converges sufficiently close to the exact solution. More specifically, suppose that x_m is the approximate solution produced at the mth iteration step. Then iterative methods continue the process until the residual vector

$$r_m = b - Ax_m \tag{1.16}$$

satisfies $\|r_m\|/\|b\| < \tau$, where $\|\cdot\|$ is a norm of a column vector and τ is a user-defined small

positive number; typically the 2-norm is used as the norm, but some iterative methods, e.g., the conjugate gradient method [14], use different norms such as the A-norm, and in practice $\tau = 10^{-6}$ is sufficiently small for accurate solutions. Iterative methods do not guarantee convergence in a fixed number of iteration steps, but they often produce accurate solutions much earlier than direct methods.

The matrix A of Eq. (1.15) constructed by the finite-difference method is typically very large (often with more than 10 million rows and columns for 3D problems) but extremely sparse (with at most 13 nonzero elements per row). For such an extremely large and sparse matrix, iterative methods are usually preferred to direct methods, because direct methods require too much computer memory [15].

Among many kinds of iterative methods, we use Krylov subspace methods [16], which are known as one of the most efficient class of iterative methods. The Krylov subspace of dimension m generated by A and r_0 is

$$\mathcal{K}_m(A, r_0) = \text{span}\left\{r_0, Ar_0, A^2 r_0, \ldots, A^{m-1} r_0\right\}, \tag{1.17}$$

where r_0 is the initial residual vector of Eq. (1.16) for an initial guess solution x_0. Krylov subspace methods find the "best" mth approximate solution x_m of Eq. (1.15) in the space $x_0 + \mathcal{K}_m(A, r_0)$. Each Krylov subspace method is distinguished from others by its own criterion for determining the best approximate solution. In general, however, all Krylov subspace methods find better and better approximate solutions as m increases because the search space becomes larger and larger, i.e., $x_0 + \mathcal{K}_m(A, r_0) \subseteq x_0 + \mathcal{K}_{m+1}(A, r_0)$.

Like direct methods, there are Krylov subspace methods specialized for matrices with specific structures and properties, such as real-symmetric, complex-Hermitian, or positive-definite matrices. Unfortunately, our matrix constructed from Maxwell's equations does not satisfy any of these properties: it is complex, nonsymmetric, and indefinite. Therefore, we need to rely on Krylov subspace methods that can handle the most generic matrices.

In this dissertation, we use the biconjugate gradient (BiCG) [17, 18], quasi-minimal residual (QMR) [19], and generalized minimal residual (GMRES) [20] methods,[1] which are such Krylov subspace methods for generic matrices. All the three methods use the 2-norm in

[1] We implement BiCG and QMR using the matrix-vector multiplication routine of the portable, extensible toolkit for scientific computation (PETSc) [21], and use GMRES that is provided in PETSc.

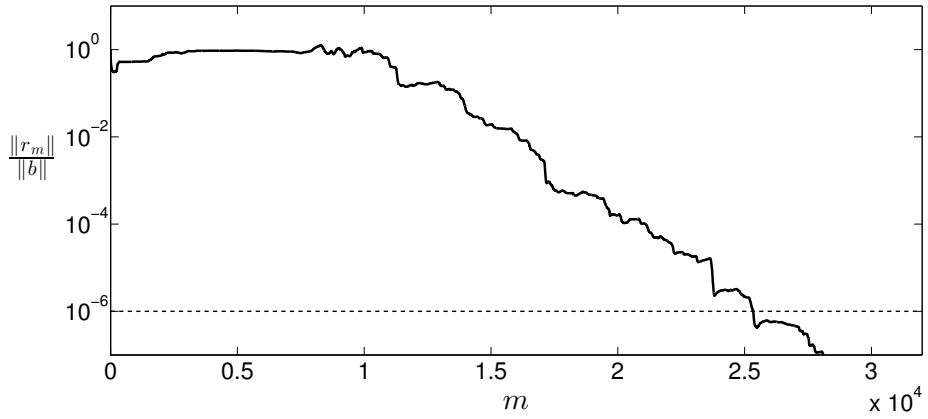
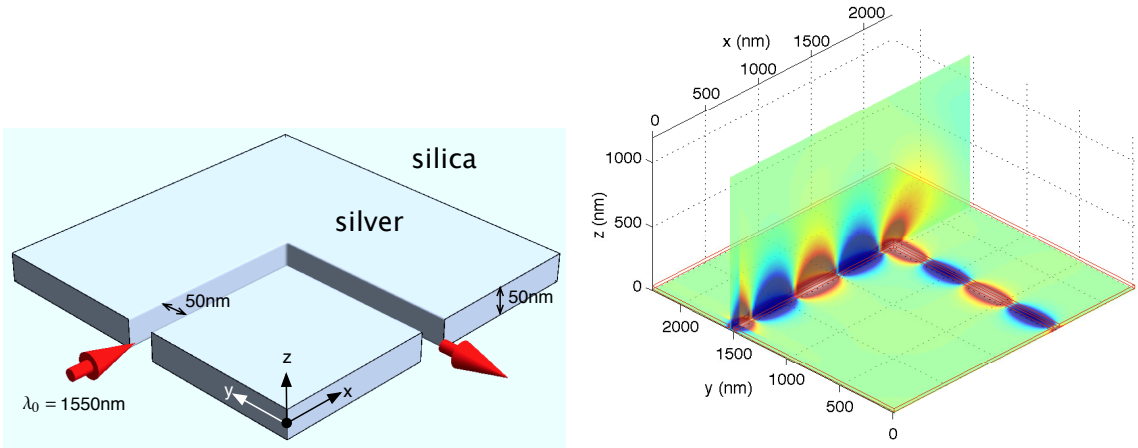


Figure 1.3: A typical convergence plot of QMR. A benchmark problem described in Fig. 1.4 is solved by QMR. The solid line plots $\|r_m\|/\|b\|$ versus m. The dashed line indicates a typical tolerance value $\tau = 10^{-6}$. Approximate solutions satisfy $\|r_m\|/\|b\| < \tau$ after about 25,000 iteration steps.

testing $\|r_m\|/\|b\| < \tau$. We use $x_0 = 0$ as an initial guess solution throughout the dissertation to create Krylov subspaces of Eq. (1.17). A typical convergence plot of $\|r_m\|/\|b\|$ versus m for QMR is shown in Fig. 1.3.

The three Krylov subspace methods have different characteristics. GMRES leads to convergence in the least number of iteration steps, but its each iteration step takes gradually more computation time and memory. Therefore GMRES is ideal for theoretical analysis (in Sec. 3.2), but impractical for large 3D problems. In contrast, BiCG and QMR are suitable for large 3D problems, because they consume constant computation time and memory over iteration steps for a given matrix A. Between the two methods, we use QMR to investigate the impact of our techniques developed in Chs. 2 and 3 on convergence speed for 3D problems (in Secs. 2.2, 2.4, and 3.3), because its $\|r_m\|/\|b\|$ decreases more stably without much oscillation than BiCG's. However, BiCG is a better method to use in practice (in Ch. 4) than QMR, because r_m for evaluating $\|r_m\|/\|b\| < \tau$ is obtained as a byproduct in BiCG whereas it should be calculated explicitly as Eq. (1.16) in QMR.



(a) Slot waveguide bend (Slot)

(b) Distribution of Re $\{H_z\}$

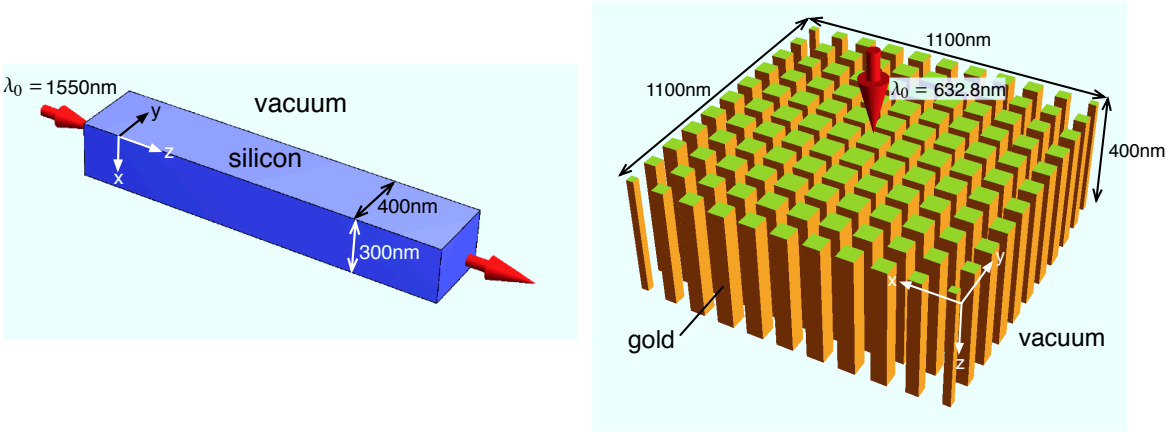
Figure 1.4: Benchmark problem "Slot": wave propagation through a plasmonic slot waveguide bend. In (a), the structure of the bend is illustrated. A narrow, 90°-bent slot waveguide is formed in a thin silver (Ag) film immersed in a silica (SiO_2) background. The vacuum wavelength and size of the structure are indicated in the figure. The red arrows specify the directions of wave propagation. In numerical simulation, all the x-, y-, z-normal boundaries of the simulation domain are covered by PML. In (b), Re $\{H_z\}$ calculated by the FDFD method is plotted on two planes: the horizontal $z = 0$ plane bisecting the film thickness, and the vertical $y =$ (const.) plane containing the central axis of the input port. Red and blue indicate Re $\{H_z\} > 0$ and Re $\{H_z\} < 0$. Only the $z \geq 0$ portion is drawn by virtue of mirror symmetry, and the PML regions are excluded. The sharp transition from blue to red near $x = 0$ is due to the **J** source plane there. The electric permittivities of silver [22] and silica [23] at $\lambda_0 = 1550$ nm are $\varepsilon_{\mathrm{Ag}} = (-129 - i3.28)\varepsilon_0$ and $\varepsilon_{\mathrm{SiO_2}} = 2.085\varepsilon_0$.

1.4 Benchmark problems

In this dissertation we introduce techniques to improve the convergence speed of iterative methods. To demonstrate the effectiveness of the techniques, we test them on three benchmark problems described in this section.

The first benchmark problem is to simulate wave propagation through a 90° bend of a plasmonic slot waveguide formed in a thin metal film (Fig. 1.4a). Plasmonic slot waveguides are a subject of active research in nanophotonics due to their capability of guiding light at deep-subwavelength scale [24].

We simulate the propagation of an EM wave at the telecommunication wavelength $\lambda_0 =$



(a) Dielectric waveguide (Diel) (b) Metallic pillar array (Array)

Figure 1.5: Benchmark problems "Diel" and "Array": a dielectric waveguide and metallic pillar array. The materials and sizes of the structures, the vacuum wavelengths, and the directions of wave propagation (red arrows) are indicated in the figures. In numerical simulation, all the six boundaries of the simulation domain of (a) are covered by PML. On the other hand, only the two z-normal boundaries of (b) are covered by PML, while the x- and y-normal boundaries are subject to periodic boundary conditions. The electric permittivities of silicon (Si) [23] at $\lambda_0 = 1550$ nm and gold (Au) [25] at $\lambda_0 = 632.8$ nm are $\varepsilon_{\text{Si}} = 12.09\varepsilon_0$ and $\varepsilon_{\text{Au}} = (-10.78 - i0.79)\varepsilon_0$.

1550 nm through the bend. A **J** source plane is placed near $x = 0$ to launch the fundamental mode of the waveguide. To simulate an infinitely long plasmonic slot waveguide immersed in a dielectric medium, all six boundary faces of the Cartesian simulation domain are covered by the perfectly matched layer (PML), which is discussed in detail in Ch. 2. The solution obtained by the FDFD method is displayed in Fig. 1.4b.

The second benchmark problem is to simulate wave propagation through a rectangular dielectric waveguide (Fig. 1.5a). We launch the fundamental mode in the dielectric waveguide.

The last benchmark problem is to simulate interaction between a plane wave and an array of metallic pillars (Fig. 1.5b). We launch a plane wave toward the pillars and observe how it is scattered by them; the detailed analysis is described in Ref. [26].

The number of grid cells in the finite-difference grid used to discretize each simulation domain is shown in Table 1.1, together with the grid cell size in the x-, y-, z-directions.

	Slot	Diel	Array
$N_x \times N_y \times N_z$	$192 \times 192 \times 240$	$220 \times 220 \times 320$	$220 \times 220 \times 130$
$\Delta_x, \Delta_y, \Delta_z$	$2 \sim 20\,\text{nm}$	$10\,\text{nm}$	$5, 5, 20\,\text{nm}$

Table 1.1: Specification of the finite-difference grids used for the three benchmark problems. Slot uses a nonuniform grid with smoothly varying grid cell size. The vector x of Eq. (1.15) has $3N_xN_yN_z$ elements, where the extra factor 3 accounts for the x-, y-, z-components of the E-field.

In Chs. 2 and 3 we will show that our techniques accelerate the convergence of iterative methods for all the three benchmark problems. The benchmark problems shown here are chosen deliberately to include different geometrical complexities and different materials such as dielectrics and metals. Therefore, such benchmark results should suggest that our techniques are effective for a wide range of problems.

Chapter 2

Accelerated solution by the correct choice of PML[1]

An expert is a person who has made all the mistakes that can be made in a very narrow field.

Niels H. D. Bohr (1885–1962)

The perfectly matched layer (PML) is an artificial medium initially developed by Bérenger that absorbs incident EM waves omnidirectionally with virtually no reflection [28]. Because EM waves incident upon PML does not reflect back, a domain surrounded by PML simulates an infinite space. Thus, the use of PML has been essential for simulating spatially unbounded systems, such as an infinitely long waveguide [29] or an isolated structure in an infinite vacuum region [26].

Bérenger's original PML was followed by many variants. In the FDTD method, the uniaxial PML (UPML) [30] and stretched-coordinate PML (SC-PML) [31–33] are the most popular, both resulting in similar numerical performance.[2]

In frequency-domain methods such as the FDFD method and FEM , on the other hand, UPML and SC-PML result in systems of linear equations (1.15) with different matrices A. In

[1] Reproduced in part with permission, from Ref. [27]: W. Shin and S. Fan, "Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell's equations solvers," *Journal of Computational Physics* **231**, pp. 3406–3431. Copyright 2012 Elsevier.

[2] The convolutional PML (CPML) [34] that is widely used in time-domain simulation is in essence SC-PML.

general, it is empirically known that the use of any PML leads to an ill-conditioned matrix and slows down the convergence of iterative methods to solve Eq. (1.15) [35–39]. Yet, to the best of our knowledge, no detailed study has been conducted to compare the degree of deterioration caused by different PMLs in frequency-domain numerical solvers, except Ref. [40] that briefly mentions empirical observations.

In this chapter, we demonstrate that the choice of PML significantly influences the convergence of iterative methods to solve the frequency-domain Maxwell's equations. More specifically, we show that SC-PML leads to far faster convergence than UPML in general, and the difference in convergence speed becomes extremely significant for plasmonic and nanophotonic systems where wavelengths are much longer than grid cell size. To prove the generality of this observation, we present a rigorous analysis that relates convergence speed to the condition number of the matrix.

The chapter is organized as follows. In Sec. 2.1 we review the basic formulations of UPML and SC-PML for the frequency-domain Maxwell's equations. Then, in Sec. 2.2 we demonstrate that SC-PML gives rise to much faster convergence of iterative methods than UPML for the benchmark problems. In Sec. 2.3 we show that SC-PML produces a much better-conditioned matrix than UPML. Finally, we introduce a diagonal preconditioning scheme for UPML in Sec. 2.4; the newly developed preconditioning scheme can be very useful in situations where UPML is easier to implement than SC-PML. In Sec. 2.5 we summarize the chapter and make a few remarks.

Throughout this chapter we assume that $\mu = \mu_0$; this is valid for most nanophotonic simulations. Also, we use the FDFD method in this chapter to discretize differential equations. However, the arguments we present should be equally applicable to other frequency-domain methods including FEM.

2.1 Review of SC-PML and UPML for the frequency-domain Maxwell's equations

To simulate an infinite space, one surrounds the EM system of interest with PML as illustrated in Fig. 2.1. As a result, the governing equation is modified from Eq. (1.6). For an EM

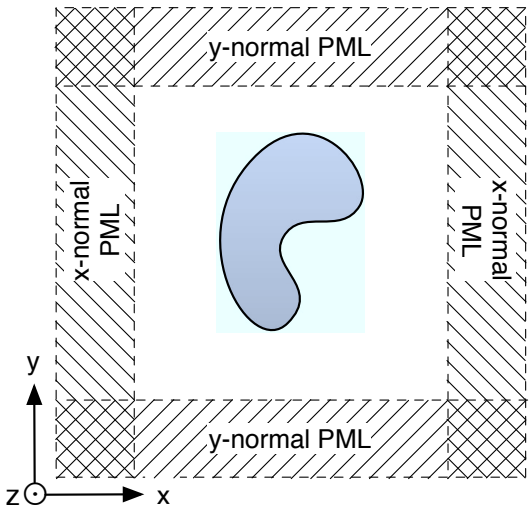


Figure 2.1: An example of an EM system surrounded by PML. In the four corner regions where the x- and y-normal PMLs overlap, waves attenuate in both directions. If the EM system is in a 3D simulation domain, PMLs can overlap up to three times. PML is either UPML or SC-PML.

system surrounded by UPML, the governing equation is the UPML equation

$$\nabla \times \overline{\overline{\mu}}_s^{-1} \nabla \times \mathbf{E} - \omega^2 \overline{\overline{\varepsilon}}_s \mathbf{E} = -i\omega \mathbf{J}, \tag{2.1}$$

where the 3×3 tensors $\overline{\overline{\varepsilon}}_s$ and $\overline{\overline{\mu}}_s$ are

$$\overline{\overline{\varepsilon}}_s = \varepsilon \begin{bmatrix} \frac{s_y s_z}{s_x} & 0 & 0 \\ 0 & \frac{s_z s_x}{s_y} & 0 \\ 0 & 0 & \frac{s_x s_y}{s_z} \end{bmatrix}, \quad \overline{\overline{\mu}}_s = \mu \begin{bmatrix} \frac{s_y s_z}{s_x} & 0 & 0 \\ 0 & \frac{s_z s_x}{s_y} & 0 \\ 0 & 0 & \frac{s_x s_y}{s_z} \end{bmatrix}. \tag{2.2}$$

On the other hand, for an EM system surrounded by SC-PML, the governing equation is the SC-PML equation

$$\nabla_s \times \mu^{-1} \nabla_s \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E} = -i\omega \mathbf{J}, \tag{2.3}$$

where

$$\nabla_s = \hat{\mathbf{x}} \frac{1}{s_x} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{1}{s_y} \frac{\partial}{\partial y} + \hat{\mathbf{z}} \frac{1}{s_z} \frac{\partial}{\partial z}. \tag{2.4}$$

In both equations, the PML scale factors s_w for $w = x, y, z$ are

$$s_w(l) = 1 - is''_w(l) = \begin{cases} 1 - i\frac{\sigma_w(l)}{\omega\varepsilon_0} & \text{inside the } w\text{-normal PML,} \\ 1 & \text{elsewhere,} \end{cases} \tag{2.5}$$

where l is the depth measured from the PML interface; $\sigma_w(l)$ is the PML loss parameter at the depth l in the w-normal PML; ε_0 is the electric permittivity of vacuum. The w-normal PML attenuates waves propagating in the w-direction. In regions such as the corners in Fig. 2.1 where multiple PMLs overlap, $s_w(l) \neq 1$ for more than one w. Also, here for simplicity we have chosen $\mathrm{Re}\{s_w(l)\} = 1$; the conclusion of this chapter, however, is equally applicable to PML with $\mathrm{Re}\{s_w(l)\} \neq 1$.

For theoretical development of PMLs, $\sigma_w(l)$ is usually assumed to be a positive constant that is independent of l. In numerical implementation of PMLs, however, $\sigma_w(l)$ gradually increases from 0 with l to prevent spurious reflection at PML interfaces. Typically, the polynomial grading scheme is adopted [6] so that

$$\sigma_w(l) = \sigma_{w,\max}\left(\frac{l}{d}\right)^m, \tag{2.6}$$

where d is the thickness of PML; $\sigma_{w,\max}$ is the maximum PML loss parameter attained at $l = d$; m is the degree of the polynomial grading, which is usually between 3 and 4. If R is the target reflection coefficient for normal incidence, the required maximum loss parameter is

$$\sigma_{w,\max} = -\frac{(m+1)\ln R}{2\eta_0 d}, \tag{2.7}$$

where $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$ is the vacuum impedance.

The modulus of $s_w(l)$ increases with l, so $|s_w(d)|$ is typically much larger than $|s_w(0)| = 1$, as can be seen in the following example. Consider a uniform finite-difference grid with grid cell size Δ. For a typical 10-layer PML with $d = 10\Delta$, $m = 4$, $R = e^{-16} \simeq 1 \times 10^{-7}$, we have $\sigma_{w,\max} = 4/\eta_0\Delta$. In the finite-difference scheme, the wavelength inside an EM medium should be at least 15Δ to approximate spatial derivatives by finite differences accurately [41].

Therefore, if the medium matched by PML is vacuum, the vacuum wavelength λ_0 corresponding to ω should satisfy $\lambda_0 \geq 15\Delta$, which implies that

$$s_w''(d) = \frac{\sigma_{w,\max}}{\omega\varepsilon_0} = \frac{\frac{4}{\eta_0\Delta}}{\frac{2\pi}{\lambda_0}c_0\varepsilon_0} = \frac{2\lambda_0}{\pi\Delta} \geq \frac{30\Delta}{\pi\Delta} \simeq 9.549, \tag{2.8}$$

where $c_0 = 1/\sqrt{\mu_0\varepsilon_0}$ is the speed of light in vacuum. Therefore, $|s_w(d)| = \sqrt{1 + s_w''(d)^2}$ is at least about 10. In nanophotonics where deep-subwavelength structures are studied, the use of $\Delta = 1\,\text{nm}$ for vacuum wavelength $\lambda_0 = 1550\,\text{nm}$ is not uncommon (see Ch. 4). In that case, $|s_w(d)|$ is nearly 1000.

Depending on the kind of PML used, we solve either Eq. (2.1) or Eq. (2.3) throughout the entire simulation domain (both inside and outside PML). Because the UPML and SC-PML equations are different, they produce different systems of linear equations (1.15), which are respectively referred to as

$$A^{\text{u}}x = b \tag{2.9}$$

and

$$A^{\text{sc}}x = b, \tag{2.10}$$

where b is common to both systems if the same $\mathbf{J}$ drives the EM fields of the two systems. We refer to A^{u} and A^{sc} as the UPML and SC-PML matrices, respectively.

In the following sections, we will see that Eq. (2.10) is much more favorable to numerical solvers than Eq. (2.9).

2.2 Convergence speed of iterative methods to solve the UPML and SC-PML equations

We apply UPML and SC-PML to each benchmark problem in Sec. 1.4 to construct two systems of linear equations (2.9) and (2.10), and compare the convergence speed of an iterative method to solve them. The iterative method used here is QMR introduced in Sec. 1.3.

Figure 2.2 shows $\|r_m\|/\|b\|$ of QMR versus the number m of iteration steps. For all the

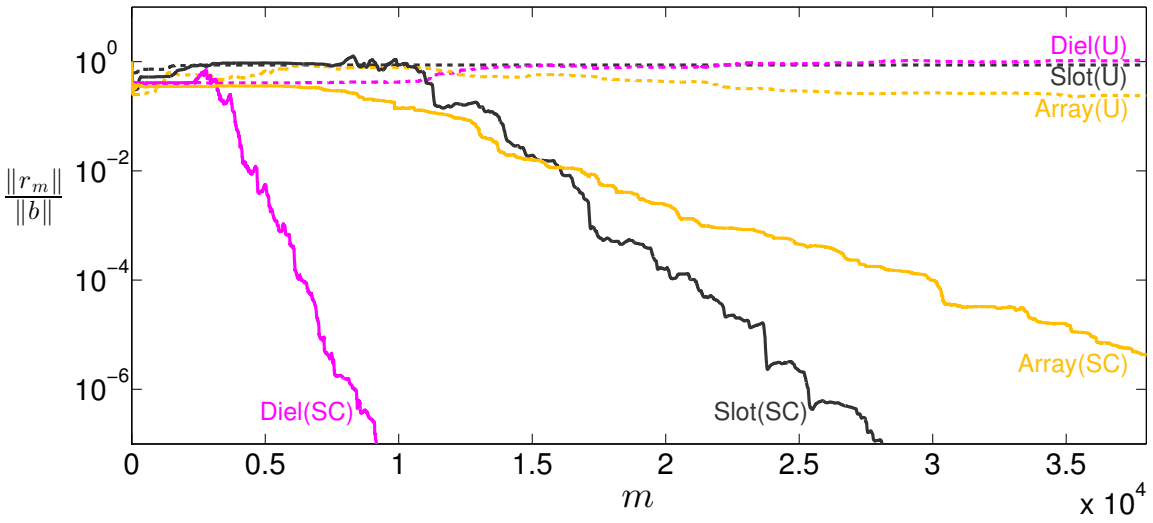


Figure 2.2: Convergence of QMR for the three benchmark problems "Slot", "Diel", and "Array" described in Sec. 1.4 surrounded by UPML (U) and SC-PML (SC). Notice that simply replacing UPML with SC-PML improves convergence speed dramatically for all the three benchmark problems.

three benchmark problems, SC-PML significantly outperforms UPML in terms of convergence speed. As mentioned at the end of Sec. 1.4, the three benchmark problems have different geometrical complexities and different materials. Therefore, Fig. 2.2 suggests that SC-PML leads to faster convergence speed than UPML for a wide range of EM systems. Moreover, the result is not specific to QMR; we have observed the same trend for other iterative methods such BiCG. Hence, we conclude that the significant difference in convergence speed originates from the intrinsic properties of UPML and SC-PML, and is independent of the kind of iterative method used.

In the next section, we relate the significantly different convergence speeds to the very different condition numbers of the UPML and SC-PML matrices.

2.3 Condition numbers of the UPML and SC-PML matrices

In this section, we present a detailed analysis of the condition numbers of the UPML and SC-PML matrices. The condition number of a matrix A is defined as

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}, \tag{2.11}$$

where $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ are the maximum and minimum singular values of A as we will review in Sec. 2.3.1. Matrices with large and small condition numbers are called ill-conditioned and well-conditioned, respectively. For convenience, we introduce notations

$$\sigma^{\mathrm{u}}_{\max} = \sigma_{\max}(A^{\mathrm{u}}), \quad \sigma^{\mathrm{u}}_{\min} = \sigma_{\min}(A^{\mathrm{u}}), \quad \kappa^{\mathrm{u}} = \frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{u}}_{\min}} \tag{2.12}$$

for the maximum and minimum singular values and the condition number of the UPML matrix A^{u}. We define $\sigma^{\mathrm{sc}}_{\max}$, $\sigma^{\mathrm{sc}}_{\min}$, and κ^{sc} similarly for the SC-PML matrix A^{sc}.

The objective of this section is to show that in general UPML produces a much worse-conditioned matrix than SC-PML, i.e., $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}} \gg 1$, provided that the two PMLs enclose the same EM system. According to Eq. (2.11), the objective is accomplished by analyzing the extreme singular values of A^{u} and A^{sc}.

All EM systems simulated in Sec. 2.2 are inhomogeneous, being composed of several different EM media. It turns out that the extreme singular values of an inhomogeneous EM system are approximately determined by the extreme singular values of the component media. Here, the extreme singular values of each component medium are defined as the extreme singular values of an infinite space filled entirely with that medium, which we refer to as the "homogeneous medium". For example, the extreme singular values of a vacuum surrounded by UPML are determined by the extreme singular values of a homogeneous vacuum and homogeneous UPML. Therefore, we study the extreme singular values of homogeneous media. Of particular interest are a homogeneous regular medium, homogeneous UPML, and homogenous SC-PML, whose extreme singular values are studied in Secs. 2.3.2 through 2.3.4.

In Sec. 2.3.5, we develop a theory based on the variational method to estimate the extreme singular values and condition numbers of inhomogeneous EM systems from the extreme singular values of the component homogeneous media. The theory predicts that

$\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}} \gg 1$. In Sec. 2.3.6, we verify the theory numerically for two inhomogeneous EM systems.

The conclusion of this section explains the results in Sec. 2.2, because A with a smaller condition number, or an ill-conditioned A, generally implies faster convergence of iterative methods to solve a system of linear equations $Ax = b$ [42Sec. 9.2]. In fact, an ill-conditioned matrix can be detrimental to direct methods as well; it is known that the LU factorization of ill-conditioned matrices tends to be inaccurate [43Sec. 6.8]. Therefore, the result in this section suggests that SC-PML should be preferable to UPML for solving the frequency-domain Maxwell's equations by both iterative and direct methods.

2.3.1 Mathematical background

For an arbitrary $A \in \mathbb{C}^{n\times n}$, one can always perform a singular value decomposition (SVD) as [44Sec. 2.5.6]

$$A = U\Sigma V^{\dagger}, \tag{2.13}$$

where $U, V \in \mathbb{C}^{n\times n}$ are unitary; $V^{\dagger}$ is the conjugate transpose of V; $\Sigma \in \mathbb{R}^{n\times n}$ is a real diagonal matrix whose diagonal elements are nonnegative. If A is nonsingular, the diagonal elements of Σ are strictly positive; the converse is also true.

The SVD can also be written as

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^{\dagger}, \tag{2.14}$$

where σ_i is the ith diagonal element of Σ; u_i and v_i are the ith column of U and V, respectively. Because U and V are unitary, each of $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_n\}$ forms an orthonormal basis of $\mathbb{C}^n$. Each σ_i is referred to as a singular value of A; u_i and v_i are the corresponding left and right singular vectors, respectively.

The maximum and minimum singular values,

$$\sigma_{\max} = \max_{1\le i\le n} \sigma_i \quad \text{and} \quad \sigma_{\min} = \min_{1\le i\le n} \sigma_i, \tag{2.15}$$

are collectively called the extreme singular values. The left and right singular vectors corresponding to $\sigma_{\max}$ are denoted by $u_{\max}$ and $v_{\max}$, and called the maximum left and right singular vectors, respectively. Similarly, the minimum left and right singular vectors are the singular vectors corresponding to $\sigma_{\min}$, and denoted by $u_{\min}$ and $v_{\min}$.

From Eq. (2.14), it follows that

$$Av_i = \sigma_i u_i \quad \text{and} \quad A^\dagger u_i = \sigma_i v_i. \tag{2.16}$$

Therefore, the singular values and vectors can be obtained by solving a Hermitian eigenvalue problem

$$H(A)\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \sigma_i \begin{bmatrix} u_i \\ v_i \end{bmatrix}, \quad \text{where } H(A) = \begin{bmatrix} 0 & A \\ A^\dagger & 0 \end{bmatrix}. \tag{2.17}$$

In Sec. 2.3.6 of this chapter, we solve Eq. (2.17) for the largest or smallest nonnegative eigenvalues by the Arnoldi Package (ARPACK) [45] to numerically calculate the extreme singular values of A.[3] ARPACK uses the Arnoldi iteration that only requires matrix-vector multiplication. For the maximum and minimum singular values of A, the matrices multiplied iteratively to vectors are $H(A)$ and $H(A)^{-1}$, respectively [47]. This means that a large system of linear equations needs to be solved repeatedly for the minimum singular value, which is extremely costly unless the LU factors of $H(A)$ are known. For this reason, all numerical calculations of the singular values and vectors in Sec. 2.3.6 are limited to two-dimensional (2D) EM systems, for which the LU factorization is easily performed.

The singular values and vectors also satisfy a different Hermitian eigenvalue equation

$$(A^\dagger A)v_i = \sigma_i^2 v_i \tag{2.18}$$

that is derived from Eq. (2.16). Because $\kappa(A^\dagger A) = \kappa(A)^2$ and $\kappa(H(A)) = \kappa(A)$, $A^\dagger A$ is much worse-conditioned than $H(A)$, so we use Eq. (2.17) rather than Eq. (2.18) to solve for the singular values numerically. Nevertheless, Eq. (2.18) turns out to be useful in the theoretical analysis in Secs. 2.3.2 through 2.3.4.

[3] The actual calculation of the extreme singular values is carried out using the MATLAB routine svds [46], which uses ARPACK internally.

The extreme singular values can also be calculated by the variational method. As a consequence of Eq. (2.14) we have

$$\sigma_{\max} = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \quad \text{and} \quad \sigma_{\min} = \min_{x\neq 0} \frac{\|Ax\|}{\|x\|}, \tag{2.19}$$

where $\|\cdot\|$ is the 2-norm of a vector. Note that the quotient $\|Ax\|/\|x\|$ is maximized to $\sigma_{\max}$ at $x = v_{\max}$ and minimized to $\sigma_{\min}$ at $x = v_{\min}$. In Sec. 2.3.5, we use the variational method to estimate the extreme singular values of inhomogeneous EM systems.

The maximum singular value of a matrix is related to a norm of the matrix. The p-norm of a matrix is defined as [44[Sec. 2.3.1]]

$$\|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \tag{2.20}$$

where $\|y\|_p = \left(\sum_i |y_i|^p\right)^{1/p}$ on the right-hand side is the p-norm of a column vector y. Comparing Eq. (2.20) for $p = 2$ with Eq. (2.19) reveals that

$$\sigma_{\max}(A) = \|A\|, \tag{2.21}$$

where the subscript 2 is omitted from $\|\cdot\|_2$ as a convention throughout this chapter.

There is an inequality that holds between the matrix p-norms [44[Corollary 2.3.2]]:

$$\|A\| \leq \sqrt{\|A\|_1 \|A\|_\infty}. \tag{2.22}$$

Because the ∞-norm satisfies $\|A\|_\infty = \|A^\top\|_1$, Eq. (2.22) implies that

$$\sigma_{\max}(A) \leq \|A\|_1 \quad \text{for symmetric } A. \tag{2.23}$$

The right-hand side of Eq. (2.23) is easily evaluated, because the 1-norm reduces to

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \left|a_{ij}\right| = (\text{the maximum absolute column sum}), \tag{2.24}$$

where a_{ij} is the (i, j)th element of A.

Finally, we note that the singular values, singular vectors, and the condition number are the properties of a matrix. Below, however, we refer to these terms as the properties of an EM system, which are understood as those of the matrix that describes the EM system. For example, "the maximum singular value of a homogeneous vacuum" means "the maximum singular value of the matrix describing a homogeneous vacuum."

2.3.2 Maximum singular values of homogeneous media

In this section, we derive approximate formulae for the maximum singular values of a homogeneous regular medium, homogeneous UPML, and homogeneous SC-PML. Here, a homogeneous medium is defined as an infinite space described by translationally invariant EM parameters; for a regular medium it means that ε is constant over all space, and for PML it means that the PML scale factors s_w for $w = x, y, z$ as well as ε are constant over all space.

For simplicity, we consider PML with only one attenuation direction, which, without loss of generality, is assumed to be the x-direction. Hence, we have $s_y = s_z = 1$ and

$$s_x = 1 - is_x'' \quad \text{with } s_x'' \gg 1, \tag{2.25}$$

where the assumption $s_x'' \gg 1$ is due to the discussion following Eq. (2.8). Equation (2.25) implies that

$$s_x \simeq -is_x'' \quad \text{and} \quad |s_x| \simeq s_x'' \gg 1. \tag{2.26}$$

We use the notations $\sigma_{\max}^{\mathrm{u}_0}$ and $\sigma_{\max}^{\mathrm{sc}_0}$ for the maximum singular values of a homogeneous UPML and SC-PML to distinguish them from $\sigma_{\max}^{\mathrm{u}}$ and $\sigma_{\max}^{\mathrm{sc}}$ of inhomogeneous EM systems defined in Eq. (2.12) and below. In addition, the maximum singular value of a homogeneous regular medium is denoted by $\sigma_{\max}^{\mathrm{r}_0}$.

Because a homogeneous EM system is spatially unbounded, discretizing the governing differential equation results in an infinitely large matrix. To avoid dealing with such an infinitely large matrix, we first examine the maximum singular values of the original differential operators used in Eqs. (1.6), (2.1), and (2.3); we take the effect of finite-difference discretization into account later. The differential operators for a homogeneous regular medium,

UPML, and SC-PML are defined as

$$T^{\mathrm{r}_0}(\mathbf{E}) = \nabla \times \mu^{-1} \nabla \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E}, \tag{2.27a}$$

$$T^{\mathrm{u}_0}(\mathbf{E}) = \nabla \times \overline{\overline{\mu}}_s^{-1} \nabla \times \mathbf{E} - \omega^2 \overline{\overline{\varepsilon}}_s \mathbf{E}, \tag{2.27b}$$

$$T^{\mathrm{sc}_0}(\mathbf{E}) = \nabla_s \times \mu^{-1} \nabla_s \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E}. \tag{2.27c}$$

Below, we refer to them collectively as T when we discuss properties that are common to all three operators. The purpose of this section is to estimate $\sigma_{\max}(T)$.

Because T is a translationally invariant operator, the composite operator $T^\dagger \circ T$ is also translationally invariant, which implies that its eigenvector, and hence the right singular vector of T, has the form [48[Sec. 2.3.2], 49[Sec. 2.6.1]]

$$\mathbf{E}_{\mathbf{k}}(\mathbf{r}) = \mathbf{F}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{2.28}$$

where $\mathbf{k}$ is real and $\mathbf{F}_{\mathbf{k}}$ is constant.

