Journal of Computational Physics 231 (2012) 3406–3431

Contents lists available at SciVerse ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell's equations solvers

Wonseok Shin, Shanhui Fan *

Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 4 September 2011
Received in revised form 9 January 2012
Accepted 11 January 2012
Available online 20 January 2012

Keywords:
Maxwell's equations
Perfectly matched layer
Finite-difference frequency-domain method
Finite-element method
Condition number
Preconditioner
Iterative methods

ABSTRACT

We show that the performance of frequency-domain solvers of Maxwell's equations is greatly affected by the kind of the perfectly matched layer (PML) used. In particular, we demonstrate that using the stretched-coordinate PML (SC-PML) results in significantly faster convergence speed than using the uniaxial PML (UPML). Such a difference in convergence behavior is explained by an analysis of the condition number of the coefficient matrices. Additionally, we develop a diagonal preconditioning scheme that significantly improves solver performance when UPML is used.

1. Introduction

The perfectly matched layer (PML) is an artificial medium initially developed by Bérenger that absorbs incident electromagnetic (EM) waves omnidirectionally with virtually no reflection [1]. 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 [2] or an isolated structure in an infinite vacuum region [3].

Bérenger's original PML was followed by many variants. In the finite-difference time-domain (FDTD) method of solving Maxwell's equations [4], the uniaxial PML (UPML) [5] and stretched-coordinate PML (SC-PML) [6–8] are the most popular, both resulting in similar numerical performance.[1]

In frequency-domain methods such as the finite-difference frequency-domain (FDFD) method and finite-element method (FEM), on the other hand, UPML and SC-PML result in the systems of linear equations

$$Ax = b \tag{1.1}$$

with different coefficient matrices A. In general, it is empirically known that the use of any PML leads to an ill-conditioned coefficient matrix and slows down the convergence of iterative methods to solve (1.1) [10–14]. 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 [15] that briefly mentions empirical observations.

* Corresponding author.
E-mail addresses: wsshin@stanford.edu (W. Shin), shanhui@stanford.edu (S. Fan).

[1] The convolutional PML (CPML) [9] that is widely used in time-domain simulation is in essence SC-PML.

0021-9991/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2012.01.013

In this paper, we demonstrate that the choice of PML significantly influences the convergence of iterative methods to solve the frequency-domain Maxwell's equations. In particular, we show that SC-PML leads to far faster convergence than UPML. We also present an analysis relating convergence speed to the condition number of the coefficient matrix.

The paper is organized as follows. In Section 2 we review the basic formulations of UPML and SC-PML for the frequency-domain Maxwell's equations. Then, in Section 3 we demonstrate that SC-PML gives rise to much faster convergence of iterative methods than UPML for realistic three-dimensional (3D) problems. In Section 4 we show that SC-PML produces a much better-conditioned coefficient matrix than UPML. Finally, we introduce a diagonal preconditioning scheme for UPML in Section 5; the newly developed preconditioning scheme can be very useful in situations where UPML is easier to implement than SC-PML.

We use the FDFD method throughout the paper to construct coefficient matrices. However, the arguments we present should be equally applicable to other frequency-domain methods including FEM.

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

In this section, we briefly review the use of PML in the frequency-domain formulation of Maxwell's equations.

Assuming a time dependence $e^{+i\omega t}$, the frequency-domain Maxwell's equations reduce to

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

where ε and μ are the electric permittivity and magnetic permeability; ω is the angular frequency; **E** and **J** are the electric field and the electric current source density, respectively. Throughout this paper, we assume that $\mu = \mu_0$, which is the magnetic permeability of a vacuum; this is valid for most nanophotonic simulations.

The FDFD method discretizes (2.1) by using finite-difference approximations of continuous spatial derivatives on a grid such as the Yee grid [16–18] to produce a system of linear equations of the form (1.1):

$$Ae = -i\omega j, \tag{2.2}$$

where e and j are column vectors that represent discretized **E** and **J**, respectively.

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 (2.1). For an EM system surrounded by UPML, the governing equation is the UPML equation

$$\nabla \times (\bar{\bar{\mu}}_s)^{-1} \nabla \times \mathbf{E} - \omega^2 \bar{\bar{\varepsilon}}_s \mathbf{E} = -i\omega \mathbf{J}, \tag{2.3}$$

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

$$\bar{\bar{\varepsilon}}_s = \varepsilon \begin{bmatrix} \frac{s_y s_z}{s_x} & 0 & 0 \\ 0 & \frac{s_z s_x}{s_y} & 0 \\ 0 & 0 & \frac{s_x s_y}{s_z} \end{bmatrix}, \qquad \bar{\bar{\mu}}_s = \mu \begin{bmatrix} \frac{s_y s_z}{s_x} & 0 & 0 \\ 0 & \frac{s_z s_x}{s_y} & 0 \\ 0 & 0 & \frac{s_x s_y}{s_z} \end{bmatrix}. \tag{2.4}$$

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

$$\nabla_s \times \mu^{-1} \nabla_s \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E} = -i\omega \mathbf{J}, \tag{2.5}$$

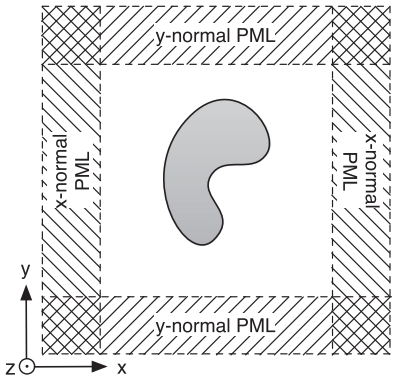


Fig. 2.1. An example of an EM system surrounded by PML. In the four corner regions where the *x*- and *y*-normal PMLs overlap, waves attenuate in both directions. If the EM system is in a 3D simulation domain, PMLs can overlap up to three times. PML is either UPML or SC-PML.

where

$$\nabla_s = \hat{\mathbf{x}}\frac{1}{s_x}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{1}{s_y}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{1}{s_z}\frac{\partial}{\partial z}. \tag{2.6}$$

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

$$s_w(l) = 1 - \mathrm{i}s''_w(l) = \begin{cases} 1 - \mathrm{i}\frac{\sigma_w(l)}{\omega\varepsilon_0} & \text{inside the } w\text{-normal PML}, \\ 1 & \text{elsewhere}, \end{cases} \tag{2.7}$$

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 a 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 paper, 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 [4] so that

$$\sigma_w(l) = \sigma_{w,\max}\left(\frac{l}{d}\right)^m, \tag{2.8}$$

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

$$\sigma_{w,\max} = -\frac{(m+1)\ln R}{2\eta_0 d}, \tag{2.9}$$

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 edge length Δ. For a typical 10-layer PML with $d = 10\Delta$, $m = 4$, $R = \mathrm{e}^{-16} \simeq 1 \times 10^{-7}$, 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 [19]. Therefore, if the medium matched by PML is a vacuum, the vacuum wavelength λ_0 corresponding to ω should satisfy $\lambda_0 \geqslant 15\Delta$, which implies that

$$s''_w(d) = \frac{\sigma_{w,\max}}{\omega\varepsilon_0} = \frac{\frac{4}{\eta_0\Delta}}{\frac{2\pi}{\lambda_0}c_0\varepsilon_0} = \frac{2\lambda_0}{\pi\Delta} \geqslant \frac{30\Delta}{\pi\Delta} \simeq 9.549, \tag{2.10}$$

where $c_0 = 1/\sqrt{\mu_0\varepsilon_0}$ is the speed of light in a 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 $\Delta = 1$ nm for a vacuum wavelength $\lambda_0 = 1550$ nm is not uncommon [20]. In these cases, $|s_w(d)|$ is nearly 1000.

Depending on the kind of PML used, we solve either (2.3) or (2.5) 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, which are respectively referred to as

$$A^{\mathrm{u}}x = b, \tag{2.11}$$

$$A^{\mathrm{sc}}x = b, \tag{2.12}$$

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 (2.12) is much more favorable to numerical solvers than (2.11).

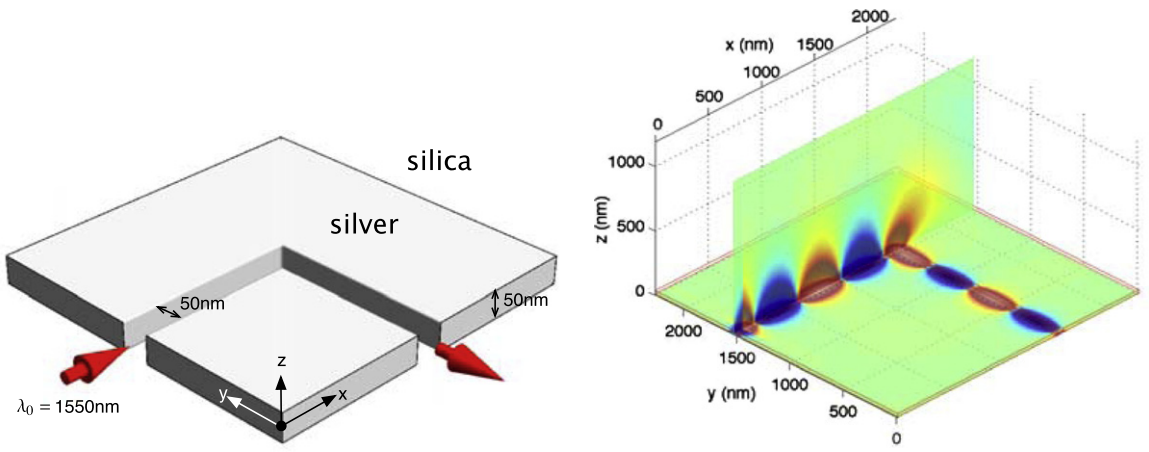
3. Convergence speed of iterative methods to solve the UPML and SC-PML equations

In this section, we apply UPML and SC-PML to realistic 3D EM systems, and compare the two PMLs in terms of convergence speed of iterative FDFD solvers.

The first EM system that we examine is a 90° bend of a slot waveguide formed in a thin metal film (Fig. 3.1(a)). Metallic slot waveguides are a subject of active research in nanophotonics due to their capability of guiding light at a deep-subwavelength scale [20].

We simulate the propagation of an EM wave at the telecommunication wavelength $\lambda_0 = 1550$ nm through the bend. A $\mathbf{J}$ source plane is placed near $x = 0$ to launch the fundamental mode of the waveguide. To simulate an infinitely long metallic slot waveguide immersed in a dielectric medium, all six boundary faces of the Cartesian simulation domain are covered by PML. The solution obtained by the FDFD method is displayed in Fig. 3.1(b).

The second EM system that we simulate is a rectangular dielectric waveguide (Fig. 3.2(a)). We launch the fundamental mode in the dielectric waveguide.



(a) Slot waveguide bend (Slot) (b) Distribution of $\mathrm{Re}\{H_z\}$

Fig. 3.1. The FDFD simulation of wave propagation through a metallic slot waveguide bend. In (a), the structure of the bend is illustrated. A narrow slot is formed between two pieces of the thin silver (Ag) film immersed in a background of silica (SiO_2). The waveguide is bent 90°. The vacuum wavelength and relevant dimensions 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 boundary faces of the Cartesian simulation domain are covered by PML. In (b), $\mathrm{Re}\{H_z\}$ 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 indicates $\mathrm{Re}\{H_z\} > 0$, and blue indicates $\mathrm{Re}\{H_z\} < 0$. Only the $z \geqslant 0$ 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 thin orange lines slightly above the $z = 0$ plane outline the two metal pieces. The electric permittivities of silver [21] and silica [22] at $\lambda_0 = 1550$ nm are $\varepsilon_{Ag} = (-129 - i3.28)\varepsilon_0$ and $\varepsilon_{SiO_2} = 2.085\varepsilon_0$, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The last system is an array of metallic pillars (Fig. 3.2(b)). We launch a plane wave toward the pillars and observe how it is scattered by them; the detailed analysis is described in [3].

For each of the three EM systems, we construct two systems of linear equations by the FDFD method: one with UPML and the other with SC-PML. The number of the grid cells in the finite-difference grid used to discretize each EM system is shown in Table 3.1, together with the grid edge lengths in the *x*-, *y*-, *z*-directions.

The constructed systems of linear equations are solved by the quasi-minimal residual (QMR) iterative method [24].[2] At the *i*th step of the QMR iteration, an approximate solution x_i is generated. As i increases, x_i eventually converges to the exact solution of the system of linear equations $Ax = b$. We assume that convergence is achieved when the residual vector

$$r_i = b - Ax_i \tag{3.1}$$

satisfies $\|r_i\|/\|b\| < \tau$, where $\|\cdot\|$ is the 2-norm of a vector and τ is a user-defined small positive number. In practice, $\tau = 10^{-6}$ is sufficient for accurate solutions.

Fig. 3.3 shows $\|r_i\|/\|b\|$ versus the iteration step i for the three EM systems, each simulated with the two different types of PMLs. For all three EM systems, SC-PML significantly outperforms UPML in terms of convergence speed.

The three EM systems tested above are chosen deliberately to include geometries with different degrees of complexities, and different materials such as dielectrics and metals. Therefore, Fig. 3.3 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 behavior for other iterative methods, such as the biconjugate gradient (BiCG) method [26]. 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 the iterative method used.

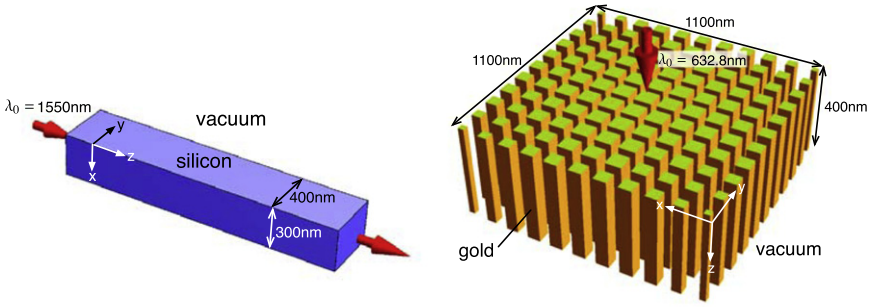
In the next section, we relate the significantly different convergence speeds to the very different condition numbers of the UPML and SC-PML matrices.

4. 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

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}, \tag{4.1}$$

[2] The large-scale matrix–vector multiplication required in the QMR algorithm is implemented by the PETSc library [25] with double-precision floating point numbers.



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

Fig. 3.2. Two additional EM systems for which the convergence of QMR is tested. The materials and dimensions 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 Cartesian 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 the periodic boundary conditions so that the metallic pillars do not extend into PML. The electric permittivities of silicon (Si) [22] at $\lambda_0 = 1550$ nm and gold (Au) [23] at $\lambda_0 = 632.8$ nm are $\varepsilon_{Si} = 12.09\varepsilon_0$ and $\varepsilon_{Au} = (-10.78 - i0.79)\varepsilon_0$, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3.1
The specification of the finite-difference grids used for the three simulated EM systems described in Figs. 3.1 and 3.2. The number of grid cells in each EM system is $N_xN_yN_z$, which results in $3N_xN_yN_z$ of unknowns in a column vector x, where the extra factor 3 accounts for the three Cartesian components of the E-field. Slot uses a nonuniform grid with smoothly varying grid edge lengths.

	Slot	Diel	Array
$N_x \times N_y \times N_z$	$192 \times 192 \times 240$	$220 \times 220 \times 320$	$220 \times 220 \times 130$
Δ_x, Δ_y, Δ_z (nm)	2–20	10	5,5,20

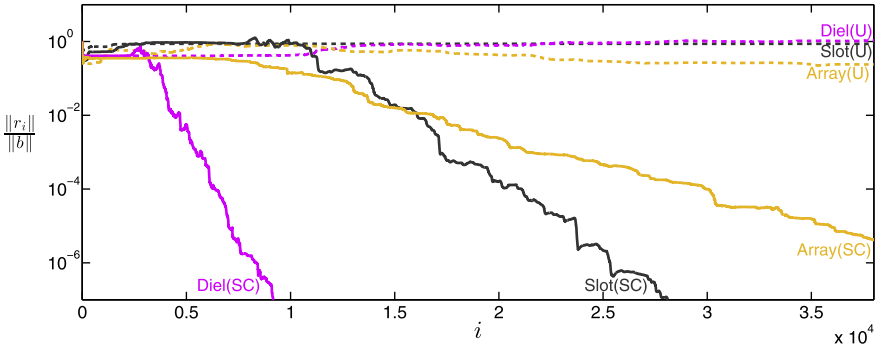


Fig. 3.3. Convergence of QMR for the metallic slot waveguide (Slot), rectangular dielectric waveguide (Diel), and the metallic pillar array (Array), combined with UPML (U) and SC-PML (SC). Notice that simply replacing UPML with SC-PML improves convergence dramatically for all the three EM systems.

where $\sigma_{max}(A)$ and $\sigma_{min}(A)$ are the maximum and minimum singular values of A as we will review in Section 4.1. Matrices with large and small condition numbers are called ill-conditioned and well-conditioned, respectively. For convenience, we introduce notations

$$\sigma^{u}_{max} = \sigma_{max}(A^{u}), \quad \sigma^{u}_{min} = \sigma_{min}(A^{u}), \quad \kappa^{u} = \frac{\sigma^{u}_{max}}{\sigma^{u}_{min}} \tag{4.2}$$

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

The objective of this section is to show that in general UPML produces a much worse-conditioned coefficient matrix than SC-PML, i.e., $\kappa^{u}/\kappa^{sc} \gg 1$, provided that the two PMLs enclose the same EM system. According to (4.1), the objective is accomplished by analyzing the extreme singular values of A^{u} and A^{sc}.

All EM systems simulated in Section 3 are inhomogeneous, being composed of several different EM media. For each component medium, we can associate a corresponding infinite space that is filled homogeneously with the medium. For

example, for an EM system of a vacuum surrounded by UPML, we can imagine an infinite space filled either with a vacuum or with UPML homogeneously. It turns out that the extreme singular values of an inhomogeneous EM system are strongly related to the extreme singular values of the homogeneous EM media constituting the inhomogeneous EM system. Of particular interest are the homogeneous regular medium, homogeneous UPML, and homogenous SC-PML. The maximum and minimum singular values of the three homogeneous media are studied in Sections 4.2–4.4.

In Section 4.5, we develop a theory based on a 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 $\kappa^{u}/\kappa^{sc} \gg 1$. In Section 4.6, we verify the theory numerically for two inhomogeneous EM systems.

The conclusion of this section explains the results in Section 3, because a smaller condition number of A generally implies faster convergence of iterative methods to solve a system of linear equations $Ax = b$ [27]. In fact, an ill-conditioned coefficient matrix can be detrimental to direct methods as well; it is known that the LU factorization of ill-conditioned matrices tends to be inaccurate [28]. 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.

4.1. Mathematical background

For an arbitrary $A \in \mathbb{C}^{n\times n}$, one can always perform a singular value decomposition (SVD) as [29]

$$A = U\Sigma V^{\dagger}, \tag{4.3}$$

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

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^{\dagger}, \tag{4.4}$$

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 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,

$$\sigma_{\max} = \max_{1\leqslant i\leqslant n} \sigma_i \ \text{ and } \ \sigma_{\min} = \min_{1\leqslant i\leqslant n} \sigma_i, \tag{4.5}$$

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 (4.4), it follows that

$$Av_i = \sigma_i u_i \ \text{ and } \ A^{\dagger}u_i = \sigma_i v_i. \tag{4.6}$$

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

$$H(A)\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \sigma_i \begin{bmatrix} u_i \\ v_i \end{bmatrix}, \ \text{where } H(A) = \begin{bmatrix} 0 & A \\ A^{\dagger} & 0 \end{bmatrix}. \tag{4.7}$$

In this paper, we solve (4.7) for the largest or smallest nonnegative eigenvalues by the Arnoldi Package (ARPACK) [30] 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 a vector are $H(A)$ and $H(A)^{-1}$, respectively [32]. 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 Section 4 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

$$(A^{\dagger}A)v_i = \sigma_i^2 v_i \tag{4.8}$$

that is derived from (4.6). 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 (4.7) rather than (4.8) to solve for the singular values numerically. Nevertheless, (4.8) turns out to be useful in the theoretical analysis in Sections 4.2, 4.3, and 4.4.

[3] The actual calculation of the extreme singular values is carried out using the MATLAB routine svds [31], which uses ARPACK internally.

The extreme singular values can also be calculated by a variational method. As a consequence of (4.4) we have

$$\sigma_{\max} = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \quad \text{and} \quad \sigma_{\min} = \min_{x\neq 0} \frac{\|Ax\|}{\|x\|}, \tag{4.9}$$

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 Section 4.5, 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 the norm of the matrix. The p-norm of a matrix is defined as [29]

$$\|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \tag{4.10}$$

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 (4.10) for $p = 2$ with (4.9) reveals that

$$\sigma_{\max}(A) = \|A\|, \tag{4.11}$$

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

There is an inequality that holds between the matrix p-norms [29]:

$$\|A\| \leqslant \sqrt{\|A\|_1 \|A\|_\infty}. \tag{4.12}$$

Because the ∞-norm satisfies $\|A\|_\infty = \|A^{\mathrm{T}}\|_1$, (4.12) implies that

$$\sigma_{\max}(A) \leqslant \|A\|_1 \quad \text{for symmetric } A. \tag{4.13}$$

The right-hand side of (4.13) is easily evaluated, because the 1-norm reduces to

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}| = (\text{the maximum absolute column sum}), \tag{4.14}$$

where a_{ij} is the (i,j) 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 coefficient matrix that describes the EM system. For example, "the maximum singular value of a homogeneous vacuum" means "the maximum singular value of the coefficient matrix describing a homogeneous vacuum."

4.2. Maximum singular values of homogeneous media

In this section, we investigate 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

$$s_x = 1 - is_x'' \quad \text{with } s_x'' \gg 1, \tag{4.15}$$

where the assumption $s_x'' \gg 1$ is due to the discussion following (2.10). Eq. (4.15) implies that

$$s_x \simeq -is_x'' \text{ and } |s_x| \simeq s_x'' \gg 1. \tag{4.16}$$

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

Because the homogeneous EM system is spatially unbounded, discretizing the governing differential equation results in the coefficient matrix of an infinite size. To avoid dealing with an infinitely large matrix, we first examine the maximum singular values of the original differential operators used in (2.1), (2.3), and (2.5); we take the effect of finite-difference discretization into account later. The differential operators for the homogeneous regular medium, UPML, and SC-PML are

$$T_0^{\mathrm{r}}(\mathbf{E}) = \nabla \times \mu^{-1} \nabla \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E}, \tag{4.17a}$$

$$T_0^{\mathrm{u}}(\mathbf{E}) = \nabla \times (\bar{\bar{\mu}}_s)^{-1} \nabla \times \mathbf{E} - \omega^2 \bar{\bar{\varepsilon}}_s \mathbf{E}, \tag{4.17b}$$

$$T_0^{\mathrm{sc}}(\mathbf{E}) = \nabla_s \times \mu^{-1} \nabla_s \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E}, \tag{4.17c}$$

respectively. Below, we refer to them as T when we discuss properties that are common to all three operators.