By applying T^{r_0}, T^{u_0}, and T^{sc_0} to $\mathbf{E}_{\mathbf{k}}$, we obtain

$$T^{\mathrm{r}_0}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k} \times \mu^{-1} \mathbf{k} \times \mathbf{E}_{\mathbf{k}} - \omega^2 \varepsilon \mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{r}_0} \mathbf{E}_{\mathbf{k}}, \tag{2.29a}$$

$$T^{\mathrm{u}_0}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k} \times \overline{\overline{\mu}}_s^{-1} \mathbf{k} \times \mathbf{E}_{\mathbf{k}} - \omega^2 \overline{\overline{\varepsilon}}_s \mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{u}_0} \mathbf{E}_{\mathbf{k}}, \tag{2.29b}$$

$$T^{\mathrm{sc}_0}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k}_s \times \mu^{-1} \mathbf{k}_s \times \mathbf{E}_{\mathbf{k}} - \omega^2 \varepsilon \mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{sc}_0} \mathbf{E}_{\mathbf{k}}, \tag{2.29c}$$

where $\mathbf{k}_s = \hat{\mathbf{x}}(k_x/s_x) + \hat{\mathbf{y}}(k_y/s_y) + \hat{\mathbf{z}}(k_z/s_z)$ with $s_y = s_z = 1$; $T_{\mathbf{k}}^{\mathrm{r}_0}$, $T_{\mathbf{k}}^{\mathrm{u}_0}$, and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ are 3×3 matrices operating on the vector $[E_{\mathbf{k},x}\ E_{\mathbf{k},y}\ E_{\mathbf{k},z}]^\top$. To facilitate computation, without loss of generality, we choose a coordinate system such that $\mathbf{k}$ lies in the xy-plane. (We recall that the attenuation direction of PML is $\hat{\mathbf{x}}$.) Then,

$$T_{\mathbf{k}}^{\mathrm{r}_0} = \begin{bmatrix} \frac{k_y^2}{\mu} - \omega^2\varepsilon & -\frac{k_x k_y}{\mu} & 0 \\ -\frac{k_x k_y}{\mu} & \frac{k_x^2}{\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \end{bmatrix}, \tag{2.30a}$$

$$T_{\mathbf{k}}^{\mathrm{u}_0} = \begin{bmatrix} \frac{k_y^2}{s_x\mu} - \frac{\omega^2\varepsilon}{s_x} & -\frac{k_x k_y}{s_x\mu} & 0 \\ -\frac{k_x k_y}{s_x\mu} & \frac{k_x^2}{s_x\mu} - s_x\omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{s_x\mu} + \frac{s_x k_y^2}{\mu} - s_x\omega^2\varepsilon \end{bmatrix}, \tag{2.30b}$$

$$T_{\mathbf{k}}^{\mathrm{sc}_0} = \begin{bmatrix} \frac{k_y^2}{\mu} - \omega^2\varepsilon & -\frac{k_x k_y}{s_x\mu} & 0 \\ -\frac{k_x k_y}{s_x\mu} & \frac{k_x^2}{s_x^2\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{s_x^2\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \end{bmatrix}. \tag{2.30c}$$

Note that Eq. (2.30) are the $\mathbf{k}$-space representations of T^{r_0}, T^{u_0}, and T^{sc_0}. Below, we refer to them collectively as $T_{\mathbf{k}}$ when we discuss properties that are common to all three matrices. We note that the quantity we want to estimate, i.e., $\sigma_{\max}(T)$, satisfies

$$\sigma_{\max}(T) = \max_{\mathbf{k}} \sigma_{\max}(T_{\mathbf{k}}). \tag{2.31}$$

By solving Eq. (2.18) with $A = T_{\mathbf{k}}$, we easily obtain one singular value $\sigma_{\mathbf{k},3}$ of $T_{\mathbf{k}}$ corresponding to a singular vector $[0\ 0\ 1]^\top$:

$$\sigma_{\mathbf{k},3}^{\mathrm{r}_0} = \left| \frac{k_x^2}{\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \right|, \quad \sigma_{\mathbf{k},3}^{\mathrm{u}_0} = \left| \frac{k_x^2}{s_x\mu} + \frac{s_x k_y^2}{\mu} - s_x\omega^2\varepsilon \right|, \quad \sigma_{\mathbf{k},3}^{\mathrm{sc}_0} = \left| \frac{k_x^2}{s_x^2\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \right|. \tag{2.32}$$

The subscript 3 of $\sigma_{\mathbf{k},3}$ indicates that the singular value is produced from the $(3,3)$th element of $T_{\mathbf{k}}$.

To estimate $\sigma_{\max}(T)$, we find lower and upper bounds of $\sigma_{\max}(T)$ using Eqs. (2.30) through (2.32). From Eq. (2.31) and the definition (2.15) of the maximum singular value, we have

$$\sigma_{\max}(T) = \max_{\mathbf{k}} \sigma_{\max}(T_{\mathbf{k}}) \geq \max_{\mathbf{k}} \sigma_{\mathbf{k},3}. \tag{2.33}$$

Also, from Eq. (2.31) and the inequality (2.23) we have

$$\sigma_{\max}(T) = \max_{\mathbf{k}} \sigma_{\max}(T_{\mathbf{k}}) \leq \max_{\mathbf{k}} \|T_{\mathbf{k}}\|_1. \tag{2.34}$$

Therefore $\sigma_{\max}(T)$ satisfies

$$\max_{\mathbf{k}} \sigma_{\mathbf{k},3} \leq \sigma_{\max}(T) \leq \max_{\mathbf{k}} \|T_{\mathbf{k}}\|_1. \tag{2.35}$$

Below we show that the lower and upper bound in the above inequality are approximately the same, and therefore we find a good estimate of $\sigma_{\max}(T)$.

We first consider the lower bound in Eq. (2.35). From Eq. (2.32) it is obvious that $\sigma_{\mathbf{k},3}$ increases with $|k_x|$ and $|k_y|$. For a continuous medium, k_x and k_y are unbounded, and so is $\sigma_{\mathbf{k},3}$. On a finite-difference grid with uniform grid cell size Δ, however, the maximum wavenumber in each Cartesian direction is the Nyquist wavenumber [41[Sec. 3.2], 49[Sec. 4.2]]

$$k_{\max} = \frac{\pi}{\Delta}, \tag{2.36}$$

and therefore $\sigma_{\mathbf{k},3}$ is bounded from above; here we note that $|k_y| \leq k_{\max}$ in 2D but $|k_y| \leq \sqrt{2}k_{\max}$ in 3D, because in 3D k_y actually contains the y- and z-components of the wavevector due to the special choice of our coordinate system made in the discussion following Eq. (2.29). Furthermore, when $k_{\max}$ is used to maximize $\sigma_{\mathbf{k},3}$, it turns out that we can ignore ω^2 terms in Eq. (2.32) because Δ is typically far smaller than a wavelength. As a result, for the three T the first inequality of Eq. (2.35) can be written approximately as

$$\sigma_{\max}^{\mathrm{r}_0} \gtrsim \frac{2k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \gtrsim \frac{|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \gtrsim \frac{k_{\max}^2}{\mu} \quad \text{in 2D}, \tag{2.37a}$$

$$\sigma_{\max}^{\mathrm{r}_0} \gtrsim \frac{3k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \gtrsim \frac{2|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \gtrsim \frac{2k_{\max}^2}{\mu} \quad \text{in 3D}, \tag{2.37b}$$

where we use the inequality (2.26) for further approximation.

Next, we consider the upper bound in Eq. (2.35). Calculating $\max_{\mathbf{k}} \|T_{\mathbf{k}}\|_1$ using Eq. (2.24), we obtain

$$\sigma_{\max}^{\mathrm{r}_0} \leq \frac{2k_{\max}^2}{\mu} + \omega^2|\varepsilon|, \quad \sigma_{\max}^{\mathrm{u}_0} \leq \frac{k_{\max}^2}{|s_x|\mu} + \frac{|s_x|k_{\max}^2}{\mu} + |s_x|\omega^2|\varepsilon|, \quad \sigma_{\max}^{\mathrm{sc}_0} \leq \frac{k_{\max}^2}{|s_x|^2\mu} + \frac{k_{\max}^2}{\mu} + \omega^2|\varepsilon| \tag{2.38}$$

in 2D, and

$$\sigma_{\max}^{\mathrm{r}_0} \leq \frac{3k_{\max}^2}{\mu} + \omega^2|\varepsilon|, \quad \sigma_{\max}^{\mathrm{u}_0} \leq \frac{k_{\max}^2}{|s_x|\mu} + \frac{2|s_x|k_{\max}^2}{\mu} + |s_x|\omega^2|\varepsilon|, \quad \sigma_{\max}^{\mathrm{sc}_0} \leq \frac{k_{\max}^2}{|s_x|^2\mu} + \frac{2k_{\max}^2}{\mu} + \omega^2|\varepsilon| \tag{2.39}$$

in 3D. Using the inequality (2.26) and ignoring the ω^2 terms again, we obtain

$$\sigma_{\max}^{\mathrm{r}_0} \lesssim \frac{2k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \lesssim \frac{|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \lesssim \frac{k_{\max}^2}{\mu} \quad \text{in 2D,} \tag{2.40a}$$

$$\sigma_{\max}^{\mathrm{r}_0} \lesssim \frac{3k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \lesssim \frac{2|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \lesssim \frac{2k_{\max}^2}{\mu} \quad \text{in 3D.} \tag{2.40b}$$

Because the approximate lower and upper bounds indicated in Eqs. (2.37) and (2.40) are the same for each of $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$, we have

$$\sigma_{\max}^{\mathrm{r}_0} \simeq \frac{2k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \simeq \frac{|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{k_{\max}^2}{\mu} \quad \text{in 2D,} \tag{2.41a}$$

$$\sigma_{\max}^{\mathrm{r}_0} \simeq \frac{3k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \simeq \frac{2|s_x|k_{\max}^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{2k_{\max}^2}{\mu} \quad \text{in 3D,} \tag{2.41b}$$

and therefore

$$\sigma_{\max}^{\mathrm{u}_0} \simeq \frac{|s_x|}{2}\sigma_{\max}^{\mathrm{r}_0} \quad \text{and} \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{1}{2}\sigma_{\max}^{\mathrm{r}_0} \quad \text{in 2D,} \tag{2.42a}$$

$$\sigma_{\max}^{\mathrm{u}_0} \simeq \frac{2|s_x|}{3}\sigma_{\max}^{\mathrm{r}_0} \quad \text{and} \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{2}{3}\sigma_{\max}^{\mathrm{r}_0} \quad \text{in 3D.} \tag{2.42b}$$

The result indicates a large contrast in magnitude between the maximum singular values of a homogeneous UPML and SC-PML: $\sigma_{\max}^{\mathrm{u}_0}$ is much larger than $\sigma_{\max}^{\mathrm{r}_0}$, whereas $\sigma_{\max}^{\mathrm{sc}_0}$ is smaller than $\sigma_{\max}^{\mathrm{r}_0}$.

We note that each estimate in Eq. (2.41) is realized by the corresponding $\sigma_{\mathbf{k},3}$ in Eq. (2.32) with appropriate $\mathbf{k}$; in 2D for example, the estimate of $\sigma_{\max}^{\mathrm{r}_0}$ is achieved by $\sigma_{\mathbf{k},3}^{\mathrm{r}_0}$ for $\mathbf{k}$ such that $|k_x| = |k_y| = k_{\max}$, and the estimates of $\sigma_{\max}^{\mathrm{u}_0}$ and $\sigma_{\max}^{\mathrm{sc}_0}$ are achieved by $\sigma_{\mathbf{k},3}^{\mathrm{u}_0}$ and $\sigma_{\mathbf{k},3}^{\mathrm{sc}_0}$ for $\mathbf{k}$ such that $k_x = 0$ and $k_y = \pm k_{\max}$. Therefore, in 2D, one of $\mathbf{k} = \pm[\hat{\mathbf{x}}k_{\max} \pm \hat{\mathbf{y}}k_{\max}]$ is an approximate wavevector of the maximum right singular vector corresponding to $\sigma_{\max}^{\mathrm{r}_0}$, and one of $\mathbf{k} = \pm\hat{\mathbf{y}}k_{\max}$ is an approximate wavevector of the maximum right singular vectors corresponding to $\sigma_{\max}^{\mathrm{u}_0}$ and $\sigma_{\max}^{\mathrm{sc}_0}$.

So far, when deriving the estimates of $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$, we have incorporated the effect of the finite-difference grid by simply imposing the upper bound $k_{\max}$ on wavevectors. The exact derivation that takes into account the finite-difference approximation of spatial

derivatives can be found in Appendix B. The results are

$$\sigma_{\max}^{\mathrm{r}_0} \simeq \frac{2(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \simeq \frac{|s_x|(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{(2/\Delta)^2}{\mu} \quad \text{in 2D,} \tag{2.43a}$$

$$\sigma_{\max}^{\mathrm{r}_0} \simeq \frac{3(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \simeq \frac{2|s_x|(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{2(2/\Delta)^2}{\mu} \quad \text{in 3D.} \tag{2.43b}$$

We note that the exact results in Eq. (2.43) differ from the approximate results in Eq. (2.41) by only a factor of $(2/\pi)^2$. Thus the approximate results presented in this section, which are simpler to derive, are in fact rather accurate. In particular, the main conclusion Eq. (2.42) of this section, which is obtained from the approximate results, turns out to hold for the exact results (2.43) as well.

2.3.3 Minimum singular values of homogeneous media

In this section, we examine the minimum singular values of a homogeneous regular medium, homogeneous UPML, and homogeneous SC-PML denoted by $\sigma_{\min}^{\mathrm{u}_0}$, $\sigma_{\min}^{\mathrm{sc}_0}$, and $\sigma_{\min}^{\mathrm{r}_0}$, respectively. Here, in addition to the assumptions $s_x = 1 - is_x''$ and $s_y = s_z = 1$ made about the PML scale factors in Sec. 2.3.2, we assume that the media have no gain, i.e., $\varepsilon'' \geq 0$ in $\varepsilon = \varepsilon' - i\varepsilon''$.

As in the previous section, here we also use the $\mathbf{k}$-space representations $T_{\mathbf{k}}^{\mathrm{r}_0}$, $T_{\mathbf{k}}^{\mathrm{u}_0}$, and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ of Eq. (2.30). We find $\sigma_{\min}^{\mathrm{r}_0}$, $\sigma_{\min}^{\mathrm{u}_0}$, and $\sigma_{\min}^{\mathrm{sc}_0}$ as the minima of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ over $\mathbf{k}$, respectively.

First, we derive the conditions for $T_{\mathbf{k}}^{\mathrm{r}_0}$, $T_{\mathbf{k}}^{\mathrm{u}_0}$, and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ to be singular. $T_{\mathbf{k}}^{\mathrm{r}_0}$ is singular when $\det(T_{\mathbf{k}}^{\mathrm{r}_0}) = -\omega^2\varepsilon(k_x^2/\mu + k_y^2/\mu - \omega^2\varepsilon)^2 = 0$, or equivalently

$$k_x^2 + k_y^2 = \omega^2\mu\varepsilon. \tag{2.44}$$

Similarly, $T_{\mathbf{k}}^{\mathrm{u}_0}$ and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ are singular when

$$\frac{k_x^2}{s_x^2} + k_y^2 = \omega^2\mu\varepsilon. \tag{2.45}$$

Now, suppose that ε is positive ($\varepsilon' > 0$, $\varepsilon'' = 0$). We see that Eq. (2.44) is satisfied by infinitely many real $\mathbf{k}$ lying on a circle in the $\mathbf{k}$-space, and Eq. (2.45) is satisfied by only

two real $\mathbf{k}$, i.e., $\mathbf{k} = \pm\hat{\mathbf{y}}\omega\sqrt{\mu\varepsilon}$, because s_x^2 has a nonzero imaginary part. Since a singular matrix has 0 as a singular value as pointed out in Sec. 2.3.1, each of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ is zero for some real $\mathbf{k}$, which implies that

$$\sigma_{\min}^{\mathrm{u}_0} = \sigma_{\min}^{\mathrm{sc}_0} = \sigma_{\min}^{\mathrm{r}_0} = 0 \quad \text{for positive } \varepsilon. \tag{2.46}$$

On the other hand, in cases where ε is either negative ($\varepsilon' < 0$, $\varepsilon'' = 0$) or complex ($\varepsilon'' > 0$), $T_{\mathbf{k}}^{\mathrm{r}_0}$, $T_{\mathbf{k}}^{\mathrm{u}_0}$, and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ are nonsingular for all real $\mathbf{k}$, because no real $\mathbf{k}$ satisfies Eq. (2.44) or (2.45). Therefore, we have

$$\sigma_{\min}^{\mathrm{u}_0}, \sigma_{\min}^{\mathrm{sc}_0}, \sigma_{\min}^{\mathrm{r}_0} > 0 \quad \text{for negative or complex } \varepsilon. \tag{2.47}$$

From Eqs. (2.46) and (2.47), we conclude that the minimum singular values of homogeneous media with positive ε (e.g., dielectrics and PMLs matching dielectrics) are always less than the minimum singular values of homogeneous media with negative ε or complex ε with $\varepsilon'' \geq 0$ (e.g., metals and PMLs matching metals).

2.3.4 Minimum singular values of homogeneous media with $\varepsilon > 0$ in a bounded domain

In Sec. 2.3.3, we have shown that the minimum singular values of a homogeneous regular medium, UPML, and SC-PML are all zero for $\varepsilon > 0$. The result has been obtained for homogeneous media in an infinite space. However, simulation domains are always bounded. In this section, we show that the minimum singular values of homogeneous media deviate from 0 in a bounded domain, even if $\varepsilon > 0$. We also compare the amount of deviation for different homogeneous media.

Throughout this section, we use the notation $c = 1/\sqrt{\mu\varepsilon}$; note that $c > 0$ because ε is assumed positive in this section.

For simplicity, suppose that the bounded domain in the xy-plane is a rectangle whose sides in the x- and y-directions are L_x and L_y, respectively. We impose periodic boundary conditions on the x- and y-boundaries of the bounded domain. Then, k_x and k_y are limited

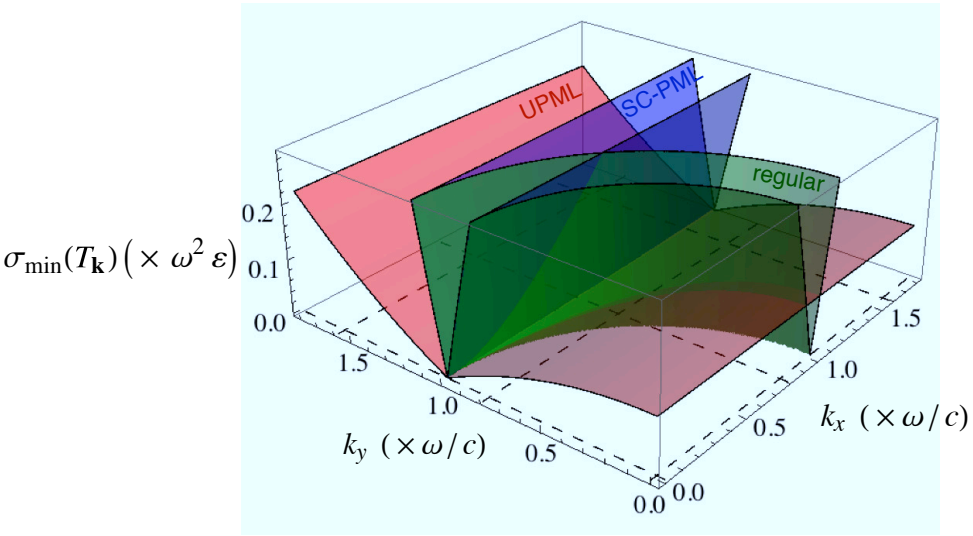


Figure 2.3: The 3D plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$ as functions of k_x and k_y. The three functions are drawn in a portion of the $\mathbf{k}$-space where the functions are close to zeros. The surface of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ is below that of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for all $\mathbf{k}$ displayed in the figure except $\mathbf{k} = \pm\hat{\mathbf{y}}(\omega/c)$ where the two surfaces are both zero. The surface of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$, on the other hand, is neither consistently below nor above the other two. The dashed lines in the $\sigma_{\min}(T_{\mathbf{k}}) = 0$ plane indicate $k_x \in K_x$ and $k_y \in K_y$, so the intersections of the dashed lines correspond to $\mathbf{k} \in K$. The rectangular simulation domain that quantizes k_x and k_y is a square of side length $L = 1.273\lambda_0$, where λ_0 is the vacuum wavelength corresponding to ω. The specific value of L is chosen so that no quantized $\mathbf{k}$ is at the zeros of the three functions. A PML scale factor $s_x = 1 - i10$ is used.

to the quantized values in the sets

$$K_x = \left\{ \frac{2\pi n_x}{L_x} : n_x \in \mathbb{Z}^+ \right\} \quad \text{and} \quad K_y = \left\{ \frac{2\pi n_y}{L_y} : n_y \in \mathbb{Z}^+ \right\}, \tag{2.48}$$

respectively, where $\mathbb{Z}^+$ is the set of nonnegative integers; due to mirror symmetry of a homogeneous UPML and SC-PML, it is sufficient to consider $k_x \geq 0$ and $k_y \geq 0$. For later use we also define the set of all quantized $\mathbf{k}$:

$$K = \left\{ \hat{\mathbf{x}} k_x + \hat{\mathbf{y}} k_y : k_x \in K_x, k_y \in K_y \right\}. \tag{2.49}$$

When there is no $\mathbf{k} \in K$ satisfying Eqs. (2.44) and (2.45), all of $\sigma_{\min}^{\mathrm{r}_0}$, $\sigma_{\min}^{\mathrm{u}_0}$, and $\sigma_{\min}^{\mathrm{sc}_0}$ deviate from 0 for a bounded domain, but by different amounts. Figure 2.3 shows $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ in a portion of the $\mathbf{k}$-space where they are close to zero. It shows that $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0}) < \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for all displayed $\mathbf{k}$ except $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ for which both $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are zero. Therefore, in general we expect $\min_{\mathbf{k}\in K} \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0}) < \min_{\mathbf{k}\in K} \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$, or equivalently $\sigma_{\min}^{\mathrm{u}_0} < \sigma_{\min}^{\mathrm{sc}_0}$. On the other hand, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$ can be either above, between, or below $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ in the figure. Hence, $\sigma_{\min}^{\mathrm{r}_0} = \min_{\mathbf{k}\in K} \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{r}_0})$ can be either less than, between, or greater than $\sigma_{\min}^{\mathrm{u}_0}$ and $\sigma_{\min}^{\mathrm{sc}_0}$, depending on the size of the bounded domain.

We now estimate an upper bound of $\sigma_{\min}^{\mathrm{u}_0}/\sigma_{\min}^{\mathrm{sc}_0}$ for a bounded domain. For that purpose, we examine the plots of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ in Fig. 2.3 in more detail. Figure 2.4a displays the same $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ shown in Fig. 2.3, but as a contour plot over an extended range of k_x. In Fig. 2.4a, we notice the following important features of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$:

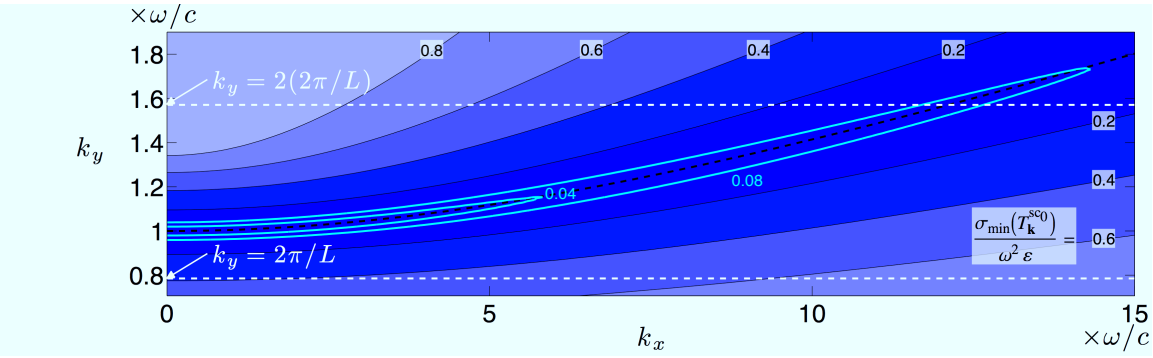
First, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ has a global minimum of zero at $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ due to the argument leading to Eq. (2.46); accordingly, the contours in the vicinity of the global minimum point form enclosing curves (cyan contours in Fig. 2.4a).

Second, the surface of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ has a "valley", where $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ is close to zero, along a curve in the $k_x k_y$-plane. The shape of the curve can be derived from Eq. (2.45), which describes the condition for $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ to be singular. Because of Eq. (2.26), the condition (2.45) is approximated by

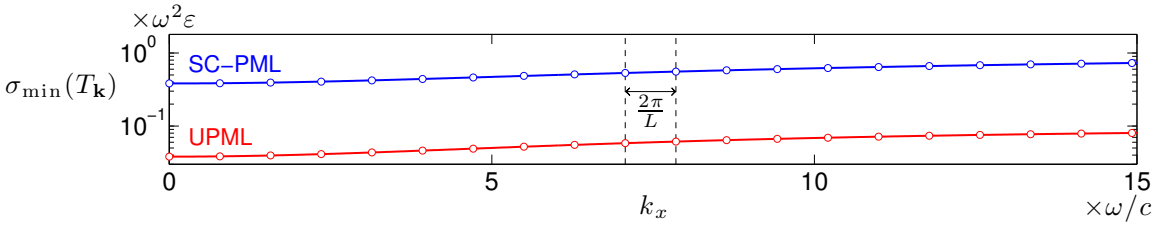
$$-\frac{k_x^2}{s_x''^2} + k_y^2 = \frac{\omega^2}{c^2}. \tag{2.50}$$

Hence, for $\mathbf{k}$ satisfying Eq. (2.50), $T_{\mathbf{k}}^{\mathrm{sc}_0}$ is *nearly* singular and has a close-to-zero singular value. Equation (2.50) thus describes the bottom of the valley of the $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ surface. The curve described by Eq. (2.50), which is a hyperbola that is indicated by a black dashed line in Fig. 2.4a, agrees well with the actual location of the bottom of the valley as can be seen from the contour plot.

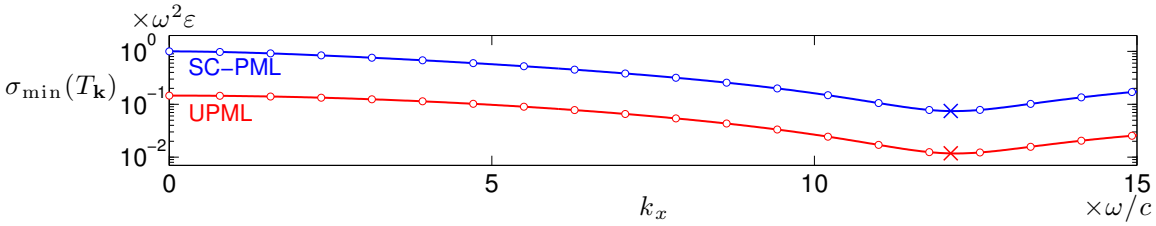
Third, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ varies much more slowly in k_x than in k_y; note that the scale of the k_y axis in Fig. 2.4a is exaggerated. This can be shown mathematically by examining Eq. (2.30c). We notice that interchanging k_x/s_x and k_y only swaps the $(1,1)$th and $(2,2)$th elements of the matrix and therefore does not change the singular values of $T_{\mathbf{k}}^{\mathrm{sc}_0}$. Hence, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ is



(a) Contour plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$



(b) $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for $k_y = (2\pi/L)$



(c) $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for $k_y = 2(2\pi/L)$

Figure 2.4: (a) The 2D contour plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$. The values of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})/\omega^2\varepsilon$ are overlaid on the corresponding solid contours; two cyan contours are drawn in addition to black contours to demonstrate that the contours are closed at large k_x's. The black dashed line is a hyperbola whose equation is Eq. (2.50), and describes the location of the valley very well. At the $k_y = 2\pi/L$ and $k_y = 2(2\pi/L)$ cross sections indicated by the two white dashed lines, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are plotted in (b) and (c). The horizontal axes are drawn using the same scale as that of (a), and the vertical axes are in a logarithmic scale. Note that the functions are minimized at $k_x = 0$ in (b), and around the "x" marks in (c). The horizontal locations of the small circles on the plots correspond to quantized k_x. All parameters are the same as those used in Fig. 2.3.

a symmetric function of k_x/s_x and k_y, and thus it has a stronger dependence on k_y than k_x since $|s_x| \gg 1$.

We do not display the contour plot of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$. However, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ also exhibits the three features described above.

Motivated by the third observation above, we derive an approximate upper bound of $\sigma_{\min}^{\mathrm{u}_0}/\sigma_{\min}^{\mathrm{sc}_0}$. Suppose that $k_y^{\mathrm{u}} \in K_y$ and $k_y^{\mathrm{sc}} \in K_y$ are the y-components of the quantized $\mathbf{k}$'s at which $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are minimized, respectively. Then, from the definitions of $\sigma_{\min}^{\mathrm{u}_0}$ and $\sigma_{\min}^{\mathrm{sc}_0}$ for a bounded domain, we have

$$\begin{aligned}\frac{\sigma_{\min}^{\mathrm{u}_0}}{\sigma_{\min}^{\mathrm{sc}_0}} &= \frac{\min_{k_x\in K_x}\min_{k_y\in K_y}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\min_{k_x\in K_x}\min_{k_y\in K_y}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})} = \frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})_{k_y=k_y^{\mathrm{u}}}}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})_{k_y=k_y^{\mathrm{sc}}}} \\ &\le \frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})_{k_y=k_y^{\mathrm{sc}}}}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})_{k_y=k_y^{\mathrm{sc}}}} \le \max_{k_y\in K_y}\left\{\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})}\right\}. \end{aligned} \tag{2.51}$$

Therefore, to estimate an upper bound of $\sigma_{\min}^{\mathrm{u}_0}/\sigma_{\min}^{\mathrm{sc}_0}$, we estimate

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})} \tag{2.52}$$

for all k_y. Because $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are slowly varying functions of k_x, we use the approximation

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})} \simeq \frac{\min_{k_x\ge 0}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\min_{k_x\ge 0}\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})} \tag{2.53}$$

to estimate Eq. (2.52).

We estimate the right-hand side of Eq. (2.53) for $k_y < \omega/c$ first. To visualize the general behaviors of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for such k_y, in Fig. 2.4b we plot them along the lower white dashed line of Fig. 2.4a. Figure 2.4b indicates that $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are minimized at $k_x = 0$ for $k_y < \omega/c$. In Appendix C.1 we show that in the limit of $s_x'' \gg 1$, which is the numerically relevant situation, $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are indeed minimized at $k_x = 0$ for all $k_y < \omega/c$. Therefore, we have

$$\min_{k_x\ge 0}\sigma_{\min}(T_{\mathbf{k}}) \simeq \sigma_{\min}(T_{\mathbf{k}})_{k_x=0} \quad \text{for } T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{u}_0}, T_{\mathbf{k}}^{\mathrm{sc}_0} \quad \text{for } k_y < \frac{\omega}{c}. \tag{2.54}$$

Since $T_{\mathbf{k}}^{\mathrm{u}_0}$ and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ of Eq. (2.30) are diagonalized for $k_x = 0$, the right-hand side of Eq. (2.54) is easily calculated as

$$\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{u}_0}\big)_{k_x=0} = \frac{1}{\mu|s_x|}\left(\frac{\omega^2}{c^2} - k_y^2\right) \quad \text{and} \quad \sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{sc}_0}\big)_{k_x=0} = \frac{1}{\mu}\left(\frac{\omega^2}{c^2} - k_y^2\right). \tag{2.55}$$

Combining Eq. (2.55) with Eqs. (2.53) and (2.54), we obtain

$$\frac{\min_{k_x\in K_x}\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{u}_0}\big)}{\min_{k_x\in K_x}\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{sc}_0}\big)} \simeq \frac{1}{|s_x|} \quad \text{for } k_y < \frac{\omega}{c}. \tag{2.56}$$

Next, we consider $k_y > \omega/c$. Such k_y is indicated by the upper white dashed line in Fig. 2.4a, along which $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are plotted in Fig. 2.4c. As seen in Fig. 2.4c, at such a given k_y the minima of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ occur in the valley, with the location of the minima very well-approximated by $k_x = s_x''[k_y^2 - \omega^2/c^2]^{1/2}$ (see Eq. (2.50)); this is shown more rigorously in Appendix C.1 for $s_x'' \gg 1$. Therefore, we have

$$\min_{k_x\geq 0}\sigma_{\min}(T_{\mathbf{k}}) \simeq \sigma_{\min}(T_{\mathbf{k}})_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}} \quad \text{for } T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{u}_0}, T_{\mathbf{k}}^{\mathrm{sc}_0} \quad \text{for } k_y > \frac{\omega}{c}. \tag{2.57}$$

By evaluating the right-hand side of Eq. (2.57) approximately, in Appendix C.2 we show that

$$\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{u}_0}\big)_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}} \simeq 2\omega^2\varepsilon\frac{k_y^2-\omega^2/c^2}{(s_x''^2+1)k_y^2-\omega^2/c^2}, \tag{2.58a}$$

$$\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{sc}_0}\big)_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}} \simeq \frac{2}{s_x''}\omega^2\varepsilon\frac{k_y^2-\omega^2/c^2}{2k_y^2-\omega^2/c^2}. \tag{2.58b}$$

The two "x" marks drawn at $k_x = s_x''\sqrt{k_y^2 - \frac{\omega^2}{c^2}}$ in Fig. 2.4c indicate the values determined by Eq. (2.58). The good agreement of the marks with the actual minima in the figure validates Eq. (2.58).

Combining Eq. (2.58) with Eqs. (2.53) and (2.57), we obtain

$$\frac{\min_{k_x\in K_x}\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{u}_0}\big)}{\min_{k_x\in K_x}\sigma_{\min}\big(T_{\mathbf{k}}^{\mathrm{sc}_0}\big)} \simeq \frac{2s_x''k_y^2 - s_x''\omega^2/c^2}{(s_x''^2+1)k_y^2-\omega^2/c^2} \quad \text{for } k_y > \frac{\omega}{c}. \tag{2.59}$$

For $k_y > \omega/c$, the right-hand side of Eq. (2.59) is an increasing function of k_y^2, so its maximum is attained at $k_y = \infty$. Hence, $\sigma_{\min}^{\mathrm{u}_0}/\sigma_{\min}^{\mathrm{sc}_0}$ is bounded from above as

$$\frac{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)}{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)} \lesssim \frac{2s_x''}{s_x''^2+1} \quad \text{for } k_y > \frac{\omega}{c}. \tag{2.60}$$

Lastly, we consider the case where k_y is very close to ω/c; such a case occurs either when $L_y \gg \lambda$ so that k_y is quantized very finely, or when $L_y \simeq m\lambda$ for some integer m so that $2\pi m/L_y \simeq \omega/c$, where $\lambda = 2\pi c/\omega$ is the wavelength in the medium described by ε. In this case, the minima of $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)$ and $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)$ over $k_x \in K$ occur in the vicinity of the global minimum point $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$. Hence, both minima are very close to zeros, and the approximation (2.53) we have used to derive Eqs. (2.56) and (2.60) may not be accurate enough. Therefore, we provide a direct estimate of Eq. (2.52) for $k_y \simeq \omega/c$.

For $k_y \simeq \omega/c$, suppose that $k_x^{\mathrm{u}} \in K_x$ and $k_x^{\mathrm{sc}} \in K_x$ are the x-components of the quantized $\mathbf{k}$'s at which $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)$ and $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)$ are minimized, respectively. Then, Eq. (2.52) satisfies

$$\frac{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)}{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)} = \frac{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_x^{\mathrm{u}}}}{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_x^{\mathrm{sc}}}} \le \frac{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_x^{\mathrm{sc}}}}{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_x^{\mathrm{sc}}}} \quad \text{for } k_y \simeq \frac{\omega}{c}. \tag{2.61}$$

It is reasonable to assume that k_x^{sc} is very close to $k_x = 0$, which is the x-component of the global minimum point. In Appendix C.3, we derive the lowest-order approximation of $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)/\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)$ around $k_x = 0$ and $k_y = \omega/c$. The result is

$$\frac{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_x^{\mathrm{sc}}}}{\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_x^{\mathrm{sc}}}} \simeq \frac{1}{|s_x|} \quad \text{for } k_y \simeq \frac{\omega}{c}. \tag{2.62}$$

From Eqs. (2.61) and (2.62), we have

$$\frac{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)}{\min_{k_x\in K_x}\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)} \lesssim \frac{1}{|s_x|} \quad \text{for } k_y \simeq \frac{\omega}{c}. \tag{2.63}$$

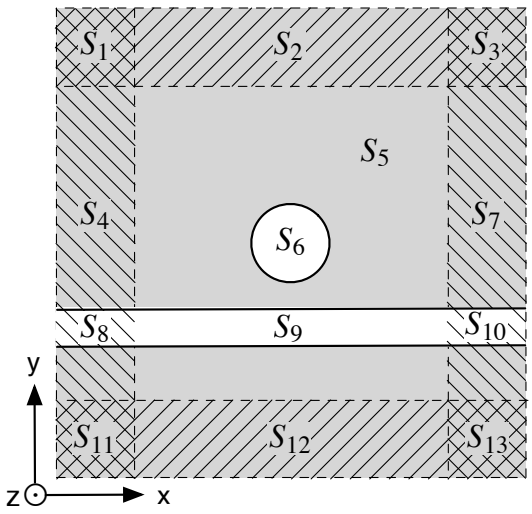


Figure 2.5: An example of an inhomogeneous EM system. The hypothetical system has a dielectric cavity (S_6) side-coupled to a dielectric waveguide (S_9) immersed in a background metal (S_5). The system is composed of several subdomains S_i, each of which is filled with a homogeneous medium. We define S_i as a domain excluding its boundary.

Combining Eqs. (2.56), (2.60), (2.63) with Eq. (2.51), we conclude that $\sigma_{\min}^{\mathrm{u}_0}/\sigma_{\min}^{\mathrm{sc}_0}$ is approximately bounded from above as

$$\frac{\sigma_{\min}^{\mathrm{u}_0}}{\sigma_{\min}^{\mathrm{sc}_0}} \lesssim \max\left\{\frac{1}{|s_x|}, \frac{2s_x''}{s_x''^2+1}, \frac{1}{|s_x|}\right\} \simeq \frac{2}{|s_x|}. \tag{2.64}$$

In summary, the minimum singular values of a homogeneous regular medium, UPML, and SC-PML for positive ε are all zero as shown in Eq. (2.46), but for a bounded domain they deviate from 0. When such deviation occurs, $\sigma_{\min}^{\mathrm{u}_0}$ is much smaller than $\sigma_{\min}^{\mathrm{sc}_0}$ as Eq. (2.64) describes, but $\sigma_{\min}^{\mathrm{r}_0}$ can be either less than, between, or greater than $\sigma_{\min}^{\mathrm{u}_0}$ and $\sigma_{\min}^{\mathrm{sc}_0}$.

2.3.5 Variational method to estimate extreme singular values of inhomogeneous EM systems

In this section, we provide general estimates of the extreme singular values of EM systems surrounded by either UPML or SC-PML using the variational method. An example of such EM systems is illustrated in Fig. 2.5. Because the EM system consists of several regular media

and PML, we refer to it as an *inhomogeneous* EM system to distinguish it from the homogeneous EM systems examined in the previous sections.

We estimate the extreme singular values of an inhomogeneous EM system using the variational method introduced in Eq. (2.19), and express them in terms of the extreme singular values of the homogeneous media examined in Secs. 2.3.2 through 2.3.4. Using the estimates, we show that

$$\frac{\sigma_{\max}^{\mathrm{u}}}{\sigma_{\max}^{\mathrm{sc}}} \gg 1, \tag{2.65}$$

$$\frac{\sigma_{\min}^{\mathrm{u}}}{\sigma_{\min}^{\mathrm{sc}}} \lesssim 1, \tag{2.66}$$

and therefore

$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma_{\max}^{\mathrm{u}}}{\sigma_{\max}^{\mathrm{sc}}} \frac{\sigma_{\min}^{\mathrm{sc}}}{\sigma_{\min}^{\mathrm{u}}} \gg 1. \tag{2.67}$$

The inequality (2.67) indicates that A^{u} is much worse-conditioned than A^{sc}.

As inferred from the discussion following Eq. (2.19), estimation of the extreme singular values by the variational method is closely related to estimation of the corresponding extreme right singular vectors. We use the notations $v_{\max}^{\mathrm{u}}, v_{\min}^{\mathrm{u}}$ and $v_{\max}^{\mathrm{sc}}, v_{\min}^{\mathrm{sc}}$ to refer to the extreme right singular vectors of A^{u} and A^{sc}.

A typical inhomogeneous EM system is composed of a few homogeneous subdomains S_i as illustrated in Fig. 2.5. At least one of the EM parameters of each subdomain is different from the corresponding parameter of the neighboring subdomains. We assume that all PML regions in the system have the same constant PML scale factors in their attenuation directions w, i.e.,

$$s_w(l) = s_0 = 1 - is_0'' \quad \text{and} \quad s_0'' \gg 1. \tag{2.68}$$

First, we estimate the maximum singular value of an inhomogeneous EM system. From Eq. (2.19), the maximum singular value $\sigma_{\max} = \sigma_{\max}(A)$ is the maximum of the quotient $r(x) = \|Ax\|/\|x\|$ over all x, where A is either A^{u} or A^{sc}. We consider the maximum of $r(x)$ over x whose nonzero elements are confined in a specific homogeneous subdomain S_i. Suppose that $x|_{S_i}$ is a column vector that has the same elements as x inside S_i and zeros

outside. We define

$$\sigma_{\max}|_{S_i} = \max_x r\big(x|_{S_i}\big). \tag{2.69}$$

Then, by the definition of $\sigma_{\max}$ we have

$$\sigma_{\max} \geq \max_i \sigma_{\max}|_{S_i}. \tag{2.70}$$

In addition, we have[4]

$$\begin{aligned}\sigma_{\max}^2 &= \max_x \frac{\|Ax\|^2}{\|x\|^2} \simeq \max_x \frac{\big\|\sum_i Ax|_{S_i}\big\|^2}{\big\|\sum_i x|_{S_i}\big\|^2} \simeq \max_x \frac{\sum_i \big\|Ax|_{S_i}\big\|^2}{\sum_i \big\|x|_{S_i}\big\|^2} \\ &= \max_x \Big(\textstyle\sum_i \rho_i(x) r\big(x|_{S_i}\big)^2\Big) \leq \max_x \Big(\textstyle\sum_i \rho_i(x) \big(\sigma_{\max}|_{S_i}\big)^2\Big),\end{aligned} \tag{2.71}$$

where

$$\rho_i(x) = \frac{\big\|x|_{S_i}\big\|^2}{\sum_j \big\|x|_{S_j}\big\|^2}. \tag{2.72}$$

Because $\sum_i \rho_i(x) = 1$, $\sum_i \rho_i(x)(\sigma_{\max}|_{S_i})^2$ is the weighted average of $(\sigma_{\max}|_{S_i})^2$ over all i, so it is always less than or equal to $\max_i(\sigma_{\max}|_{S_i})^2$. Thus Eq. (2.71) leads to

$$\sigma_{\max}^2 \lesssim \max_x \Big(\max_i \big(\sigma_{\max}|_{S_i}\big)^2\Big) = \max_i \big(\sigma_{\max}|_{S_i}\big)^2. \tag{2.73}$$

The two inequalities (2.70) and (2.73) dictate

$$\sigma_{\max} \simeq \max_i \sigma_{\max}|_{S_i}. \tag{2.74}$$

[4]Here we use four equalities $\|x\|^2 \simeq \big\|\sum_i x|_{S_i}\big\|^2 = \sum_i \big\|x|_{S_i}\big\|^2$ and $\|Ax\|^2 \simeq \big\|\sum_i Ax|_{S_i}\big\|^2 \simeq \sum_i \big\|Ax|_{S_i}\big\|^2$. Out of the four equalities, only $\big\|\sum_i x|_{S_i}\big\|^2 = \sum_i \big\|x|_{S_i}\big\|^2$ is exact because the elements of $x|_{S_i}$ are zeros at the boundary of S_i by definition (See the caption of Fig. 2.5) so that $x|_{S_i}$ is orthogonal to $x|_{S_j}$ for $i \neq j$. The other three equalities are approximate, because x and $\sum_i x|_{S_i}$ are different at the boundaries of the subdomains, and $Ax|_{S_i}$ is not necessarily orthogonal to $Ax|_{S_j}$ for neighboring S_i and S_j. Still, the approximations hold as long as the elements of a vector at the boundaries of the subdomains contribute negligibly to the norm of the vector.

Therefore, the maximum singular value of an inhomogeneous EM system can be approximated by the largest of the maximum singular values of the homogeneous subdomains constituting the inhomogeneous system. Accordingly, the maximum right singular vector $v_{\max}$ tends to be concentrated in a specific subdomain $S_i = S$ for which $\sigma_{\max} \simeq \sigma_{\max}|_S$.

Because $A\,x|_{S_i} = A_i\,x|_{S_i}$, where A_i is the operator for the homogeneous medium used in S_i, $\sigma_{\max}|_{S_i}$ is approximated as[5]

$$\sigma_{\max}|_{S_i} \simeq \begin{cases} \sigma_{\max}^{\mathrm{r}_0} & \text{for } S_i \text{ outside the PML region,} \\ \sigma_{\max}^{\mathrm{u}_0} & \text{for } S_i \text{ inside the UPML region,} \\ \sigma_{\max}^{\mathrm{sc}_0} & \text{for } S_i \text{ inside the SC-PML region.} \end{cases} \tag{2.75}$$

Here, we ignore the subdomains S_i where PMLs overlap (e.g., the four corners in Fig. 2.5), simply because they typically do not interact with incident waves strongly; we will see in Sec. 2.3.6 that this assumption is consistent with direct numerical calculations. Note that $\sigma_{\max}|_{S_i}$'s in Eq. (2.75) are independent of ε, because $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$ do not depend on ε as shown in Eq. (2.43).

We apply Eq. (2.75) to Eq. (2.74) for $A = A^{\mathrm{u}}$ and $A = A^{\mathrm{sc}}$ separately to estimate $\sigma_{\max}^{\mathrm{u}}$ and $\sigma_{\max}^{\mathrm{sc}}$. An inhomogeneous EM system consists of regular media and UPML for $A = A^{\mathrm{u}}$, and of regular media and SC-PML for $A = A^{\mathrm{sc}}$. Therefore, we have

$$\sigma_{\max}^{\mathrm{u}} \simeq \max\left\{\sigma_{\max}^{\mathrm{r}_0}, \sigma_{\max}^{\mathrm{u}_0}\right\} = \sigma_{\max}^{\mathrm{u}_0}, \tag{2.76}$$

$$\sigma_{\max}^{\mathrm{sc}} \simeq \max\left\{\sigma_{\max}^{\mathrm{r}_0}, \sigma_{\max}^{\mathrm{sc}_0}\right\} = \sigma_{\max}^{\mathrm{r}_0}, \tag{2.77}$$

where the magnitudes of $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$ are compared using Eq. (2.42). Equations (2.76) and (2.77) imply that $v_{\max}^{\mathrm{u}}$ and $v_{\max}^{\mathrm{sc}}$ tend to be concentrated in the UPML region and the region of regular media, respectively.

[5]For $\sigma_{\max}|_{S_i}$ to be approximated well by one of $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$, the subdomain S_i needs to be sufficiently large, because each homogeneous medium studied in Sec. 2.3.2 is assumed to fill an infinite space. However, for 2D problems for example, as described in the discussion following Eq. (2.41) the maximum right singular vectors $\mathbf{E}_{\mathbf{k}}$ of the three homogeneous media in Sec. 2.3.2 have $|\mathbf{k}| = \sqrt{2}k_{\max}$ or $|\mathbf{k}| = k_{\max}$, which correspond to the wavelengths $\sqrt{2}\Delta$ or 2Δ that are much smaller than the usual size of a subdomain. Hence, S_i is in effect an infinite space when the maximum singular value is concerned, which justifies the approximation (2.75).

From Eqs. (2.76), (2.77), and (2.42), we obtain

$$\frac{\sigma_{\max}^{\mathrm{u}}}{\sigma_{\max}^{\mathrm{sc}}} \simeq \frac{\sigma_{\max}^{\mathrm{u}_0}}{\sigma_{\max}^{\mathrm{r}_0}} \simeq \frac{|s_0|}{2} \quad \text{in 2D,} \tag{2.78a}$$

$$\frac{\sigma_{\max}^{\mathrm{u}}}{\sigma_{\max}^{\mathrm{sc}}} \simeq \frac{\sigma_{\max}^{\mathrm{u}_0}}{\sigma_{\max}^{\mathrm{r}_0}} \simeq \frac{2|s_0|}{3} \quad \text{in 3D,} \tag{2.78b}$$

which prove Eq. (2.65).

Next, we estimate the minimum singular value of an inhomogeneous EM system. Defining $\sigma_{\min} = \sigma_{\min}(A)$ and $\sigma_{\min}|_{S_i} = \min_x r(x|_{S_i})$, and following a process similar to Eqs. (2.70) through (2.73) except that now we minimize instead of maximize, we obtain

$$\sigma_{\min} \simeq \min_i \sigma_{\min}|_{S_i}, \tag{2.79}$$

which is a result that parallels Eq. (2.74). Therefore, the minimum singular value of an inhomogeneous EM system can be approximated by the smallest of the minimum singular values of the homogeneous subdomains constituting the inhomogeneous system. Accordingly, the minimum right singular vector $v_{\min}$ tends to be concentrated in a specific subdomain $S_i = S$ for which $\sigma_{\min} \simeq \sigma_{\min}|_S$.

Below, we make one more assumption. We assume that at least one of the PML subdomains (e.g., S_8 or S_{10} in Fig. 2.5) is adjacent to, and hence matches, a dielectric (as opposed to metallic) subdomain. This assumption is not very restrictive, because after all, as seen in the benchmark problems in Sec. 1.4, the purpose of using PML is to simulate situations where there are waves propagating out of the simulation domain; such outgoing waves are supported only in the presence of a dielectric adjacent to PML.

With this additional assumption, when looking for the smallest of $\sigma_{\min}|_{S_i}$'s in Eq. (2.79), we can ignore subdomains made of metals or lossy materials, because such materials always have larger minimum singular values than lossless dielectrics as shown in Sec. 2.3.3. Then, in Eq. (2.79) we only need to consider subdomains D_j made of dielectrics and subdomains P_k made of either UPML or SC-PML that match such dielectrics. For these subdomains, we

have

$$\sigma_{\min}\big|_{S_i} \simeq \begin{cases} \sigma_{\min}^{\mathrm{r}_0}\big|_{D_j} & \text{for } S_i = D_j, \\ \sigma_{\min}^{\mathrm{u}_0}\big|_{P_k} & \text{for } S_i = P_k \text{ inside the UPML region}, \\ \sigma_{\min}^{\mathrm{sc}_0}\big|_{P_k} & \text{for } S_i = P_k \text{ inside the SC-PML region}, \end{cases} \tag{2.80}$$

where $\sigma_{\min}^{\mathrm{r}_0}\big|_{D_j}$, $\sigma_{\min}^{\mathrm{u}_0}\big|_{P_k}$, and $\sigma_{\min}^{\mathrm{sc}_0}\big|_{P_k}$ are the minimum singular values of the three homogeneous media in a *bounded* domain examined in Sec. 2.3.4; the bounded domain in this case is either P_k or D_j.

We apply Eq. (2.80) to Eq. (2.79) for $A = A^{\mathrm{u}}$ and $A = A^{\mathrm{sc}}$ separately to estimate $\sigma_{\min}^{\mathrm{u}}$ and $\sigma_{\min}^{\mathrm{sc}}$. An inhomogeneous EM system consists of regular media and UPML for $A = A^{\mathrm{u}}$, and of regular media and SC-PML for $A = A^{\mathrm{sc}}$. Therefore, we have

$$\sigma_{\min}^{\mathrm{u}} \simeq \min\left\{\min_j \sigma_{\min}^{\mathrm{r}_0}\big|_{D_j}, \min_k \sigma_{\min}^{\mathrm{u}_0}\big|_{P_k}\right\} = \min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{u}_0}\big|_{P}\right\}, \tag{2.81}$$

$$\sigma_{\min}^{\mathrm{sc}} \simeq \min\left\{\min_j \sigma_{\min}^{\mathrm{r}_0}\big|_{D_j}, \min_k \sigma_{\min}^{\mathrm{sc}_0}\big|_{P_k}\right\} = \min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}\right\}, \tag{2.82}$$

where D, P, and P' are the subdomains that minimize $\sigma_{\min}^{\mathrm{r}_0}\big|_{D_j}$, $\sigma_{\min}^{\mathrm{u}_0}\big|_{P_k}$, and $\sigma_{\min}^{\mathrm{sc}_0}\big|_{P_k}$, respectively. Equations (2.81) and (2.82) imply that both $\nu_{\min}^{\mathrm{u}}$ and $\nu_{\min}^{\mathrm{sc}}$ tend to be concentrated in either a dielectric or a dielectric-matching PML. Whether they are in a dielectric or PML, however, depends on the magnitude of $\sigma_{\min}^{\mathrm{r}_0}\big|_{D}$ relative to $\sigma_{\min}^{\mathrm{u}_0}\big|_{P}$ and $\sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}$.

For the same subdomain P_k, $(\sigma_{\min}^{\mathrm{u}_0}\big|_{P_k})/(\sigma_{\min}^{\mathrm{sc}_0}\big|_{P_k}) \ll 1$ according to Eq. (2.64). Hence, we have

$$\frac{\sigma_{\min}^{\mathrm{u}_0}\big|_{P}}{\sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}} \le \frac{\sigma_{\min}^{\mathrm{u}_0}\big|_{P'}}{\sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}} \ll 1, \tag{2.83}$$

which results in

$$\frac{\sigma_{\min}^{\mathrm{u}}}{\sigma_{\min}^{\mathrm{sc}}} \simeq \frac{\min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{u}_0}\big|_{P}\right\}}{\min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}\right\}} \le \frac{\min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}\right\}}{\min\left\{\sigma_{\min}^{\mathrm{r}_0}\big|_{D}, \sigma_{\min}^{\mathrm{sc}_0}\big|_{P'}\right\}} = 1. \tag{2.84}$$

The inequality (2.84) directly leads to Eq. (2.66).

From Eqs. (2.78) and (2.84), we conclude that

$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma_{\max}^{\mathrm{u}}}{\sigma_{\max}^{\mathrm{sc}}} \frac{\sigma_{\min}^{\mathrm{sc}}}{\sigma_{\min}^{\mathrm{u}}} \gtrsim \frac{|s_0|}{2} \quad \text{in 2D}, \tag{2.85a}$$

$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{sc}}_{\max}} \frac{\sigma^{\mathrm{sc}}_{\min}}{\sigma^{\mathrm{u}}_{\min}} \gtrsim \frac{2|s_0|}{3} \quad \text{in 3D.} \tag{2.85b}$$

Therefore, the condition number of an inhomogeneous EM system surrounded by UPML is much larger than the condition number of the same EM system surrounded by SC-PML in general.

We end this section with two remarks. First, Eq. (2.84) does not necessarily mean that $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$ is close to 1. For example, consider a case where $\sigma^{\mathrm{r}_0}_{\min}|_D$ is greater than both $\sigma^{\mathrm{u}_0}_{\min}|_P$ and $\sigma^{\mathrm{sc}_0}_{\min}|_{P'}$ in Eqs. (2.81) and (2.82). Such a case leads to

$$\frac{\sigma^{\mathrm{u}}_{\min}}{\sigma^{\mathrm{sc}}_{\min}} \simeq \frac{\sigma^{\mathrm{u}_0}_{\min}|_P}{\sigma^{\mathrm{sc}_0}_{\min}|_{P'}} \leq \frac{\sigma^{\mathrm{u}_0}_{\min}|_{P'}}{\sigma^{\mathrm{sc}_0}_{\min}|_{P'}} \lesssim \frac{2}{|s_0|}, \tag{2.86}$$

where the last inequality is from Eq. (2.64). The inequality (2.86) demonstrates that $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$ can be much smaller than 1 indeed. This further implies that it is possible to have

$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{sc}}_{\max}} \frac{\sigma^{\mathrm{sc}}_{\min}}{\sigma^{\mathrm{u}}_{\min}} \gtrsim \frac{|s_0|^2}{4} \quad \text{in 2D,} \tag{2.87a}$$

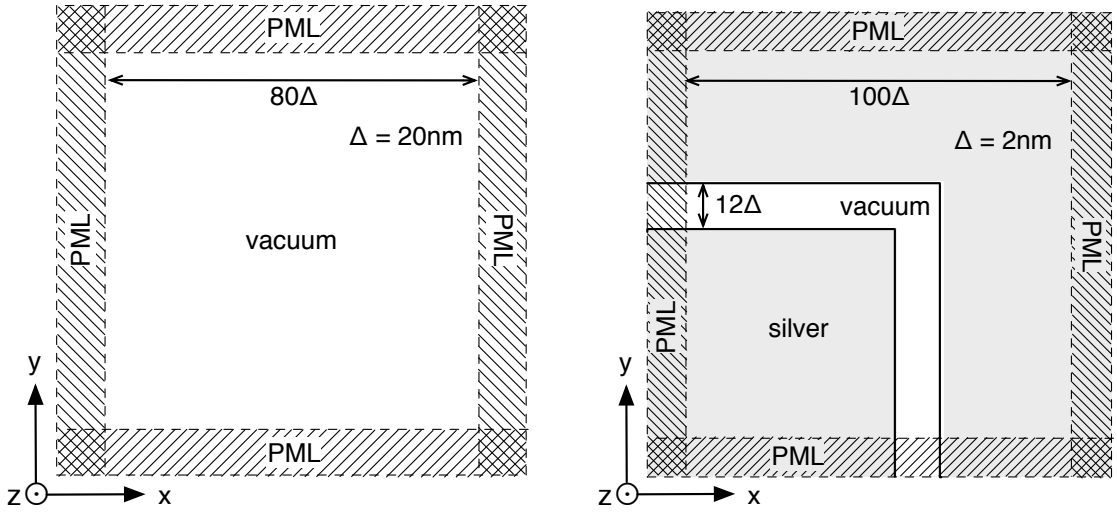
$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{sc}}_{\max}} \frac{\sigma^{\mathrm{sc}}_{\min}}{\sigma^{\mathrm{u}}_{\min}} \gtrsim \frac{|s_0|^2}{3} \quad \text{in 3D,} \tag{2.87b}$$

which predict much larger $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}$ than is expected from Eq. (2.85).

Second, as shown in Eq. (2.85), $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}$ increases with $|s_0|$. Therefore, in nanophotonics where $|s_0|$ can exceed 1000 as mentioned in Sec. 2.1, we expect the ratio between the condition numbers of the UPML and SC-PML matrices to be very large. Especially, when Eq. (2.87) holds, $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}$ can be of the order of 10^5.

2.3.6 Numerical validation

In this section, we numerically validate the analysis in Sec. 2.3.5. We consider two 2D EM systems as examples: a vacuum surrounded by PML (Fig. 2.6a), and a metal-dielectric-metal (MDM) waveguide bend surrounded by PML (Fig. 2.6b). For these two EM systems, we numerically calculate their extreme singular values as well as the corresponding extreme right singular vectors. We compare the behavior of these quantities to the discussions in the



(a) Vacuum (b) MDM waveguide bend

Figure 2.6: Two inhomogeneous EM systems whose extreme singular values and condition numbers are numerically calculated: (a) a vacuum surrounded by PML, and (b) a metal-dielectric-metal waveguide bend surrounded by PML. The grid cell sizes Δ of the uniform grids used to discretize Maxwell's equations are indicated in the figures. Relevant dimensions of the structures are displayed in terms of Δ. All PMLs are 10Δ thick. For both EM systems, the vacuum wavelength $\lambda_0 = 1550\,\text{nm}$ is used. In (b), the electric permittivity of silver [22] at λ_0 is $\varepsilon_{\text{Ag}} = (-129 - i3.28)\varepsilon_0$.

previous sections.

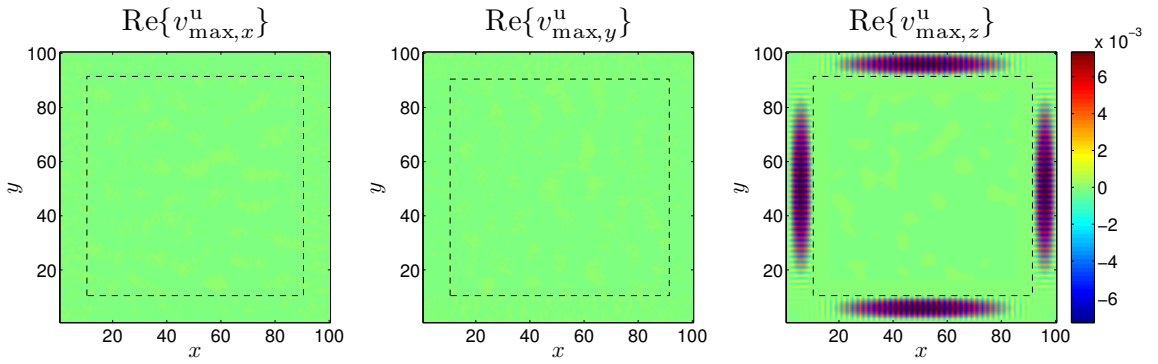
We first examine the system in Fig. 2.6a. Here, we use a constant PML loss parameter. With $\Delta = 20\,\text{nm}$, $d = 10\Delta$, $m = 0$, and $R = e^{-16} \simeq 1 \times 10^{-7}$ in Eqs. (2.6) and (2.7), the PML scale factor of Eq. (2.5) is

$$s_w(l) = s_0 = 1 - i9.868 \tag{2.88}$$

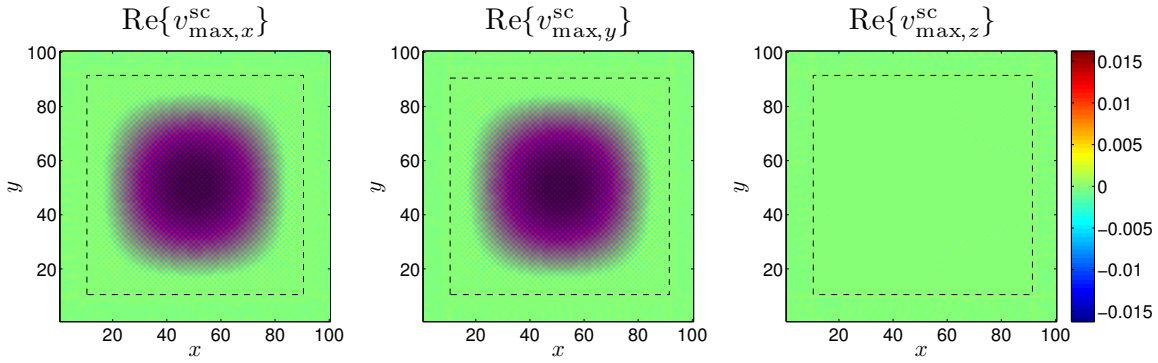
in each attenuation direction w.

Table 2.1 compares numerically calculated $\sigma^{\text{u}}_{\text{max}}$ and $\sigma^{\text{sc}}_{\text{max}}$ with their estimates derived in Eqs. (2.76) and (2.77). The agreement is very good with errors only about 0.1 ~ 0.2%. As a result, $\sigma^{\text{u}}_{\text{max}}/\sigma^{\text{sc}}_{\text{max}}$ is also estimated very accurately by $\sigma^{\text{u}_0}_{\text{max}}/\sigma^{\text{r}_0}_{\text{max}} \simeq |s_0|/2 = 4.959$, and thus Eq. (2.78) is validated.

We visualize numerically calculated $v^{\text{u}}_{\text{max}}$ and $v^{\text{sc}}_{\text{max}}$ in Fig. 2.7. Note that the figure plots



(a) v_{max} of the vacuum surrounded by UPML



(b) v_{max} of the vacuum surrounded by SC-PML

Figure 2.7: The maximum right singular vectors (a) $v^{\mathrm{u}}_{\mathrm{max}}$ of the vacuum surrounded by UPML, and (b) $v^{\mathrm{sc}}_{\mathrm{max}}$ of the vacuum surrounded by SC-PML. The real parts of the x-, y-, z-components of $v^{\mathrm{u}}_{\mathrm{max}}$ and $v^{\mathrm{sc}}_{\mathrm{max}}$ are displayed. Outside the dashed boxes are PMLs matching the vacuum, and both UPML and SC-PML are constructed with a constant PML loss parameter. Note that $v^{\mathrm{u}}_{\mathrm{max}}$ is concentrated in the UPML region, whereas $v^{\mathrm{sc}}_{\mathrm{max}}$ is concentrated in the vacuum region. Also notice the high-frequency oscillation of both the maximum right singular vectors. The numbers along the horizontal and vertical axes in each plot indicate the x- and y-indices of the grid cells.

	$\sigma^{\mathrm{u}}_{\max}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{sc}}_{\max}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{u}}_{\max}/\sigma^{\mathrm{sc}}_{\max}$
Numerical	9.896×10^{-2}	1.998×10^{-2}	4.953
Estimated	9.919×10^{-2}	2.000×10^{-2}	4.959

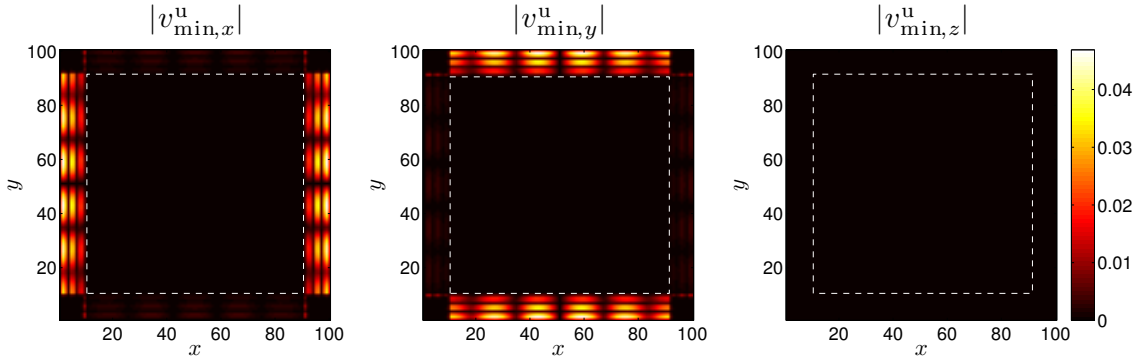
Table 2.1: The maximum singular values $\sigma^{\mathrm{u}}_{\max}$ and $\sigma^{\mathrm{sc}}_{\max}$ of the vacua surrounded by UPML and SC-PML, respectively, along with the ratio $\sigma^{\mathrm{u}}_{\max}/\sigma^{\mathrm{sc}}_{\max}$. Notice the excellent agreement between the estimates and numerically calculated values. The numerically calculated maximum singular values are obtained by solving Eq. (2.17) so that $\|Av_{\max}-\sigma_{\max}u_{\max}\|/\|u_{\max}\| < 10^{-11}$ for $A = A^{\mathrm{u}}, A^{\mathrm{sc}}$. The estimates of the maximum singular values are evaluated using $\sigma^{\mathrm{u}_0}_{\max}$ and $\sigma^{\mathrm{r}_0}_{\max}$ in Eq. (2.43) with $s_x = s_0$. The unit μ_0^{-1}/nm^2 of the singular values is the normalization factor used in our numerical solver.

the real parts of the x-, y-, z-components of $v_{\max}$; because $v_{\max}$ is the solution of Eq. (1.14) for the electric current source density $j = (i\sigma_{\max}/\omega)u_{\max}$, the x-, y-, z-components of $v_{\max}$ are well-defined as the Cartesian components of the solution E-field.

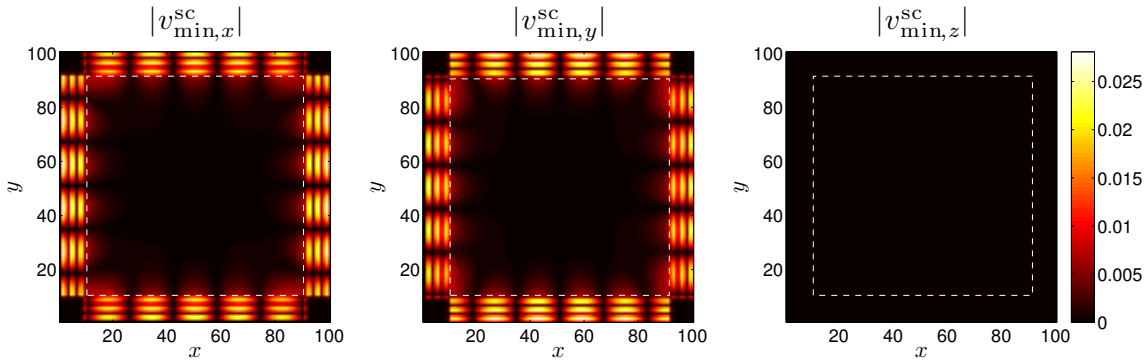
Figure 2.7 shows that $v^{\mathrm{u}}_{\max}$ is concentrated in the UPML region, whereas $v^{\mathrm{sc}}_{\max}$ is concentrated in the vacuum region. This is exactly what we expect from the discussion of Eqs. (2.76) and (2.77). Moreover, $v^{\mathrm{u}}_{\max}$ and $v^{\mathrm{sc}}_{\max}$ are indeed quite similar to the maximum right singular vectors of a homogeneous UPML and regular medium described in the discussion following Eq. (2.42). Notice that both $v^{\mathrm{u}}_{\max}$ and $v^{\mathrm{sc}}_{\max}$ exhibit fast spatial oscillations, but the oscillations have different wavevectors $\mathbf{k}$. For $v^{\mathrm{u}}_{\max}$, the dominant wavevector in each UPML section is normal to the attenuation direction, and the wavelength is 2Δ. Thus, in the x-normal UPML section for example, the dominant wavevector of $v^{\mathrm{u}}_{\max}$ is $\mathbf{k} = \pm\hat{\mathbf{y}}(2\pi/2\Delta)$. On the other hand, the dominant wavevector of $v^{\mathrm{sc}}_{\max}$ is $\mathbf{k} = \pm[\hat{\mathbf{x}}(2\pi/2\Delta) \pm \hat{\mathbf{y}}(2\pi/2\Delta)]$. These are exactly the wavevectors of the maximum right singular vectors of the homogeneous UPML and regular medium described in the discussion following Eq. (2.42).

We now examine the minimum singular values of the same system of Fig. 2.6a. Table 2.2 displays numerically calculated $\sigma^{\mathrm{u}}_{\min}$ and $\sigma^{\mathrm{sc}}_{\min}$ as well as the ratio between the two. The ratio is clearly less than 1, validating Eq. (2.84). Note that we do not have the estimates of the minimum singular values in the table, because in Sec. 2.3.5 we have provided only a general bound of the ratio $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$, but not detailed estimates of the individual minimum singular values.

Notice that $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$ in Table 2.2 is in fact close to $2/|s_0| = 0.2016$. This is consistent



(a) v_{min} of the vacuum surrounded by UPML



(b) v_{min} of the vacuum surrounded by SC-PML

Figure 2.8: The minimum right singular vectors (a) $v^{\mathrm{u}}_{\mathrm{min}}$ of the vacuum surrounded by UPML, and (b) $v^{\mathrm{sc}}_{\mathrm{min}}$ of the vacuum surrounded by SC-PML. The absolute values of the x-, y-, z-components of $v^{\mathrm{u}}_{\mathrm{min}}$ and $v^{\mathrm{sc}}_{\mathrm{max}}$ are displayed. Note that both the minimum right singular vectors are concentrated in the PML region. The numbers along the horizontal and vertical axes in each plot indicate the x- and y-indices of the grid cells.

	$\sigma^{\mathrm{u}}_{\min}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{sc}}_{\min}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$
Numerical	4.181×10^{-7}	1.975×10^{-6}	0.2117

Table 2.2: The minimum singular values $\sigma^{\mathrm{u}}_{\min}$ and $\sigma^{\mathrm{sc}}_{\min}$ of the vacua surrounded by UPML and SC-PML, respectively, along with the ratio $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$. Note that $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}\le 1$ as expected from Eq. (2.84). The numerically calculated minimum singular values are obtained by solving Eq. (2.17) so that $\|Av_{\min}-\sigma_{\min}u_{\min}\|/\|u_{\min}\|<10^{-11}$ for $A=A^{\mathrm{u}},A^{\mathrm{sc}}$. The unit μ_0^{-1}/nm^2 of the singular values is the normalization factor used in our numerical solver.

with $v^{\mathrm{u}}_{\min}$ and $v^{\mathrm{sc}}_{\min}$ shown in Fig. 2.8, where we plot the absolute values of the complex elements of each singular vector. We see that $v^{\mathrm{u}}_{\min}$ is concentrated in the UPML region, and $v^{\mathrm{sc}}_{\min}$ is concentrated in the SC-PML region. According to the discussion of Eqs. (2.81) and (2.82), this corresponds to a case where $\sigma^{\mathrm{r}_0}_{\min}|_D$ is greater than both $\sigma^{\mathrm{u}_0}_{\min}|_P$ and $\sigma^{\mathrm{sc}_0}_{\min}|_{P'}$. Then, $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$ satisfies Eq. (2.86) in addition to Eq. (2.84), which explains why $\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$ is close to the upper bound $2/|s_0|$ in Eq. (2.86). However, we note that $v^{\mathrm{u}}_{\min}$ and $v^{\mathrm{sc}}_{\min}$ are not always concentrated in the PML region; for the same system, it is actually possible to make them concentrated in the region of regular media (vacuum in the present case) by changing the wavelength or the size of the simulation domain.

Combining the results in Table 2.1 and 2.2, we obtain $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}=23.40\gg 1$, which is consistent with our conclusion in Sec. 2.3.5.

As a second example, we investigate the MDM waveguide bend in Fig. 2.6b. In this case, to be consistent with the typical use of PML in numerical simulations, we use a graded PML loss parameter $\sigma_w(l)$ rather than a constant one. With $\Delta=2\,\mathrm{nm}$, $d=10\Delta$, $m=4$, and $R=e^{-16}\simeq 1\times10^{-7}$ in Eqs. (2.6) and (2.7), the PML scale factor of Eq. (2.5) is

$$s_w(l)=s_0(l)=1-i493.4\left(\frac{l}{d}\right)^4 \tag{2.89}$$

in each attenuation direction w. Note that $|s_w(d)|$, which is the the maximum of $|s_w(l)|$, has increased from about 10 in Eq. (2.88) to about 500 in Eq. (2.89); the significant increase in $|s_w(d)|$ is due to two factors: the use of the graded PML loss parameter, and the reduction of Δ from 20 nm to 2 nm. Therefore, as discussed at the end of Sec. 2.3.5, we expect much larger $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}$ for this system than for the first example analyzed above.

Table 2.3 shows the numerically calculated extreme singular values of A^{u} and A^{sc} for the

	$\sigma^{\mathrm{u}}_{\max}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{sc}}_{\max}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{u}}_{\max}/\sigma^{\mathrm{sc}}_{\max}$
Numerical	5.167×10^2	2.001	258.2
Estimated	4.934×10^2	2.000	246.7

(a) Maximum singular values of the MDM waveguide bends

	$\sigma^{\mathrm{u}}_{\min}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{sc}}_{\min}(\times\mu_0^{-1}/\mathrm{nm}^2)$	$\sigma^{\mathrm{u}}_{\min}/\sigma^{\mathrm{sc}}_{\min}$
Numerical	2.095×10^{-6}	4.739×10^{-6}	0.4420

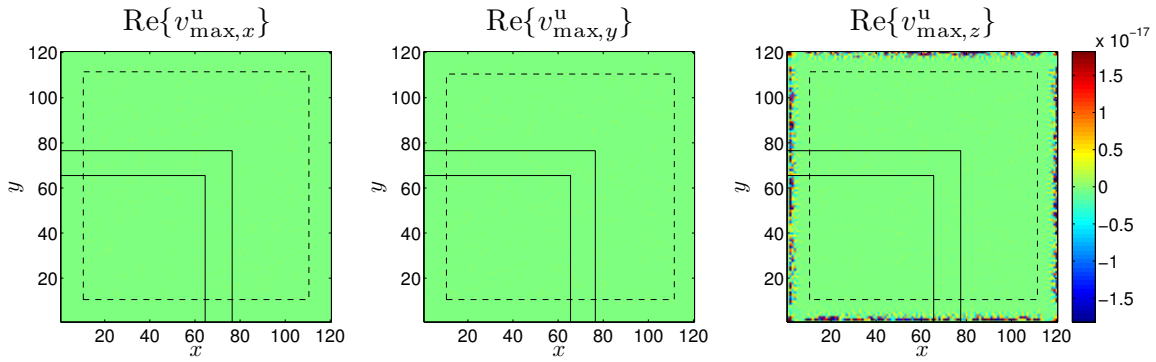
(b) Minimum singular values of the MDM waveguide bends

Table 2.3: The extreme singular values of the MDM waveguide bends surrounded by UPML and SC-PML. The extreme singular values are calculated by solving Eq. (2.17) so that $\|Av_i - \sigma_i u_i\|/\|u_i\| < 10^{-11}$ for $A = A^{\mathrm{u}}, A^{\mathrm{sc}}$. In (a), the estimates are evaluated using $\sigma^{\mathrm{u}_0}_{\max}$ and $\sigma^{\mathrm{r}_0}_{\max}$ in Eq. (2.43) with $s_x = s_0(d)$. Notice that $\sigma^{\mathrm{u}}_{\max}/\sigma^{\mathrm{sc}}_{\max}$ is much larger than it is in Table 2.1. The unit μ_0^{-1}/nm^2 of the singular values is the normalization factor used in our numerical solver.