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 [33,34]

$$\mathbf{E}_{\mathbf{k}}(\mathbf{r}) = \mathbf{F}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{4.18}$$

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

By applying T_0^{r}, T_0^{u}, and T_0^{sc} to $\mathbf{E}_{\mathbf{k}}$, we obtain

$$T_0^{\mathrm{r}}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k}\times\mu^{-1}\mathbf{k}\times\mathbf{E}_{\mathbf{k}} - \omega^2\varepsilon\mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{r}_0}\mathbf{E}_{\mathbf{k}}, \tag{4.19a}$$

$$T_0^{\mathrm{u}}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k}\times(\bar{\bar{\mu}}_s)^{-1}\mathbf{k}\times\mathbf{E}_{\mathbf{k}} - \omega^2\bar{\bar{\varepsilon}}_s\mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{u}_0}\mathbf{E}_{\mathbf{k}}, \tag{4.19b}$$

$$T_0^{\mathrm{sc}}(\mathbf{E}_{\mathbf{k}}) = -\mathbf{k}_s\times\mu^{-1}\mathbf{k}_s\times\mathbf{E}_{\mathbf{k}} - \omega^2\varepsilon\mathbf{E}_{\mathbf{k}} \equiv T_{\mathbf{k}}^{\mathrm{sc}_0}\mathbf{E}_{\mathbf{k}}, \tag{4.19c}$$

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 $[E_{\mathbf{k},x}\ \ E_{\mathbf{k},y}\ \ E_{\mathbf{k},z}]^{\mathrm{T}}$. 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,

$$T_{\mathbf{k}}^{\mathrm{r}_0} = \begin{bmatrix} \frac{k_y^2}{\mu} - \omega^2\varepsilon & -\frac{k_xk_y}{\mu} & 0 \\ -\frac{k_xk_y}{\mu} & \frac{k_x^2}{\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \end{bmatrix}, \tag{4.20a}$$

$$T_{\mathbf{k}}^{\mathrm{u}_0} = \begin{bmatrix} \frac{k_y^2}{s_x\mu} - \frac{\omega^2\varepsilon}{s_x} & -\frac{k_xk_y}{s_x\mu} & 0 \\ -\frac{k_xk_y}{s_x\mu} & \frac{k_x^2}{s_x\mu} - s_x\omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{s_x\mu} + \frac{s_xk_y^2}{\mu} - s_x\omega^2\varepsilon \end{bmatrix}, \tag{4.20b}$$

$$T_{\mathbf{k}}^{\mathrm{sc}_0} = \begin{bmatrix} \frac{k_y^2}{\mu} - \omega^2\varepsilon & -\frac{k_xk_y}{s_x\mu} & 0 \\ -\frac{k_xk_y}{s_x\mu} & \frac{k_x^2}{s_x^2\mu} - \omega^2\varepsilon & 0 \\ 0 & 0 & \frac{k_x^2}{s_x^2\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon \end{bmatrix}. \tag{4.20c}$$

Note that (4.20) are the $\mathbf{k}$-space representations of T_0^{r}, T_0^{u}, and T_0^{sc}. Below, we refer to them as $T_{\mathbf{k}}$ when we discuss properties that are common to all three matrices.

By solving (4.8) with $A = T_{\mathbf{k}}$, we easily obtain one singular value $\sigma_{\mathbf{k},3}$ of $T_{\mathbf{k}}$ corresponding to a singular vector $[0\,0\,1]^{\mathrm{T}}$:

$$\sigma_{\mathbf{k},3}^{\mathrm{r}_0} = \left|\frac{k_x^2}{\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon\right|,\quad \sigma_{\mathbf{k},3}^{\mathrm{u}_0} = \left|\frac{k_x^2}{s_x\mu} + \frac{s_xk_y^2}{\mu} - s_x\omega^2\varepsilon\right|,\quad \sigma_{\mathbf{k},3}^{\mathrm{sc}_0} = \left|\frac{k_x^2}{s_x^2\mu} + \frac{k_y^2}{\mu} - \omega^2\varepsilon\right|. \tag{4.21}$$

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

By the definition of the maximum singular value (4.5), $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$ have the corresponding quantities in (4.21) as their lower bounds. To find the maximum lower bounds, we maximize the right-hand sides of (4.21).

For a continuous medium, k_x and k_y are unbounded, and so are the maximum singular values according to (4.21). In a finite-difference grid with uniform edge length Δ, however, the maximum wavenumber in each Cartesian direction is [19,34]

$$k_{\max} = \frac{\pi}{\Delta}. \tag{4.22}$$

Furthermore, when $k_{\max}$ is used to maximize the right-hand sides of (4.21), it turns out that we can ignore ω^2 terms because Δ is typically far smaller than the wavelength in the PML regions. As a result,

$$\sigma_{\max}^{\mathrm{r}_0} \gtrsim \frac{2k_{\max}^2}{\mu},\quad \sigma_{\max}^{\mathrm{u}_0} \gtrsim \frac{|s_x|k_{\max}^2}{\mu},\quad \sigma_{\max}^{\mathrm{sc}_0} \gtrsim \frac{k_{\max}^2}{\mu}. \tag{4.23}$$

Next, we derive upper bounds of $\sigma_{\max}^{\mathrm{r}_0}$, $\sigma_{\max}^{\mathrm{u}_0}$, and $\sigma_{\max}^{\mathrm{sc}_0}$. The inequality (4.13) dictates that

$$\sigma_{\max}(T) = \max_{\mathbf{k}} \sigma_{\max}(T_{\mathbf{k}}) \leqslant \max_{\mathbf{k}} \|T_{\mathbf{k}}\|_1. \tag{4.24}$$

Calculating $\|T_{\mathbf{k}}\|_1$ according to (4.14), we have

$$\sigma_{\max}^{\mathrm{r}_0} \leqslant \frac{2k_{\max}^2}{\mu} + \omega^2|\varepsilon|,\quad \sigma_{\max}^{\mathrm{u}_0} \leqslant \frac{k_{\max}^2}{|s_x|\mu} + \frac{|s_x|k_{\max}^2}{\mu} + |s_x|\omega^2|\varepsilon|,\quad \sigma_{\max}^{\mathrm{sc}_0} \leqslant \frac{k_{\max}^2}{|s_x|^2\mu} + \frac{k_{\max}^2}{\mu} + \omega^2|\varepsilon|. \tag{4.25}$$

Using (4.16) and ignoring the ω^2 terms again, we obtain

$$\sigma_{\max}^{\mathrm{r}_0} \lesssim \frac{2k_{\max}^2}{\mu},\quad \sigma_{\max}^{\mathrm{u}_0} \lesssim \frac{|s_x|k_{\max}^2}{\mu},\quad \sigma_{\max}^{\mathrm{sc}_0} \lesssim \frac{k_{\max}^2}{\mu}. \tag{4.26}$$

Because the approximate lower and upper bounds indicated in (4.23) and (4.26) are the same for each of $\sigma^{r_0}_{max}$, $\sigma^{u_0}_{max}$, and $\sigma^{sc_0}_{max}$, we have

$$\sigma^{r_0}_{max} \simeq \frac{2k^2_{max}}{\mu}, \quad \sigma^{u_0}_{max} \simeq \frac{|s_x|k^2_{max}}{\mu}, \quad \sigma^{sc_0}_{max} \simeq \frac{k^2_{max}}{\mu}, \tag{4.27}$$

and therefore

$$\sigma^{u_0}_{max} \simeq \frac{|s_x|}{2}\sigma^{r_0}_{max} \quad \text{and} \quad \sigma^{sc_0}_{max} \simeq \frac{1}{2}\sigma^{r_0}_{max}. \tag{4.28}$$

The result indicates a large contrast between the maximum singular values of the homogeneous UPML and SC-PML; $\sigma^{u_0}_{max}$ is much larger than $\sigma^{r_0}_{max}$, whereas $\sigma^{sc_0}_{max}$ is smaller than $\sigma^{r_0}_{max}$.

We note that each estimate in (4.27) is realized by the corresponding $\sigma_{\mathbf{k},3}$ in (4.21) with appropriate $\mathbf{k}$; the estimate of $\sigma^{r_0}_{max}$ is achieved for $\mathbf{k}$ such that $|k_x| = |k_y| = k_{max}$, and the estimates of $\sigma^{u_0}_{max}$ and $\sigma^{sc_0}_{max}$ are achieved for $\mathbf{k}$ such that $k_x = 0$ and $k_y = \pm k_{max}$. Therefore, $\mathbf{k} = \pm[\hat{\mathbf{x}}k_{max} \pm \hat{\mathbf{y}}k_{max}]$ is an approximate wavevector of the maximum right singular vector corresponding to $\sigma^{r_0}_{max}$, and $\mathbf{k} = \pm\hat{\mathbf{y}}k_{max}$ is an approximate wavevector of the maximum right singular vectors corresponding to $\sigma^{u_0}_{max}$ and $\sigma^{sc_0}_{max}$.

So far, when deriving the estimates of $\sigma^{r_0}_{max}$, $\sigma^{u_0}_{max}$, and $\sigma^{sc_0}_{max}$, we have incorporated the effect of the finite-difference grid by simply imposing the upper bound k_{max} on wavevectors. By considering the finite-difference approximations of T^r_0, T^u_0, T^{sc}_0 in (4.17), we can obtain the following exact estimates:

$$\sigma^{r_0}_{max} \simeq \frac{2(2/\Delta)^2}{\mu}, \quad \sigma^{u_0}_{max} \simeq \frac{|s_x|(2/\Delta)^2}{\mu}, \quad \sigma^{sc_0}_{max} \simeq \frac{(2/\Delta)^2}{\mu}. \tag{4.29}$$

We note that the exact results in (4.29) differ from the approximate results in (4.27) 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 (4.28) of this section, which is obtained from the approximate results, turns out to hold for the exact results (4.29) as well.

4.3. Minimum singular values of homogeneous media

In this section, we investigate the minimum singular values of a homogeneous regular medium, homogeneous UPML, and homogeneous SC-PML denoted by $\sigma^{u_0}_{min}$, $\sigma^{sc_0}_{min}$, and $\sigma^{r_0}_{min}$, 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 Section 4.2, we assume that the media have no gain, i.e., $\varepsilon = \varepsilon' - i\varepsilon''$ satisfies $\varepsilon'' \geqslant 0$.

As in the previous section, here we also use the $\mathbf{k}$-space representations $T^{r_0}_{\mathbf{k}}$, $T^{u_0}_{\mathbf{k}}$, and $T^{sc_0}_{\mathbf{k}}$ of (4.20). We find $\sigma^{r_0}_{min}$, $\sigma^{u_0}_{min}$, and $\sigma^{sc_0}_{min}$ as the minima of $\sigma_{min}(T^{r_0}_{\mathbf{k}})$, $\sigma_{min}(T^{u_0}_{\mathbf{k}})$, and $\sigma_{min}(T^{sc_0}_{\mathbf{k}})$ over $\mathbf{k}$, respectively.

First, we derive the conditions for $T^{r_0}_{\mathbf{k}}$, $T^{u_0}_{\mathbf{k}}$, and $T^{sc_0}_{\mathbf{k}}$ to be singular. $T^{r_0}_{\mathbf{k}}$ is singular when $\det(T^{r_0}_{\mathbf{k}}) = -\omega^2\varepsilon(k^2_x/\mu + k^2_y/\mu - \omega^2\varepsilon)^2 = 0$, or equivalently

$$k^2_x + k^2_y = \omega^2\mu\varepsilon. \tag{4.30}$$

Similarly, $T^{u_0}_{\mathbf{k}}$ and $T^{sc_0}_{\mathbf{k}}$ are singular when

$$\frac{k^2_x}{s^2_x} + k^2_y = \omega^2\mu\varepsilon. \tag{4.31}$$

Now, suppose that ε is positive ($\varepsilon' > 0$, $\varepsilon'' = 0$). We see that (4.30) is satisfied by infinitely many real $\mathbf{k}$ lying on a circle in the $\mathbf{k}$-space, and (4.31) is satisfied by only two real $\mathbf{k}$, i.e., $\mathbf{k} = \pm\hat{\mathbf{y}}\omega\sqrt{\mu\varepsilon}$, because s^2_x has a nonzero imaginary part. Since a singular matrix has 0 as a singular value as pointed out in Section 4.1, each of $\sigma_{min}(T^{r_0}_{\mathbf{k}})$, $\sigma_{min}(T^{u_0}_{\mathbf{k}})$, and $\sigma_{min}(T^{sc_0}_{\mathbf{k}})$ is zero for some real $\mathbf{k}$, which implies that

$$\sigma^{u_0}_{min} = \sigma^{sc_0}_{min} = \sigma^{r_0}_{min} = 0 \quad \text{for positive } \varepsilon. \tag{4.32}$$

On the other hand, in cases where ε is either negative ($\varepsilon' < 0$, $\varepsilon'' = 0$) or complex ($\varepsilon'' > 0$), $T^{r_0}_{\mathbf{k}}$, $T^{u_0}_{\mathbf{k}}$, and $T^{sc_0}_{\mathbf{k}}$ are nonsingular for all real $\mathbf{k}$, because no real $\mathbf{k}$ satisfies (4.30) or (4.31). Therefore, we have

$$\sigma^{u_0}_{min}, \sigma^{sc_0}_{min}, \sigma^{r_0}_{min} > 0 \quad \text{for negative or complex } \varepsilon. \tag{4.33}$$

From (4.32) and (4.33), we conclude that the minimum singular values of the homogeneous media with positive ε (e.g., dielectrics and PMLs matching dielectrics) are always less than the minimum singular values of the homogeneous media with other ε satisfying $\varepsilon'' \geqslant 0$ (e.g., metals and PMLs matching metals).