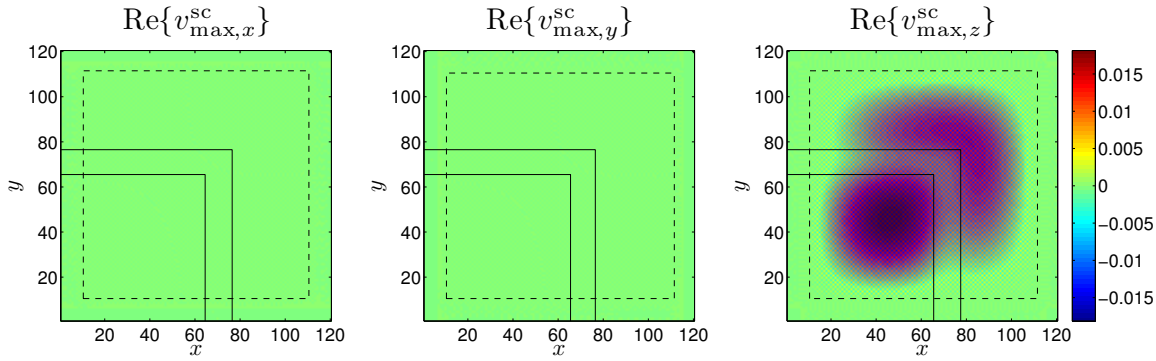
MDM waveguide bend. From the table, we confirm that both Eqs. (2.78) and (2.84) are satisfied. Also, we have much larger $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}}$ for this example than for the first example; for the present system, we have $\kappa^{\mathrm{u}}/\kappa^{\mathrm{sc}} = 584.2$.

In Table 2.3a, to estimate $\sigma^{\mathrm{u}}_{\max}$ as derived in Eq. (2.76), we have used $\sigma^{\mathrm{u}_0}_{\max}$ of Eq. (2.43). Strictly speaking, Eq. (2.43) is applicable only for UPML with a *constant* PML loss parameter. However, each UPML subdomain with a graded PML loss parameter can be treated as a stack of UPML subdomains, each of which has a constant PML loss parameter. In such a stack, the outermost UPML subdomain, which is closest to the edge of the simulation domain and described by the PML scale factor $s_0(d)$, has the largest $\sigma^{\mathrm{u}_0}_{\max}$. Hence, we use $\sigma^{\mathrm{u}_0}_{\max}$ in Eq. (2.43) with $s_x = s_0(d)$ as an estimate of $\sigma^{\mathrm{u}}_{\max}$ in Table 2.3a. The estimate agrees quite well with numerically calculated $\sigma^{\mathrm{u}}_{\max}$. Accordingly, $v^{\mathrm{u}}_{\max}$ is expected to be concentrated in the outermost layers of the graded UPML subdomains.

Figure 2.9 displays $v^{\mathrm{u}}_{\max}$ and $v^{\mathrm{sc}}_{\max}$ for the MDM waveguide bend. As discussed above, $v^{\mathrm{u}}_{\max}$ is indeed concentrated in the outermost UPML region, and $v^{\mathrm{sc}}_{\max}$ is also concentrated in the region of regular media as expected. In addition, both $v^{\mathrm{u}}_{\max}$ and $v^{\mathrm{sc}}_{\max}$ exhibit the same fast spatial oscillation as seen in the first example.

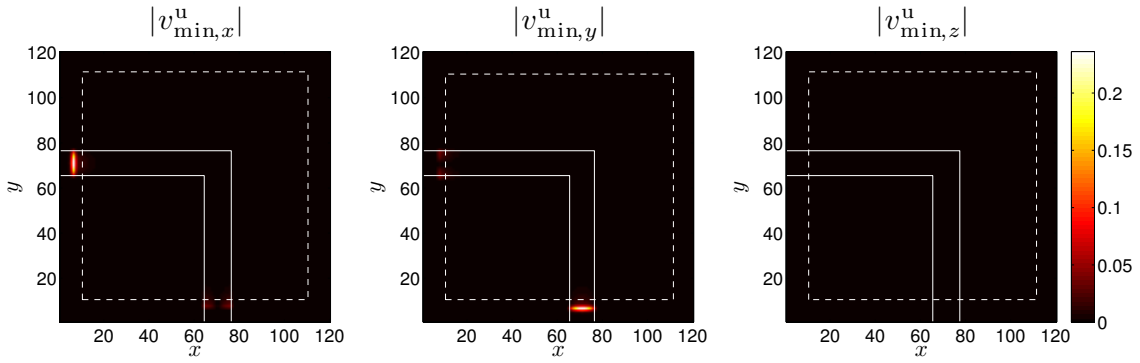


(a) v_{max} of the MDM waveguide bend surrounded by UPML

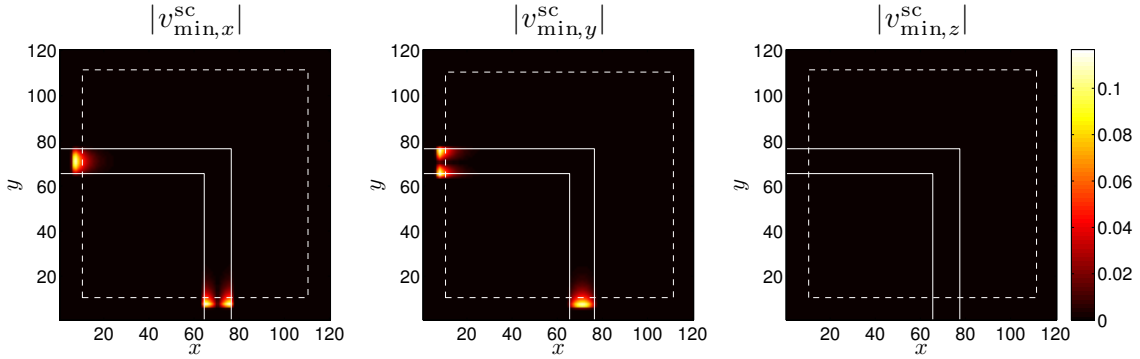


(b) v_{max} of the MDM waveguide bend surrounded by SC-PML

Figure 2.9: The maximum right singular vectors (a) $v^{\mathrm{u}}_{\mathrm{max}}$ of the MDM waveguide bend surrounded by UPML, and (b) $v^{\mathrm{sc}}_{\mathrm{max}}$ of the same waveguide bend surrounded by SC-PML. The real parts of the x-, y-, z-components of $v^{\mathrm{u}}_{\mathrm{max}}$ and $v^{\mathrm{sc}}_{\mathrm{max}}$ are displayed. Outside the dashed boxes are PMLs, and both UPML and SC-PML are constructed with graded PML loss parameters. The solid lines indicate the silver-vacuum interfaces; between the solid lines is vacuum. Note that $v^{\mathrm{u}}_{\mathrm{max}}$ is squeezed toward the boundary of the simulation domain where the PML loss parameters are maximized, whereas $v^{\mathrm{sc}}_{\mathrm{max}}$ is concentrated in the region of regular media. Also notice the high-frequency oscillation of both the maximum right singular vectors. The numbers along the horizontal and vertical axes in each plot indicate the x- and y-indices of the grid cells.



(a) v_{min} of the MDM waveguide bend surrounded by UPML



(b) v_{min} of the MDM waveguide bend surrounded by SC-PML

Figure 2.10: The minimum right singular vectors (a) $v^{\mathrm{u}}_{\mathrm{min}}$ of the MDM waveguide bend surrounded by UPML, and (b) $v^{\mathrm{sc}}_{\mathrm{min}}$ of the same waveguide bend surrounded by SC-PML. The absolute values of the x-, y-, z-components of $v^{\mathrm{u}}_{\mathrm{min}}$ and $v^{\mathrm{sc}}_{\mathrm{max}}$ are displayed. Note that the nonzero elements of both the minimum right singular vectors are mostly confined in the dielectric sections in the PML region. The numbers along the horizontal and vertical axes in each plot indicate the x- and y-indices of the grid cells.

We also display $v^{\mathrm{u}}_{\mathrm{min}}$ and $v^{\mathrm{sc}}_{\mathrm{min}}$ for the MDM waveguide bend in Fig. 2.10. Both the minimum right singular vectors are concentrated in the slot region, where the electric permittivity is ε_0. This follows the prediction in Sec. 2.3.5 that the minimum right singular vectors tend to be concentrated in neither metals nor PMLs matching metals, but in either dielectrics or PMLs matching dielectrics.

In summary of this section, all of the detailed predictions made in Sec. 2.3.5 about the behavior of the extreme singular values, extreme right singular vectors, and the condition numbers are validated numerically.

2.4 Diagonal preconditioning scheme for UPML

Our results in Secs. 2.2 and 2.3 strongly indicate that SC-PML is superior to UPML in solving the frequency-domain Maxwell's equations by iterative methods. However, there are cases where one would like to use UPML for practical reasons. For example, in FEM, UPML is easier to implement than SC-PML, because UPML is described by the same finite-element equation as regular media, whereas SC-PML is not [35, 50].

To use UPML in iterative solvers of the frequency-domain Maxwell's equations, one needs to accelerate convergence. For this purpose, Ref. [38] suggested to avoid overlap of UPMLs at the corners of the simulation domain, even though some reflection occurs at the corners as a result. The primary assumption in Ref. [38] was that the factors $s_{w_1}s_{w_2}/s_{w_3}$ in Eq. (2.2), which become especially large in overlapping UPML regions, resulted in an ill-conditioned matrix. However, the arguments in Sec. 2.3.5 show that even without overlap of UPMLs the matrix is still quite ill-conditioned. In addition, Figs. 2.7 and 2.8 illustrate that the extreme right singular vectors do not reside in the overlapping UPML regions, and thus at least for some EM systems, overlap of UPMLs is not the origin of the large condition number of the UPML matrix.

Reference [39] reported enhanced convergence speed achieved by using an approximate inverse preconditioner to the UPML matrix. However, the approximate inverse preconditioner requires solving an additional optimization problem, which can be time-consuming for large 3D EM systems.

In this section, we introduce a simple diagonal preconditioning scheme for the UPML

matrix to achieve accelerated convergence of iterative methods. We first explore the relation between the UPML matrix and SC-PML matrix in Sec. 2.4.1. Based on this relation, in Sec. 2.4.2 we devise the left and right diagonal preconditioners for the UPML matrix, and apply the preconditioners to the benchmark problem "Slot" described in Sec. 1.4 to demonstrate the effectiveness of the preconditioning scheme.

2.4.1 Relation between UPML and SC-PML

In this section, we relate the EM fields in a system surrounded by UPML with those in the same system surrounded by SC-PML. Both PMLs are assumed to have the same and constant PML scale factors.

Suppose that the SC-PML equation (2.3) has $\mathbf{E}^{\text{sc}}$ as the solution for a given electric current source density $\mathbf{J}^{\text{sc}}$. With straightforward substitution, we can show that the following E-field and electric current source density satisfy the UPML equation (2.1):

$$\mathbf{E}^{\text{u}} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \mathbf{E}^{\text{sc}}, \quad \mathbf{J}^{\text{u}} = \begin{bmatrix} s_y s_z & 0 & 0 \\ 0 & s_z s_x & 0 \\ 0 & 0 & s_x s_y \end{bmatrix} \mathbf{J}^{\text{sc}}. \tag{2.90}$$

The transformations in Eq. (2.90) can also be derived by applying the coordinate transformation of Maxwell's equations introduced in Ref. [51Appendix]. It is also interesting to note that the transformation for $\mathbf{E}$ in Eq. (2.90) predicts the discontinuity of the normal component of the E-field at the UPML interface as described in Ref. [6Sec. 7.5.2].

We note that the transformation for $\mathbf{E}$ in Eq. (2.90) was derived earlier in Refs. [52, 53]. However, the transformation for $\mathbf{J}$ in Eq. (2.90) has been mostly ignored so far, because the electric current source is usually placed *outside PML* where the transformation has no effect.

The transformations (2.90) can be written in terms of matrices and column vectors as

$$e^{\text{u}} = S_l e^{\text{sc}}, \quad j^{\text{u}} = S_a j^{\text{sc}}. \tag{2.91}$$

In the FDFD method, S_l and S_a are diagonal matrices whose diagonal elements are the length scale factors s_w and area scale factors $s_{w_1} s_{w_2}$, respectively.

Now, we relate A^{u} and A^{sc} using Eq. (2.91). Recall the systems of linear equations (2.9) and (2.10). In the present notation, they are

$$A^{\mathrm{u}} e^{\mathrm{u}} = -i\omega j^{\mathrm{u}}, \tag{2.92}$$

$$A^{\mathrm{sc}} e^{\mathrm{sc}} = -i\omega j^{\mathrm{sc}}, \tag{2.93}$$

considering Eq. (1.14). Substituting Eq. (2.91) in Eq. (2.92), we obtain

$$\left(S_a^{-1} A^{\mathrm{u}} S_l\right) e^{\mathrm{sc}} = -i\omega j^{\mathrm{sc}}. \tag{2.94}$$

Comparing Eq. (2.93) with Eq. (2.94), we conclude that

$$A^{\mathrm{sc}} = S_a^{-1} A^{\mathrm{u}} S_l. \tag{2.95}$$

We emphasize that the simple relation Eq. (2.95) between A^{u} and A^{sc} holds only for PMLs with constant PML scale factors; if the scale factors were not constant, the transformation in Ref. [51Appendix] would not transform the SC-PML equation into the UPML equation.

2.4.2 Scale-factor-preconditioned UPML

In actual numerical simulations where PMLs are implemented with graded PML loss parameters, the equality in Eq. (2.95) does not hold by the reason explained at the end of Sec. 2.4.1. Nevertheless, the right-hand side of Eq. (2.95) proposes a preconditioning scheme for the UPML matrix, which we refer to as the "scale-factor preconditioning scheme." In this preconditioning scheme, instead of solving the discretized UPML equation (2.9) directly, we first solve

$$\left(S_a^{-1} A^{\mathrm{u}} S_l\right) y = S_a^{-1} b \tag{2.96}$$

for y, and then recover the solution x of Eq. (2.9) as

$$x = S_l y. \tag{2.97}$$

The scale-factor preconditioning scheme does not change the kind of PML used in the

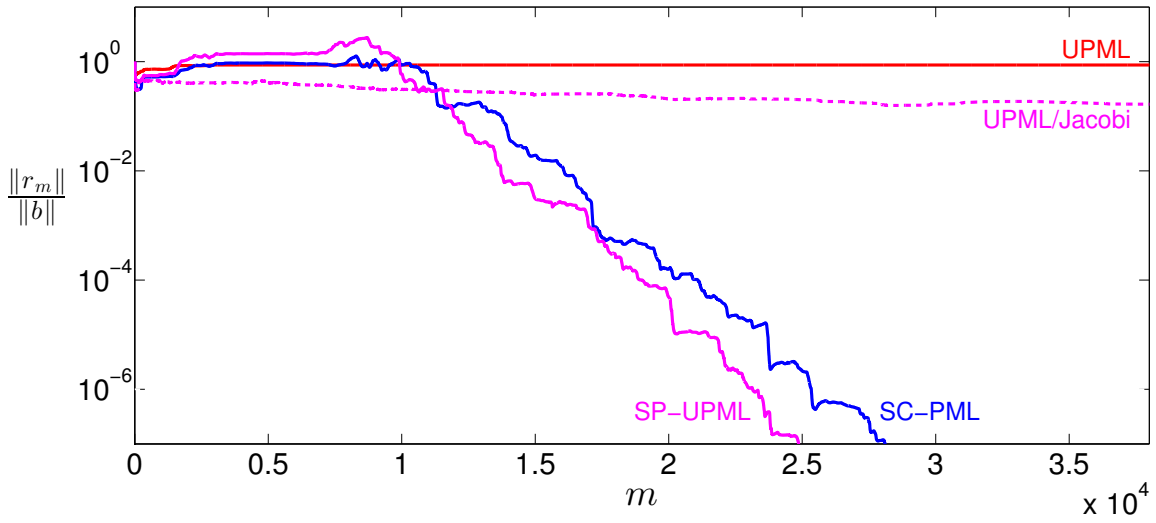


Figure 2.11: Convergence of QMR for the UPML equation, SC-PML equation, SP-UPML equation, and the UPML equation preconditioned by the Jacobi preconditioner. The examined EM system is the benchmark problem "Slot" described in Sec. 1.4, so the plots for the UPML and SC-PML equations are identical to the corresponding plots in Fig. 2.2. The solid and dashed magenta lines are for the UPML equation preconditioned by some preconditioners. Note that the convergence for the SP-UPML equation is as fast as that for the SC-PML equation, which shows the effectiveness of the scale-factor preconditioning scheme. On the other hand, the Jacobi preconditioning scheme barely improves the convergence for the UPML equation.

EM system from UPML; the solution x obtained from Eqs. (2.96) and (2.97) is exactly the solution of the discretized UPML equation (2.9). Even so, we refer to the implementation of UPML with the scale-factor preconditioning scheme as the "scale-factor-preconditioned UPML" (SP-UPML).

The SP-UPML matrix, $A^{\mathrm{sp}} = S_a^{-1} A^{\mathrm{u}} S_l$, is not equal to A^{sc} when S_a and S_l are constructed for graded PML loss parameters. However, we can expect it to have similar characteristics as A^{sc}, and therefore to be much better-conditioned than A^{u} itself. Hence, the discretized SP-UPML equation (2.96) can be much more favorable to numerical solvers than the discretized UPML equation.

As a numerical test, we solve the discretized SP-UPML equation by QMR for the the benchmark problem "Slot", which was solved also in Sec. 2.2. The convergence behavior for SP-UPML is depicted in Fig. 2.11, together with those for UPML and SC-PML. The figure

demonstrates that SP-UPML performs as well as SC-PML; in fact, it achieves slightly faster convergence than SC-PML.

To highlight the effectiveness of the scale-factor preconditioning scheme, we also plot $\|r_m\|/\|b\|$ for the UPML equation preconditioned by the conventional Jacobi preconditioner in Fig. 2.11. The system of linear equations for the Jacobi-preconditioned UPML equation is

$$P_{\text{jac}}^{-1} A^{\text{u}} x = P_{\text{jac}}^{-1} b, \tag{2.98}$$

where the Jacobi preconditioner P_{jac} is a diagonal matrix with the same diagonal elements as A^{u}. The Jacobi preconditioning scheme makes convergence for UPML slightly faster, but does not accelerate it as much as our proposed scale-factor preconditioning scheme.

The scale-factor preconditioning scheme also has a few advantages over the approximate inverse preconditioning scheme used in Ref. [39]. First, the scale-factor preconditioners S_a and S_l^{-1} are determined analytically using the PML scale factors, and do not require solving additional optimization problems. Second, the scale-factor preconditioners are diagonal, so they are much faster to apply and more efficient to store than any approximate inverse preconditioners.

2.5 Summary and remarks

SC-PML is more favorable to numerical solvers of the frequency-domain Maxwell's equations than UPML. For iterative solvers, SC-PML induces much faster convergence than UPML. For direct solvers, SC-PML promises more accurate solutions than UPML because it produces much better-conditioned matrices; the better-conditioned matrices also explain the faster convergence of iterative solvers for SC-PML.

Nevertheless, there are cases where UPML is easier to implement than SC-PML. In such cases, the scale-factor preconditioning scheme, which makes the UPML equation similar to the SC-PML equation, proves to be useful. This preconditioning scheme is much more effective than the conventional Jacobi preconditioning scheme and more efficient than the approximate inverse preconditioning scheme.

For numerical demonstrations, we constructed matrices by the FDFD method throughout the chapter, but we emphasize that the conclusions of this chapter are not limited to a specific method of discretizing the frequency-domain Maxwell's equations. For example, the condition number analysis in Sec. 2.3 was in essence estimation of the extreme singular values of the differential operators for homogeneous media. The scale-factor preconditioning scheme in Sec. 2.4 resulted from relating the UPML and SC-PML equations before discretization. None of these approaches depend on the FDFD method.

In particular, our conclusion should hold for the finite-element method of discretizing Maxwell's equations. In the major results, the only modification for FEM is that the scale-factor preconditioners S_a and S_l^{-1} in Sec. 2.4 may not be diagonal but can have up to 3 nonzero elements per row, because the edge elements in FEM are not necessarily in the Cartesian directions. This could make construction of the preconditioners somewhat more complex in FEM than in the FDFD method, but the existence of the preconditioners is still guaranteed. We can further make the preconditioners diagonal if, in 2D for example, we use a hybrid mesh that consists of rectangular elements inside PML and triangular elements outside PML.

Chapter 3

Accelerated solution by engineering the eigenvalue distribution[1]

Essentially, all models are wrong, but some are useful.

George E. P. Box (1919–2013)

In plasmonic and nanophotonic systems, objects often have deep-subwavelength feature size. When the frequency-domain Maxwell's equations for such systems are discretized, wavelengths are typically much longer than grid cell size Δ. The discretized equations are referred to as being in the "low-frequency regime," which will be defined more rigorously in Sec. 3.1.

In this chapter, we develop a technique that is effective in the low-frequency regime. For simplicity, we assume nonmagnetic materials (i.e., $\mu = \mu_0$) and solve the equation

$$\nabla \times \nabla \times \mathbf{E} - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J} \tag{3.1}$$

that is modified from Eq. (1.6).

[1]Reproduced in part with permission, from W. Shin and S. Fan, "Accelerated solution of the frequency-domain Maxwell's equations by engineering the eigenvalue distribution of the operator," submitted to *Optics Express* for publication. Unpublished work copyright 2013 OSA.

In the low-frequency regime it is well-known that convergence is quite slow when iterative methods are directly applied to solve Eq. (3.1). The huge null space of the operator $\nabla \times (\nabla \times \quad)$ was shown to be the origin of the slow convergence [54, 55], and several techniques to improve the convergence speed have been developed.

The first class of techniques is based on the Helmholtz decomposition, which decomposes the E-field as $\mathbf{E} = \Psi + \nabla\varphi$, where Ψ is a divergence-free vector field and φ is a scalar field [54–60]. Because $\nabla \cdot \Psi = 0$, Eq. (3.1) is written as

$$-\nabla^2\Psi - \omega^2\mu_0\varepsilon(\Psi + \nabla\varphi) = -i\omega\mu_0\mathbf{J}, \tag{3.2}$$

where the operator $\nabla \times (\nabla \times \quad)$, which has a huge null space, is replaced with the negative Laplacian $-\nabla^2$, which is positive-definite for appropriate boundary conditions and thus has the smallest possible null space. However, these techniques either solve an extra equation for the extra unknown φ at every iteration step [54–57], which can be time-consuming, or increase the number of the rows and columns of the matrix by about 33% [58–60], which requires more memory.

The second class of techniques utilizes the charge-free condition

$$\nabla \cdot (\varepsilon\mathbf{E}) = 0. \tag{3.3}$$

The condition (3.3) holds at every source-free (i.e., $\mathbf{J} = 0$) position, where Eq. (3.1) can be modified to

$$\nabla \times \nabla \times \mathbf{E} + s\nabla\left[\nabla \cdot \left((\varepsilon/\varepsilon_0)\mathbf{E}\right)\right] - \omega^2\mu_0\varepsilon\mathbf{E} = 0 \tag{3.4}$$

for an arbitrary constant s; note that the right-hand side is 0 because $\mathbf{J} = 0$. In this class of techniques, Eqs. (3.1) and (3.4) are solved at positions with and without sources, respectively.

Reference [61] applied the above technique with $s = +1$ to boundary value problems described in Ref. [62] and achieved accelerated convergence. Such boundary value problems satisfied $\mathbf{J} = 0$ everywhere, so Eq. (3.4) was solved throughout the entire simulation domain.

However, Ref. [61] did not conduct a detailed comparison of convergence speed between different values of s. It also did not report whether its technique leads to accelerated convergence for problems with sources, even though many problems have nonzero electric current

sources **J** inside the simulation domain. Reference [1] applied the technique with $s = +1$ to problems with sources, but only in order to suppress spurious modes rather than to accelerate convergence.

In this chapter, we develop a modification of Eq. (3.1) that improves convergence speed even if electric current sources **J** exist inside the simulation domain.[2] Unlike the previous technique that made the modification only at source-free positions, our technique modifies Eq. (3.1) everywhere including positions with sources. For the modification, we utilize the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0, \quad \text{or} \quad \nabla \cdot (\varepsilon \mathbf{E}) = \frac{i}{\omega} \nabla \cdot \mathbf{J}, \tag{3.5}$$

which can be derived by taking the divergence of Eq. (3.1). When Eq. (3.5) is manipulated appropriately and then added to Eq. (3.1), we obtain

$$\nabla \times \nabla \times \mathbf{E} + s \nabla \left[\varepsilon^{-1} \nabla \cdot (\varepsilon \mathbf{E}) \right] - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J} + s \frac{i}{\omega} \nabla \left[\varepsilon^{-1} \nabla \cdot \mathbf{J} \right] \tag{3.6}$$

for a constant s. The modified equation (3.6) is the equation to solve in this chapter.

The solution E-field of Eq. (3.6) is the same as the solution of the original equation (3.1) regardless of the value of s, because the solution of Eq. (3.1) always satisfies Eq. (3.5). However, the choice of s affects the convergence speed of iterative methods significantly. In this chapter, we demonstrate that $s = -1$ induces faster convergence speed than other values of s by comparing the convergence behavior of iterative methods for $s = -1, 0, +1$; the latter two values of s are of particular interest, because $s = 0$ reduces Eq. (3.6) to the original equation (3.1) and $s = +1$ is the value that Ref. [61] used in Eq. (3.4), which is similar to Eq. (3.6).

We also show that the difference in convergence behavior results from the different eigenvalue distributions of the operators for different s. There are many general mathematical studies about the dependence of the convergence behavior on the eigenvalue distribution [42[Sec. 9.2], 63–69]. Our aim here is instead to provide an intuitive understanding of the convergence behavior specifically for the operator of Eq. (3.6). For this purpose, we visualize the residual vector and residual polynomial at each iteration step. As a result, we find that

[2] At the final stage of our work, we were made aware of a related work by M. Kordy, E. Cherkaev, and P. Wannamaker, "Schelkunoff potential for electromagnetic field: proof of existence and uniqueness" (to be published), where an equation similar to our Eq. (3.6) with $s = -1$ was developed.

convergence speed deteriorates substantially for $s = 0$ because the operator has eigenvalues clustered near zero, and for $s = +1$ because the operator is strongly indefinite.

The rest of this chapter is organized as follows. In Sec. 3.1 we investigate the eigenvalue distribution of the operator in Eq. (3.6) for $s = 0, -1, +1$ for a simple homogeneous system. We also define the low-frequency regime rigorously in the section. In Sec. 3.2, we relate the eigenvalue distribution with the convergence behavior of an iterative method. In Sec. 3.3, we solve Eq. (3.6) for the benchmark problems described in Sec. 1.4 to compare the convergence behavior of an iterative method for the three values of s. In Sec. 3.4 we summarize the chapter and make a few remarks.

3.1 Eigenvalue distribution of the operator for a homogeneous system

In this section, we consider the operator in Eq. (3.6) for a homogeneous system and show that the properties of the eigenvalue distribution of the operator strongly depend on the value of s. The impact of s on the eigenvalue distribution has been studied in detail in the literature of the deflation method (also known as the penalty method) [70–72]. Here we only highlight those aspects that are important for the present study.

For a homogeneous system where ε is constant, Eq. (3.6) is simplified to

$$\nabla \times \nabla \times \mathbf{E} + s\nabla(\nabla \cdot \mathbf{E}) - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J} + s\frac{i}{\omega\varepsilon}\nabla(\nabla \cdot \mathbf{J}), \tag{3.7}$$

where the operator

$$T = \nabla \times (\nabla \times \quad) + s\nabla(\nabla \cdot \quad) - \omega^2 \mu_0 \varepsilon \tag{3.8}$$

is Hermitian for real ε. Because ε is constant in this section, the eigenvalue distribution of T is shifted from the eigenvalue distribution of a Hermitian operator

$$T_0 = \nabla \times (\nabla \times \quad) + s\nabla(\nabla \cdot \quad) \tag{3.9}$$

by a constant $-\omega^2 \mu_0 \varepsilon$. In the low-frequency regime such shift is negligible, and thus the

	$s = 0$	$s < 0$	$s > 0$
multiplicity of $\lambda = 0$	very high	low	
definiteness of T_0	positive-semidefinite		indefinite

Table 3.1: Properties of the eigenvalue distributions of T_0 for different s. Depending on the sign of s, T_0 has very different eigenvalue distributions in terms of the multiplicity of the eigenvalue 0 and the definiteness of T_0

eigenvalue distribution of T_0 approximates that of T very well. Hence, we examine the eigenvalue distribution of T_0 below to investigate the eigenvalue distribution of T.

In Appendix D, we show that $\mathbf{F}_\mathbf{k} e^{-i\mathbf{k}\cdot\mathbf{r}}$ with

$$\mathbf{F}_\mathbf{k} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \begin{bmatrix} k_z \\ 0 \\ -k_x \end{bmatrix}, \begin{bmatrix} -k_y \\ k_x \\ 0 \end{bmatrix} \tag{3.10}$$

are the three eigenfunctions of both $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ for each wavevector $\mathbf{k}$. We also show in the same appendix that the corresponding three eigenvalues are

$$\lambda = 0, \ |\mathbf{k}|^2, \ |\mathbf{k}|^2 \tag{3.11}$$

for $\nabla \times (\nabla \times \quad)$, and

$$\lambda = -|\mathbf{k}|^2, \ 0, \ 0 \tag{3.12}$$

for $\nabla(\nabla \cdot \quad)$. Therefore, T_0 has

$$\lambda = -s|\mathbf{k}|^2, \ |\mathbf{k}|^2, \ |\mathbf{k}|^2 \tag{3.13}$$

as three eigenvalues for each wavevector $\mathbf{k}$.

Equation (3.13) indicates that the eigenvalue distribution of T_0 is greatly affected by the value of s. Specifically, the multiplicity of the eigenvalue 0 depends critically on whether s is 0 or not: for $s = 0$ T_0 has a very high multiplicity of the eigenvalue 0 because Eq. (3.13) has 0 as an eigenvalue for every $\mathbf{k}$, whereas for $s \neq 0$ T_0 does not have such a high multiplicity of the eigenvalue 0. The definiteness of T_0 also depends on the value of s: for $s \leq 0$ T_0

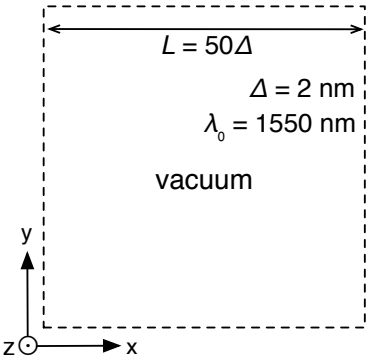
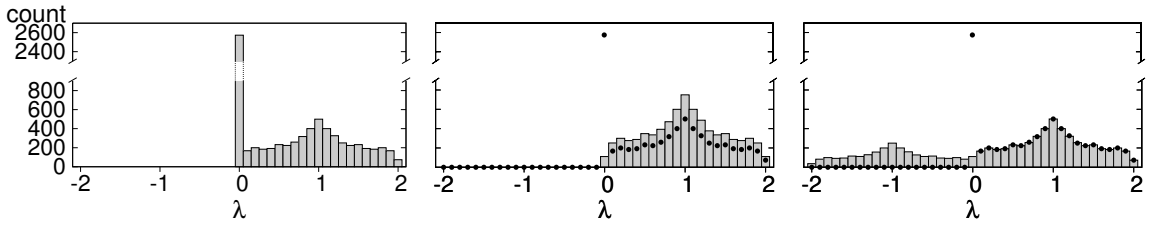


Figure 3.1: A 2D square domain filled with vacuum ($\varepsilon = \varepsilon_0$) for which the eigenvalue distribution of T is calculated numerically for $s = 0, -1, +1$. The domain is homogeneous in the z-direction, whereas its x- and y-boundaries are subject to periodic boundary conditions. The square domain is discretized on a finite-difference grid with cell size $\Delta = 2\,\text{nm}$. The domain is composed of 50×50 grid cells, which lead to 7500 eigenvalues in total. A vacuum wavelength $\lambda_0 = 1550\,\text{nm}$, which puts the system in the low-frequency regime, is assumed for the electric current source to be used in Sec. 3.2.

is positive-semidefinite because the three eigenvalues in Eq. (3.13) are always nonnegative, whereas for $s > 0$ T_0 is indefinite because Eq. (3.13) has both positive and negative numbers as eigenvalues. The different properties of the eigenvalue distributions of T_0 for $s = 0$, $s < 0$, and $s > 0$ are summarized in Table 3.1.

The above description of the eigenvalue distributions of T_0 should approximately hold for the eigenvalue distributions of T as well in the low-frequency regime, as mentioned in the discussion of Eqs. (3.8) and (3.9). Moreover, even though T is a differential operator defined in an infinite space, it turns out that the description also applies to the matrix A discretized from T that is defined in a spatially bounded simulation domain.

To demonstrate, we numerically calculate the eigenvalues of A for a 2D system shown in Fig. 3.1, a square domain filled with vacuum. The domain is discretized on a finite-difference grid with $N_x \times N_y = 50 \times 50$ cells and cell size $\Delta = 2\,\text{nm}$. Therefore, the matrix A for each s has $3N_xN_y = 7500$ rows and columns, where the extra factor 3 accounts for the three Cartesian components of the E-field. We choose ω corresponding to the vacuum wavelength $\lambda_0 = 1550\,\text{nm}$, which puts the system in the low-frequency regime as will be seen at the end of this section. The matrices A are constructed for three values of s: 0, −1, and +1, each of



(a) $s = 0$ (b) $s = -1$ (c) $s = +1$

Figure 3.2: The eigenvalue distribution of A discretized from T for (a) $s = 0$, (b) $s = -1$, and (c) $s = +1$ for the vacuum-filled domain illustrated in Fig. 3.1. All 7500 eigenvalues λ of A are calculated for each s and categorized into 41 intervals in the horizontal axis that represents the range of the eigenvalues; the unit of the horizontal axis is nm^{-2}. The height of the column on each interval represents the number of the eigenvalues in the interval. In (b) and (c), the black dots indicate the eigenvalue distribution for $s = 0$ shown in (a). The vertical axes are broken due to the extremely tall column at $\lambda \simeq 0$ in (a). The local maxima at $\lambda = \pm 1\,\mathrm{nm}^{-2}$ are the Van Hove singularities [73Ch. 8] arising from the lattice structure imposed by the finite-difference grid.

which represents each category of s in Table 3.1.

The distributions of the numerically calculated eigenvalues of A for $s = 0, -1, +1$ are shown as three plots in Fig. 3.2. In each plot, the horizontal axis represents eigenvalues, and it is divided into 41 intervals $t_{-20}, \ldots, t_0, \ldots, t_{20}$ where $t_0 \ni 0$. The height of the column on each interval corresponds to the number of the eigenvalues in the interval.

The eigenvalue distributions of A shown in Fig. 3.2 agree well with the description of the eigenvalues of T_0 in Table 3.1: the very tall column on t_0 in Fig. 3.2a indicates the very high multiplicity of $\lambda \simeq 0$ for $s = 0$, and the eigenvalues distributed over $t_{j<0}$ and $t_{j>0}$ in Fig. 3.2c indicate a strongly indefinite operator for $s = +1$. In addition, the height of the column on t_0 in Fig. 3.2a is about 2500, or one third of the total number of eigenvalues, which agrees with Eq. (3.13) for $s = 0$ where one of the three eigenvalues is 0 for each $\mathbf{k}$; the columns on $t_{j>0}$ are about 1.5 times taller in Fig. 3.2b than in Fig. 3.2a, which also agrees with Eq. (3.13) where the number of $|\mathbf{k}|^2$ increases from two for $s = 0$ to three for $s = -1$.

We end the section by providing a quantitative definition of the low-frequency regime. Suppose that A_0 is the matrix discretized from T_0 of Eq. (3.9). For $s = 0$, the eigenvalues of

A_0 range from 0 to $8/\Delta_{\min}^2$, where $\Delta_{\min}$ is the minimum grid cell size;[3] note that the range agrees with Fig. 3.2a. The eigenvalue distribution of A is the shifted eigenvalue distribution of A_0 by $-\omega^2\mu_0\varepsilon$. The low-frequency regime is where the magnitude of the shift is so small that A has an almost identical eigenvalue distribution as A_0. Therefore, the condition for the low-frequency regime is

$$\omega^2\mu_0|\varepsilon| \ll 8/\Delta_{\min}^2. \tag{3.14}$$

Equation (3.14) is consistent with the condition introduced in Ref. [54], but here we provide a condition that is based on a more accurate estimate of the maximum eigenvalue of A_0. We can rewrite Eq. (3.14) in terms of the vacuum wavelength λ_0 as

$$\lambda_0/\Delta_{\min} \gg \pi\sqrt{|\varepsilon_r|/2}, \tag{3.15}$$

where $\varepsilon_r = \varepsilon/\varepsilon_0$ is the relative electric permittivity. The system described in Fig. 3.1 satisfies Eq. (3.15), so it is in the low-frequency regime.

3.2 Impact of the eigenvalue distribution on the convergence behavior of GMRES

In this section, we explain how the different eigenvalue distributions for different values of s examined in Sec. 3.1 influence the convergence behavior of an iterative method to solve Eq. (3.7).

For each of $s = -1, 0, +1$, we discretize Eq. (3.7) using the FDFD method for the system illustrated in Fig. 3.1 with an x-polarized electric dipole current source placed at the center of the simulation domain. We then solve the discretized equation by an iterative method to observe the convergence behavior. The iterative method to use in this section is GMRES introduced in Sec. 1.3; we use GMRES without restart because the system is sufficiently small.

Figure 3.3 shows $\|r_m\|/\|b\|$ versus the number m of iteration steps for the three values of s. As can be seen in the figure, the convergence behavior of GMRES is quite different for

[3] To obtain $8/\Delta_{\min}^2$, take the first equation of Eq. (2.43a) and then multiply the extra factor $\mu = \mu_0$ to account for the difference between Eqs. (1.6) and (3.1).

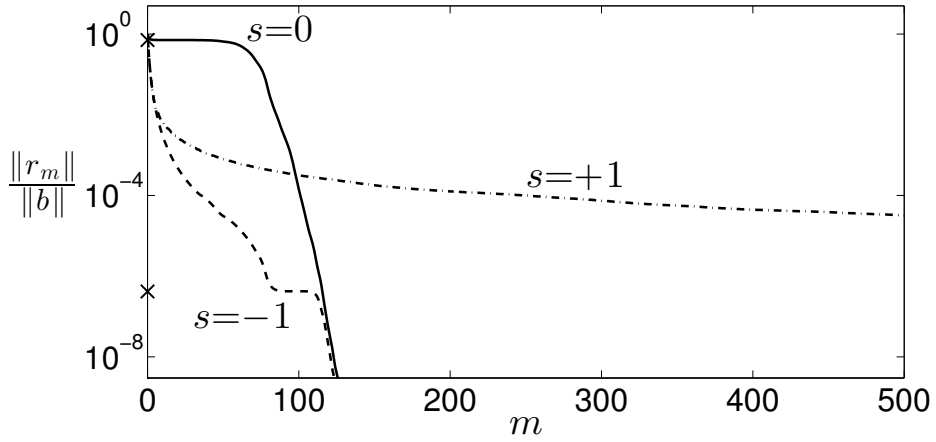


Figure 3.3: Convergence behavior of GMRES for the vacuum-filled domain illustrated in Fig. 3.1. Three systems of linear equations discretized from Eq. (3.7) for $s = 0, -1, +1$ are solved by GMRES. In the iteration process of GMRES for each s, we plot the relative residual norm $\|r_m\|/\|b\|$ at each iteration step m. Notice that for $s = 0$ the relative residual norm stagnates initially; for $s = -1$ it stagnates around $m = 100$; for $s = +1$ it does not stagnate, but decreases very slowly. The upper and lower "X" marks on the vertical axis indicate the values around which our theory expects $\|r_m\|/\|b\|$ to stagnate for $s = 0$ and $s = -1$, respectively.

different s, with $s = -1$ far more superior than the other two choices of s.

The overall trend of the convergence behavior shown in Fig. 3.3 is consistent with the mathematical theories of iterative methods. For example, the convergence stagnates initially for $s = 0$, and according to Ref. [65] this is typical behavior of GMRES for a matrix with many eigenvalues close to 0 such as our A for $s = 0$ (see Fig. 3.2a). Also, the convergence is very slow for $s = +1$, and Ref. [42[Sec. 9.2]] argues that in general the Krylov subspace methods converge much more slowly for indefinite matrices such as our A for $s = +1$ (see Fig. 3.2c) than for definite matrices. In this section we provide a more intuitive explanation for the convergence behavior by using the residual polynomial.

We first review the residual polynomial of GMRES briefly. Suppose that $\mathcal{P}_m$ is the set of all polynomials $\tilde{p}_m$ of degree at most m such that

$$\tilde{p}_m(0) = 1. \tag{3.16}$$

For each $\tilde{p}_m \in \mathcal{P}_m$, we can define a column vector

$$\tilde{r}_m \equiv \tilde{p}_m(A) r_0. \tag{3.17}$$

At the mth iteration step of GMRES, it turns out that the residual vector r_m of Eq. (1.16) is the $\tilde{r}_m$ with the smallest 2-norm. We refer to the $\tilde{p}_m$ for $\tilde{r}_m = r_m$ as the residual polynomial p_m. Therefore, from Eq. (3.17) we have

$$r_m = p_m(A) r_0. \tag{3.18}$$

Below, we show how the eigenvalue distribution of A influences p_m at each iteration step and hence influences the convergence behavior of GMRES. The matrix $A \in \mathbb{C}^{n \times n}$ for our homogeneous system described in Fig. 3.1 is Hermitian because it is discretized from the Hermitian operator T of Eq. (3.7). Hence, the eigendecomposition of A is

$$A = V \Lambda V^\dagger, \tag{3.19}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \tag{3.20}$$

with real eigenvalues λ_i and the corresponding normalized eigenvectors v_i, and $V^\dagger$ is the conjugate transpose of V; note that V is unitary, i.e., $V^\dagger V = I$. Substituting Eq. (3.19) in Eq. (3.17), we obtain

$$\tilde{r}_m = V \tilde{p}_m(\Lambda) V^\dagger r_0. \tag{3.21}$$

We then define column vectors

$$z_m \equiv V^\dagger (r_m / \|b\|) \quad \text{and} \quad \tilde{z}_m \equiv V^\dagger (\tilde{r}_m / \|b\|), \tag{3.22}$$

whose ith elements, which are referred to as z_{mi} and $\tilde{z}_{mi}$ below, are the projections of $r_m/\|b\|$ and $\tilde{r}_m/\|b\|$ onto the direction of the ith eigenvector v_i. From Eqs. (3.21) and (3.22) we

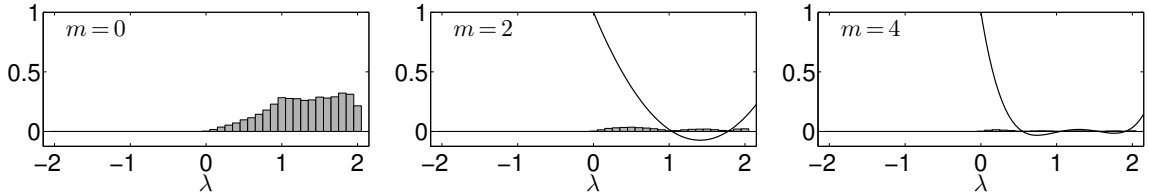


Figure 3.4: Initial evolution of $r_m/\|b\|$ for $s = -1$. Relative residual vectors $r_m/\|b\|$ are visualized at three iteration steps $m = 0, 2, 4$. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. Notice that all the columns almost vanish only after four iteration steps. In the plots for $m = 2$ and $m = 4$, the residual polynomials p_m are also plotted as solid curves; note that they always satisfy the condition (3.16).

obtain

$$\tilde{z}_m = \tilde{p}_m(\Lambda)z_0 = \begin{bmatrix} \tilde{p}_m(\lambda_1) & & \\ & \ddots & \\ & & \tilde{p}_m(\lambda_n) \end{bmatrix} z_0, \tag{3.23}$$

which can be written element-by-element as

$$\tilde{z}_{mi} = \tilde{p}_m(\lambda_i)z_{0i}. \tag{3.24}$$

Because $\|\tilde{z}_m\| = \|\tilde{r}_m\|/\|b\|$, GMRES minimizes $\|\tilde{z}_m\|$ to $\|z_m\|$ when it minimizes $\|\tilde{r}_m\|$ to $\|r_m\|$ at the mth iteration step.

According to Eq. (3.24), $|\tilde{z}_{mi}|$ is minimized to 0 when $\tilde{p}_m$ has λ_i as a root. Thus, the most ideal $\tilde{p}_m$ has all the n eigenvalues of A as its roots, because it reduces $\|\tilde{z}_m\|$ to 0. However, $\tilde{p}_m$ has at most m roots, and m, which is the number of iteration steps, is typically far less than n. Therefore, $\tilde{p}_m$ needs to have its roots optimally placed near the eigenvalues to minimize $\|\tilde{z}_m\|$. Hence, the eigenvalue distribution of A greatly influences the convergence behavior of GMRES.

We now seek to understand the convergence behavior of GMRES for the different choices of s. We begin with $s = -1$. In Fig. 3.4 we plot $r_m/\|b\|$ for $s = -1$ as bar graphs at the first few iteration steps. The horizontal axis in each plot represents eigenvalues. We divide the range of eigenvalues into the same 41 intervals $t_{-20}, \ldots, t_0, \ldots, t_{20}$ used in Fig. 3.2; note that $t_0 \ni 0$. The height of the column on each interval is the norm of the projection of $r_m/\|b\|$

onto the space spanned by the eigenvectors whose corresponding eigenvalues are contained in the interval. More specifically, the height of the column on t_j after m iteration steps is

$$h_{mj} = \left[\sum_{\lambda_i \in t_j} z_{mi}^2\right]^{1/2}. \tag{3.25}$$

Note that $[\sum_j h_{mj}^2]^{1/2} = \|r_m\|/\|b\|$, and thus the sum of the squares of the column heights is a direct measure of convergence.

A few properties of $r_m/\|b\|$ for $s = -1$ shown in Fig. 3.4 are readily predicted from the corresponding eigenvalue distribution of the matrix A presented in Fig. 3.2b. For instance, A has no eigenvalues in $t_{j<0}$, and therefore $r_m/\|b\|$ has components only in $t_{j\geq 0}$ throughout the iteration process as demonstrated in Fig. 3.4. Also, A has very few eigenvalues in t_0, and thus $r_0/\|b\|$ has a very weak component in t_0 as can be seen in the $m = 0$ plot in Fig. 3.4.

Now, we relate $r_m/\|b\|$ with the residual polynomial to explain the convergence behavior of GMRES for $s = -1$. The residual polynomial $p_m(\lambda)$, which is obtained by solving a least squares problem, is also plotted in Fig. 3.4 at each iteration step. As the iteration proceeds, the residual polynomial in Fig. 3.4 has more and more roots, but only in $t_{j\geq 0}$, because the eigenvalues exist only in $t_{j\geq 0}$ and the roots of residual polynomials should stay close to the eigenvalues as mentioned in the discussion following Eq. (3.24). Also, as Eq. (3.24) predicts, the columns in each plot of Fig. 3.4 almost vanish at the roots of the residual polynomial. Therefore, all the columns quickly shrink as the number of the roots of the residual polynomial increases in the iteration process of GMRES. The fast reduction of the column heights provides visualization of the fast convergence of GMRES for $s = -1$ shown in Fig. 3.3.

Next, we examine the convergence behavior for $s = 0$. Figure 3.5 shows $r_m/\|b\|$ for $s = 0$ at the first few iteration steps. Note that $r_0/\|b\|$ has a tall column on t_0 because A has many eigenvalues in t_0 as shown in Fig. 3.2a. Also, the tall column on t_0 persists during the initial period of the iteration process.

To explain the above convergence behavior for $s = 0$, we show that for a nearly positive-definite matrix the column on t_0 is persistent during the initial period of the iteration process of GMRES in general. For that purpose, we compare the three polynomials $\tilde{p}_m \in \mathcal{P}_m$ shown in Fig. 3.6. The three $\tilde{p}_m$ are chosen as candidates for the residual polynomial p_m for a nearly

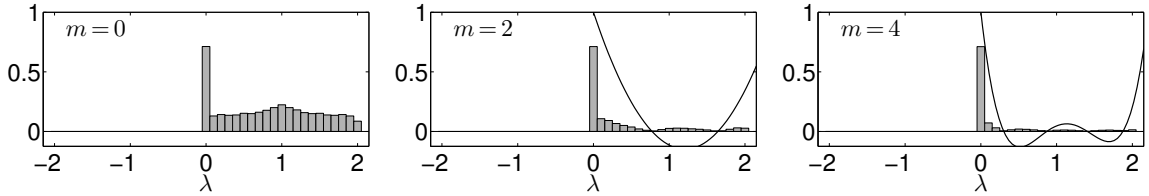


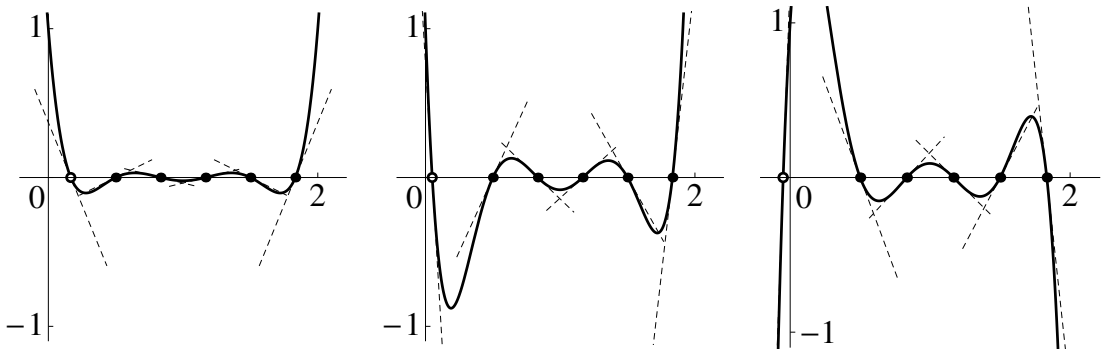
Figure 3.5: Initial evolution of $r_m/\|b\|$ for $s = 0$. Relative residual vectors $r_m/\|b\|$ are visualized at three iteration steps $m = 0, 2, 4$. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. Notice that most columns almost vanish only after four iteration steps, except for the very persistent column at $\lambda \simeq 0$. In the plots for $m = 2$ and $m = 4$, the residual polynomials p_m are also plotted as solid curves; note that they always satisfy the condition (3.16).

positive-definite matrix, and therefore the roots of the polynomials are placed in $t_{j\geq 0}$ according to the discussion following Eq. (3.24). The three $\tilde{p}_m$ have the same roots except for their smallest roots: $\tilde{p}_m$ in Fig. 3.6a does not have its smallest root in t_0, whereas $\tilde{p}_m$ in Figs. 3.6b and 3.6c do. Note that the latter two $\tilde{p}_m$ can shrink the column on t_0 more effectively than the first $\tilde{p}_m$ according to Eq. (3.24).

However, the slopes at the roots of the latter two $\tilde{p}_m$ are steeper than the slopes at the corresponding roots of the first $\tilde{p}_m$ as shown in Fig. 3.6. In Appendix E, we prove rigorously that the slopes of $\tilde{p}_m$ at all roots indeed increase as the smallest root decreases in magnitude. In general, $\tilde{p}_m$ with steeper slopes at the roots oscillates with larger amplitudes around the horizontal axis because it varies faster around the axis; compare the amplitudes of oscillation in Fig. 3.6a with those in Figs. 3.6b and 3.6c. The increased amplitudes of oscillation amplify $|\tilde{z}_{mi}|$ overall according to Eq. (3.24), and thus $\|\tilde{z}_m\|$ as well.

In other words, shrinking the column on t_0 (by placing the smallest root of $\tilde{p}_m$ in t_0) is achieved only at the penalty of amplifying the columns on $t_{j>0}$. This penalty is too heavy when the columns on $t_{j>0}$ constitute a considerable portion of $\|\tilde{z}_m\|$. Therefore, roots of residual polynomials are not placed in t_0 until the columns on $t_{j>0}$ become quite small, which results in the persistence of the column on t_0 during the initial period of the iteration process.

Because the height of the column on t_0 remains almost the same at the initial iteration steps of GMRES, h_{00} of Eq. (3.25), which is the initial height of this column, provides an



(a) Reference $\tilde{p}_m$ (b) Smallest root made smaller (c) Smallest root made smaller and negative

Figure 3.6: Impact of the magnitude of the smallest root of a polynomial $\tilde{p}_m \in \mathcal{P}_m$ on the oscillation amplitudes of $\tilde{p}_m$. Three $\tilde{p}_m$ of degree 6 are shown. In each figure, a solid line represents a polynomial; an open dot on the horizontal axis indicates the smallest root; solid dots indicate the other roots; dashed lines show the slopes of the polynomial at the roots. The three polynomials have the same roots except for their smallest roots: the smallest root in (a) becomes smaller positive and negative roots in (b) and (c), respectively. Notice that the slopes at all roots in (a) become steeper in (b) and (c) as the smallest root decreases in magnitude, and as a result the amplitudes of oscillation of $\tilde{p}_m$ around the horizontal axis increase.

approximate lower bound of $\|z_m\| = \|r_m\|/\|b\|$ during the initial period of the iteration process. A more accurate lower bound is calculated as the norm of $r_0/\|b\|$ projected onto the eigenspace of the eigenvalue closest to 0. For our example system, for $s = 0$ the calculated lower bound is 0.707. Note that $\|r_m\|/\|b\|$ for $s = 0$ indeed stagnates initially at this value in Fig. 3.3. For $s = -1$ the calculated lower bound is 4.16×10^{-7}, at which $\|r_m\|/\|b\|$ also stagnates as shown in Fig. 3.3. However, this value is much smaller than the lower bound for $s = 0$, because for $s = -1$ the initial height of the column on t_0 is almost negligible as shown in the $m = 0$ plot in Fig. 3.4. In fact, the value is smaller than the conventional tolerance $\tau = 10^{-6}$ mentioned below Eq. (1.16), so the stagnation does not deteriorate the convergence speed for $s = -1$.

Lastly, we examine the convergence behavior for $s = +1$. Figure 3.7 shows $r_m/\|b\|$ for $s = +1$ at some first ($m = 0, 4, 7, 11$) and later ($m = 120, 140$) iteration steps. Because the matrix A for $s = +1$ has both positive and negative eigenvalues as indicated in Fig. 3.2c, $r_m/\|b\|$ has components in both $t_{j>0}$ and $t_{j<0}$, but in the present example the components of $r_m/\|b\|$

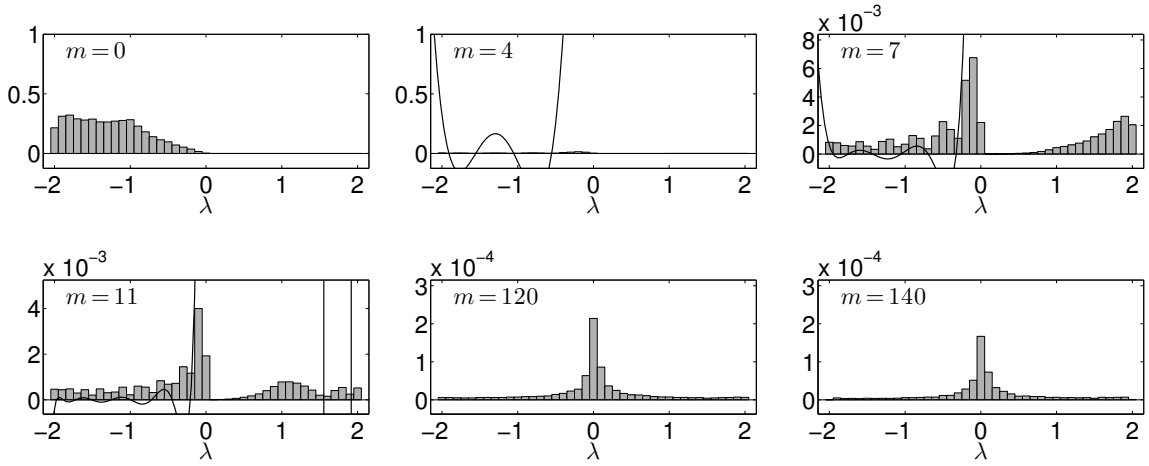
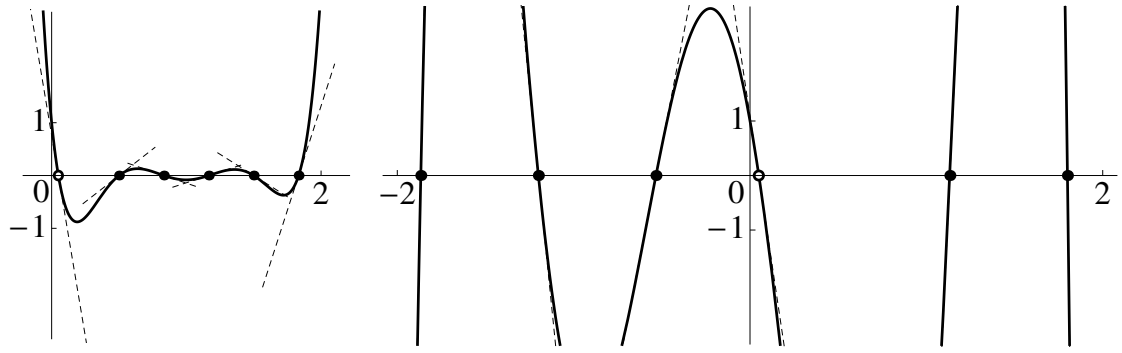


Figure 3.7: Evolution of $r_m/\|b\|$ for $s = +1$. Relative residual vectors $r_m/\|b\|$ are visualized at iteration steps $m = 0, 4, 7$ in the first row and at $m = 11, 120, 140$ in the second row. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. The vertical scale of the plot is magnified as the iteration proceeds. Notice that the column at $\lambda \simeq 0$ is very persistent during the later period of the iteration process ($m = 120, 140$). In the plots for $m = 4, 7, 11$, the residual polynomials p_m are also plotted as solid curves; the residual polynomials are not plotted for $m = 100$ and $m = 120$ because they have too many roots.

are concentrated in $t_{j<0}$ initially ($m = 0$ plot in Fig. 3.7). Thus GMRES begins with the roots of residual polynomials placed in $t_{j<0}$ ($m = 4$ plot in Fig. 3.7). However, such residual polynomials have large values in $t_{j>0}$, so they amplify the initially very small components of $r_m/\|b\|$ in $t_{j>0}$ according to Eq. (3.24), and eventually we reach a point where the components of $r_m/\|b\|$ in $t_{j>0}$ and $t_{j<0}$ become comparable ($m = 7$ plot in Fig. 3.7). Afterwards, GMRES places the roots of residual polynomials in both $t_{j>0}$ and $t_{j<0}$ so that the components of $r_m/\|b\|$ in both regions are reduced.

We note that the convergence behavior for $s = +1$ is initially quite similar to that for $s = -1$ because $r_0/\|b\|$ for $s = +1$ has components concentrated in $t_{j<0}$ and only a very weak component in t_0. Therefore, $\|r_m\|/\|b\|$ reduces quickly for $s = +1$ without stagnation during the initial period of the iteration process as shown in Fig. 3.3.

During the later period of the iteration process, however, the reduction of $\|r_m\|/\|b\|$ for $s = +1$ slows down significantly, and eventually $s = +1$ produces the slowest convergence among the three values of s as shown in Fig. 3.3. The slow reduction of $\|r_m\|/\|b\|$ is due to



(a) $\tilde{p}_m$ for a nearly positive-definite matrix

(b) $\tilde{p}_m$ for a strongly indefinite matrix

Figure 3.8: Candidates for the residual polynomials for (a) a nearly positive-definite matrix and (b) strongly indefinite matrix. In each figure, a solid line represents a polynomial $\tilde{p}_m \in \mathcal{P}_m$; an open dot on the horizontal axis indicates the smallest-magnitude root; solid dots indicate the other roots; dashed lines show the slopes of the polynomial at the roots. The two polynomials have the same smallest-magnitude root ζ_0, and thus have approximately the same slope $-1/\zeta_0$ at their smallest-magnitude roots. Note that for both $\tilde{p}_m$ the slopes get steeper at the roots further away from the median of the roots. Hence, the slopes of most dashed lines are gentler than $1/|\zeta_0|$ in (a) and steeper than $1/|\zeta_0|$ in (b). As a result, $\tilde{p}_m$ in (b) has larger amplitudes of oscillation around the horizontal axis than $\tilde{p}_m$ in (a).

the very persistent column on t_0: comparing the plots for $m = 120$ and $m = 140$ in Fig. 3.7 shows that the column barely reduces for 20 iteration steps.

We have shown earlier that the column on t_0 is quite persistent for a nearly positive-definite matrix. The argument relied on the properties proved in Appendix E about a polynomial $\tilde{p}_m \in \mathcal{P}_m$ with only positive roots. We can easily extend the proof in the appendix to $\tilde{p}_m$ with both positive and negative roots, and then show that the column on t_0 is persistent also for a strongly indefinite matrix, which explains the slow convergence for $s = +1$ described above. However, the explanation is insufficient to explain why the convergence is *much* slower for $s = +1$ than for $s = 0$ as indicated in Fig. 3.3.

Here, we show that the column on t_0 is in fact even more persistent for a strongly indefinite matrix than for a nearly positive-definite matrix. For that purpose, we compare the two polynomials $\tilde{p}_m \in \mathcal{P}_m$ shown in Fig. 3.8. As can be seen from the locations of their roots, they are candidates for the residual polynomials for different matrices: $\tilde{p}_m$ shown in Fig. 3.8a is appropriate for a nearly positive-definite matrix (referred to as A_{def} below), and $\tilde{p}_m$ shown in

Fig. 3.8b is appropriate for a strongly indefinite matrix (referred to as A_{ind} below). Moreover, we choose these two $\tilde{p}_m$ to have the same smallest-magnitude root ζ_0 in t_0. Being elements of $\mathcal{P}_m$, both $\tilde{p}_m$ satisfy Eq. (3.16). Hence, we have $|\tilde{p}'_m(\zeta_0)| \simeq 1/|\zeta_0|$ for both $\tilde{p}_m$, where $\tilde{p}'_m$ is the first derivative of $\tilde{p}_m$.

Now, we note that $|\tilde{p}'_m|$ evaluated at a root of $\tilde{p}_m$ tends to decrease as the root gets closer to the median of the roots; see Appendix F for a more rigorous explanation. Hence, $|\tilde{p}'_m| \leq 1/|\zeta_0|$ tends to hold at most roots of $\tilde{p}_m$ for A_{def}, because ζ_0 is one of the farthest roots from the median of the roots. On the other hand, $|\tilde{p}'_m| \geq 1/|\zeta_0|$ tends to hold at most roots of $\tilde{p}_m$ for A_{ind}, because ζ_0 is one of the closest roots to the median of the roots. Therefore, $\tilde{p}_m$ for A_{ind} has much steeper slopes at most roots than $\tilde{p}_m$ for A_{def} in general, and thus has larger amplitudes of oscillation around the horizontal axis, as demonstrated in Fig. 3.8.

Combined with Eq. (3.24), the above argument shows that shrinking the column on t_0 (by placing the smallest-magnitude root of $\tilde{p}_m$ in t_0) increases $\|z_m\|$ much more for a strongly indefinite matrix than for a nearly positive-definite matrix. Therefore, the column on t_0 should be much more persistent for a strongly indefinite matrix than for a nearly positive-definite matrix in general, which explains the much slower convergence for $s = +1$ than for $s = 0$ in Fig. 3.3.

In summary of this section, we have shown that $s = -1$ produces the most superior convergence behavior; $s = 0$ induces stagnation during the initial period of the iteration process due to the high multiplicity of eigenvalues near zero; $s = +1$ leads to the slowest convergence overall due to the strongly indefinite matrix. We have provided a graphical explanation of the difference in the convergence behavior of GMRES, for which a strong theoretical basis exists, using a simple system of a homogeneous dielectric medium.

The arguments provided in this section can be easily extended to show that $s = -1$ is indeed optimal among all values in general. Compared with the case of $s = -1$, according to Eq. (3.13), for $s > 0$ A is always more strongly indefinite and therefore the convergence should be slower; for $-1 < s < 0$ A has more eigenvalues clustered near zero and thus the initial stagnation period should be longer; for $s < -1$ A has a wider eigenvalue value range, so the condition number of A should be larger and the convergence should be slower as seen in Ch. 2. In other words, $s = -1$ is the value that leaves the matrix A nearly positive-definite while removing the eigenvalues clustered near zero sufficiently without increasing

	Slot	Diel	Array
λ_0	1550 nm	1550 nm	632.8 nm
$\Delta_{\min}$	2 nm	10 nm	5 nm
$\max \lvert\varepsilon_r\rvert$	129.0	12.09	10.81

Table 3.2: Benchmark problems' parameters used in Eq. (3.15). When substituted in Eq. (3.15), these parameters prove that all the three benchmark problems described in Sec. 1.4 are in the low-frequency regime.

the condition number. With separate numerical experiments we have verified that $s = -1$ is indeed superior to values other than $s = 0$ and $s = +1$ as well.

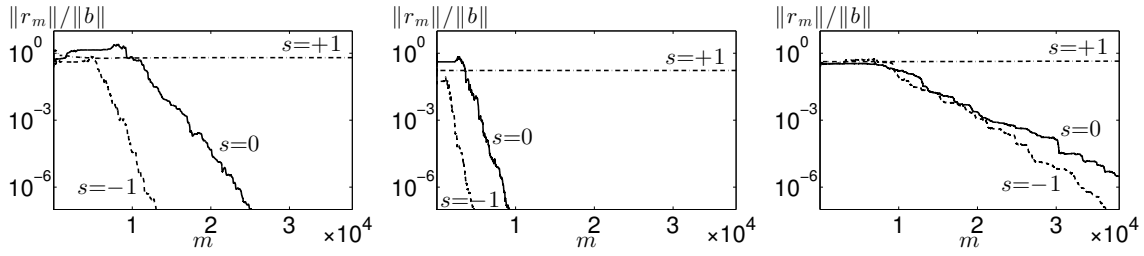
In the next section we will see that the difference in the convergence behavior for different choices of s is in fact quite general in practical situations.

3.3 Convergence behavior of QMR for 3D inhomogeneous systems

In this section, we solve Eq. (3.6) for 3D inhomogeneous systems of practical interest by an iterative method, and demonstrate that $s = -1$ still induces faster convergence than $s = 0$ and $s = +1$. We note that the systems examined in this section are inhomogeneous and have complex ε in general. The analyses in Secs. 3.1 and 3.2, therefore, do not hold strictly here. Nevertheless, we will see that the analyses for the homogeneous system in the previous sections provide insight in understanding the convergence behavior for more general systems examined in this section.

The three 3D inhomogeneous systems we consider are the benchmark problems described in Sec. 1.4. To prevent reflection of EM waves from boundaries, we enclose each system by SC-PML, because SC-PML produces a much better-conditioned matrix than the more commonly used UPML as shown in Ch. 2. For each system, we construct three systems of linear equations $Ax = b$ corresponding to $s = -1, 0, +1$ by the FDFD method. Considering the parameters summarized in Table 3.2 and the condition (3.15), all the three systems are in the low-frequency regime.

The constructed systems of linear equations are solved by QMR introduced in Sec. 1.3.



(a) Slot (b) Diel (c) Array

Figure 3.9: Convergence behavior of QMR for the three benchmark problems "Slot", "Diel", and "Array" described in Sec. 1.4 for $s = -1, 0, +1$. For all the three problems QMR converges fastest for $s = -1$, whereas it barely converges for $s = +1$.

GMRES that was used in Sec. 3.2 to solve a 2D problem is not suitable for 3D problems here because it requires too much memory as explained in Sec. 1.3.

Figure 3.9 shows the convergence behavior of QMR for the three systems. Note that for all three systems the choice of $s = -1$ results in the fastest convergence, and the choice of $s = +1$ barely leads to convergence. As mentioned at the end of Sec. 1.4, the three benchmark problems have different geometrical complexities and different materials. Therefore, Fig. 3.9 suggests that the superiority of $s = -1$ over both $s = 0$ and $s = +1$ is quite general.

Even though both the iterative method and the systems in this section are significantly different from those in the previous section, the convergence behaviors are very similar. We explain the resemblance as follows.

The matrix for an inhomogeneous system is actually not much different from that for a homogenous system in many cases. Indeed, most inhomogeneous systems consist of several homogeneous subdomains. Inside each homogeneous subdomain of such an inhomogeneous system, the differential operator in Eq. (3.6) for the inhomogeneous system is the same as the differential operator (3.8) for a homogeneous system, whereas at the interfaces between the subdomains it is not. Nevertheless, the number of finite-difference grid points assigned at the interfaces is usually much smaller than that of the grid points assigned inside the homogeneous subdomains. Therefore most rows of the matrix for the inhomogeneous system should be the same as those for a homogeneous system discretized on the same grid.

In addition, the differential operator (3.8) for a homogeneous system is nearly Hermitian in the low-frequency regime even if ε is complex, because it is approximated well by the

Hermitian operator (3.9).

Hence, the matrix for an inhomogeneous system is actually quite similar to the nearly Hermitian matrix for a homogeneous system. Because QMR reduces to GMRES for Hermitian matrices in exact arithmetic [74[Secs. 2.4, 3.3]], it is reasonable that the convergence behavior of QMR for an inhomogeneous system is similar to that of GMRES for a homogeneous system.

The matrix for an inhomogeneous system deviates more from that for a homogeneous system as the number of homogeneous subdomains increases, because then the number of grid points assigned at the interfaces between homogeneous subdomains increases. Therefore, we can expect that the convergence behavior for an inhomogeneous system would deviate from that for a homogeneous system as the number of homogeneous subdomains increases. Such deviation is demonstrated in Fig. 3.9c, where the system has many metallic pillars; note that the convergence behavior for $s = -1$ is no longer very different from that for $s = 0$ in this case.

3.4 Summary and remarks

We have introduced a new method to accelerate the convergence of iterative solvers of the frequency-domain Maxwell's equations in the low-frequency regime. The method solves a new equation that is modified from the original Maxwell's equations using the continuity equation.

The operator of the newly formulated equation does not have the high multiplicity of near-zero eigenvalues that makes the convergence stagnate for the original operator. At the same time, the new operator is nearly positive-definite, so it avoids the long-term slow convergence that indefinite operators suffer from.

In this chapter, we have considered only nonmagnetic materials ($\mu = \mu_0$). For magnetic materials ($\mu \neq \mu_0$), we note that a similar equation

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} + s \nabla \left[(\mu\varepsilon)^{-1} \nabla \cdot (\varepsilon \mathbf{E}) \right] - \omega^2 \varepsilon \mathbf{E} = -i\omega \mathbf{J} + s \frac{i}{\omega} \nabla \left[(\mu\varepsilon)^{-1} \nabla \cdot \mathbf{J} \right], \tag{3.26}$$

which can also be formulated from Maxwell's equations and the continuity equation, can be

used instead of Eq. (3.6) to accelerate the convergence of iterative methods. We leave the discussion on the optimal value of s in this equation for future work.

Because our method achieves accelerated convergence by formulating a new equation, it can be easily combined with other acceleration techniques such as preconditioning and better iterative methods.

Chapter 4

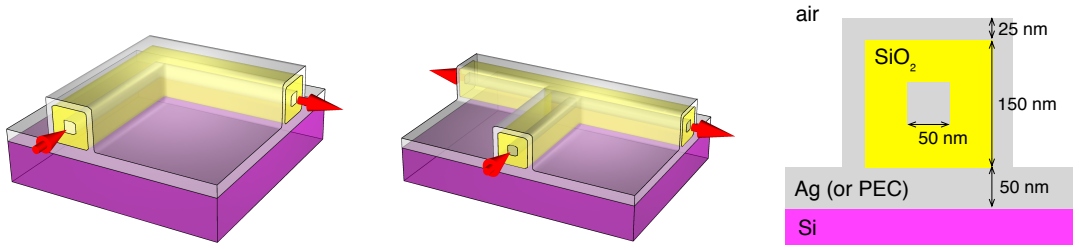
Design of plasmonic coaxial waveguide bends and splitters by the FDFD method[1]

There is no abstract art. You must always start with something. Afterward you can remove all traces of reality.

Pablo Picasso (1881–1973)

Routing of light in arbitrary directions inside a submicron-scale volume is one of the most basic functions sought after in nanophotonics [75–77]. Plasmonic waveguides, despite Ohmic loss inherent in metals, have therefore been considered important components of nanophotonics due to their capability of guiding light through deep-subwavelength mode areas [78–83]. A natural question arose as to whether basic waveguide components such as sharp 90° bends and T-splitters can be constructed in plasmonic waveguides in a simple and compact manner. In 2D metal-dielectric-metal (MDM) waveguides, it was numerically demonstrated that these components can bend and split input power almost perfectly without introducing additional reflection and radiation loss on top of the inherent Ohmic loss of the straight waveguide [84].

[1]Reproduced in part with permission, from W. Shin et al., "Broadband sharp 90-degree bends and T-splitters in plasmonic coaxial waveguides," submitted to *Nano Letters* for publication. Unpublished work copyright 2013 American Chemical Society.



(a) Bend (b) Splitter (c) Waveguide cross section

Figure 4.1: Structures of a sharp 90° bend and T-splitter in plasmonic coaxial waveguides. (a) and (b) show the structures of a bend and T-splitter with the propagation directions of light indicated by red arrows. (c) shows the cross section of the reference plasmonic coaxial waveguide. Silver is used as the metal, but it is substituted by PEC in order to illustrate some of the physics.

In realistic 3D plasmonic waveguides, however, it turns out to be significantly more difficult to design sharp 90° bends and T-splitters without additional loss into undesirable channels such as reflection and radiation. For example, plasmonic slot waveguides [24], which are 3D analogs of the 2D MDM waveguides, suffer from substantial reflection of about 16 % when bent sharply by 90° for near-infrared wavelengths [85]. The reflection can be suppressed by rounding the sharp corner and hence increasing the bending radius of curvature slightly [85, 86], but there always exists loss into radiation and surface wave [87]. These extra loss channels also induce unwanted crosstalk between optical components; such crosstalk can be especially detrimental in densely integrated optical circuits, where optical components are placed close to each other [88].

V-grooves are another type of plasmonic waveguides for which nearly perfect transmission of optical waves through sharp 90° bends was reported [89]. However, the lossless transmission through the bends in the V-grooves is narrowbanded because it relies on an interference phenomenon [90]. Moreover, unless the taper angles of the V-grooves are very narrow, the modes of the V-grooves may not be at deep-subwavelength scale [79, 91, 92].

In this chapter, we propose to use a different type of plasmonic waveguides, namely

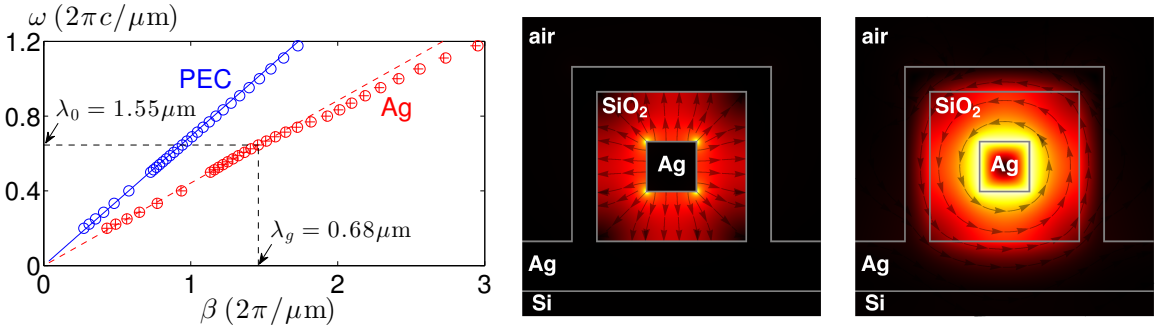
plasmonic coaxial waveguides, for implementing sharp 90° bends and T-splitters. The plasmonic coaxial waveguides have been studied both theoretically [93, 94] and experimentally [95], and they have been applied to achieve a variety of novel functions such as deep-subwavelength focusing [96, 97], enhanced transmission [98–101], and negative refraction [102–104]. In addition to this repertoire of applications, we demonstrate numerically that the plasmonic coaxial waveguides are also useful for building sharp 90° bends and T-splitters that experience nearly no loss other than the inherent Ohmic loss of the straight waveguide itself over a broad range of wavelengths, including the telecommunication wavelength of 1.55 μm. The structures of a bend and T-splitter are described in Figs. 4.1a and 4.1b.