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

In Section 4.3, we have shown that the minimum singular values of the 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 the 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 to the quantized values in the sets

$$K_x = \left\{ \frac{2\pi n_x}{L_x} : n_x \in \mathbb{Z}^+ \right\} \quad \text{and} \quad K_y = \left\{ \frac{2\pi n_y}{L_y} : n_y \in \mathbb{Z}^+ \right\}, \tag{4.34}$$

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

$$K = \{\hat{\mathbf{x}} k_x + \hat{\mathbf{y}} k_y : k_x \in K_x, k_y \in K_y\}. \tag{4.35}$$

When there is no $\mathbf{k} \in K$ satisfying (4.30) and (4.31), all of $\sigma_{\min}^{r_0}$, $\sigma_{\min}^{u_0}$, and $\sigma_{\min}^{sc_0}$ deviate from 0 for a bounded domain, but by different amounts. Fig. 4.1 shows $\sigma_{\min}(T_{\mathbf{k}}^{r_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ in a portion of the $\mathbf{k}$-space where they are close to zero. It shows that $\sigma_{\min}(T_{\mathbf{k}}^{u_0}) < \sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ for all displayed $\mathbf{k}$ except $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ for which both are zero. Therefore, in general we expect $\min_{\mathbf{k}\in K}\sigma_{\min}(T_{\mathbf{k}}^{u_0}) < \min_{\mathbf{k}\in K}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$, or equivalently $\sigma_{\min}^{u_0} < \sigma_{\min}^{sc_0}$. On the other hand, $\sigma_{\min}(T_{\mathbf{k}}^{r_0})$ can be either above or below each of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ in the figure. Hence, $\sigma_{\min}^{r_0} = \min_{\mathbf{k}\in K}\sigma_{\min}(T_{\mathbf{k}}^{r_0})$ can be either less or greater than each of $\sigma_{\min}^{u_0}$ and $\sigma_{\min}^{sc_0}$, depending on the size of the bounded domain.

We now estimate an upper bound of $\sigma_{\min}^{u_0}/\sigma_{\min}^{sc_0}$ for a bounded domain. For that purpose, we examine the plots of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ in Fig. 4.1 in more detail. Fig. 4.2(a) displays the same $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ shown in Fig. 4.1, but as a contour plot over an extended range of k_x. In Fig. 4.2(a), we notice the following important features of $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$:

First, $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ has a global minimum of zero at $\mathbf{k} = \hat{\mathbf{y}}(\omega/c)$ due to the argument following (4.31); accordingly, the contours in the vicinity of the global minimum point form enclosing curves (cyan contours in Fig. 4.2(a)).

Second, the surface of $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ has a "valley", where $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ is close to zero, along a curve in the k_xk_y-plane. The shape of the curve can be derived from (4.31), which describes the condition for $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ to be singular. Because of (4.16), the condition (4.31) is approximated by

$$-\frac{k_x^2}{s_x''^2} + k_y^2 = \frac{\omega^2}{c^2}. \tag{4.36}$$

Hence, for $\mathbf{k}$ satisfying (4.36), $T_{\mathbf{k}}^{sc_0}$ is *nearly* singular and has a close-to-zero singular value. Eq. (4.36) thus describes the bottom of the valley of the $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ surface. The curve described by (4.36), which is a hyperbola that is indicated by a black dashed line in Fig. 4.2(a), agrees well with the actual location of the bottom of the valley as can be seen from the contour plot.

Third, $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ varies much more slowly in k_x than in k_y; note that the scale of the k_y axis in Fig. 4.2(a) is exaggerated. This can be shown mathematically by examining (4.20c). We notice that interchanging k_x/s_x and k_y only swaps the (1, 1) and

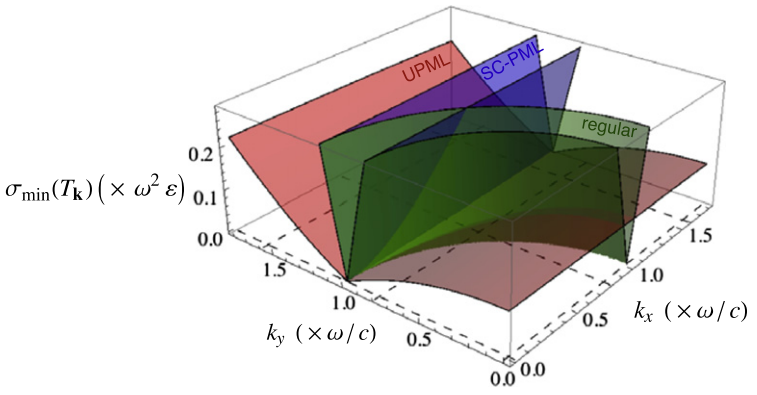


Fig. 4.1. The 3D plot of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$, $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$, and $\sigma_{\min}(T_{\mathbf{k}}^{r_0})$ as functions of k_x and k_y. The three functions are drawn in a portion of the $\mathbf{k}$-space where the functions are close to zeros. The surface of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ is below that of $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ 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 $\sigma_{\min}(T_{\mathbf{k}}^{r_0})$, on the other hand, is neither consistently below or 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. The PML scale factor $s_x = 1 - i10$ is used.

(2, 2) elements of the matrix and does not change the singular values of $T^{\mathrm{sc}_0}_{\mathbf{k}}$. Hence, $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})$ is 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 $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})$. However, $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})$ also exhibits the three features described above.

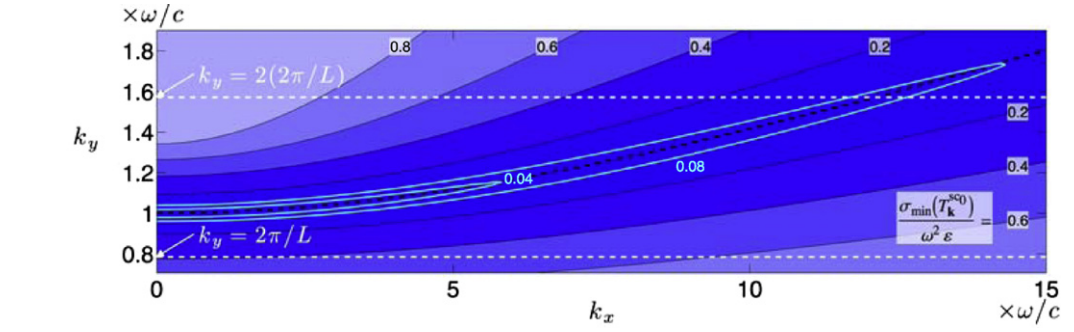
Motivated by the third observation above, we derive an approximate upper bound of $\sigma^{\mathrm{u}_0}_{\min}/\sigma^{\mathrm{sc}_0}_{\min}$. Suppose that $k^{\mathrm{u}}_y \in K_y$ and $k^{\mathrm{sc}}_y \in K_y$ are the y-components of the quantized $\mathbf{k}$'s at which $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})$ and $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})$ are minimized, respectively. Then, from the definitions of $\sigma^{\mathrm{u}_0}_{\min}$ and $\sigma^{\mathrm{sc}_0}_{\min}$ for a bounded domain, we have

$$\frac{\sigma^{\mathrm{u}_0}_{\min}}{\sigma^{\mathrm{sc}_0}_{\min}} = \frac{\min_{k_x\in K_x}\min_{k_y\in K_y}\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})}{\min_{k_x\in K_x}\min_{k_y\in K_y}\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})} = \frac{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})_{k_y=k^{\mathrm{u}}_y}}{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})_{k_y=k^{\mathrm{sc}}_y}} \leqslant \frac{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})_{k_y=k^{\mathrm{sc}}_y}}{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})_{k_y=k^{\mathrm{sc}}_y}}$$
$$\leqslant \max_{k_y\in K_y}\left\{\frac{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})}{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})}\right\}. \tag{4.37}$$

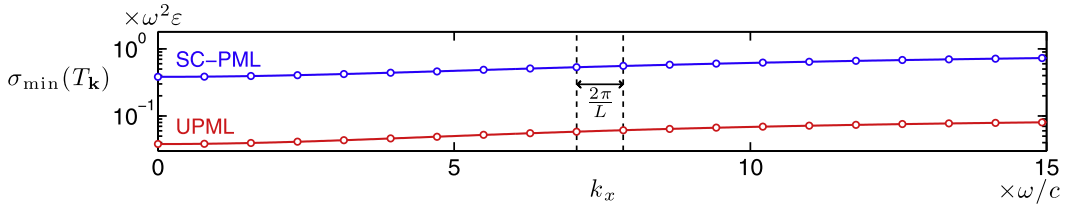
Therefore, to estimate an upper bound of $\sigma^{\mathrm{u}_0}_{\min}/\sigma^{\mathrm{sc}_0}_{\min}$, we estimate

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})}{\min_{k_x\in K_x}\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})} \tag{4.38}$$

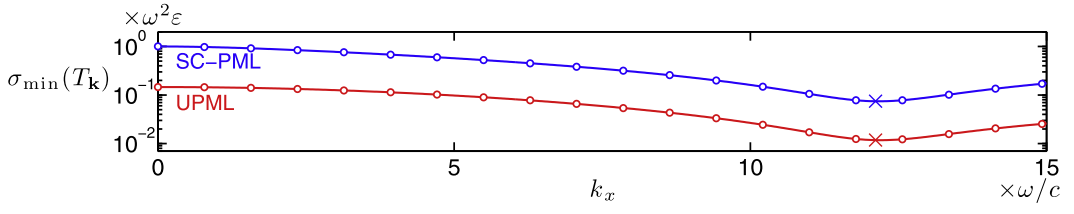
for all k_y. Because $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})$ and $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})$ are slowly varying functions of k_x, we use the approximation



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



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



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

Fig. 4.2. (a) The 2D contour plot of $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})$. The values of $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})/\omega^2\varepsilon$ 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 (4.36), 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, $\sigma_{\min}(T^{\mathrm{u}_0}_{\mathbf{k}})$ and $\sigma_{\min}(T^{\mathrm{sc}_0}_{\mathbf{k}})$ 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. 4.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{u_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})}\simeq\frac{\min_{k_x\geqslant 0}\sigma_{\min}(T_{\mathbf{k}}^{u_0})}{\min_{k_x\geqslant 0}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})} \tag{4.39}$$

to estimate (4.38).

We estimate the right-hand side of (4.39) for $k_y < \omega/c$ first. To visualize the general behaviors of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ for such k_y, in Fig. 4.2(b) we plot them along the lower white dashed line of Fig. 4.2(a). Fig. 4.2(b) indicates that $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ are minimized at $k_x = 0$ for $k_y < \omega/c$. In Appendix A we show that in the limit of $s_x'' \gg 1$, which is the numerically relevant situation, $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ are indeed minimized at $k_x = 0$ for all $k_y < \omega/c$. Therefore, we have

$$\min_{k_x\geqslant 0}\sigma_{\min}(T_{\mathbf{k}})\simeq\sigma_{\min}(T_{\mathbf{k}})_{k_x=0}\quad\text{for } T_{\mathbf{k}}=T_{\mathbf{k}}^{u_0},T_{\mathbf{k}}^{sc_0}\quad\text{for } k_y<\frac{\omega}{c}. \tag{4.40}$$

Since $T_{\mathbf{k}}^{u_0}$ and $T_{\mathbf{k}}^{sc_0}$ of (4.20) are diagonalized for $k_x = 0$, the right-hand side of (4.40) is easily calculated as

$$\sigma_{\min}(T_{\mathbf{k}}^{u_0})_{k_x=0}=\frac{1}{\mu|s_x|}\left(\frac{\omega^2}{c^2}-k_y^2\right)\quad\text{and}\quad\sigma_{\min}(T_{\mathbf{k}}^{sc_0})_{k_x=0}=\frac{1}{\mu}\left(\frac{\omega^2}{c^2}-k_y^2\right). \tag{4.41}$$

Combining (4.41) with (4.39) and (4.40), we obtain

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{u_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})}\simeq\frac{1}{|s_x|}\quad\text{for } k_y<\frac{\omega}{c}. \tag{4.42}$$

Next, we consider $k_y > \omega/c$. Such k_y is indicated by the upper white dashed line in Fig. 4.2(a), along which $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ are plotted in Fig. 4.2(c). As seen in Fig. 4.2(c), at such a given k_y the minima of $\sigma_{\min}(T_{\mathbf{k}}^{u_0})$ and $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$ 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 (4.36)); this is shown more rigorously in Appendix A for $s_x'' \gg 1$. Therefore, we have

$$\min_{k_x\geqslant 0}\sigma_{\min}(T_{\mathbf{k}})\simeq\sigma_{\min}(T_{\mathbf{k}})_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}}\quad\text{for } T_{\mathbf{k}}=T_{\mathbf{k}}^{u_0},T_{\mathbf{k}}^{sc_0}\quad\text{for } k_y>\frac{\omega}{c}. \tag{4.43}$$

By evaluating the right-hand side of (4.43) approximately, in Appendix C we show that

$$\sigma_{\min}(T_{\mathbf{k}}^{u_0})_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}}\simeq 2\omega^2\varepsilon\frac{k_y^2-\omega^2/c^2}{(s_x''^2+1)k_y^2-\omega^2/c^2}, \tag{4.44a}$$

$$\sigma_{\min}(T_{\mathbf{k}}^{sc_0})_{k_x=s_x''\sqrt{k_y^2-\frac{\omega^2}{c^2}}}\simeq\frac{2}{s_x''}\omega^2\varepsilon\frac{k_y^2-\omega^2/c^2}{2k_y^2-\omega^2/c^2}. \tag{4.44b}$$

The two "x" marks drawn at $k_x = s_x''\sqrt{k_y^2 - \frac{\omega^2}{c^2}}$ in Fig. 4.2(c) indicate the values determined by (4.44). The good agreement of the marks with the actual minima in the figure validates (4.44).

Combining (4.44) with (4.39) and (4.43), we obtain

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{u_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})}\simeq\frac{2s_x''k_y^2-s_x''\omega^2/c^2}{(s_x''^2+1)k_y^2-\omega^2/c^2}\quad\text{for } k_y>\frac{\omega}{c}. \tag{4.45}$$

For $k_y > \omega/c$, the right-hand side of (4.45) is an increasing function of k_y^2, so its maximum is attained at $k_y = \infty$. Hence, $\sigma_{\min}^{u_0}/\sigma_{\min}^{sc_0}$ is bounded from above as

$$\frac{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{u_0})}{\min_{k_x\in K_x}\sigma_{\min}(T_{\mathbf{k}}^{sc_0})}\lesssim\frac{2s_x''}{s_x''^2+1}\quad\text{for } k_y>\frac{\omega}{c}. \tag{4.46}$$

Combining (4.42) and (4.46) with (4.37), we conclude that $\sigma_{\min}^{u_0}/\sigma_{\min}^{sc_0}$ is approximately bounded from above as

$$\frac{\sigma_{\min}^{u_0}}{\sigma_{\min}^{sc_0}}\lesssim\max\left\{\frac{1}{|s_x|},\frac{2s_x''}{s_x''^2+1}\right\}\simeq\frac{2}{|s_x|}, \tag{4.47}$$

where (4.16) is used in the last approximation. The inequality (4.47) implies that $\sigma_{\min}^{u_0}$ is much smaller than $\sigma_{\min}^{sc_0}$ for a bounded domain.

In summary, the minimum singular values of the homogeneous regular medium, UPML, and SC-PML for positive ε are all zero as shown in (4.32), but for a bounded domain they deviate from 0. When such deviation occurs, $\sigma_{\min}^{u_0}$ is much smaller than $\sigma_{\min}^{sc_0}$ as (4.47) describes, but $\sigma_{\min}^{r_0}$ can be either less or greater than each of $\sigma_{\min}^{u_0}$ and $\sigma_{\min}^{sc_0}$.

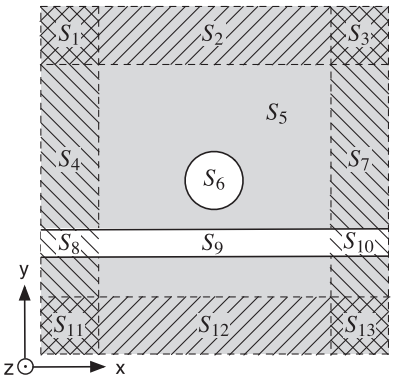


Fig. 4.3. An example of an inhomogeneous EM system. The hypothetical system has a dielectric cavity (S_6) side-coupled to a dielectric waveguide (S_9) immersed in a background metal (S_5). The system is composed of several subdomains S_i, each of which is filled with a homogeneous medium. We define S_i as a domain excluding its boundary.

4.5. Variational method to estimate the extreme singular values and condition numbers 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. An example of such EM systems is illustrated in Fig. 4.3. Because the EM system consists of regular media and PML, we refer to it as an *inhomogeneous* EM system to distinguish 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 (4.9), and express them in terms of the extreme singular values of the homogeneous media examined in Sections 4.2, 4.3, and 4.4. Using the estimates, we show that

$$\frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{sc}}_{\max}} \gg 1, \tag{4.48}$$

$$\frac{\sigma^{\mathrm{u}}_{\min}}{\sigma^{\mathrm{sc}}_{\min}} \lesssim 1, \tag{4.49}$$

and therefore

$$\frac{\kappa^{\mathrm{u}}}{\kappa^{\mathrm{sc}}} = \frac{\sigma^{\mathrm{u}}_{\max}}{\sigma^{\mathrm{sc}}_{\max}} \frac{\sigma^{\mathrm{sc}}_{\min}}{\sigma^{\mathrm{u}}_{\min}} \gg 1. \tag{4.50}$$

The inequality (4.50) indicates that A^{u} is much worse-conditioned than A^{sc}.

As inferred from the discussion following (4.9), 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 $v^{\mathrm{u}}_{\max}$, $v^{\mathrm{u}}_{\min}$ and $v^{\mathrm{sc}}_{\max}$, $v^{\mathrm{sc}}_{\min}$ 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. 4.3. 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.,

$$s_w(l) = s_0 = 1 - is_0'' \quad \text{and} \quad s_0'' \gg 1. \tag{4.51}$$

First, we estimate the maximum singular value of an inhomogeneous EM system. From (4.9), 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:

$$\sigma_{\max}|_{S_i} = \max_x r(x|_{S_i}), \tag{4.52}$$

where $x|_{S_i}$ is a column vector that has the same elements as x inside S_i and zeros outside. Then, by the definition of $\sigma_{\max}$ we have

$$\sigma_{\max} \geqslant \max_i \sigma_{\max}|_{S_i}. \tag{4.53}$$

In addition, we have[4]

[4] Here we use four equalities $\|x\|^2 \simeq \|\sum_i x|_{S_i}\|^2 = \sum_i \|x|_{S_i}\|^2$ and $\|Ax\|^2 \simeq \|\sum_i Ax|_{S_i}\|^2 \simeq \sum_i \|Ax|_{S_i}\|^2$. Out of the four equalities, only $\|\sum_i x|_{S_i}\|^2 = \sum_i \|x|_{S_i}\|^2$ is exact because the elements of $x|_{S_i}$ are zeros at the boundary of S_i by definition (See the caption of Fig. 4.3) 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.

$$\sigma_{\max}^2 = \max_x \frac{\|Ax\|^2}{\|x\|^2} \simeq \max_x \frac{\|\sum_i Ax|_{S_i}\|^2}{\|\sum_i x|_{S_i}\|^2} \simeq \max_x \frac{\sum_i \|Ax|_{S_i}\|^2}{\sum_i \|x|_{S_i}\|^2} = \max_x \left(\sum_i \rho_i(x) r(x|_{S_i})^2 \right)$$
$$\leqslant \max_x \left(\sum_i \rho_i(x) (\sigma_{\max}|_{S_i})^2 \right), \tag{4.54}$$

where

$$\rho_i(x) = \frac{\|x|_{S_i}\|^2}{\sum_j \|x|_{S_j}\|^2}. \tag{4.55}$$

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 (4.54) leads to

$$\sigma_{\max}^2 \lesssim \max_x \left(\max_i (\sigma_{\max}|_{S_i})^2 \right) = \max_i \left(\sigma_{\max}|_{S_i} \right)^2. \tag{4.56}$$

The two inequalities (4.53) and (4.56) dictate

$$\sigma_{\max} \simeq \max_i \sigma_{\max}|_{S_i}. \tag{4.57}$$

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

$$\sigma_{\max}|_{S_i} \simeq \begin{cases} \sigma_{\max}^{r_0} & \text{for } S_i \text{ outside the PML region,} \\ \sigma_{\max}^{u_0} & \text{for } S_i \text{ inside the UPML region,} \\ \sigma_{\max}^{sc_0} & \text{for } S_i \text{ inside the SC-PML region.} \end{cases} \tag{4.58}$$

Here, we ignore S_i's with overlapping PMLs (e.g., the four corners in Fig. 4.3), simply because they typically do not interact with incident waves strongly; we will see in Section 4.6 that this assumption is consistent with direct numerical calculations. Note that $\sigma_{\max}|_{S_i}$'s in (4.58) are independent of ε, because $\sigma_{\max}^{r_0}$, $\sigma_{\max}^{u_0}$, and $\sigma_{\max}^{sc_0}$ do not depend on ε as shown in (4.29).

We apply (4.58) to (4.57) for $A = A^u$ and $A = A^{sc}$ separately to estimate $\sigma_{\max}^u$ and $\sigma_{\max}^{sc}$. The inhomogeneous EM system consists of regular media and UPML for $A = A^u$, and of regular media and SC-PML for $A = A^{sc}$. Therefore, we have

$$\sigma_{\max}^u \simeq \max\{\sigma_{\max}^{r_0}, \sigma_{\max}^{u_0}\} = \sigma_{\max}^{u_0}, \tag{4.59}$$
$$\sigma_{\max}^{sc} \simeq \max\{\sigma_{\max}^{r_0}, \sigma_{\max}^{sc_0}\} = \sigma_{\max}^{r_0}, \tag{4.60}$$

where the magnitudes of $\sigma_{\max}^{r_0}$, $\sigma_{\max}^{u_0}$, and $\sigma_{\max}^{sc_0}$ are compared using (4.28). Eqs. (4.59) and (4.60) imply that $v_{\max}^u$ and $v_{\max}^{sc}$ tend to be concentrated in the UPML region and the region of regular media, respectively.

From (4.59), (4.60), and (4.28), we obtain

$$\frac{\sigma_{\max}^u}{\sigma_{\max}^{sc}} \simeq \frac{\sigma_{\max}^{u_0}}{\sigma_{\max}^{r_0}} \simeq \frac{|s_0|}{2}, \tag{4.61}$$

which proves (4.48).

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 (4.53)–(4.56) except that now we minimize instead of maximize, we obtain

$$\sigma_{\min} \simeq \min_i \sigma_{\min}|_{S_i}, \tag{4.62}$$

which is a result parallel to (4.57). 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. 4.3) 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 examples in Section 3, the purpose of using PML is to simulate situations where there are waves

[5] For $\sigma_{\max}|_{S_i}$ to be approximated well by one of $\sigma_{\max}^{r_0}, \sigma_{\max}^{u_0}$, and $\sigma_{\max}^{sc_0}$, the subdomain S_i needs to be sufficiently large, because each homogeneous medium studied in Section 4.2 is assumed to fill an infinite space. However, as described in the discussion following (4.27), the maximum right singular vectors $\mathbf{E}_\mathbf{k}$ of the three homogeneous media in Section 4.2 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 (4.58).

propagating out of the simulation domain; such outgoing waves are supported only in the presence of a dielectric matched by PML.