This chapter is organized as follows. In Sec. 4.1 we examine the properties of straight coaxial waveguides. In Sec. 4.2 we demonstrate that sharp 90° bends in plasmonic coaxial waveguides can bend input waves almost perfectly. In Sec. 4.3 we show that T-splitters in plasmonic coaxial waveguides can be designed to split input waves nearly perfectly. In Sec. 4.4 we summarize the chapter and make a few remarks.

We note that the numerical simulation of wave propagation through bends and splitters in this chapter is accomplished efficiently by the 3D FDFD method combined with the two acceleration techniques developed in Chs. 2 and 3. BiCG introduced in Sec. 1.3 is used as an iterative method to solve the system of linear equations Eq. (1.15) constructed by the FDFD method.

4.1 Properties of plasmonic coaxial waveguides

In this section we examine the properties of plasmonic coaxial waveguides. Unlike most of the previous works [93–97, 99–104], we use coaxial waveguides with square rather than circular cross sections because they are easier to fabricate using lithography-based fabrication techniques. The cross section of the "reference waveguide" examined in this chapter is described in Fig. 4.1c. The waveguide is placed on top of a silicon (Si) substrate, and the space between inner and outer coaxial metals is filled with silica (SiO_2). Throughout this chapter, we use silver (Ag) as the metal. We also use the perfect electric conductor (PEC) in order to illustrate some of the physics. The choice of the metal used will be specified explicitly for each numerical result.



(a) Dispersion relation (b) Transverse E-field (c) Transverse H-field

Figure 4.2: Properties of the reference coaxial waveguide. (a) shows the dispersion relations of the fundamental modes of the reference waveguides made of silver and PEC. The data points are obtained by the 2D FDFD mode solver. The blue solid line is the light line in silica, and its alignment with the blue open dots proves the accuracy of the mode solver results. The red open dots and cross hairs are for the waveguides with 25 nm-thick and infinitely thick outer silver layers, respectively, and their coincidence proves the effectiveness of the 25 nm-thick outer silver layer. The red dashed line connects the origin and the data point for λ_0 = 1.55 μm, for which the guide wavelength is λ_g = 0.68 μm, and its alignment with the red open dots shows that the mode is quasi-TEM. The propagation length [24] for λ_0 = 1.55 μm is L_p = 6.82 μm. (b) and (c) show the magnitudes (colors) and directions (arrows) of the transverse E- and H-fields of the fundamental mode of the reference waveguide made of silver for λ_0 = 1.55 μm; the longitudinal components are not shown because they are much smaller than the transverse components. The dielectric constants of silicon [23], silica [23], and silver [25] are taken from tabulated data.

To study the properties of the reference waveguide, we use the 2D FDFD mode solver to solve the waveguide mode equation numerically for each vacuum wavelength λ_0 [24, 105]. The FDFD method allows us to use tabulated dielectric constants of dispersive materials such as silver [25] directly, including both the real and imaginary parts. For accurate yet efficient solution, we use nonuniform spatial grids whose cells are as small as 1 nm inside the waveguide region and as large as 20 nm further outside. To simulate infinitely long waveguides, we surround the entire simulation domain by SC-PML, which is much superior to the more commonly used UPML as shown in Ch. 2.

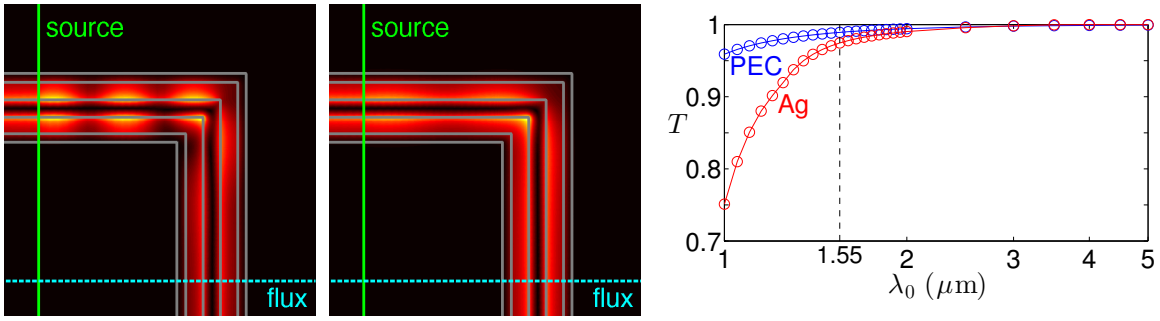
The fundamental mode of the reference waveguide made of silver is a quasi-transverse-electromagnetic (quasi-TEM) mode. The quasi-TEM mode inherits many desirable properties of the true TEM mode of the same coaxial waveguide made of PEC [24, 93]. First,

it has a nearly linear dispersion relation as shown in Fig. 4.2a. Therefore, the waveguide has a nearly constant group velocity over a broad range of wavelengths, and has no cutoff wavelength; the latter means that it can guide light with wavelengths much larger than the cross-sectional dimensions of the coaxial waveguide. Second, the E- and H-fields of the quasi-TEM mode are tightly confined between the two metals, as shown in Figs. 4.2b and 4.2c for a vacuum wavelength $\lambda_0 = 1.55\,\mu m$. Remarkably, we find that only 25 nm-thick outer metal layer is sufficient for confining the fields within the waveguide, because the opposite electric charges and currents carried by the inner and outer metals cancel the fields outside the coaxial waveguide efficiently. Note that the area of the silica region, where most of the fields are confined, is less than $(1.55\,\mu m/10)^2$, which is at deep-subwavelength scale for the vacuum wavelength $\lambda_0 = 1.55\,\mu m$. In fact, because of the above mentioned field cancellation effect, the area of confinement stays almost the same even if λ_0 increases from 1.55 μm, so the confinement becomes even stronger for longer wavelengths.

4.2 Performance of sharp 90° bends

In this section we examine the performance of the sharp 90° bend in the reference waveguide made of silver shown in Fig. 4.1a. For each vacuum wavelength λ_0 between 1 μm and 5 μm, we excite the fundamental mode of the waveguide by an electric current source plane located 0.5 μm before the bend; in reality the fundamental mode can be excited by coupling the lowest-order transverse magnetic (TM_{01}) mode of optical fibers [106] or the quasi-TEM mode of plasmonic coaxial lasers [107]. We then measure the transmitted power through a flux plane located 0.5 μm after the bend.

Figures. 4.3a and 4.3b show the solutions of Maxwell's equations for such measurement for $\lambda_0 = 1\,\mu m$ and $\lambda_0 = 1.55\,\mu m$; notice that for $\lambda_0 = 1\,\mu m$ a strong standing wave pattern is formed in the input waveguide by the interference between the input and reflected waves, whereas for $\lambda_0 = 1.55\,\mu m$ the pattern is diminished significantly, which means that most of input power is transmitted through the bend without reflection. We perform a similar measurement of the power transmitted over $(0.5 + 0.5)\,\mu m$ in a straight waveguide. The ratio of the measurement in the bend with respect to the measurement in the straight waveguide is the transmittance of the bend. Such a definition of transmittance is intended to capture loss



(a) $\lambda_0 = 1\,\mu\text{m}$ (b) $\lambda_0 = 1.55\,\mu\text{m}$ (c) Transmittance through the bend

Figure 4.3: Performance of the sharp 90° bend. (a) and (b) plot the amplitude of the H-field component normal to the plane containing the axes of the input and output waveguides for $\lambda_0 = 1\,\mu\text{m}$ and $\lambda_0 = 1.55\,\mu\text{m}$. Silver is used as the metal. The distances from the center of the junction to the electric current source plane (green) and the flux measurement plane (cyan) are both 0.5 μm. Notice the strong standing wave pattern in the input waveguide in (a) due to significant reflection of the input wave at the junction. (c) shows the transmittance spectra through the bend. The transmittance is higher for the PEC waveguide than the silver waveguide, but both converge to 100 % as the vacuum wavelength λ_0 increases. For the silver waveguide, $T = 97.5\,\%$ is obtained at $\lambda_0 = 1.55\,\mu\text{m}$. The nearly perfect transmission is achieved for $\lambda_0 \geq 1.55\,\mu\text{m}$, a broad range of wavelengths; note that the horizontal axis is on a logarithmic scale.

that is added on top of the propagation loss of the straight waveguide; the additional loss can be due to additional Ohmic loss introduced by the bend, or due to reflection and radiation loss at the bend.

The measured transmittance is shown as a spectrum in Fig. 4.3c. It shows that the transmittance approaches 100 % as the wavelength increases. Especially, a transmittance of 97.5 % is achieved for $\lambda_0 = 1.55\,\mu\text{m}$ without any optimization of the geometry of the bend. This remarkable phenomenon can be explained as follows. For wavelengths much larger than the cross-sectional dimensions of the waveguide, the quasi-static approximation applies, and hence the junction between the input and output waveguides at the bend can be accurately modeled as a junction between two transmission lines [108]. Because the input and output waveguides have the same cross-sectional shape, the two transmission lines have the same characteristic impedance. Therefore, due to the impedance matching, the transmittance should be 100 % in the quasi-static limit. The perfect transmission in the quasi-static

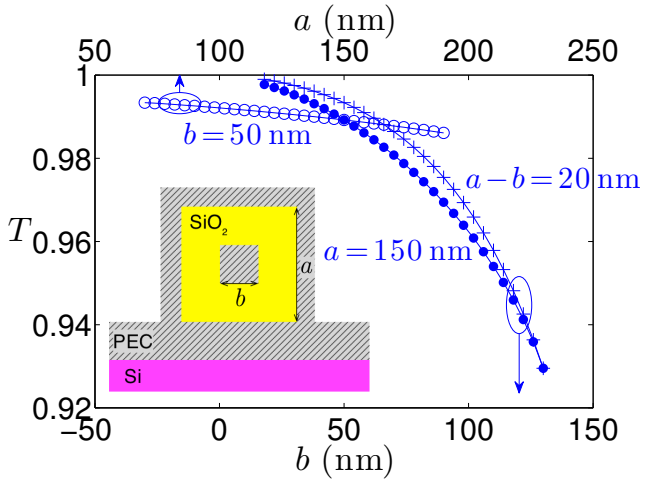
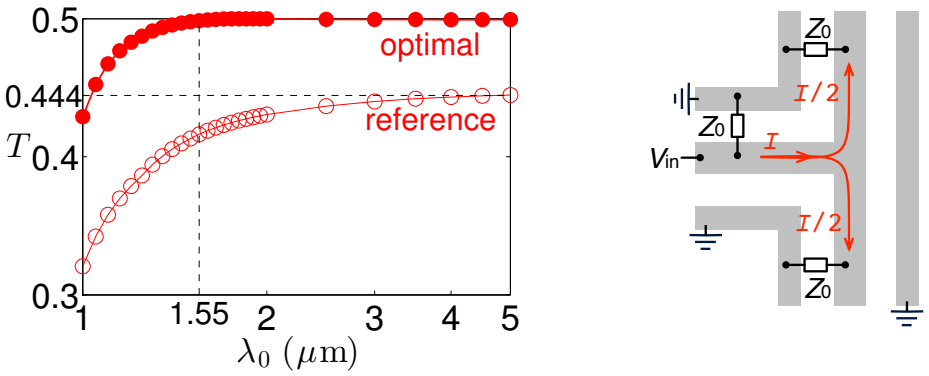


Figure 4.4: Performance of the sharp 90° bend in the PEC coaxial waveguide with various cross-sectional dimensions. The parameters a and b are indicated in the inset figure. The transmittance is measured while varying a for b = 50 nm (open dots), varying b for a = 150 nm (solid dots), and varying b for $a - b$ = 20 nm (cross hairs). Note that the high transmittance for b = 50 nm is not very sensitive to the variation of a. On the other hand, the variation of b affects the transmittance significantly.

limit also occurs in a bend in the reference waveguide made of PEC as shown in Fig. 4.3c, which is consistent with the result in the classical microwave literature [109].

We note that the nearly perfect transmission is a broadband phenomenon achieved from λ_0 = 1.55 μm to mid-infrared, because the quasi-static approximation applies to any sufficiently long wavelengths. The closed structure of the coaxial waveguide that prohibits coupling with other leakage channels such as radiation is crucial for the nearly perfect transmission; in other conventional 3D plasmonic waveguides that are open to the leakage channels, the transmission does not become perfect in the quasi-static limit because additional loss into the leakage channels is unavoidable.

Even though the quasi-static approximation holds in general when the cross-sectional dimensions of the coaxial waveguide are much smaller than the wavelength, we find that the size of the inner metal is the most critical among all cross-sectional dimensions. In Fig. 4.4, we measure the transmittance through the PEC coaxial waveguide bend for a vacuum wavelength λ_0 = 1.55 μm as varying cross-sectional dimensions of the waveguide while maintaining one of the following three parameters the same: a, b, and $a - b$ (see the inset of



(a) Transmittance (b) Current flow

Figure 4.5: Performance of the T-splitter. (a) shows the transmittance into one of the two output waveguides of the splitter shown in Fig. 4.1b. Silver is used as the metal. Both the output waveguides are the reference waveguides, while the input waveguide is either the reference or the optimal (see Fig.4.6) waveguide. For the reference input waveguide the transmittance reaches 44.4 % as the vacuum wavelength λ_0 increases, whereas for the optimal input waveguide it reaches the ideal 50.0 % over a broad range of wavelengths of $\lambda_0 \geq 1.55\,\mu$m; note that the horizontal axis is on a logarithmic scale. (b) shows the flow of electric current in the T-junction, in which the three reference waveguides are modeled as three transmission lines with the same characteristic impedance Z_0. When the voltage difference V_{in} is applied between the inner and outer metals, the current I is launched in the input waveguide and equally divided into the two output waveguides at the junction. Note that the three seemingly separate outer metal pieces shown in the figure are actually a single connected outer metal layer.

Fig. 4.4). The figure shows that the transmittance stays high as long as b is much smaller than the wavelength, whereas it decreases fast as b increases even if the other two parameters are at deep-subwavelength scale. The strong impact of b, i.e., the size of the inner metal, on the transmittance can be explained by the electromagnetic field distributions of the waveguide mode shown in Figs. 4.2b and 4.2c: the fields are strongest near the inner metal, so the size of the inner metal is likely to affect transmission properties significantly.

4.3 Performance of T-splitters

In this section we examine the performance of the T-splitter in the reference waveguide made of silver shown in Fig. 4.1b. By the same method used for the bend in Sec. 4.2, we measure the

transmittance through one of the two output waveguides. An ideal splitter should transmit 50 % of the input power into each output waveguide. However, for the splitter shown in Fig. 4.1b for which the input and output waveguides have the same cross-sectional shape, it turns out that the transmittance reaches only 44.4 % even for long wavelengths as shown in Fig. 4.5a.

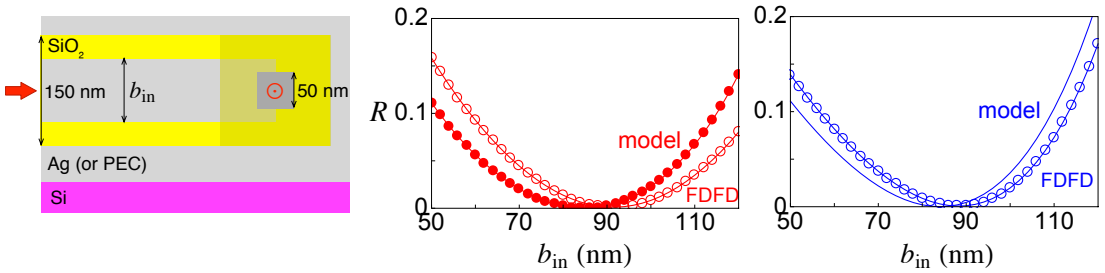
This asymptotic value of the transmittance can also be explained by modeling the T-junction as a junction between three transmission lines in the quasi-static limit. Because all the three waveguides connected at the T-junction have the same cross-sectional shape, the characteristic impedance Z_{in} of the input transmission line and Z_{out} of each of the two output transmission lines have the same value Z_0, i.e., $Z_{\text{in}} = Z_{\text{out}} = Z_0$. The two output transmission lines form a parallel combination of impedances with respect to the input transmission line, because the current flowing through the input transmission line is equally divided into the two output transmission lines as indicated in Fig. 4.5b. Therefore, the load impedance seen by the input transmission line is $Z_L = Z_{\text{out}}/2$, and the reflectance is calculated as

$$R = \frac{|Z_L - Z_{\text{in}}|^2}{|Z_L + Z_{\text{in}}|^2} = \frac{|Z_0/2 - Z_0|^2}{|Z_0/2 + Z_0|^2} = \frac{1}{9}. \tag{4.1}$$

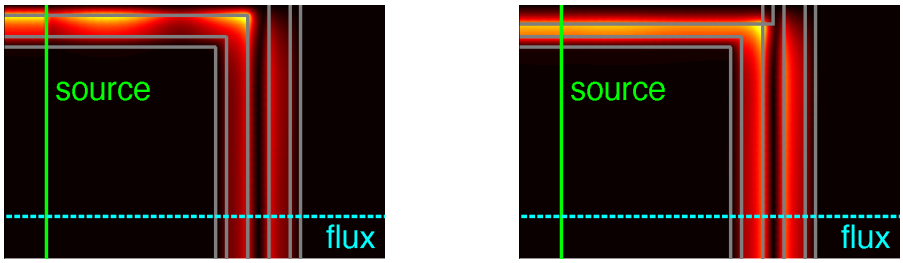
The remaining $1-R$ portion of the input power that is not reflected should be equally divided into each output transmission line, because no leakage channels such as radiation are coupled at the junction. Therefore, the transmittance into each output transmission line should be $T = (1 - R)/2 = 4/9 = 44.4\,\%$ in the quasi-static limit, which is exactly the asymptotic value in Fig. 4.5a.

Equation (4.1) suggests one way of eliminating the reflection, which is to *decrease* Z_{in} from Z_0 to $Z_0/2$. We note that for the T-splitter in the 2D MDM waveguide the recipe was opposite, i.e., to *increase* Z_{in}, because the two output waveguides formed a series rather a parallel combination of impedances [84]. It is important to emphasize that the plasmonic coaxial waveguides and 2D MDM waveguides have very different topology in their transmission line models for the T-splitters, despite close connection between these two classes of waveguides in terms of the modal properties of the straight waveguides. [94].

To eliminate the reflection, we decrease Z_{in} gradually by increasing the size b_{in} of the inner metal of the input waveguide as described in Fig. 4.6a, and measure the reflectance



(a) Side view (b) Ag (c) PEC



(d) b_{in} = 50 nm (e) b_{in} = 90 nm

Figure 4.6: Optimization of the T-splitter. (a) shows the structure of the T-splitter on the vertical cross section containing the axis of the input waveguide. The inner metal size b_{in} of the input waveguide is increased from the original size 50 nm to eliminate the reflection of the input wave at the T-junction. The red arrow and bullseye indicate the propagation direction of light. (b) shows the spectra of the actual reflectance in the 3D FDFD solution of Maxwell's equations (open dots) and the reflectance predicted by the transmission line model (solid dots) for the T-splitter in the silver coaxial waveguide. The actual reflectance almost vanishes for b_{in} = 90 nm, but the transmission line model slightly underestimates the optimal b_{in}. (c) shows similar spectra for the T-splitter in the PEC coaxial waveguide, but the prediction by the transmission line model is made using the analytic formula of the characteristic impedance of the PEC coaxial waveguide. (d) and (e) plot the amplitude of the H-field component normal to the plane containing the axes of the input and output waveguides of the splitters with the reference (b_{in} = 50 nm) and optimal (b_{in} = 90 nm) input waveguides. Only a half of the splitter is shown by virtue of the mirror symmetry. Silver is used as the metal, and the distances from the center of the junction to the electric current source plane (green) and the flux measurement plane (cyan) are both 0.5 μm. Notice the strong standing wave pattern in (d) due to significant reflection of the input wave at the T-junction.

for $\lambda_0 = 1.55\,\mu\text{m}$. Fig. 4.6b shows that indeed the reflection almost vanishes for $b_{\text{in}} = 90\,\text{nm}$, for which the reflectance is only 0.17 %. Figs. 4.6d and 4.6e show the solutions of Maxwell's equations for $b_{\text{in}} = 50\,\text{nm}$ and $b_{\text{in}} = 90\,\text{nm}$ for $\lambda_0 = 1.55\,\mu\text{m}$. Notice the presence of a strong standing wave pattern in the input waveguide in Fig. 4.6d and the absence of such a standing wave pattern in Fig. 4.6e, which confirms that the choice of $b_{\text{in}} = 90\,\text{nm}$ indeed results in nearly perfect transmission.

Even though the splitter is optimized for $\lambda_0 = 1.55\,\mu\text{m}$, the same splitter turns out to exhibit nearly perfect transmission over a broad range of wavelengths spanning from $\lambda_0 = 1.55\,\mu\text{m}$ to mid-infrared as shown in Fig. 4.5a. The broadband perfect transmission is also explained by the transmission line model as follows. Because the mode of each plasmonic coaxial waveguide is quasi-TEM, its characteristic impedance is nearly independent of wavelength. Therefore, for sufficiently long wavelengths for which the transmission line model holds, Eq. (4.1) predicts the nearly constant transmittance, which is $T = 50\,\%$ for the optimized splitter.

To assess the validity of the transmission line model more quantitatively, we predict the reflectance by substituting numerically calculated impedances Z in the formula for R in Eq. (4.1); Z is calculated as V/I, where the voltage difference V between the inner and outer metals and the electric current I flowing through the inner metal are calculated by numerically integrating the E- and H-fields of the fundamental mode of the waveguide [85, 110]. In Fig. 4.6b, the predicted reflectance is compared with the actual reflectance in the 3D FDFD solution of Maxwell's equations. In Fig. 4.6c we make a similar comparison between the predicted and actual reflectances for the splitter in the PEC coaxial waveguide, for which the analytic expression for Z exists [111[Sec. 3.2.4], 112–114]. For both the silver and PEC coaxial waveguides, the prediction from the transmission line model correctly describes the overall trend of the actual reflectance as a function of b_{in}, but we find that the model slightly underestimates the optimal b_{in} for vanishing reflectance.

In addition to its dependence on the impedances of the input and output waveguides, the reflectance also depends on the detailed shape of the junction between the input and output waveguides: it turns out that the reflectance does not reduce to zero unless the thicker inner metal of the input waveguide extends into the junction and reaches the central axis of the output waveguide as shown in Fig. 4.6a. Such dependence on the detailed shape of

the junction is of course not incorporated in the simple transmission line model above. Following the literature of microwave circuits one could develop a more sophisticated model that includes additional lumped circuit elements [115[Sec. 9.6], 116] to describe the properties of the junction more accurately. Nevertheless, as we can see here the simple model above does provide very important guidance for the overall design of the T-splitter.

4.4 Summary and remarks

We have proposed and analyzed compact and realistic designs of sharp 90° bends and T-splitters in plasmonic coaxial waveguides. The bends and splitters transmit optical waves nearly perfectly without parasitic reflection and radiation, so they help to minimize undesirable crosstalk in integrated optical circuits. Also, because the nearly perfect transmission occurs over a broad range of wavelengths, the performance of our bends and splitters should be tolerant of fabrication errors, thermal expansion, and wavelength detuning. The proposed designs can be fabricated either by standard lithography-based fabrication methods, or by more sophisticated methods utilizing silver nanowire bends [117, 118] and dielectric-coated silver nanowires [119–121]. Another interesting approach is to use highly doped semiconductor nanowires instead of metals as conductive inner pieces of plasmonic coaxial waveguides; because the semiconductor nanowires can be grown into various structures such as bends [122], branches [123, 124], and combs [125], they can be useful for building networks of plasmonic coaxial waveguides.

Chapter 5

Conclusion and final remarks

When earnest labor brings you fame and glory,
And all earth's noblest ones upon you smile,
Remember that life's longest, grandest story
Fills but a moment in earth's little while:
"This, too, shall pass away."

Lanta W. Smith (1856–1939)

The finite-difference frequency-domain (FDFD) method is a numerical method of solving Maxwell's equations with several attractive features compared to other methods. As a frequency-domain method, it provides an easier way to analyze steady states and systems with dispersive materials than time-domain methods such as the finite-difference time-domain (FDTD) method. In addition, thanks to the straightforwardness of the finite-difference scheme, discretization of Maxwell's equations into a large system of linear equations $Ax = b$ is conceptually much simpler to understand and easier to implement in the FDFD method than in other frequency-domain methods such as the finite element method and method of moments.

When implemented naïvely, however, the FDFD method faces the problem of the slow convergence of iterative methods to solve $Ax = b$. In this dissertation, we have developed two techniques to greatly enhance the convergence speed. The first technique is to better-condition the matrix A by using an appropriate perfectly matched layer (PML), i.e., SC-PML

or SP-UPML, instead of the more commonly used UPML. The second technique is to remove the near-zero eigenvalues of A by utilizing the continuity equation. Combining the two techniques, we have achieved a dramatic increase in convergence speed; for example, the benchmark problem "Slot" originally took 6 hours to solve with 1024 CPU cores, but after applying the techniques it took only 10 minutes with 128 CPU cores, resulting in a nearly 300-fold speedup. Beyond solving a few benchmark problems, we have demonstrated the efficiency of our 3D FDFD solver in a practical situation by using it in designing nearly perfect bends and splitters in plasmonic coaxial waveguides.

The use of iterative methods in our 3D FDFD solver is somewhat unconventional from the perspective of the numerical linear algebra community. When the problem becomes larger, the numerical linear algebra community is willing to use more computation time per-iteration rather than more iteration steps. For example, they try to keep the number of iteration steps below 100 even for a very large problem, as shown in Refs. [126, 127]. Usually this is possible only with very sophisticated algorithms or rather dense preconditioners, both of which are in general not easy to implement in parallel computing environment.

On the contrary, we carry out the iteration process for a very large number of iteration steps; for example, to solve "Slot" we use about 12,000 iteration steps; this is less than 0.1 % of the total number of unknowns ($3N_xN_yN_z \simeq 27$ million), but it is still an enormous number to the numerical linear algebra community. Nevertheless, we solve this large 3D problem very efficiently within several minutes thanks to highly parallel computing environment; note that our two techniques achieve accelerated convergence by modifying the differential equation before discretization, and therefore the discretized equation remains very sparse and highly suitable for parallel computing environment. We think that our approach of using extremely lightweight per-iteration computation for an unconventionally large number of iteration steps on highly parallelized hardware is one practical way of solving large and complex problems using iterative methods.

Recently, graphics processing units (GPUs) have emerged as a powerful parallel computing platform. A GPU has a huge number of computing cores but a very limited amount of on-chip memory. Using iterative methods, our 3D FDFD solver consumes a very small amount of memory compared to other solvers using direct methods. Especially, if the solver is implemented with the matrix-free formulation, which performs the curl operation on Yee's

finite-difference grid itself without ever constructing a matrix, then our 3D FDFD solver becomes extremely memory-efficient, and hence it is well-suited to be implemented on GPUs. A preliminary implementation shows that a single GPU can drive our solver as fast as 256 CPU cores. Numerical simulation of complex photonic devices on a desktop using GPUs is certainly foreseeable.

We will be excited to see the FDFD method become popular and contribute to the advancement of photonics technology.

Appendix A

First-order perturbation method for nondegenerate singular values of symmetric matrices

In Appendix C.2, the singular values of symmetric matrices are calculated by a perturbation method, which we describe in this appendix. The overall derivation is very similar to the derivation of the widely used perturbation method for the nondegenerate eigenvalues of Hermitian matrices, for which we refer readers to Ref. [128].

For a symmetric matrix $A \in \mathbb{C}^{n\times n}$ such that $A^\top = A$, its SVD is known to reduce to

$$A = V^* \Sigma V^\dagger, \tag{A.1}$$

where V^* is the complex conjugate of V. In other words, $U = V^*$ in the original singular value decomposition of Eq. (2.13) and $u_i = v_i^*$ in Eq. (2.14). The decomposition (A.1) is called the Takagi factorization or the symmetric SVD [129[Corollary 4.4.4], 130, 131].

Suppose that $A^{(0)} \in \mathbb{C}^{n\times n}$ is a symmetric matrix whose Takagi factorization in the form (2.14) is

$$A^{(0)} = \sum_{r=1}^{n} \sigma_r^{(0)} v_r^{(0)*} v_r^{(0)\dagger}. \tag{A.2}$$

We consider a symmetric matrix A that is perturbed from $A^{(0)}$:

$$A = A^{(0)} + \delta A^{(1)}, \tag{A.3}$$

where δ is a small number that characterizes the strength of the perturbation. We seek to calculate the singular values of A, whose Takagi factorization is written as

$$A = \sum_{r=1}^{n} \sigma_r v_r^* v_r^\dagger. \tag{A.4}$$

We assume that the singular values of A and $A^{(0)}$ are both nondegenerate. Then, for any singular value σ_r of A, the corresponding right singular vector v_r is unique up to an arbitrary phase factor $e^{i\theta_r}$ with θ_r real [129$^{\text{Theorem 7.3.5}}$], because v_r is the unit eigenvector corresponding to a distinct eigenvalue σ_r^2 of the Hermitian eigenvalue problems (2.18);[1] the same is true for $v_r^{(0)}$ corresponding to $\sigma_r^{(0)}$ of $A^{(0)}$. As a result,

$$(\sigma_r, v_r) \to \left(\sigma_r^{(0)}, e^{i\phi_r} v_r^{(0)}\right) \quad \text{for some real } \phi_r \text{ as } \delta \to 0 \tag{A.5}$$

because $A \to A^{(0)}$ as $\delta \to 0$. The nondegeneracy constraint is important in obtaining Eq. (A.5); without this constraint, in cases where $\sigma_q^{(0)} = \sigma_r^{(0)}$ for $q \neq r$, v_r converges to a unit vector in span $\left\{v_q^{(0)}, v_r^{(0)}\right\}$ instead.

For the perturbed matrix A, we want to express its pth singular value σ_p to first order in δ. Noting that $\left\{v_1^{(0)}, \ldots, v_n^{(0)}\right\}$ is an orthonormal basis of $\mathbb{C}^n$, we expand the corresponding right singular vector v_p as

$$v_p = \sum_{r=1}^{n} c_r v_r^{(0)}. \tag{A.6}$$

[1] The phase factor $e^{i\theta_r}$ is arbitrary for the general SVD, but in fact it is not for the Takagi factorization [130]; the equality in Eq. (A.4) cannot be maintained for real σ_r if v_r is scaled by a factor of $e^{i\theta_r}$, unless $e^{i\theta_r} = \pm 1$. The only exception arises when $\sigma_r = 0$, whose corresponding right singular vector v_r can be freely scaled by any phase factor. Unfortunately, we have to deal with such an exceptional case in Appendix C.2, so we allow the freedom to vary the phase factor of v_r.

From Eq. (A.5), we see that $v_p \simeq e^{i\phi_p} v_p^{(0)}$ for small δ. Thus, to lowest order in δ,

$$c_r = \begin{cases} e^{i\phi_p} O(1) = O(1) & \text{for } r = p, \\ O(\delta) & \text{for } r \neq p. \end{cases} \tag{A.7}$$

From Eq. (A.4) we have $\sigma_p v_p^* = A v_p$. Substituting Eqs. (A.3) and (A.6) into this, we obtain

$$\sigma_p \sum_{r=1}^{n} c_r^* v_r^{(0)*} = \sum_{r=1}^{n} c_r (A^{(0)} + \delta A^{(1)}) v_r^{(0)}. \tag{A.8}$$

Subsequent multiplication of $v_p^{(0)\top}$ to Eq. (A.8) produces

$$c_p^* \sigma_p = c_p \sigma_p^{(0)} + \sum_{r=1}^{n} \delta c_r \left(v_p^{(0)\top} A^{(1)} v_r^{(0)} \right), \tag{A.9}$$

where Eq. (A.2) is used to obtain the first term of the right-hand side.

Now, because of Eq. (A.7), all terms in the sum in Eq. (A.9) are in the order of δ^2 unless $r = p$. Hence,

$$c_p^* \sigma_p = c_p \left[\sigma_p^{(0)} + \delta \left(v_p^{(0)\top} A^{(1)} v_p^{(0)} \right) \right] + O(\delta^2), \tag{A.10}$$

or equivalently

$$\sigma_p - \frac{c_p}{c_p^*} \left[\sigma_p^{(0)} + \delta \left(v_p^{(0)\top} A^{(1)} v_p^{(0)} \right) \right] = O(\delta^2). \tag{A.11}$$

By taking the modulus of Eq. (A.11) and using the triangle inequality, we obtain

$$-\left|O(\delta^2)\right| \;\leq\; \sigma_p - \left| \sigma_p^{(0)} + \delta \left(v_p^{(0)\top} A^{(1)} v_p^{(0)} \right) \right| \;\leq\; \left|O(\delta^2)\right|, \tag{A.12}$$

where $\left|\sigma_p\right| = \sigma_p$ and $\left|c_p / c_p^*\right| = 1$ are used. Therefore, we have

$$\sigma_p = \left| \sigma_p^{(0)} + \delta \left(v_p^{(0)\top} A^{(1)} v_p^{(0)} \right) \right| + O(\delta^2), \tag{A.13}$$

which is the expression of the perturbed singular value σ_p in terms of the original singular value $\sigma_p^{(0)}$, original singular vector $v_p^{(0)}$, perturbation matrix $A^{(1)}$, and the perturbation strength δ.

Appendix B

Maximum singular values of homogeneous media accounting for finite-difference approximation

In this appendix, we derive Eq. (2.43) in Sec. 2.3.2, considering the finite-difference approximation of the spatial derivatives used in $T = T^{\mathrm{r}_0}, T^{\mathrm{u}_0}, T^{\mathrm{sc}_0}$ of Eq. (2.27).

When T is applied, $\mathbf{E}_{\mathbf{k}}(\mathbf{r}) = \mathbf{F}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}$ of Eq. (2.28) is differentiated spatially by two curl operators. The first curl operator generates an H-field from $\mathbf{E}_{\mathbf{k}}$ as indicated in Eq. (1.1a) for $\mathbf{M} = 0$. The generated H-field, which is denoted by $\mathbf{H}_{\mathbf{k}}(\mathbf{r}) = \mathbf{G}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}$ here, is differentiated by the second curl operator as shown in Eq. (1.1b). In this double curl operation, the p-components of $\mathbf{E}_{\mathbf{k}}$ and $\mathbf{H}_{\mathbf{k}}$ are differentiated as

$$\frac{\partial E_{\mathbf{k},p}(\mathbf{r})}{\partial w} = -ik_w E_{\mathbf{k},p}(\mathbf{r}) \tag{B.1}$$

and

$$\frac{\partial H_{\mathbf{k},p}(\mathbf{r})}{\partial w} = -ik_w H_{\mathbf{k},p}(\mathbf{r}) \tag{B.2}$$

for $p, w = x, y, z$.

On the interlaced E-field grid and H-field grid of Yee's grid [11] with uniform cell size Δ,

the finite-difference method approximates Eqs. (B.1) and (B.2) by

$$\frac{E_{\mathbf{k},p}(\mathbf{r}_e + \hat{\mathbf{w}}\Delta) - E_{\mathbf{k},p}(\mathbf{r}_e)}{\Delta} = \frac{e^{-ik_w\Delta} - 1}{\Delta} E_{\mathbf{k},p}(\mathbf{r}_e) \equiv f(k_w) E_{\mathbf{k},p}(\mathbf{r}_e) \tag{B.3}$$

and

$$\frac{H_{\mathbf{k},p}(\mathbf{r}_h) - H_{\mathbf{k},p}(\mathbf{r}_h - \hat{\mathbf{w}}\Delta)}{\Delta} = \frac{1 - e^{ik_w\Delta}}{\Delta} H_{\mathbf{k},p}(\mathbf{r}_h) \equiv b(k_w) H_{\mathbf{k},p}(\mathbf{r}_h), \tag{B.4}$$

where $\hat{\mathbf{w}}$ is the unit vector in the w-direction; $\mathbf{r}_e$ and $\mathbf{r}_h$ are the grid points in the E-field grid and H-field grid.

Using Eqs. (B.3) and (B.4), the $\mathbf{k}$-space representations of the finite-difference approximations of T^{r_0}, T^{u_0}, and T^{sc_0} are

$$T_{\mathbf{k}}^{\mathrm{r}_0} = \begin{bmatrix} -\frac{b(k_y)f(k_y)}{\mu} - \omega^2\varepsilon & \frac{b(k_y)f(k_x)}{\mu} & 0 \\ \frac{b(k_x)f(k_y)}{\mu} & -\frac{b(k_x)f(k_x)}{\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & -\frac{b(k_x)f(k_x)}{\mu} - \frac{b(k_y)f(k_y)}{\mu} - \omega^2\varepsilon \end{bmatrix}, \tag{B.5a}$$

$$T_{\mathbf{k}}^{\mathrm{u}_0} = \begin{bmatrix} -\frac{b(k_y)f(k_y)}{s_x\mu} - \frac{\omega^2\varepsilon}{s_x} & \frac{b(k_y)f(k_x)}{s_x\mu} & 0 \\ \frac{b(k_x)f(k_y)}{s_x\mu} & -\frac{b(k_x)f(k_x)}{s_x\mu} - s_x\omega^2\varepsilon & 0 \\ 0 & 0 & -\frac{b(k_x)f(k_x)}{s_x\mu} - \frac{s_x b(k_y)f(k_y)}{\mu} - s_x\omega^2\varepsilon \end{bmatrix}, \tag{B.5b}$$

$$T_{\mathbf{k}}^{\mathrm{sc}_0} = \begin{bmatrix} -\frac{b(k_y)f(k_y)}{\mu} - \omega^2\varepsilon & \frac{b(k_y)f(k_x)}{s_x\mu} & 0 \\ \frac{b(k_x)f(k_y)}{s_x\mu} & -\frac{b(k_x)f(k_x)}{s_x^2\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & -\frac{b(k_x)f(k_x)}{s_x^2\mu} - \frac{b(k_y)f(k_y)}{\mu} - \omega^2\varepsilon \end{bmatrix}. \tag{B.5c}$$

Note that Eq. (B.5) reduces to Eq. (2.30) as $\Delta \to 0$ because $f(k_w), b(k_w) \to -ik_w$. This means that we obtain Eq. (B.5) by simply replacing k_w's in Eq. (2.30) with either $if(k_w)$ or $ib(k_w)$ depending on the situation. Accordingly, we can follow the same procedure as in Sec. 2.3.2 to obtain estimates of the maximum singular values from Eq. (B.5). In 2D, the only difference is that $|if(k_w)|, |ib(k_w)| \in [0, 2/\Delta]$ in Eq. (B.5) whereas $|k_w| \in [0, k_{\max}]$ in Eq. (2.30). Therefore, we substitute $2/\Delta$ for $k_{\max}$ in Eq. (2.41) to obtain

$$\sigma_{\max}^{\mathrm{r}_0} \simeq \frac{2(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{u}_0} \simeq \frac{|s_x|(2/\Delta)^2}{\mu}, \quad \sigma_{\max}^{\mathrm{sc}_0} \simeq \frac{(2/\Delta)^2}{\mu}, \tag{B.6}$$

which is Eq. (2.43a). In 3D we still have $|if(k_x)|, |ib(k_x)| \in [0, 2/\Delta]$ and $|k_x| \in [0, k_{\max}]$, but

have $\left|if(k_y)\right|, \left|ib(k_y)\right| \in [0, 2\sqrt{2}/\Delta]$ and $\left|k_y\right| \in [0, \sqrt{2}k_{\max}]$ due to the special choice of our coordinate system made in the discussion following Eq. (2.29). Considering this difference we can easily derive Eq. (2.43b) as well.

Appendix C

Lengthy derivations of various formulae in Sec. 2.3.4

In this appendix, we derive various formulae used in Sec. 2.3.4 to estimate the minimum singular values of homogeneous UPML and SC-PML with $\varepsilon > 0$ in a bounded domain.

C.1 k_x's minimizing $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for a given k_y

In this section, we derive k_x's minimizing $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ for a given k_y; the derived k_x's are used in Eqs. (2.54) and (2.57) in Sec. 2.3.4. The assumptions $k_x \geq 0$, $k_y \geq 0$, and $\varepsilon > 0$ made in Sec. 2.3.4 apply here.

We first consider $k_y < \omega/c$. For such k_y, we show that $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are increasing functions of k_x, and therefore they are minimized at $k_x = 0$. To that end, we examine the analytic formulae of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ as functions of k_x and k_y.

The analytic formulae of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ are quite complex, so we use approximations of $T_{\mathbf{k}}^{\mathrm{u}_0}$ and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ to simplify these formulae. Because of Eq. (2.26), $T_{\mathbf{k}}^{\mathrm{u}_0}$ and $T_{\mathbf{k}}^{\mathrm{sc}_0}$ of Eq. (2.30) are approximated to

$$\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0} = \begin{bmatrix} -\frac{k_y^2-\omega^2/c^2}{is_x''\mu} & \frac{k_xk_y}{is_x''\mu} & 0 \\ \frac{k_xk_y}{is_x''\mu} & -\frac{k_x^2+s_x''^2\omega^2/c^2}{is_x''\mu} & 0 \\ 0 & 0 & -\frac{k_x^2+s_x''^2(\omega^2/c^2-k_y^2)}{is_x''\mu} \end{bmatrix}, \tag{C.1a}$$

$$\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0} = \begin{bmatrix} \frac{k_y^2-\omega^2/c^2}{\mu} & \frac{k_x k_y}{is_x''\mu} & 0 \\ \frac{k_x k_y}{is_x''\mu} & -\frac{k_x^2+s_x''^2\omega^2/c^2}{s_x''^2\mu} & 0 \\ 0 & 0 & -\frac{k_x^2+s_x''^2(\omega^2/c^2-k_y^2)}{s_x''^2\mu} \end{bmatrix}, \tag{C.1b}$$

where $c = 1/\sqrt{\mu\varepsilon}$ is substituted.

Now, we examine the singular values of $\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}$. The singular value of $\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}$ corresponding to the singular vector $[0\ 0\ 1]^\top$ is

$$\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0} = \frac{1}{s_x''\mu}\left|k_x^2 + s_x''^2\left(\frac{\omega^2}{c^2} - k_y^2\right)\right|, \tag{C.2}$$

which is an increasing function of k_x for $k_y < \omega/c$.

The remaining two singular values of $\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}$ corresponding to the singular vectors of the form $[a\ b\ 0]^\top$ are

$$\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0} = \frac{\sqrt{f_1 - f_2}}{\sqrt{2}s_x''\mu}, \quad \tilde{\sigma}_{\mathbf{k},2}^{\mathrm{u}_0} = \frac{\sqrt{f_1 + f_2}}{\sqrt{2}s_x''\mu}, \tag{C.3}$$

where

$$\begin{aligned} f_1 &= \left(k_x^2 + k_y^2 + s_x''^2\frac{\omega^2}{c^2}\right)^2 + \frac{\omega^2}{c^2}\left(\frac{\omega^2}{c^2} - 2(s_x''^2+1)k_y^2\right), \\ f_2 &= \left(k_x^2 + k_y^2 + (s_x''^2-1)\frac{\omega^2}{c^2}\right)\left[\left(k_x^2 + k_y^2 - (s_x''^2+1)\frac{\omega^2}{c^2}\right)^2 + 4k_x^2(s_x''^2+1)\frac{\omega^2}{c^2}\right]^{1/2}. \end{aligned} \tag{C.4}$$

Between the two singular values, we are only interested in $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$, the smaller of the two.

To prove that $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ is an increasing function of k_x, we show by straightforward algebra that the first derivative of $f_1 - f_2$ with respect to k_x is nonnegative. The derivative turns out to be

$$\frac{\partial}{\partial k_x}(f_1 - f_2) = \frac{-f_3 + f_4}{f_5}, \tag{C.5}$$

where

$$f_3 = 2k_x f_6\left(f_7 + f_6^2 + 2\frac{\omega^2}{c^2}f_6\right), \quad f_4 = 4k_x\left(f_6 + \frac{\omega^2}{c^2}\right)f_5, \quad f_5 = f_6\sqrt{f_7}, \tag{C.6}$$

and

$$f_6 = k_x^2 + k_y^2 + (s_x''^2 - 1)\frac{\omega^2}{c^2}, \quad f_7 = \left[k_x^2 + k_y^2 - (s_x''^2 + 1)\frac{\omega^2}{c^2}\right]^2 + 4k_x^2(s_x''^2 + 1)\frac{\omega^2}{c^2}. \tag{C.7}$$

We note that f_3, f_4, and f_5 are all positive because f_6 and f_7 are positive. In addition, the numerator $-f_3 + f_4$ in Eq. (C.5) is nonnegative because

$$-f_3^2 + f_4^2 = 64k_x^2k_y^2(s_x''^2 + 1)\frac{\omega^4}{c^4}\left[k_x^2 + s_x''^2\left(\frac{\omega^2}{c^2} - k_y^2\right)\right]f_6^2 \tag{C.8}$$

is nonnegative. Therefore, for $k_y < \omega/c$, Eq. (C.5) is nonnegative and $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ is an increasing function of k_x.

So far, we have shown that $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ and $\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}$ are increasing functions of k_x for a given $k_y < \omega/c$. Thus, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}) = \min\left\{\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}, \tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}\right\}$ is also an increasing function of k_x. Since we are considering $k_x \geq 0$, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0})$ is minimized at $k_x = 0$.

We can follow a similar procedure to prove that $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0})$ is an increasing function of k_x for $k_y < \omega/c$. Very briefly, the singular value of $T_{\mathbf{k}}^{\mathrm{sc}_0}$ corresponding to the singular vector $[0\ 0\ 1]^\top$ is

$$\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{sc}_0} = \frac{1}{s_x''^2\mu}\left|k_x^2 + s_x''^2\left(\frac{\omega^2}{c^2} - k_y^2\right)\right|, \tag{C.9}$$

which is an increasing function of k_x for $k_y < \omega/c$. The smaller of the remaining two singular values of $T_{\mathbf{k}}^{\mathrm{sc}_0}$ corresponding to the singular vectors of the form $[a\ b\ 0]^\top$ is

$$\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{sc}_0} = \frac{\sqrt{g_1 - g_2}}{\sqrt{2}s_x''^2\mu}, \tag{C.10}$$

where

$$g_1 = (k_x^2 + s_x''^2k_y^2)^2 + 2s_x''^2\frac{\omega^2}{c^2}\left[k_x^2 + s_x''^2\left(\frac{\omega^2}{c^2} - k_y^2\right)\right], \tag{C.11a}$$

$$g_2 = (k_x^2 + s_x''^2k_y^2)\left[k_x^4 + 2k_x^2s_x''^2\left(k_y^2 + \frac{2\omega^2}{c^2}\right) + s_x''^4\left(k_y^2 - \frac{2\omega^2}{c^2}\right)^2\right]^{1/2}. \tag{C.11b}$$

Then we have

$$\frac{\partial}{\partial k_x}(g_1 - g_2) = \frac{-g_3 + g_4}{g_5}, \tag{C.12}$$

where

$$g_3 = 4k_x\left[(k_x^2 + s_x''^2 k_y^2)^2 + s_x''^2\frac{\omega^2}{c^2}\left(3k_x^2 + s_x''^2\left(\frac{2\omega^2}{c^2} - k_y^2\right)\right)\right], \tag{C.13a}$$

$$g_4 = 4k_x\left(k_x^2 + s_x''^2\left(k_y^2 + \frac{\omega^2}{c^2}\right)\right)g_5, \tag{C.13b}$$

$$g_5 = \left[k_x^4 + 2k_x^2 s_x''^2\left(\frac{2\omega^2}{c^2} + k_y^2\right) + s_x''^4\left(k_y^2 - \frac{2\omega^2}{c^2}\right)^2\right]^{1/2}. \tag{C.13c}$$

In Eq. (C.12), the denominator g_5 is positive. In addition, the numerator $-g_3 + g_4$ is nonnegative because g_3, g_4, and

$$-g_3^2 + g_4^2 = 128k_x^2 k_y^2 s_x''^6\frac{\omega^4}{c^4}\left(k_x^2 + s_x''^2\left(\frac{\omega^2}{c^2} - k_y^2\right)\right) \tag{C.14}$$

are nonnegative. Therefore, $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{sc}_0}$ and $\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{sc}_0}$ are increasing functions of k_x for a given $k_y < \omega/c$, which implies that $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0}) = \min\left\{\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{sc}_0}, \tilde{\sigma}_{\mathbf{k},3}^{\mathrm{sc}_0}\right\}$ is also an increasing function of k_x. Since we are considering $k_x \geq 0$, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0})$ is minimized at $k_x = 0$.

Next, we consider $k_y > \omega/c$, and show that $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0})$ are minimized at $k_x = k_{x0} \equiv s_x''[k_y^2 - \omega^2/c^2]^{1/2}$ for such k_y. First of all, we see that $\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}$ and $\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{sc}_0}$ of Eqs. (C.2) and (C.9) are minimized at $k_x = k_{x0}$. In addition, since Eqs. (C.8) and (C.14) are negative for $k_x < k_{x0}$ and positive for $k_x > k_{x0}$, $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ and $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{sc}_0}$ are minimized at $k_x = k_{x0}$. Therefore, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}) = \min\left\{\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}, \tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}\right\}$ and $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0}) = \min\left\{\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{sc}_0}, \tilde{\sigma}_{\mathbf{k},3}^{\mathrm{sc}_0}\right\}$ are minimized at $k_x = k_{x0}$.

In summary, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0})$ and $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0})$ are minimized at $k_x = 0$ for $k_y < \omega/c$, and at $k_x = k_{x0}$ for $k_y > \omega/c$. Because $\widetilde{T}_{\mathbf{k}} = \widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}, \widetilde{T}_{\mathbf{k}}^{\mathrm{sc}_0}$ are good approximations of $T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{u}_0}, T_{\mathbf{k}}^{\mathrm{sc}_0}$, we have

$$\min_{k_x\geq 0}\sigma_{\min}(T_{\mathbf{k}}) \simeq \min_{k_x\geq 0}\sigma_{\min}(\widetilde{T}_{\mathbf{k}}) = \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}\right)_{k_x=0} \simeq \sigma_{\min}(T_{\mathbf{k}})_{k_x=0} \quad \text{for } k_y < \frac{\omega}{c} \tag{C.15}$$

and

$$\min_{k_x\geq 0}\sigma_{\min}(T_{\mathbf{k}}) \simeq \min_{k_x\geq 0}\sigma_{\min}(\widetilde{T}_{\mathbf{k}}) = \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}\right)_{k_x=k_{x0}} \simeq \sigma_{\min}(T_{\mathbf{k}})_{k_x=k_{x0}} \quad \text{for } k_y > \frac{\omega}{c}, \tag{C.16}$$

which are Eqs. (2.54) and (2.57), respectively.

C.2 Estimates of $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_{x0}}$ and $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_{x0}}$ for a given $k_y > \omega/c$

In this section, we derive the estimates of $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_{x0}}$ and $\sigma_{\min}\left(T_{\mathbf{k}}^{\mathrm{sc}_0}\right)_{k_x=k_{x0}}$ for a given $k_y > \omega/c$, where $k_{x0} = s_x''\sqrt{k_y^2 - \frac{\omega^2}{c^2}}$, using the perturbation method developed in Appendix A; the estimates are used in Eq. (2.58) in Sec. 2.3.4. The specific assumption $k_y > \omega/c$ as well as the assumptions $k_x \geq 0$, $k_y \geq 0$, $\varepsilon > 0$ made in Sec. 2.3.4 apply here.

Suppose that the given k_y is $k_{y0} > \omega/c$. Then

$$k_{x0} = s_x''\sqrt{k_{y0}^2 - \frac{\omega^2}{c^2}}. \tag{C.17}$$

Also define

$$\mathbf{k}_0 = \hat{\mathbf{x}}k_{x0} + \hat{\mathbf{y}}k_{y0}. \tag{C.18}$$

Then, the left-hand sides of Eq. (2.58) are $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{u}_0})$ and $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{sc}_0})$, which we evaluate below.

We approximate $\sigma_{\min}(T_{\mathbf{k}_0})$ for $T_{\mathbf{k}_0} = T_{\mathbf{k}_0}^{\mathrm{u}_0}, T_{\mathbf{k}_0}^{\mathrm{sc}_0}$ to first order in a small perturbation parameter δ. The perturbed quantity in $T_{\mathbf{k}_0}$ is the real part of s_x. We write s_x of Eq. (2.25) as

$$s_x = -is_x''(1+\delta), \tag{C.19}$$

where

$$\delta = \frac{i}{s_x''}. \tag{C.20}$$

Because $|\delta| \ll 1$ due to Eq. (2.26), the approximation of $\sigma_{\min}(T_{\mathbf{k}_0})$ to first order in δ should be an accurate estimate of $\sigma_{\min}(T_{\mathbf{k}_0})$.

We first derive the approximation of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{u}_0})$. One singular value of $T_{\mathbf{k}_0}^{\mathrm{u}_0}$ corresponding to the singular vector $[0\ 0\ 1]^\top$ is $\sigma_{\mathbf{k}_0,3}^{\mathrm{u}_0}$, which is $\sigma_{\mathbf{k},3}^{\mathrm{u}_0}$ of Eq. (2.32) for $\mathbf{k} = \mathbf{k}_0$. Because Eq. (C.19) implies

$$\frac{1}{s_x^2} = -\frac{1}{s_x''^2(1+\delta)^2} = -\frac{1}{s_x''^2}(1-2\delta) + O(\delta^2), \tag{C.21}$$

we have

$$\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,3} = |s_x| \left| -\frac{k_{x0}^2}{s_x''^2 \mu}(1-2\delta) + \frac{k_{y0}^2}{\mu} - \omega^2\varepsilon \right| + O(\delta^2) = 2|\delta||s_x| \left(\frac{k_{y0}^2}{\mu} - \omega^2\varepsilon \right) + O(\delta^2), \quad \text{(C.22)}$$

where k_{x0} is expressed in terms of k_{y0} using Eq. (C.17). Substituting Eq. (C.20) in Eq. (C.22) leads to

$$\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,3} = \frac{2\sqrt{s_x''^2+1}}{s_x''} \left(\frac{k_{y0}^2}{\mu} - \omega^2\varepsilon \right) + O(\delta^2). \quad \text{(C.23)}$$

The remaining two singular values of $T^{\mathrm{u}_0}_{\mathbf{k}_0}$ correspond to the singular vectors of the form $[a \;\; b \;\; 0]^\top$. Therefore, we can derive the two singular values by applying the perturbation method established in Appendix A to the top-left 2×2 block of $T^{\mathrm{u}_0}_{\mathbf{k}_0}$. Using Eq. (C.19) and

$$\frac{1}{s_x} = -\frac{1}{is_x''(1+\delta)} = -\frac{1}{is_x''}(1-\delta) + O(\delta^2), \quad \text{(C.24)}$$

we approximate the top-left 2×2 block of $T^{\mathrm{u}_0}_{\mathbf{k}}$ of Eq. (2.30b) for $\mathbf{k} = \mathbf{k}_0$ as

$$A = \begin{bmatrix} \frac{k_{y0}^2}{s_x\mu} - \frac{\omega^2\varepsilon}{s_x} & -\frac{k_{x0}k_{y0}}{s_x\mu} \\ -\frac{k_{x0}k_{y0}}{s_x\mu} & \frac{k_{x0}^2}{s_x\mu} - s_x\omega^2\varepsilon \end{bmatrix} \simeq \begin{bmatrix} \left(-\frac{k_{y0}^2}{is_x''\mu} + \frac{\omega^2\varepsilon}{is_x''}\right)(1-\delta) & \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) \\ \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) & -\frac{k_{x0}^2}{is_x''\mu}(1-\delta) + is_x''(1+\delta)\omega^2\varepsilon \end{bmatrix}. \quad \text{(C.25)}$$

Following the notations in Appendix A, Eq. (C.25) is decomposed as

$$A \simeq A^{(0)} + \delta A^{(1)} = \begin{bmatrix} -\frac{k_{x0}^2}{is_x''^3\mu} & \frac{k_{x0}k_{y0}}{is_x''\mu} \\ \frac{k_{x0}k_{y0}}{is_x''\mu} & \frac{is_x''k_{y0}^2}{\mu} \end{bmatrix} + \delta \begin{bmatrix} \frac{k_{x0}^2}{is_x''^3\mu} & -\frac{k_{x0}k_{y0}}{is_x''\mu} \\ -\frac{k_{x0}k_{y0}}{is_x''\mu} & \frac{k_{x0}^2}{is_x''\mu} + is_x''\omega^2\varepsilon \end{bmatrix}, \quad \text{(C.26)}$$

where $A^{(0)}$ and $A^{(1)}$ are simplified using Eq. (C.17).

We obtain the two singular values $\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,1}$ and $\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,2}$ of $T^{\mathrm{u}_0}_{\mathbf{k}_0}$ from A. However, since eventually we are interested in $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}_0})$, we focus on the smaller of the two, which is denoted by $\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,1}$. Because δ is small, it is reasonable to assume that the smaller singular value of A is the one perturbed from the smaller singular value of $A^{(0)}$, which is denoted by $\sigma_1^{(0)}$. Thus, we estimate $\sigma^{\mathrm{u}_0}_{\mathbf{k}_0,1}$ as the perturbation of $\sigma_1^{(0)}$. In fact, $\sigma_1^{(0)} = 0$ since $\det(A^{(0)}) = 0$.

The right singular vector $v_1^{(0)}$ corresponding to $\sigma_1^{(0)}$ is calculated by solving the eigenvalue problem $(A^{(0)\dagger}A^{(0)})v_1^{(0)} = \sigma_1^{(0)}v_1^{(0)}$ as described in Eq. (2.18). The result is

$$v_1^{(0)} = \frac{1}{\sqrt{k_{x0}^2/s_x''^2 + s_x''^2 k_{y0}^2}} \begin{bmatrix} -is_x'' k_{y0}^2 \\ -ik_{x0}^2/s_x'' \end{bmatrix}. \tag{C.27}$$

Using Eqs. (C.26) and (C.27) in Eq. (A.13), we obtain

$$\sigma_{\mathbf{k}_0,1}^{\mathrm{u}_0} = \left|\sigma_1^{(0)} + \delta\left(v_1^{(0)\top}A^{(1)}v_1^{(0)}\right)\right| + O(\delta^2) = 2\omega^2\varepsilon\frac{k_{y0}^2 - \omega^2\mu\varepsilon}{(s_x''^2+1)k_{y0}^2 - \omega^2\mu\varepsilon} + O(\delta^2), \tag{C.28}$$

where Eqs. (C.17), (C.19), and (C.20) are used to simplify the result.

Taking the ratio between Eqs. (C.23) and (C.28), we can easily see that $\sigma_{\mathbf{k}_0,1}^{\mathrm{u}_0} < \sigma_{\mathbf{k}_0,3}^{\mathrm{u}_0}$ in the leading order. Therefore, we conclude that

$$\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{u}_0}) = 2\omega^2\varepsilon\frac{k_{y0}^2 - \omega^2\mu\varepsilon}{(s_x''^2+1)k_{y0}^2 - \omega^2\mu\varepsilon} + O(\delta^2), \tag{C.29}$$

which is Eq. (2.58a).

Next, we derive the approximation of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{sc}_0})$. The overall procedure is very similar to the derivation of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathrm{u}_0})$. One singular value of $T_{\mathbf{k}_0}^{\mathrm{sc}_0}$ corresponding to the singular vector $[0\ 0\ 1]^\top$ is $\sigma_{\mathbf{k}_0,3}^{\mathrm{sc}_0}$, which is $\sigma_{\mathbf{k},3}^{\mathrm{sc}_0}$ of Eq. (2.32) for $\mathbf{k} = \mathbf{k}_0$. Using Eq. (C.21), we obtain

$$\sigma_{\mathbf{k}_0,3}^{\mathrm{sc}_0} = \left|-\frac{k_{x0}^2}{s_x''^2\mu}(1-2\delta) + \frac{k_{y0}^2}{\mu} - \omega^2\varepsilon\right| + O(\delta^2) = 2|\delta|\left(\frac{k_{y0}^2}{\mu} - \omega^2\varepsilon\right) + O(\delta^2), \tag{C.30}$$

where k_{x0} is expressed in terms of k_{y0} using Eq. (C.17). Substituting Eq. (C.20) in Eq. (C.30) results in

$$\sigma_{\mathbf{k}_0,3}^{\mathrm{sc}_0} = \frac{2}{s_x''}\left(\frac{k_{y0}^2}{\mu} - \omega^2\varepsilon\right) + O(\delta^2). \tag{C.31}$$

The remaining two singular values of $T_{\mathbf{k}_0}^{\mathrm{sc}_0}$ correspond to the singular vectors of the form $[a\ b\ 0]^\top$. Therefore, we derive the two singular values by applying the perturbation method of Appendix A to the top-left 2×2 block of $T_{\mathbf{k}_0}^{\mathrm{sc}_0}$. Using Eqs. (C.21) and (C.24), the top-left

2×2 block of $T_{\mathbf{k}}^{\mathrm{sc}_0}$ of Eq. (2.30c) for $\mathbf{k} = \mathbf{k}_0$ is approximated as

$$A = \begin{bmatrix} \frac{k_{y0}^2}{\mu} - \omega^2\varepsilon & -\frac{k_{x0}k_{y0}}{s_x\mu} \\ -\frac{k_{x0}k_{y0}}{s_x\mu} & \frac{k_{x0}^2}{s_x^2\mu} - \omega^2\varepsilon \end{bmatrix} \simeq \begin{bmatrix} \frac{k_{y0}^2}{\mu} - \omega^2\varepsilon & \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) \\ \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) & -\frac{k_{x0}^2}{s_x''^2\mu}(1-2\delta) - \omega^2\varepsilon \end{bmatrix}, \tag{C.32}$$

which is decomposed as

$$A \simeq A^{(0)} + \delta A^{(1)} = \begin{bmatrix} \frac{k_{x0}^2}{s_x''^2\mu} & \frac{k_{x0}k_{y0}}{is_x''\mu} \\ \frac{k_{x0}k_{y0}}{is_x''\mu} & -\frac{k_{y0}^2}{\mu} \end{bmatrix} + \delta \begin{bmatrix} 0 & -\frac{k_{x0}k_{y0}}{is_x''\mu} \\ -\frac{k_{x0}k_{y0}}{is_x''\mu} & \frac{2k_{x0}^2}{s_x''^2\mu} \end{bmatrix}, \tag{C.33}$$

where $A^{(0)}$ and $A^{(1)}$ are simplified using Eq. (C.17).

We solve the eigenvalue problem $(A^{(0)\dagger}A^{(0)})v_1^{(0)} = \sigma_1^{(0)}v_1^{(0)}$ for $\sigma_1^{(0)} = 0$ to obtain

$$v_1^{(0)} = \frac{1}{\sqrt{k_{x0}^2/s_x''^2 + k_{y0}^2}} \begin{bmatrix} k_{y0} \\ -ik_{x0}/s_x'' \end{bmatrix}. \tag{C.34}$$

Using Eqs. (C.33) and (C.34) in Eq. (A.13), we obtain the singular value of A perturbed from $\sigma_1^{(0)}$:

$$\sigma_{\mathbf{k}_0,1}^{\mathrm{sc}_0} = \left|\sigma_1^{(0)} + \delta\left(v_1^{(0)\top}A^{(1)}v_1^{(0)}\right)\right| + O(\delta^2) = \frac{2}{s_x''}\omega^2\varepsilon\frac{k_{y0}^2 - \omega^2\mu\varepsilon}{2k_{y0}^2 - \omega^2\mu\varepsilon} + O(\delta^2), \tag{C.35}$$

where Eqs. (C.17), (C.19), and (C.20) are used to simplify the result.

Taking the ratio between Eqs. (C.31) and (C.35), we can easily see that $\sigma_{\mathbf{k}_0,1}^{\mathrm{sc}_0} < \sigma_{\mathbf{k}_0,3}^{\mathrm{sc}_0}$ in the leading order. Therefore, we conclude that

$$\sigma_{\min}\left(T_{\mathbf{k}_0}^{\mathrm{sc}_0}\right) = \frac{2}{s_x''}\omega^2\varepsilon\frac{k_{y0}^2 - \omega^2\mu\varepsilon}{2k_{y0}^2 - \omega^2\mu\varepsilon} + O(\delta^2), \tag{C.36}$$

which is Eq. (2.58b).

C.3 Lowest-order approximation of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})/\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ around $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$

In this section, we derive the lowest-order approximation of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})/\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})$ around $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$, or equivalently around $k_x = 0$ and $k_y = \omega/c$; the approximation is used in Eq. (2.62) in Sec. 2.3.4.

First, we derive the lowest-order approximation of $[\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})]^2$. According to Eq. (2.18), the squares of the singular values of $T_{\mathbf{k}}^{\mathrm{u}_0}$ are the eigenvalues of $(T_{\mathbf{k}}^{\mathrm{u}_0})^\dagger T_{\mathbf{k}}^{\mathrm{u}_0}$. From $T_{\mathbf{k}}^{\mathrm{u}_0}$ of Eq. (2.30b), the eigenvalue of $(T_{\mathbf{k}}^{\mathrm{u}_0})^\dagger T_{\mathbf{k}}^{\mathrm{u}_0}$ corresponding to the eigenvector $[0\ 0\ 1]$ is

$$(\sigma_{\mathbf{k},3}^{\mathrm{u}_0})^2 = \frac{1}{\mu^2}\left[\frac{k_x^2}{1-is_x''} + (1-is_x'')\left(k_y^2 - \frac{\omega^2}{c^2}\right)\right]\left[\frac{k_x^2}{1+is_x''} + (1+is_x'')\left(k_y^2 - \frac{\omega^2}{c^2}\right)\right], \tag{C.37}$$

which is the square of $\sigma_{\mathbf{k},3}^{\mathrm{u}_0}$ in Eq. (2.32). We expand Eq. (C.37) into a Taylor series with respect to two variables $\Delta_{k_x} = k_x - 0$ and $\Delta_{k_y} = k_y - \omega/c$. The lowest-order-term representation of the series, shown in the order of ascending powers of Δ_{k_x}, is

$$\begin{aligned}(\sigma_{\mathbf{k},3}^{\mathrm{u}_0})^2 = &\left[\frac{4(s_x''^2+1)}{\mu^2}\frac{\omega^2}{c^2}\Delta_{k_y}^2 + O(\Delta_{k_y}^3)\right] + \left[-\frac{4}{\mu^2}\frac{s_x''^2-1}{s_x''^2+1}\frac{\omega}{c}\Delta_{k_y} + O(\Delta_{k_y}^2)\right]\Delta_{k_x}^2 \\ &+ \left[\frac{1}{(s_x''^2+1)\mu^2}\right]\Delta_{k_x}^4.\end{aligned} \tag{C.38}$$

The remaining two eigenvalues of $(T_{\mathbf{k}}^{\mathrm{u}_0})^\dagger T_{\mathbf{k}}^{\mathrm{u}_0}$ correspond to the eigenvectors of the form $[a\ b\ 0]^\top$. Solving the eigenvalue equation directly, we obtain the smaller of the two:

$$(\sigma_{\mathbf{k},1}^{\mathrm{u}_0})^2 = \frac{1}{2(s_x''^2+1)\mu^2}\left[f^{\mathrm{u}_0}(k_x,k_y) - \sqrt{g^{\mathrm{u}_0}(k_x,k_y)}\right], \tag{C.39}$$

where

$$f^{\mathrm{u}_0}(k_x,k_y) = (k_x^2+k_y^2)^2 + 2\left((s_x''^2-1)k_x^2 - k_y^2\right)\frac{\omega^2}{c^2} + \left((s_x''^2+1)^2+1\right)\frac{\omega^4}{c^4} \tag{C.40}$$

and

$$g^{\mathrm{u}_0}(k_x,k_y) = (k_x^2+k_y^2)^4 + 4(k_x^2+k_y^2)^2\left((s_x''^2-1)k_x^2-k_y^2\right)\frac{\omega^2}{c^2} + 2\left[(3s_x''^4-2s_x''^2+2)k_x^4 + 2\left((s_x''^2+1)^2+1\right)k_x^2k_y^2 - \left((s_x''^2+1)^2-3\right)k_y^4\right]\frac{\omega^4}{c^4} + 4s_x''^2(s_x''^2+2)\left((s_x''^2-1)k_x^2+k_y^2\right)\frac{\omega^6}{c^6} + s_x''^4(s_x''^2+2)^2\frac{\omega^8}{c^8}. \tag{C.41}$$

We expand Eq. (C.39) into a Taylor series with respect to Δ_{k_x} and Δ_{k_y}. The lowest-order-term representation of the series, shown in the order of ascending powers of Δ_{k_x}, is

$$(\sigma_{\mathbf{k},1}^{\mathrm{u}_0})^2 = \left[\frac{4}{(1+s_x''^2)\mu^2}\frac{\omega^2}{c^2}\Delta_{k_y}^2 + O(\Delta_{k_y}^3)\right] + \left[-\frac{4}{\mu^2}\frac{s_x''^2-1}{(s_x''^2+1)^3}\frac{\omega}{c}\Delta_{k_y} + O(\Delta_{k_y}^2)\right]\Delta_{k_x}^2 + \left[\frac{1}{(s_x''^2+1)^3\mu^2} + O(\Delta_{k_y})\right]\Delta_{k_x}^4 + O(\Delta_{k_x}^5). \tag{C.42}$$

Because $\Delta_{k_x} \simeq 0$ and $\Delta_{k_y} \simeq 0$, the $O(\Delta_{k_x}^n)$ and $O(\Delta_{k_y}^n)$ terms in Eqs. (C.38) and (C.42) can be ignored. Then, we realize that $(\sigma_{\mathbf{k},1}^{\mathrm{u}_0})^2 \simeq (\sigma_{\mathbf{k},3}^{\mathrm{u}_0})^2/(s_x''^2+1)^2$ and thus $(\sigma_{\mathbf{k},1}^{\mathrm{u}_0})^2 \ll (\sigma_{\mathbf{k},3}^{\mathrm{u}_0})^2$. Therefore, the lowest-order approximation of $[\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})]^2$ around $k_x = 0$ and $k_y = \omega/c$ is

$$[\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})]^2 \simeq \left[\frac{4}{(1+s_x''^2)\mu^2}\frac{\omega^2}{c^2}\Delta_{k_y}^2\right] - \left[\frac{4}{\mu^2}\frac{s_x''^2-1}{(s_x''^2+1)^3}\frac{\omega}{c}\Delta_{k_y}\right]\Delta_{k_x}^2 + \left[\frac{1}{(s_x''^2+1)^3\mu^2}\right]\Delta_{k_x}^4. \tag{C.43}$$

The lowest-order approximation of $[\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})]^2$ can be derived similarly by expanding $(\sigma_{\mathbf{k},3}^{\mathrm{sc}_0})^2$ and $(\sigma_{\mathbf{k},1}^{\mathrm{sc}_0})^2$ into Taylor series and choosing the smaller of the two:

$$[\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})]^2 \simeq \left[\frac{4}{\mu^2}\frac{\omega^2}{c^2}\Delta_{k_y}^2\right] + \left[-\frac{4}{\mu^2}\frac{s_x''^2-1}{(s_x''^2+1)^2}\frac{\omega}{c}\Delta_{k_y}\right]\Delta_{k_x}^2 + \left[\frac{1}{(s_x''^2+1)^2\mu^2}\right]\Delta_{k_x}^4. \tag{C.44}$$

We realize that Eq. (C.43) is proportional to Eq. (C.44) with the proportionality factor $1/(1+s_x''^2)$. Therefore, around $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ we have

$$\frac{\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})}{\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{sc}_0})} \simeq \frac{1}{\sqrt{1+s_x''^2}} = \frac{1}{|s_x|}, \tag{C.45}$$

which is Eq. (2.62).

Appendix D

Eigenvalues and eigenfunctions of $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$

Using the **k**-space representations of the operators, in this section we derive the eigenvalues Eq. (3.11) of $\nabla \times (\nabla \times \quad)$ and Eq. (3.12) of $\nabla(\nabla \cdot \quad)$ as well as their corresponding eigenfunctions.

Because both $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ are translationally invariant, their eigenfunctions have the form [48[Sec. 2.3.2], 49[Sec. 2.6.1]]

$$\mathbf{F} = \mathbf{F}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{D.1}$$

where $\mathbf{r}$ represents position, $\mathbf{k} = \hat{\mathbf{x}}k_x + \hat{\mathbf{y}}k_y + \hat{\mathbf{z}}k_z$ is a real constant wavevector, and $\mathbf{F}_{\mathbf{k}} = \hat{\mathbf{x}}F^x_{\mathbf{k}} + \hat{\mathbf{y}}F^y_{\mathbf{k}} + \hat{\mathbf{z}}F^z_{\mathbf{k}}$ is a $\mathbf{k}$-dependent vector.

We reformulate the eigenvalue equations $\nabla \times (\nabla \times \mathbf{F}) = \lambda\mathbf{F}$ and $\nabla(\nabla \cdot \mathbf{F}) = \lambda\mathbf{F}$ by substituting Eq. (D.1) for $\mathbf{F}$. Then, the eigenvalue equation for $\nabla \times (\nabla \times \quad)$ is

$$\begin{bmatrix} k_y^2 + k_z^2 & -k_x k_y & -k_x k_z \\ -k_y k_x & k_z^2 + k_x^2 & -k_y k_z \\ -k_z k_x & -k_z k_y & k_x^2 + k_y^2 \end{bmatrix} \begin{bmatrix} F^x_{\mathbf{k}} \\ F^y_{\mathbf{k}} \\ F^z_{\mathbf{k}} \end{bmatrix} = \lambda \begin{bmatrix} F^x_{\mathbf{k}} \\ F^y_{\mathbf{k}} \\ F^z_{\mathbf{k}} \end{bmatrix}, \tag{D.2}$$

and the eigenvalue equation for $\nabla(\nabla \cdot \quad)$ is

$$-\begin{bmatrix} k_x^2 & k_x k_y & k_x k_z \\ k_y k_x & k_y^2 & k_y k_z \\ k_z k_x & k_z k_y & k_z^2 \end{bmatrix} \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix} = \lambda \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix}. \tag{D.3}$$

By solving Eqs. (D.2) and (D.3) for a given $\mathbf{k}$, we obtain

$$\lambda = 0, \quad |\mathbf{k}|^2, \quad |\mathbf{k}|^2, \tag{D.4}$$

which is Eq. (3.11), as the eigenvalues of $\nabla \times (\nabla \times \quad)$, and

$$\lambda = -|\mathbf{k}|^2, \quad 0, \quad 0, \tag{D.5}$$

which is Eq. (3.12), as the eigenvalues of $\nabla(\nabla \cdot \quad)$, and Eq. (D.1) with

$$\mathbf{F}_{\mathbf{k}} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \begin{bmatrix} k_z \\ 0 \\ -k_x \end{bmatrix}, \begin{bmatrix} -k_y \\ k_x \\ 0 \end{bmatrix} \tag{D.6}$$

as the eigenfunctions corresponding to both Eqs. (D.4) and (D.5).

We note from Eqs. (D.4) and (D.5) that $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ are positive-semidefinite and negative-semidefinite, respectively.

Appendix E

Effect of the smallest root of $\tilde{p}_m \in \mathcal{P}_m$ on the slopes at the roots

In this section, we show that the slopes at the roots of a polynomial $\tilde{p}_m \in \mathcal{P}_m$ with all positive roots become steeper when the smallest root decreases in magnitude. This behavior is illustrated in Fig. 3.6.