With this additional assumption, when looking for the smallest of $\sigma_{\min}|_{S_i}$'s in (4.62), 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 Section 4.3. Then, in (4.62) 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

$$\sigma_{\min}|_{S_i} \simeq \begin{cases} \sigma^{r_0}_{\min}|_{D_j} & \text{for } S_i = D_j, \\ \sigma^{u_0}_{\min}|_{P_k} & \text{for } S_i = P_k \text{ inside the UPML region}, \\ \sigma^{sc_0}_{\min}|_{P_k} & \text{for } S_i = P_k \text{ inside the SC-PML region}, \end{cases} \tag{4.63}$$

where $\sigma^{r_0}_{\min}|_{D_j}$, $\sigma^{u_0}_{\min}|_{P_k}$, and $\sigma^{sc_0}_{\min}|_{P_k}$ are the minimum singular values of the three homogeneous media in a *bounded* domain examined in Section 4.4; the bounded domain in this case is either P_k or D_j.

We apply (4.63) to (4.62) for $A = A^{u}$ and $A = A^{sc}$ separately to estimate $\sigma^{u}_{\min}$ and $\sigma^{sc}_{\min}$. The inhomogeneous EM system consists of regular media and UPML for $A = A^{u}$, and of regular media and SC-PML for $A = A^{sc}$. Therefore, we have

$$\sigma^{u}_{\min} \simeq \min\left\{ \min_j \sigma^{r_0}_{\min}|_{D_j}, \min_k \sigma^{u_0}_{\min}|_{P_k} \right\} = \min\left\{ \sigma^{r_0}_{\min}|_{D}, \sigma^{u_0}_{\min}|_{P} \right\}, \tag{4.64}$$

$$\sigma^{sc}_{\min} \simeq \min\left\{ \min_j \sigma^{r_0}_{\min}|_{D_j}, \min_k \sigma^{sc_0}_{\min}|_{P_k} \right\} = \min\left\{ \sigma^{r_0}_{\min}|_{D}, \sigma^{sc_0}_{\min}|_{P'} \right\}, \tag{4.65}$$

where D, P, and P' are the subdomains that minimize $\sigma^{r_0}_{\min}|_{D_j}$, $\sigma^{u_0}_{\min}|_{P_k}$, and $\sigma^{sc_0}_{\min}|_{P_k}$, respectively. Eqs. (4.64) and (4.65) imply that both $v^{u}_{\min}$ and $v^{sc}_{\min}$ 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 $\sigma^{r_0}_{\min}|_{D}$ relative to $\sigma^{u_0}_{\min}|_{P}$ and $\sigma^{sc_0}_{\min}|_{P'}$.

For the same subdomain P_k, $(\sigma^{u_0}_{\min}|_{P_k})/(\sigma^{sc_0}_{\min}|_{P_k}) \ll 1$ according to (4.47). Hence, we have

$$\frac{\sigma^{u_0}_{\min}|_{P}}{\sigma^{sc_0}_{\min}|_{P'}} \leqslant \frac{\sigma^{u_0}_{\min}|_{P'}}{\sigma^{sc_0}_{\min}|_{P'}} \ll 1, \tag{4.66}$$

which results in

$$\frac{\sigma^{u}_{\min}}{\sigma^{sc}_{\min}} \simeq \frac{\min\left\{\sigma^{r_0}_{\min}|_{D}, \sigma^{u_0}_{\min}|_{P}\right\}}{\min\left\{\sigma^{r_0}_{\min}|_{D}, \sigma^{sc_0}_{\min}|_{P'}\right\}} \leqslant \frac{\min\left\{\sigma^{r_0}_{\min}|_{D}, \sigma^{sc_0}_{\min}|_{P'}\right\}}{\min\left\{\sigma^{r_0}_{\min}|_{D}, \sigma^{sc_0}_{\min}|_{P'}\right\}} = 1. \tag{4.67}$$

The inequality (4.67) directly leads to (4.49).

From (4.61) and (4.67), we conclude that

$$\frac{\kappa^{u}}{\kappa^{sc}} = \frac{\sigma^{u}_{\max}}{\sigma^{sc}_{\max}} \frac{\sigma^{sc}_{\min}}{\sigma^{u}_{\min}} \gtrsim \frac{|s_0|}{2}. \tag{4.68}$$

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, (4.67) does not necessarily mean that $\sigma^{u}_{\min}/\sigma^{sc}_{\min}$ is close to 1. For example, consider a case where $\sigma^{r_0}_{\min}|_{D}$ is greater than both $\sigma^{u_0}_{\min}|_{P}$ and $\sigma^{sc_0}_{\min}|_{P'}$ in (4.64) and (4.65). Such a case leads to

$$\frac{\sigma^{u}_{\min}}{\sigma^{sc}_{\min}} \simeq \frac{\sigma^{u_0}_{\min}|_{P}}{\sigma^{sc_0}_{\min}|_{P'}} \leqslant \frac{\sigma^{u_0}_{\min}|_{P'}}{\sigma^{sc_0}_{\min}|_{P'}} \lesssim \frac{2}{|s_0|}, \tag{4.69}$$

where the last inequality is from (4.47). The inequality (4.69) demonstrates that $\sigma^{u}_{\min}/\sigma^{sc}_{\min}$ can be much smaller than 1 indeed. It further implies that

$$\frac{\kappa^{u}}{\kappa^{sc}} = \frac{\sigma^{u}_{\max}}{\sigma^{sc}_{\max}} \frac{\sigma^{sc}_{\min}}{\sigma^{u}_{\min}} \gtrsim \frac{|s_0|^2}{4}, \tag{4.70}$$

which predicts much larger κ^{u}/κ^{sc} than is expected from (4.68).

Second, as shown in (4.68), κ^{u}/κ^{sc} increases with $|s_0|$. Therefore, in nanophotonics where $|s_0|$ can exceed 1000 as mentioned in Section 2, we expect the ratio between the condition numbers of the UPML and SC-PML matrices to be very large. Especially, when (4.70) holds, κ^{u}/κ^{sc} can be on the order of 10^5.

4.6. Numerical calculation of the extreme singular values and condition numbers of inhomogeneous EM systems

In this section, we numerically validate the analysis in Section 4.5. We consider two EM systems as examples: a vacuum surrounded by PML (Fig. 4.4(a)), and a metal-dielectric-metal (MDM) waveguide bend surrounded by PML (Fig. 4.4(b)). For these two EM systems, we numerically calculate their extreme singular values as well as the corresponding extreme right singular vectors. We compare the behaviors of these quantities to the discussions in the previous sections.

We first examine the system in Fig. 4.4(a). Here, we use a constant PML loss parameter. With $\Delta = 20$ nm, $d = 10\Delta$, $m = 0$, and $R = e^{-16} \simeq 1 \times 10^{-7}$ in (2.8) and (2.9), the PML scale factor of (2.7) is

$$s_w(l) = s_0 = 1 - i9.868 \tag{4.71}$$

in each attenuation direction w.

Table 4.1 compares numerically calculated σ^{u}_{max} and σ^{sc}_{max} with their estimates derived in (4.59) and (4.60). The agreement is very good with errors only about 0.1–0.2%. As a result, $\sigma^{u}_{max}/\sigma^{sc}_{max}$ is also estimated very accurately by $\sigma^{u_0}_{max}/\sigma^{r_0}_{max} \simeq |s_0|/2 = 4.959$, and thus (4.61) is validated.

We visualize numerically calculated v^{u}_{max} and v^{sc}_{max} in Fig. 4.5. Note that the figure plots the real parts of the x-, y-, z-components of v_{max}; because v_{max} is the solution of Maxwell's equations for the 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.

Fig. 4.5 shows that v^{u}_{max} is concentrated in the UPML region, whereas v^{sc}_{max} is concentrated in the vacuum region. This is exactly what we expect from the discussion of (4.59) and (4.60). Moreover, v^{u}_{max} and v^{sc}_{max} are indeed quite similar to the maximum right singular vectors of the homogeneous UPML and regular medium, respectively. Notice that both v^{u}_{max} and v^{sc}_{max} exhibit fast spatial oscillations, but the oscillations have different wavevectors $\mathbf{k}$. For v^{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^{u}_{max} is $\mathbf{k} = \pm\hat{\mathbf{y}}(2\pi/2\Delta)$. On the other hand, the dominant wavevector of v^{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 (4.27).

We now examine the minimum singular values of the same system of Fig. 4.4(a). Table 4.2 displays numerically calculated σ^{u}_{min} and σ^{sc}_{min} as well as the ratio between the two. The ratio is clearly less than 1, validating (4.67). Note that we do not have the estimates of the minimum singular values in the table, because in Section 4.5 we have provided only a general bound of the ratio $\sigma^{u}_{min}/\sigma^{sc}_{min}$, but not detailed estimates of the individual minimum singular values.

Notice that $\sigma^{u}_{min}/\sigma^{sc}_{min}$ in Table 4.2 is in fact close to $2/|s_0| = 0.2016$. This is consistent with v^{u}_{min} and v^{sc}_{min} shown in Fig. 4.6, where we plot the absolute values of the complex elements of each singular vector. We see that v^{u}_{min} is concentrated in the UPML region, and v^{sc}_{min} is concentrated in the SC-PML region. According to the discussion of (4.64) and (4.65), this corresponds to a case where $\sigma^{r_0}_{min}|_D$ is greater than both $\sigma^{u_0}_{min}|_P$ and $\sigma^{sc_0}_{min}|_{P'}$. Then, $\sigma^{u}_{min}/\sigma^{sc}_{min}$ satisfies (4.69) in addition to (4.67), which explains why $\sigma^{u}_{min}/\sigma^{sc}_{min}$ is close to the upper bound $2/|s_0|$ in (4.69). However, we note that v^{u}_{min} and v^{sc}_{min} are not always concentrated in the PML region; for the same system, it is actually possible to change the wavelength or the size of the simulation domain so that they are concentrated in the region of regular media.

Combining the results in Tables 4.1 and 4.2, we obtain $\kappa^{u}/\kappa^{sc} = 23.40 \gg 1$, which is consistent with our conclusion in Section 4.5.

As a second example, we investigate the MDM waveguide bend in Fig. 4.4(b). To be consistent with the typical use of PML in numerical simulations, we use a graded PML loss parameter $\sigma_w(l)$. With $\Delta = 2$ nm, $d = 10\Delta$, $m = 4$, and $R = e^{-16} \simeq 1 \times 10^{-7}$ in (2.8) and (2.9), the PML scale factor of (2.7) is

$$s_w(l) = s_0(l) = 1 - i493.4\left(\frac{l}{d}\right)^4 \tag{4.72}$$

in each attenuation direction w. Note that $|s_w(d)|$, which is the maximum of $|s_w(l)|$, has increased from about 10 in (4.71) to about 500 in (4.72); the significant increase in $|s_w(d)|$ is due to two factors: the use of the graded PML loss parameter, and the

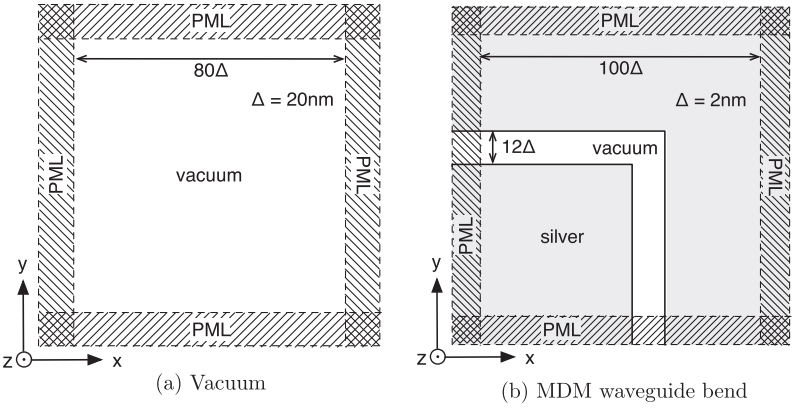
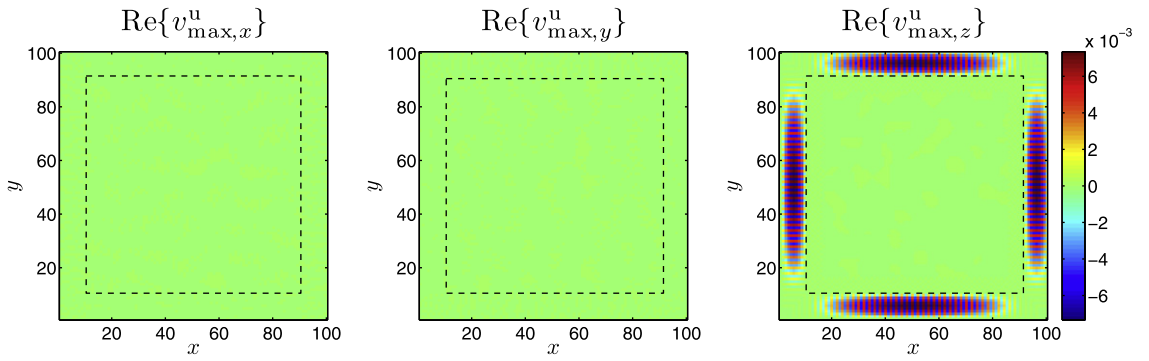


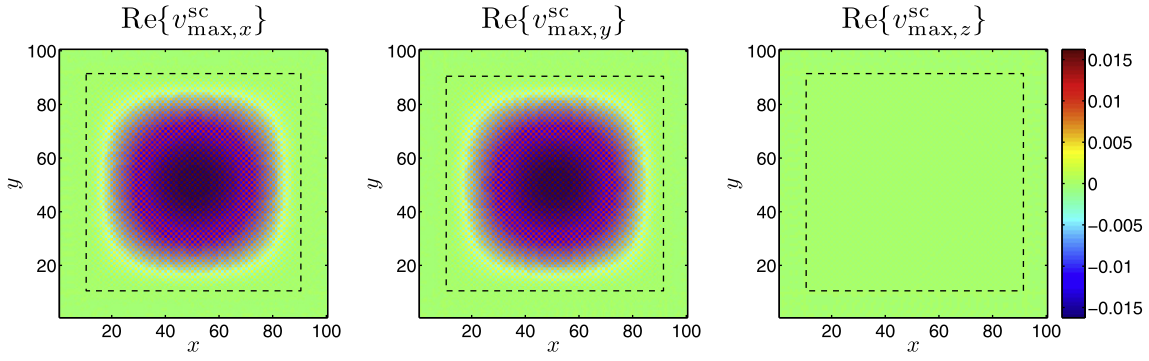
Fig. 4.4. Two inhomogeneous EM systems whose extreme singular values and condition numbers are numerically calculated: (a) a vacuum surrounded by PML, and (b) a metal-dielectric-metal waveguide bend surrounded by PML. The edge lengths Δ of the uniform grids used to discretize Maxwell's equations are 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 $\lambda_0 = 1550$ nm is used. In (b), the electric permittivity of silver [21] at λ_0 is $\varepsilon_{Ag} = (-129 - i3.28)\varepsilon_0$.

Table 4.1
The maximum singular values σ^{u}_{max} and σ^{sc}_{max} of the vacua surrounded by UPML and SC-PML, respectively, along with the ratio $\sigma^{u}_{max}/\sigma^{sc}_{max}$. Notice the excellent agreement between the estimates and numerically calculated values. The numerically calculated maximum singular values are obtained by solving (4.7) so that $\|Av_{max} - \sigma_{max}u_{max}\|/\|u_{max}\| < 10^{-11}$ for $A = A^{u}, A^{sc}$. The estimates of the maximum singular values are evaluated using $\sigma^{u_0}_{max}$ and $\sigma^{r_0}_{max}$ in (4.29) 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.

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



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



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

Fig. 4.5. The maximum right singular vectors (a) v^{u}_{max} of the vacuum surrounded by UPML, and (b) v^{sc}_{max} of the vacuum surrounded by SC-PML. The real parts of the *x*-, *y*-, *z*-components of v^{u}_{max} and v^{sc}_{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^{u}_{max} is concentrated in the UPML region, whereas v^{sc}_{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 points.

reduction of Δ from 20 nm to 2 nm. Therefore, as discussed at the end of Section 4.5, we expect much larger κ^{u}/κ^{sc} for this system than for the first example analyzed above.

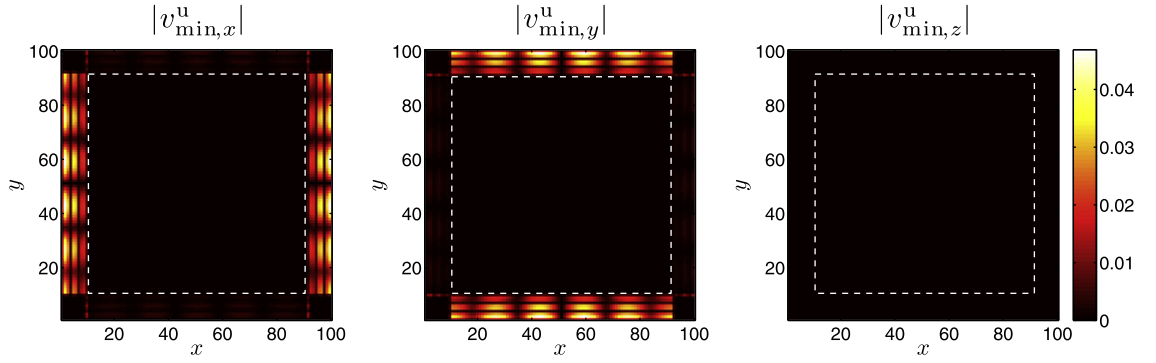
Table 4.3 shows the numerically calculated extreme singular values of A^{u} and A^{sc} for the MDM waveguide bend. From the table, we confirm that both (4.61) and (4.67) are satisfied. Also, we have much larger κ^{u}/κ^{sc} for this example than for the first example; for the present system, we have $\kappa^{u}/\kappa^{sc} = 584.2$.

In Table 4.3a, to estimate σ^{u}_{max} as derived in (4.59), we have used $\sigma^{u_0}_{max}$ of (4.29). Strictly speaking, (4.29) is applicable only for UPML with a *constant* PML loss parameter. However, each UPML subdomain with a graded PML loss parameter can be thought 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 $\sigma^{u_0}_{max}$. Hence, we use $\sigma^{u_0}_{max}$ in (4.29) with $s_x = s_0(d)$ as an estimate of σ^{u}_{max} in Table 4.3a. The estimate agrees quite well with numerically calculated σ^{u}_{max}. Accordingly, v^{u}_{max} is expected to be concentrated in the outermost layers of the graded UPML subdomains.

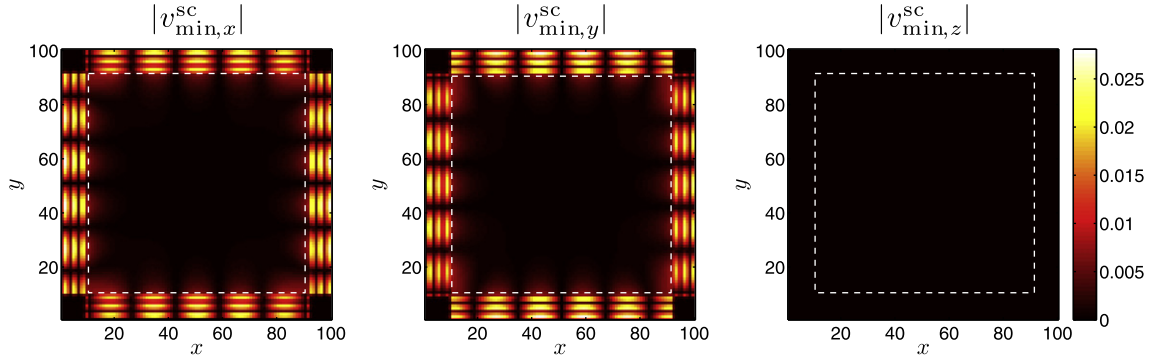
Fig. 4.7 displays v^{u}_{max} and v^{sc}_{max} for the MDM waveguide bend. As discussed above, v^{u}_{max} is indeed concentrated in the outermost UPML region, and v^{sc}_{max} is also concentrated in the region of regular media as expected. In addition, both v^{u}_{max} and v^{sc}_{max} exhibit the same fast spatial oscillation as seen in the first example.

Table 4.2
The minimum singular values σ^{u}_{min} and σ^{sc}_{min} of the vacua surrounded by UPML and SC-PML, respectively, along with the ratio $\sigma^{u}_{min}/\sigma^{sc}_{min}$. Note that $\sigma^{u}_{min}/\sigma^{sc}_{min} \leqslant 1$ as expected from (4.67). The numerically calculated minimum singular values are obtained by solving (4.7) so that $\|Av_{min} - \sigma_{min} u_{min}\|/\|u_{min}\| < 10^{-11}$ for $A = A^{u}, A^{sc}$. The unit μ_0^{-1}/nm^2 of the singular values is the normalization factor used in our numerical solver.

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



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



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

Fig. 4.6. The minimum right singular vectors (a) v^{u}_{min} of the vacuum surrounded by UPML, and (b) v^{sc}_{min} of the vacuum surrounded by SC-PML. The absolute values of the *x*-, *y*-, *z*-components of v^{u}_{min} and v^{sc}_{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 points.

We also display v^{u}_{min} and v^{sc}_{min} for the MDM waveguide bend in Fig. 4.8. Both the minimum right singular vectors are concentrated in the slot region, where the electric permittivity is ε_0. This follows the prediction in Section 4.5 that the minimum right singular vectors tend to be concentrated in either dielectrics or PMLs matching dielectrics.

In summary of this section, all of the detailed predictions made in Section 4.5 about the behaviors of the extreme singular values, extreme right singular vectors, and the condition numbers are demonstrated numerically.

5. Diagonal preconditioning scheme for the UPML equation

Our results in Sections 3 and 4 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 [10,35].

To use UPML in iterative solvers of the frequency-domain Maxwell's equations, one needs to accelerate convergence. For this purpose, [13] 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 [13] was that the factors $s_{w_1} s_{w_2}/s_{w_3}$ in (2.4), which become especially large in overlapping UPML regions, resulted in an ill-conditioned coefficient matrix. However, the arguments in Section 4.5 show that even without overlap of UPMLs the coefficient matrix is still quite ill-conditioned. In addition, Figs. 4.5 and

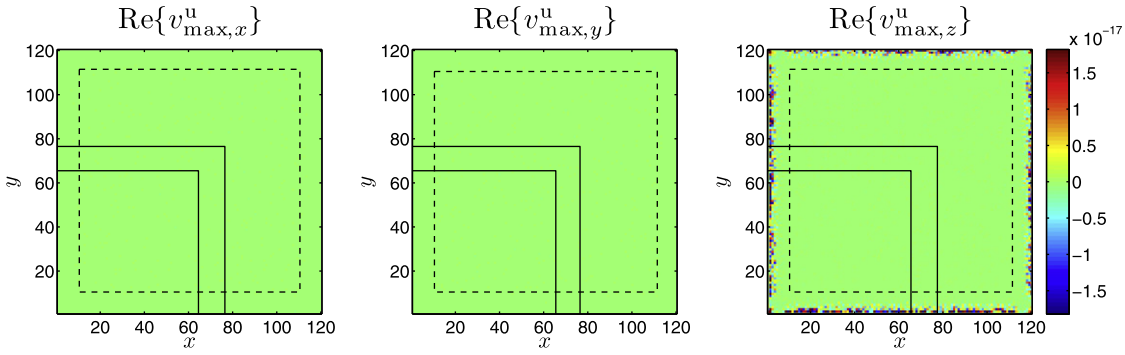
Table 4.3
The extreme singular values of the MDM waveguide bends surrounded by UPML and SC-PML. The extreme singular values are calculated by solving (4.7) so that $\|Av_i - \sigma_i u_i\|/\|u_i\| < 10^{-11}$ for $A = A^{u}, A^{sc}$. In (a), the estimates are evaluated using $\sigma^{u_0}_{max}$ and $\sigma^{r_0}_{max}$ in (4.29) with $s_x = s_0(d)$. Notice that $\sigma^{u}_{max}/\sigma^{sc}_{max}$ is much larger than it is in Table 4.1. The unit μ_0^{-1}/nm^2 of the singular values is the normalization factor used in our numerical solver.