Since $\tilde{p}_m \in \mathcal{P}_m$ satisfies the condition (3.16), it can be factored as

$$\tilde{p}_m(\zeta) = \prod_{i=1}^{d_m} \left(1 - \frac{\zeta}{\zeta_i}\right), \tag{E.1}$$

where $d_m \leq m$ is the degree of $\tilde{p}_m$ and ζ_i's are the roots of $\tilde{p}_m$. Hence, the slope of $\tilde{p}_m$ at a root ζ_k is

$$\tilde{p}'_m(\zeta_k) = -\frac{1}{\zeta_k} \prod_{i \neq k} \left(1 - \frac{\zeta_k}{\zeta_i}\right). \tag{E.2}$$

Now, suppose that $0 < \zeta_1 < \cdots < \zeta_{d_m}$. We can easily show that $|\tilde{p}'_m(\zeta_k)|$ increases for any k when ζ_1 decreases toward zero (while remaining positive) as follows. For $k = 1$, we have

$$|\tilde{p}'_m(\zeta_1)| = \frac{1}{\zeta_1} \left(1 - \frac{\zeta_1}{\zeta_2}\right) \cdots \left(1 - \frac{\zeta_1}{\zeta_{d_m}}\right), \tag{E.3}$$

which clearly increases as ζ_1 decreases to 0. For $k > 1$, we have

$$|\tilde{p}'_m(\zeta_k)| = \left(\frac{\zeta_k}{\zeta_1} - 1\right)\left[\frac{1}{\zeta_k}\prod_{i\neq 1,k}\left|1 - \frac{\zeta_k}{\zeta_i}\right|\right], \tag{E.4}$$

where the parentheses enclose the only quantity that is dependent on ζ_1. We can therefore see that $|\tilde{p}'_m(\zeta_k)|$ increases as ζ_1 decreases for $k > 1$ as well. Therefore, for a given $\tilde{p}_m \in \mathcal{P}_m$ whose roots are all positive, the slopes of $\tilde{p}_m$ at the roots become steeper if the smallest root decreases in magnitude while remaining positive. This situation is illustrated by the transition from Fig. 3.6a to Fig. 3.6b.

The slopes at the roots also become steeper when the originally positive ζ_1 is replaced by a negative value, as long as the negative value is smaller in magnitude than the original ζ_1. Replacing the originally positive ζ_1 with a negative quantity that is smaller in magnitude is equivalent to first replacing ζ_1 with a smaller positive value and then flipping its sign. Because we have already shown above that the slopes get steeper when the originally positive ζ_1 is replaced by a smaller positive value, it is sufficient to show that flipping the sign of ζ_1 makes the slopes even steeper. For a negative ζ_1, the slopes at the roots are

$$|\tilde{p}'_m(\zeta_1)| = \frac{1}{|\zeta_1|}\left(1 + \frac{|\zeta_1|}{\zeta_2}\right)\cdots\left(1 + \frac{|\zeta_1|}{\zeta_{d_m}}\right) \tag{E.5}$$

and

$$|\tilde{p}'_m(\zeta_k)| = \left(\frac{\zeta_k}{|\zeta_1|} + 1\right)\left[\frac{1}{\zeta_k}\prod_{i\neq 1,k}\left|1 - \frac{\zeta_k}{\zeta_i}\right|\right] \tag{E.6}$$

for $k > 1$. These slopes are steeper than Eqs. (E.3) and (E.4), respectively, which are the slopes for a positive ζ_1 with the same magnitude. Therefore, for a given $\tilde{p}_m \in \mathcal{P}_m$ whose roots are all positive, the slopes of $\tilde{p}_m$ at the roots become steeper if the smallest root is replaced by the one that is smaller in magnitude but negative. This situation is illustrated by the transition from Fig. 3.6a to Fig. 3.6c.

Appendix F

Trend in the slopes of a polynomial at the roots

In this section, we consider a polynomial p with all real roots, and show that the slope of p evaluated at a root closer to the median of the roots tends to be gentler than the slope evaluated at a root farther away from the median of the roots. This behavior is illustrated in Fig. 3.8.

Consider a polynomial of degree m,

$$p(\zeta) = \alpha \prod_{i=1}^{m} (\zeta - \zeta_i), \tag{F.1}$$

with $\zeta_1 < \cdots < \zeta_m$. The slope of p at a root ζ_k is

$$p'(\zeta_k) = \alpha \prod_{i \neq k} (\zeta_k - \zeta_i). \tag{F.2}$$

Now, we evaluate $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$. We first consider the case where the roots are evenly spaced, i.e., $\zeta_{i+1} - \zeta_i = (\text{const.})$, for which we have

$$\frac{|p'(\zeta_{k+1})|}{|p'(\zeta_k)|} = \frac{k!\,(m-k-1)!}{(k-1)!\,(m-k)!} = \frac{k}{m-k}. \tag{F.3}$$

Equation (F.3) is an increasing function of k for $1 \leq k \leq m-1$, and it is less than 1 for

$k < m/2$ and greater than 1 for $k > m/2$. Therefore, $|p'(\zeta_k)|$ is largest for $k = 1$ and $k = m$, and it decreases as k becomes closer to $k = \lfloor (m+1)/2 \rfloor$ and $k = \lceil (m+1)/2 \rceil$, which are the medians of the indices. In other words, for p with evenly spaced roots, the slopes of p get gentler at the roots closer to the median of the roots.

It is reasonable to expect that the above trend in the slopes also holds for p with unevenly spaced roots, unless the unevenness is too severe. To verify the expectation, we examine $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$ without assuming $\zeta_{i+1} - \zeta_i = (\text{const.})$:

$$\frac{|p'(\zeta_{k+1})|}{|p'(\zeta_k)|} = \frac{\prod_{i\neq k+1}|\zeta_{k+1}-\zeta_i|}{\prod_{i\neq k}|\zeta_k-\zeta_i|} = \prod_{i\neq k,k+1}\frac{|\zeta_{k+1}-\zeta_i|}{|\zeta_k-\zeta_i|} = \prod_{i=1}^{k-1}\left(\frac{\zeta_{k+1}-\zeta_i}{\zeta_k-\zeta_i}\right)\prod_{i=k+2}^{m}\left(\frac{\zeta_i-\zeta_{k+1}}{\zeta_i-\zeta_k}\right)$$
$$= \left[\prod_{i=1}^{k-1}\left(1+\frac{\zeta_{k+1}-\zeta_k}{\zeta_k-\zeta_i}\right)\right]\left[\prod_{i=k+2}^{m}\left(1-\frac{\zeta_{k+1}-\zeta_k}{\zeta_i-\zeta_k}\right)\right]. \tag{F.4}$$

Here, the factors within the first (second) brackets are always greater (less) than 1, so the number of factors greater (less) than 1 increases (decreases) for increasing k. Therefore, $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$ tends to be less than 1 for smaller k, and it tends to be greater than 1 for larger k. This means that as k increases $|p'(\zeta_k)|$ tends to decrease first and then tends to increase. Hence, even if the roots of p are unevenly spaced, the slopes of p tend to get gentler at the roots closer to the median of the roots.

Bibliography

[1] N. J. Champagne II, J. G. Berryman, and H. M. Buettner. "FDFD: A 3D finite-difference frequency-domain code for electromagnetic induction tomography". In: *Journal of Computational Physics* **170** (2001), pp. 830–848 (cit. on pp. ix, 61).

[2] G. Veronis and S. Fan. "Overview of Simulation Techniques for Plasmonic Devices". In: *Surface Plasmon Nanophotonics*. Ed. by M. L. Brongersma and P. Kik. Springer, 2007, pp. 169–182 (cit. on p. ix).

[3] U. S. Inan and R. A. Marshall. *Numerical Electromagnetics: The FDTD Method*. Cambridge University Press, 2011 (cit. on pp. ix, 7).

[4] J.-M. Jin. *The Finite Element Method in Electromagnetics*. 2nd ed. Wiley, 2002 (cit. on p. ix).

[5] X.-Q. Sheng and W. Song. *Essentials of Computational Electromagnetics*. Wiley, 2012 (cit. on p. ix).

[6] A. Taflove and S. C. Hagness. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*. 3rd ed. Artech House Publishers, 2005 (cit. on pp. ix, 7, 18, 54).

[7] A. Taflove. "Application of the Finite-Difference Time-Domain Method to Sinusoidal Steady-State Electromagnetic-Penetration Problems". In: *Electromagnetic Compatibility, IEEE Transactions on* **EMC-22** (1980), pp. 191–202 (cit. on p. 2).

[8] R. T. Ling. "A Finite-Difference Frequency-Domain (FD-FD) Approach To Electromagnetic Scattering Problems". In: *Journal of Electromagnetic Waves and Applications* **3** (1989), pp. 107–128 (cit. on p. 2).

[9] R. C. Rumpf. "Design and Optimization of Nano-optical Elements by Coupling Fabrication to Optical Behavior". PhD thesis. 2006 (cit. on p. 4).

[10] R. C. Rumpf, A. Tal, and S. M. Kuebler. "Rigorous electromagnetic analysis of volumetrically complex media using the slice absorption method". In: *Journal of Optical Society of America A* **24** (2007), pp. 3123–3134 (cit. on p. 4).

[11] K. Yee. "Numerical solution of inital boundary value problems involving Maxwell's equations in isotropic media". In: *Antennas and Propagation, IEEE Transactions on* **14** (1966), pp. 302–307 (cit. on pp. 7, 101).

[12] J. T. Smith. "Conservative modeling of 3-D electromagnetic fields, Part I: Properties and error analysis". In: *Geophysics* **61** (1996), pp. 1308–1318 (cit. on p. 7).

[13] J. W. Demmel. *Applied Numerical Algebra*. SIAM, 1997 (cit. on p. 8).

[14] M. R. Hestenes and E. Stiefel. "Methods of conjugate gradients for solving linear systems". In: *Journal of Research of the National Bureau of Standards* **49** (1952), pp. 409–436 (cit. on p. 9).

[15] Y. Saad. *Iterative Methods for Sparse Linear Systems*. 2nd ed. SIAM, 2003 (cit. on p. 9).

[16] V. Simoncini and D. B. Szyld. "Recent computational developments in Krylov subspace methods for linear systems". In: *Numerical Linear Algebra with Applications* **14** (2007), pp. 1–59 (cit. on p. 9).

[17] R. Fletcher. "Conjugate gradient methods for indefinite systems". In: *Numerical Analysis* (1976), pp. 73–89 (cit. on p. 9).

[18] D. A. H. Jacobs. "A generalization of the conjugate-gradient method to solve complex systems". In: *IMA Journal of Numerical Analysis* **6** (1986), pp. 447–452 (cit. on p. 9).

[19] R. W. Freund and N. M. Nachtigal. "QMR: a quasi-minimal residual method for non-Hermitian linear systems". In: *Numerische Mathematik* **60** (1991), pp. 315–339 (cit. on p. 9).

[20] Y. Saad and M. H. Schultz. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems". In: *SIAM Journal on Scientific and Statistical Computing* **7** (1986), pp. 856–869 (cit. on p. 9).

[21] S. Balay, J. Brown, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. *PETSc Web page*. http://www.mcs.anl.gov/petsc. 2011 (cit. on p. 9).

[22] P. B. Johnson and R. W. Christy. "Optical constants of the noble metals". In: *Physical Review B* **6** (1972), pp. 4370–4379 (cit. on pp. 11, 45).

[23] E. D. Palik, ed. *Handbook of Optical Constants of Solids*. Academic Press, 1985 (cit. on pp. 11, 12, 84).

[24] G. Veronis and S. Fan. "Modes of Subwavelength Plasmonic Slot Waveguides". In: *Journal of Lightwave Technology* **25** (2007), pp. 2511–2521 (cit. on pp. 11, 82, 84).

[25] D. R. Lide, ed. *CRC Handbook of Chemistry and Physics*. 88th ed. CRC Press, 2007 (cit. on pp. 12, 84).

[26] L. Verslegers, P. B. Catrysse, Z. Yu, W. Shin, Z. Ruan, and S. Fan. "Phase front design with metallic pillar arrays". In: *Optics Letters* **35** (2010), pp. 844–846 (cit. on pp. 12, 15).

[27] W. Shin and S. Fan. "Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell's equations solvers". In: *Journal of Computational Physics* **231** (2012), pp. 3406–3431 (cit. on p. 15).

[28] J.-P. Bérenger. "A perfectly matched layer for the absorption of electromagnetic waves". In: *Journal of Computational Physics* **114** (1994), pp. 185–200 (cit. on p. 15).

[29] G. Veronis and S. Fan. "Theoretical investigation of compact couplers between dielectric slab waveguides and two-dimensional metal-dielectric-metal plasmonic waveguides". In: *Optics Express* **15** (2007), pp. 1211–1221 (cit. on p. 15).

[30] Z. S. Sacks, D. M. Kingsland, R. Lee, and J.-F. Lee. "A perfectly matched anisotropic absorber for use as an absorbing boundary condition". In: *Antennas and Propagation, IEEE Transactions on* **43** (1995), pp. 1460–1463 (cit. on p. 15).

[31] W. C. Chew and W. H. Weedon. "A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates". In: *Microwave and Optical Technology Letters* **7** (1994), pp. 599–604 (cit. on p. 15).

[32] C. M. Rappaport. "Perfectly matched absorbing boundary conditions based on anisotropic lossy mapping of space". In: *Microwave and Guided Wave Letters, IEEE* **5** (1995), pp. 90–92 (cit. on p. 15).

[33] R. Mittra and U. Pekel. "A new look at the perfectly matched layer (PML) concept for the reflectionless absorption of electromagnetic waves". In: *Microwave and Guided Wave Letters, IEEE* **5** (1995), pp. 84–86 (cit. on p. 15).

[34] J. A. Roden and S. D. Gedney. "Convolution PML (CPML): An efficient FDTD implementation of the CFS-PML for arbitrary media". In: *Microwave and Optical Technology Letters* **27** (2000), pp. 334–339 (cit. on p. 15).

[35] J.-Y. Wu, D. M. Kingsland, J.-F. Lee, and R. Lee. "A comparison of anisotropic PML to Berenger's PML and its application to the finite-element method for EM scattering". In: *Antennas and Propagation, IEEE Transactions on* **45** (1997), pp. 40–50 (cit. on pp. 16, 53).

[36] Y. Y. Botros and J. L. Volakis. "A robust iterative scheme for FEM applications terminated by the perfectly matched layer (PML) absorbers". In: *Proceedings of the Fifteenth National Radio Science Conference*. 1998, pp. D11/1–D11/8 (cit. on p. 16).

[37] B. Stupfel. "A study of the condition number of various finite element matrices involved in the numerical solution of Maxwell's equations". In: *Antennas and Propagation, IEEE Transactions on* **52** (2004), pp. 3048–3059 (cit. on p. 16).

[38] P. K. Talukder, F.-J. Schmuckle, R. Schlundt, and W. Heinrich. "Optimizing the FDFD Method in Order to Minimize PML-Related Numerical Problems". In: *2007 International Microwave Symposium (IMS 2007)*. 2007, pp. 293–296 (cit. on pp. 16, 53).

[39] Y. Y. Botros and J. L. Volakis. "Preconditioned generalized minimal residual iterative scheme for perfectly matched layer terminated applications". In: *Microwave and Guided Wave Letters, IEEE* **9** (1999), pp. 45–47 (cit. on pp. 16, 53, 57).

[40] J.-M. Jin and W. C. Chew. "Combining PML and ABC for the finite-element analysis of scattering problems". In: *Microwave and Optical Technology Letters* **12** (1996), pp. 192–197 (cit. on p. 16).

[41] K. S. Kunz and R. J. Luebbers. *The Finite Difference Time Domain Method for Electromagnetics*. CRC Press, 1993 (cit. on pp. 18, 28).

[42] M. Benzi, G. H. Golub, and J. Liesen. "Numerical solution of saddle point problems". In: *Acta Numerica* **14** (2005), pp. 1–137 (cit. on pp. 22, 61, 67).

[43] B. N. Datta. *Numerical Linear Algebra and Applications*. 2nd ed. SIAM, 2010 (cit. on p. 22).

[44] G. H. Golub and C. F. Van Loan. *Matrix Computations*. 3rd ed. The Johns Hopkins University Press, 1996 (cit. on pp. 22, 24).

[45] R. B. Lehoucq, K. Maschhoff, D. C. Sorensen, and C. Yang. *ARPACK Web page*. http://www.caam.rice.edu/software/ARPACK. 2011 (cit. on p. 23).

[46] *MATLAB Web page*. http://www.mathworks.com/products/matlab. 2011 (cit. on p. 23).

[47] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide : Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998 (cit. on p. 23).

[48] J. W. Goodman. *Introduction to Fourier Optics*. 3rd ed. Roberts & Company Publishers, 2005 (cit. on pp. 26, 115).

[49] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-Time Signal Processing*. 2nd ed. Prentice Hall, 1999 (cit. on pp. 26, 28, 115).

[50] C. T. Wolfe, U. Navsariwala, and S. D. Gedney. "A parallel finite-element tearing and interconnecting algorithm for solution of the vector wave equation with PML absorbing medium". In: *Antennas and Propagation, IEEE Transactions on* **48** (2000), pp. 278–284 (cit. on p. 53).

[51] C. Kottke, A. Farjadpour, and S. G. Johnson. "Perturbation theory for anisotropic dielectric interfaces, and application to subpixel smoothing of discretized numerical methods". In: *Physical Review E* **77** (2008), art. no. 036611 (cit. on pp. 54, 55).

[52] F. L. Teixeira and W. C. Chew. "General closed-form PML constitutive tensors to match arbitrary bianisotropic and dispersive linear media". In: *Microwave and Guided Wave Letters, IEEE* **8** (1998), pp. 223–225 (cit. on p. 54).

[53] S. D. Gedney. "An anisotropic perfectly matched layer-absorbing medium for the truncation of FDTD lattices". In: *Antennas and Propagation, IEEE Transactions on* **44** (1996), pp. 1630–1639 (cit. on p. 54).

[54] G. A. Newman and D. L. Alumbaugh. "Three-dimensional induction logging problems, Part 2: A finite-difference solution". In: *Geophysics* **67** (2002), pp. 484–491 (cit. on pp. 60, 66).

[55] C. J. Weiss and G. A. Newman. "Electromagnetic induction in a generalized 3D anisotropic earth, Part 2: The LIN preconditioner". In: *Geophysics* **68** (2003), pp. 922–930 (cit. on p. 60).

[56] J. T. Smith. "Conservative modeling of 3-D electromagnetic fields, Part II: Biconjugate gradient solution and an accelerator". In: *Geophysics* **61** (1996), pp. 1319–1324 (cit. on p. 60).

[57] V. L. Druskin, L. A. Knizhnerman, and P. Lee. "New spectral Lanczos decomposition method for induction modeling in arbitrary 3-D geometry". In: *Geophysics* **64** (1999), pp. 701–706 (cit. on p. 60).

[58] E. Haber, U. M. Ascher, D. A. Aruliah, and D. W. Oldenburg. "Fast simulation of 3D electromagnetic problems using potentials". In: *Journal of Computational Physics* **163** (2000), pp. 150–171 (cit. on p. 60).

[59] J. Hou, R. Mallan, and C. Torres-Verdin. "Finite-difference simulation of borehole EM measurements in 3D anisotropic media using coupled scalar-vector potentials". In: *Geophysics* **71** (2006), G225–G233 (cit. on p. 60).

[60] R. Hiptmair, F. Krämer, and J. M. Ostrowski. "A robust Maxwell formulation for all frequencies". In: *Magnetics, IEEE Transactions on* **44** (2008), pp. 682–685 (cit. on p. 60).

[61] K. Beilenhoff, W. Heinrich, and H. L. Hartnagel. "Improved finite-difference formulation in frequency domain for three-dimensional scattering problems". In: *Microwave Theory and Techniques, IEEE Transactions on* **40** (1992), pp. 540–546 (cit. on pp. 60, 61).

[62] A. Christ and H. L. Hartnagel. "Three-dimensional finite-difference method for the analysis of microwave-device embedding". In: *Microwave Theory and Techniques, IEEE Transactions on* **35** (1987), pp. 688–696 (cit. on p. 60).

[63] A. Jennings. "Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method". In: *IMA Journal of Applied Mathematics* **20** (1977), pp. 61–72 (cit. on p. 61).

[64] A. van der Sluis and H. A. van der Vorst. "The rate of convergence of conjugate gradients". In: *Numerische Mathematik* **48** (1986), pp. 543–560 (cit. on p. 61).

[65] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer. "GMRES and the minimal polynomial". In: *BIT Numerical Mathematics* **36** (1996), pp. 664–675 (cit. on pp. 61, 67).

[66] S. Goossens and D. Roose. "Ritz and harmonic Ritz values and the convergence of FOM and GMRES". In: *Numerical Linear Algebra with Applications* **6** (1999), pp. 281–293 (cit. on p. 61).

[67] B. Beckermann and A. B. J. Kuijlaars. "Superlinear convergence of conjugate gradients". In: *SIAM Journal on Numerical Analysis* **39** (2001), pp. 300–329 (cit. on p. 61).

[68] B. Beckermann and A. B. J. Kuijlaars. "On the sharpness of an asymptotic error estimate for conjugate gradients". In: *BIT Numerical Mathematics* **41** (2001), pp. 856–867 (cit. on p. 61).

[69] O. Axelsson. "Iteration number for the conjugate gradient method". In: *Mathematics and Computers in Simulation* **61** (2003), pp. 421–435 (cit. on p. 61).

[70] J. P. Webb. "The Finite-Element Method for Finding Modes of Dielectric-Loaded Cavities". In: *Microwave Theory and Techniques, IEEE Transactions on* **33** (1985), pp. 635–639 (cit. on p. 62).

[71] F. Kikuchi. "Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism". In: *Computer Methods in Applied Mechanics and Engineering* **64** (1987), pp. 509–521 (cit. on p. 62).

[72] D. A. White and J. M. Koning. "Computing solenoidal eigenmodes of the vector Helmholtz equation: a novel approach". In: *Magnetics, IEEE Transactions on* **38** (2002), pp. 3420–3425 (cit. on p. 62).

[73] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Saunders College, 1976 (cit. on p. 65).

[74] R. W. Freund, G. H. Golub, and N. M. Nachtigal. "Iterative solution of linear systems". In: *Acta Numerica* **1** (1992), pp. 57–100 (cit. on p. 78).

[75] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade. *Photonic Crystals: Molding the Flow of Light*. 2nd ed. Princeton University Press, 2008 (cit. on p. 81).

[76] C. Manolatou, S. G. Johnson, S. Fan, P. R. Villeneuve, H. A. Haus, and J. D. Joannopoulos. "High-density integrated optics". In: *Journal of Lightwave Technology* **17** (1999), pp. 1682–1692 (cit. on p. 81).

[77] D. A. B. Miller. "Rationale and challenges for optical interconnects to electronic chips". In: *Proceedings of the IEEE* **88** (2000), pp. 728–749 (cit. on p. 81).

[78] W. L. Barnes, A. Dereux, and T. W. Ebbesen. "Surface plasmon subwavelength optics". In: *Nature* **424** (2003), pp. 824–830 (cit. on p. 81).

[79] E. Ozbay. "Plasmonics: merging photonics and electronics at nanoscale dimensions." In: *Science* **311** (2006), pp. 189–193 (cit. on pp. 81, 82).

[80] S. Maier. "Plasmonics: The promise of highly integrated optical devices". In: *IEEE Journal of Selected Topics in Quantum Electronics* **12** (2006), pp. 1671–1677 (cit. on p. 81).

[81] R. F. Oulton, V. J. Sorger, D. A. Genov, D. F. P. Pile, and X. Zhang. "A hybrid plasmonic waveguide for subwavelength confinement and long-range propagation". In: *Nature Photonics* **2** (2008), pp. 496–500 (cit. on p. 81).

[82] M. L. Brongersma and V. M. Shalaev. "The case for plasmonics". In: *Science* **328** (2010), pp. 440–441 (cit. on p. 81).

[83] D. K. Gramotnev and S. I. Bozhevolnyi. "Plasmonics beyond the diffraction limit". In: *Nature Photonics* **4** (2010), pp. 83–91 (cit. on p. 81).

[84] G. Veronis and S. Fan. "Bends and splitters in metal-dielectric-metal subwavelength plasmonic waveguides". In: *Applied Physics Letters* **87** (2005), art. no. 131102 (cit. on pp. 81, 89).

[85] W. Cai, W. Shin, S. Fan, and M. L. Brongersma. "Elements for plasmonic nanocircuits with three-dimensional slot waveguides". In: *Advanced Materials* **22** (2010), pp. 5120–5124 (cit. on pp. 82, 91).

[86] L. Liu, Z. Han, and S. He. "Novel surface plasmon waveguide for high integration". In: *Optics Express* **13** (2005), pp. 6645–6650 (cit. on p. 82).

[87] W. Shin, A. Raman, and S. Fan. "Instantaneous electric energy and electric power dissipation in dispersive media". In: *Journal of Optical Society of America B* **29** (2012), pp. 1048–1054 (cit. on p. 82).

[88] G. Veronis and S. Fan. "Crosstalk between three-dimensional plasmonic slot waveguides". In: *Optics Express* **16** (2008), pp. 2129–2140 (cit. on p. 82).

[89] D. F. P. Pile and D. K. Gramotnev. "Plasmonic subwavelength waveguides: next to zero losses at sharp bends". In: *Optics Letters* **30** (2005), pp. 1186–1188 (cit. on p. 82).

[90] V. S. Volkov, S. I. Bozhevolnyi, E. Devaux, and T. W. Ebbesen. "Bend loss for channel plasmon polaritons". In: *Applied Physics Letters* **89** (2006), art. no. 143108 (cit. on p. 82).

[91] D. F. P. Pile and D. K. Gramotnev. "Channel plasmon-polariton in a triangular groove on a metal surface". In: *Optics Letters* **29** (2004), pp. 1069–1071 (cit. on p. 82).

[92] E. Moreno, F. J. Garcia-Vidal, S. G. Rodrigo, L. Martin-Moreno, and S. I. Bozhevolnyi. "Channel plasmon-polaritons: modal shape, dispersion, and losses". In: *Optics Letters* **31** (2006), pp. 3447–3449 (cit. on p. 82).

[93] F. I. Baida, A. Belkhir, D. Van Labeke, and O. Lamrous. "Subwavelength metallic coaxial waveguides in the optical range: Role of the plasmonic modes". In: *Physical Review B* **74** (2006), art. no. 205419 (cit. on pp. 83, 84).

[94] P. B. Catrysse and S. Fan. "Understanding the dispersion of coaxial plasmonic structures through a connection with the planar metal-insulator-metal geometry". In: *Applied Physics Letters* **94** (2009), art. no. 231111 (cit. on pp. 83, 89).

[95] R. de Waele, S. P. Burgos, A. Polman, and H. A. Atwater. "Plasmon dispersion in coaxial waveguides from single-cavity optical transmission measurements." In: *Nano Letters* **9** (2009), pp. 2832–2837 (cit. on p. 83).

[96] A. Rusina, M. Durach, K. A. Nelson, and M. I. Stockman. "Nanoconcentration of terahertz radiation in plasmonic waveguides". In: *Optics Express* **16** (2008), pp. 18576–18589 (cit. on p. 83).

[97] A. Weber-Bargioni, A. Schwartzberg, M. Cornaglia, A. Ismach, J. J. Urban, Y. Pang, R. Gordon, J. Bokor, M. B. Salmeron, D. F. Ogletree, P. Ashby, S. Cabrini, and P. J. Schuck. "Hyperspectral nanoscale imaging on dielectric substrates with coaxial optical antenna scan probes." In: *Nano Letters* **11** (2011), pp. 1201–1207 (cit. on p. 83).

[98] A. Moreau, G. Granet, F. I. Baida, and D. Van Labeke. "Light transmission by subwavelength square coaxial aperture arrays in metallic films". In: *Optics Express* **11** (2003), pp. 1131–1136 (cit. on p. 83).

[99] F. I. Baida, D. Van Labeke, G. Granet, A. Moreau, and A. Belkhir. "Origin of the super-enhanced light transmission through a 2-D metallic annular aperture array: a study of photonic bands". In: *Applied Physics B: Lasers and Optics* **79** (2004), pp. 1–8 (cit. on p. 83).

[100] F. I. Baida. "Enhanced transmission through subwavelength metallic coaxial apertures by excitation of the TEM mode". In: *Applied Physics B: Lasers and Optics* **89** (2007), pp. 145–149 (cit. on p. 83).

[101] J. Rybczynski, K. Kempa, A. Herczynski, Y. Wang, M. J. Naughton, Z. F. Ren, Z. P. Huang, D. Cai, and M. Giersig. "Subwavelength waveguide for visible light". In: *Applied Physics Letters* **90** (2007), art. no. 021104 (cit. on p. 83).

[102] F. J. Rodríguez-Fortuño, C. García-Meca, R. Ortuño, J. Martí, and A. Martínez. "Coaxial plasmonic waveguide array as a negative-index metamaterial". In: *Optics Letters* **34** (2009), pp. 3325–3327 (cit. on p. 83).

[103] S. P. Burgos, R. de Waele, A. Polman, and H. A. Atwater. "A single-layer wide-angle negative-index metamaterial at visible frequencies". In: *Nature Materials* **9** (2010), pp. 407–412 (cit. on p. 83).

[104] R. de Waele, S. P. Burgos, H. A. Atwater, and A. Polman. "Negative refractive index in coaxial plasmon waveguides". In: *Optics Express* **18** (2010), pp. 12770–12778 (cit. on p. 83).

[105] J. A. Pereda, A. Vegas, and A. Prieto. "An improved compact 2D fullwave FDFD method for general guided wave structures". In: *Microwave and Optical Technology Letters* **38** (2003), pp. 331–335 (cit. on p. 84).

[106] S. Ramachandran, P. Kristensen, and M. F. Yan. "Generation and propagation of radially polarized beams in optical fibers". In: *Optics Letters* **34** (2009), pp. 2525–2527 (cit. on p. 85).

[107] M. Khajavikhan, A. Simic, M. Katz, J. H. Lee, B. Slutsky, A. Mizrahi, V. Lomakin, and Y. Fainman. "Thresholdless nanoscale coaxial lasers." In: *Nature* **482** (2012), pp. 204–207 (cit. on p. 85).

[108] N. Marcuvitz. *Waveguide Handbook*. The Institution of Engineering and Technology, 2009 (cit. on p. 86).

[109] E. A. Navarro, C. Wu, P. Y. Chung, and J. Litva. "Sensitivity analysis of the nonorthogonal FDTD method applied to the study of square coaxial waveguide structures". In: *Microwave and Optical Technology Letters* **8** (1995), pp. 138–140 (cit. on p. 87).

[110] J.-S. Huang, T. Feichtner, P. Biagioni, and B. Hecht. "Impedance matching and emission properties of nanoantennas in an optical nanocircuit". In: *Nano Letters* **9** (2009), pp. 1897–1902 (cit. on p. 91).

[111] B. C. Wadel. *Transmission Line Design Handbook*. Artech House, 1991 (cit. on p. 91).

[112] S. W. Conning. "The characteristic impedance of square coaxial line (correspondence)". In: *Microwave Theory and Techniques, IEEE Transactions on* **12** (1964), pp. 468–468 (cit. on p. 91).

[113] G. M. Anderson. "The calculation of the capacitance of coaxial cylinders of rectangular cross-section". In: *American Institute of Electrical Engineers, Transactions of the* **69** (1950), pp. 728–731 (cit. on p. 91).

[114] F. Bowman. *Introduction to Elliptic Functions with Applications*. Dover, 1961 (cit. on p. 91).

[115] C. G. Montgomery, R. H. Dicke, and E. M. Purcell, eds. *Principles of Microwave Circuits*. McGraw-Hill, 1948 (cit. on p. 92).

[116] R. Garg. "Coaxial Line Discontinuities". In: *Encyclopedia of RF and Microwave Engineering*. Ed. by K. Chang. John Wiley & Sons, 2005, pp. 653–658 (cit. on p. 92).

[117] A. W. Sanders, D. A. Routenberg, B. J. Wiley, Y. Xia, E. R. Dufresne, and M. A. Reed. "Observation of plasmon propagation, redirection, and fan-out in silver nanowires." In: *Nano Letters* **6** (2006), pp. 1822–1826 (cit. on p. 92).

[118] W. Wang, Q. Yang, F. Fan, H. Xu, and Z. L. Wang. "Light propagation in curved silver nanowire plasmonic waveguides". In: *Nano Letters* **11** (2011), pp. 1603–1608 (cit. on p. 92).

[119] Y. Yin, Y. Lu, Y. Sun, and Y. Xia. "Silver Nanowires Can Be Directly Coated with Amorphous Silica To Generate Well-Controlled Coaxial Nanocables of Silver/Silica". In: *Nano Letters* **2** (2002), pp. 427–430 (cit. on p. 92).

[120] X. Sun and Y. Li. "Cylindrical Silver Nanowires: Preparation, Structure, and Optical Properties". In: *Advanced Materials* **17** (2005), pp. 2626–2630 (cit. on p. 92).

[121] L.-B. Luo, S.-H. Yu, H.-S. Qian, and T. Zhou. "Large-Scale Fabrication of Flexible Silver/Cross-Linked Poly(vinyl alcohol) Coaxial Nanocables by a Facile Solution Approach". In: *Journal of the American Chemical Society* **127** (2005), pp. 2822–2823 (cit. on p. 92).

[122] C. J. Barrelet, A. B. Greytak, and C. M. Lieber. "Nanowire Photonic Circuit Elements". In: *Nano Letters* **4** (2004), pp. 1981–1985 (cit. on p. 92).

[123] D. Wang, F. Qian, C. Yang, Z. Zhong, and C. M. Lieber. "Rational Growth of Branched and Hyperbranched Nanowire Structures". In: *Nano Letters* **4** (2004), pp. 871–874 (cit. on p. 92).

[124] K. A. Dick, K. Deppert, M. W. Larsson, T. Mårtensson, W. Seifert, L. R. Wallenberg, and L. Samuelson. "Synthesis of branched 'nanotrees' by controlled seeding of multiple branching events." In: *Nature Materials* **3** (2004), pp. 380–384 (cit. on p. 92).

[125] J. Yao, H. Yan, and C. M. Lieber. "A nanoscale combing technique for the large-scale assembly of highly aligned nanowires". In: *Nature Nanotechnology* **8** (2013), pp. 329–335 (cit. on p. 92).

[126] B. Engquist and L. Ying. "Fast algorithms for high frequency wave propagation". In: *Numerical Analysis of Multiscale Problems*. Springer, 2012, pp. 127–161 (cit. on p. 94).

[127] T. Kolev and P. Vassilevski. "Parallel Auxiliary Space AMG Solver for H(div) Problems". In: *SIAM Journal on Scientific Computing* **34** (2012), A3079–A3098 (cit. on p. 94).

[128] L. D. Landau and E. M. Lifshitz. *Quantum Mechanics: Non-relativistic Theory*. 3rd ed. Vol. 3. Course of Theoretical Physics. Butterworth-Heinemann, 1977 (cit. on p. 97).

[129] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985 (cit. on pp. 97, 98).

[130] T. Takagi. "On an algebraic problem related to an analytic theorem of Carathedory and Fejer and on an allied theorem of Landau". In: *Japanese Journal of Mathematics* **1** (1924), pp. 82–93 (cit. on pp. 97, 98).

[131] A. Bunse-Gerstner and W. Gragg. "Singular value decompositions of complex symmetric matrices". In: *Journal of Computational and Applied Mathematics* **21** (1988), pp. 41–54 (cit. on p. 97).

Index

Arnoldi Package (ARPACK), 23

characteristic impedance, 91
condition number, 21, 75
constitutive equation, 2
continuity equation, 61

direct method, 8

eigenfunction, 63, 115
eigenvalue, 63, 115
electric field, 1
electric permittivity, 1
    dielectric, 31
    metal, 31

finite-difference frequency-domain (FDFD) method, 3
finite-difference method, 5
finite-difference time-domain (FDTD) method, 3
finite element method (FEM), ix, 15, 93
Fourier transform, 2

graphics processing unit (GPU), 94

Helmholtz decomposition, 60
homogeneity

homogeneous EM system, 25, 62
inhomogeneous EM system, 39, 77

iterative method, 8, 94
biconjugate gradient (BiCG), 9, 20, 83
conjugate gradient, 9
convergence of, 10, 19, 56, 66, 77, 94
generalized minimal residual (GMRES), 9, 66
Krylov subspace, 9
quasi-minimal residual (QMR), 9, 19, 76

low-frequency regime, 59, 65

magnetic field, 1
magnetic permeability, 1
magnetic material, 78
nonmagnetic material, 59
matrix and operator
composite, 26
conjugate transpose, 22, 68
Hermitian, 9, 23, 98
ill-conditioned, 3, 21
indefinite, 64, 74
Laplacian, 60
nonsingular, 22, 31
positive-definite, 9, 60, 71
positive-semidefinite, 64
singular, 31
symmetric, 9, 24, 97
translationally invariant, 25, 115
unitary, 22, 68
well-conditioned, 21
matrix decomposition

Cholesky factorization, 8
eigenvalue decomposition, 68
$LDL^\intercal$ factorization, 8
$LDM^\intercal$ factorization, 8
LU factorization, 8
singular value decomposition (SVD), 22, 97
symmetric SVD, 97
Takagi factorization, 97
Maxwell's equations, 1
frequency-domain, 1
time-domain, 2
method of moments, ix, 93

norm
∞-norm, 24
1-norm, 24
2-norm, 9
A-norm, 9
matrix norm, 24
p-norm, 24
vector norm, 8, 24
null space, 60
Nyquist wavenumber, 28

perfect electric conductor (PEC), 83
perfectly matched layer (PML), 15
convolutional PML (CPML), 15
scale-factor-preconditioned UPML (SP-UPML), 56
stretched-coordinate PML (SC-PML), 15
uniaxial PML (UPML), 15
periodic boundary condition, 64
perturbation method, 97

portable, extensible toolkit for scientific computation (PETSc), 9
preconditioner
    approximate inverse preconditioner, 53, 57
    Jacobi preconditioner, 57
    scale-factor preconditioner, 55

quasi-static approximation, 86

residual polynomial, 67
residual vector, 8, 68

singular value, 22
singular vector, 22
speed of light, 19

Taylor series, 113
telecommunication wavelength, 11, 83
transmission line, 86

vacuum impedance, 4, 18
variational method, 24, 39

waveguide
    metal-dielectric-metal (MDM), 44, 81
    plasmonic coaxial, 83
      bend, 85
      splitter, 88
    plasmonic slot, 11, 82
    rectangular dielectric, 12
    V-groove, 82