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

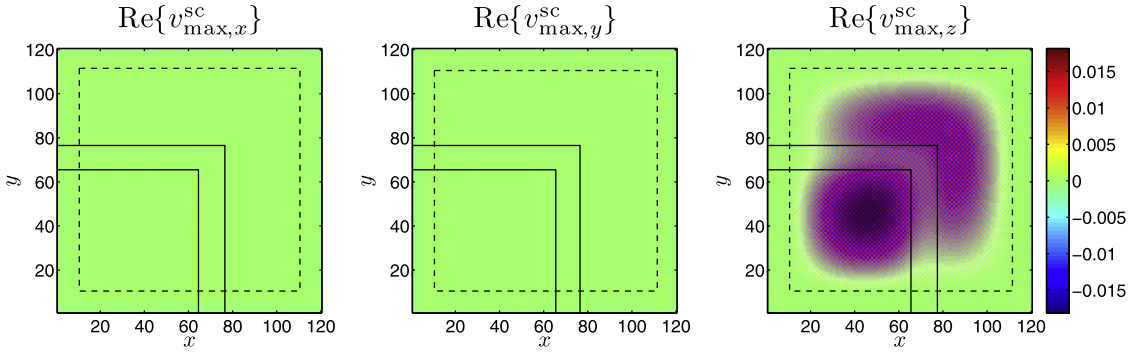
(a) Maximum singular values of the MDM waveguide bends

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

(b) Minimum singular values of the MDM waveguide bends



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



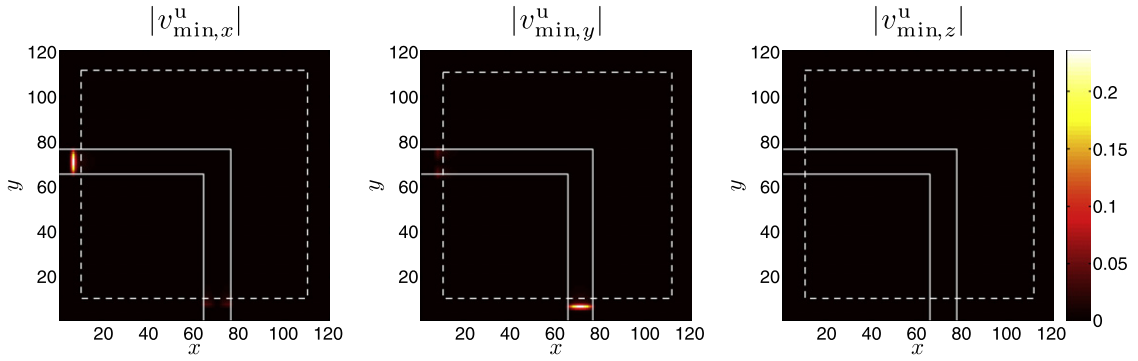
(b) v_{max} of the MDM waveguide bend surrounded by SC-PML

Fig. 4.7. The maximum right singular vectors (a) v^{u}_{max} of the MDM waveguide bend surrounded by UPML, and (b) v^{sc}_{max} of the same waveguide bend surrounded by SC-PML. The real parts of the x-, y-, z-components of v^{u}_{max} and v^{sc}_{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 a vacuum. Note that v^{u}_{max} is squeezed toward the boundary of the simulation domain where the PML loss parameters are maximized, whereas v^{sc}_{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 points.

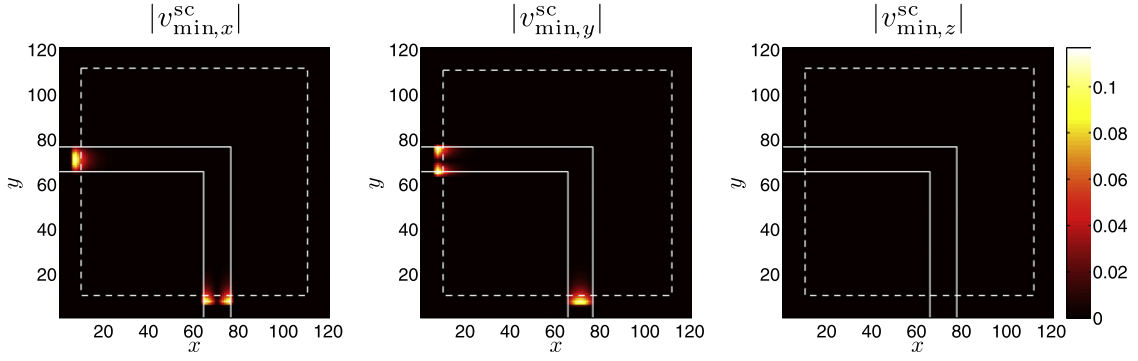
4.6 illustrate that the extreme right singular vectors do not reside in the overlapping UPML regions, and thus at least for some EM systems, overlap of UPMLs is not the reason for the large condition number of the UPML matrix.

Reference [14] 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 Section 5.1. Based on this relation, in Section 5.2 we devise the left and right diagonal preconditioners for the UPML matrix, and apply the preconditioners to the same 3D metallic slot waveguide bend examined in Section 3 to demonstrate the effectiveness of the preconditioning scheme.



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



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

Fig. 4.8. The minimum right singular vectors (a) v^{u}_{min} of the MDM waveguide bend surrounded by UPML, and (b) v^{sc}_{min} of the same waveguide bend surrounded by SC-PML. The absolute values of the *x*-, *y*-, *z*-components of v^{u}_{min} and v^{sc}_{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 points.

5.1. Relation between UPML and SC-PML with constant PML scale factors

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 Eq. (2.5) has $\mathbf{E}^{sc}$ as the solution for a given electric current source density $\mathbf{J}^{sc}$. With straightforward substitution, we can show that the following *E*-field and electric current source density satisfy the UPML Eq. (2.3):

$$\mathbf{E}^{u} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \mathbf{E}^{sc}, \quad \mathbf{J}^{u} = \begin{bmatrix} s_y s_z & 0 & 0 \\ 0 & s_z s_x & 0 \\ 0 & 0 & s_x s_y \end{bmatrix} \mathbf{J}^{sc}. \tag{5.1}$$

The transformations in (5.1) can also be derived by applying the coordinate transformation of Maxwell's equations introduced in [36]. It is also interesting to note that the transformation for **E** in (5.1) predicts the discontinuity of the normal component of the *E*-field at the UPML interface described in Section 7.5.2 of [4].

We note that the transformation for **E** in (5.1) was derived earlier in [37,38]. However, the transformation for **J** in (5.1) has been mostly ignored so far, because the electric current source is usually placed *outside PML* where the transformation has no effect.

The transformations (5.1) can be written in terms of matrices and column vectors as

$$e^{u} = S_l e^{sc}, \quad j^{u} = S_a j^{sc}. \tag{5.2}$$

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 (5.2). Recall the systems of linear Eqs. (2.11) and (2.12). In the present notation, they are

$$A^{u}e^{u} = -i\omega j^{u}, \tag{5.3}$$

$$A^{sc}e^{sc} = -i\omega j^{sc}, \tag{5.4}$$

considering (2.2). Substituting (5.2) in (5.3), we obtain

$$\left(S_a^{-1}A^{u}S_l\right)e^{sc} = -i\omega j^{sc}. \tag{5.5}$$

Comparing (5.4) with (5.5), we conclude that

$$A^{sc} = S_a^{-1}A^{u}S_l. \tag{5.6}$$

We emphasize that the simple relation (5.6) 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 [36] would not transform the SC-PML equation into the UPML equation.

5.2. Scale-factor-preconditioned UPML equation

In actual numerical simulations where PMLs are implemented with graded PML loss parameters, the equality in (5.6) does not hold by the reason explained at the end of Section 5.1. Nevertheless, the right-hand side of (5.6) suggests 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 Eq. (2.11) directly, we first solve

$$\left(S_a^{-1}A^{u}S_l\right)y = S_a^{-1}b \tag{5.7}$$

for y, and then recover the solution x of (2.11) as

$$x = S_l y. \tag{5.8}$$

The scale-factor preconditioning scheme does not change the kind of PML used in the EM system from UPML; the solution x obtained from (5.7) and (5.8) is exactly the solution of the discretized UPML Eq. (2.11). 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, $A^{sp} = S_a^{-1}A^{u}S_l$, 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 Eq. (5.7) 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 3D metallic slot waveguide bend examined in Section 3. The convergence behavior for SP-UPML is depicted in Fig. 5.1, 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 $\|r_i\|/\|b\|$ for the UPML equation preconditioned by the conventional Jacobi preconditioner in Fig. 5.1. The system of linear equations for the Jacobi-preconditioned UPML equation is

$$P_{jac}^{-1}A^{u}x = P_{jac}^{-1}b, \tag{5.9}$$

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 [14]. 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.

6. Conclusion and final 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 coefficient matrices; this also explains 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.

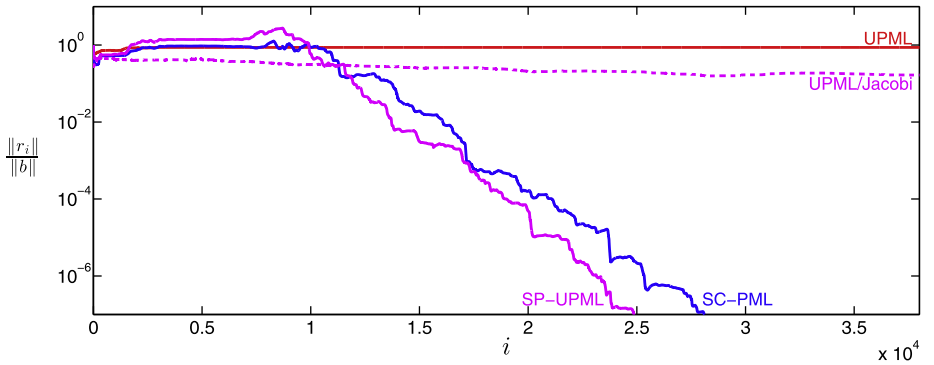


Fig. 5.1. Convergence of QMR for the UPML equation, SC-PML equation, SP-UPML equation, and the UPML equation preconditioned by the Jacobi preconditioner. The examined EM system is the metallic slot waveguide bend illustrated in Fig. 3.1, so the plots for the UPML and SC-PML equations are identical to the corresponding plots in Fig. 3.3. 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.

For numerical demonstrations, we constructed coefficient matrices by the FDFD method throughout the paper, but we emphasize that the conclusions of this paper are not limited to a specific method of discretizing the frequency-domain Maxwell's equations. For example, the condition number analysis in Section 4 was in essence estimation of the extreme singular values of the differential operators for homogeneous media. The scale-factor preconditioning scheme in Section 5 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 Section 5 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.

Acknowledgements

We thank Victor Liu and Dr. Zhichao Ruan for helpful comments. This work was supported by the AFOSR MURI program (FA9550–09–1–0704), the NSF Grant No. DMS-0968809, and the Interconnect Focus Center, funded under the Focus Center Research Program, a Semiconductor Research Corporation entity. Wonseok Shin also acknowledges the support of Samsung Scholarship Foundation.

Appendix A. Derivation of k_x minimizing $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ for a given k_y

In this section, we derive k_x used in (4.40) and (4.43) for $T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{u}_0}$. The case for $T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{sc}_0}$ can be treated similarly. The general assumptions $k_x \geqslant 0$, $k_y \geqslant 0$, $\varepsilon > 0$, and $s_x'' \gg 1$ of Section 4.4 apply here.

We first consider $k_y < \omega/c$. For such k_y, we show that $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ is an increasing function of k_x, and therefore it is minimized at $k_x = 0$. To that end, we derive the analytic formula of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ and examine its first derivative with respect to k_x.

The analytic formula of $\sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})$ is quite complex, so we use an approximation of $T_{\mathbf{k}}^{\mathrm{u}_0}$ to simplify the formula. Because of (4.16), $T_{\mathbf{k}}^{\mathrm{u}_0}$ of (4.20b) is approximated to

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

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

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 $[001]^{\mathrm{T}}$ is

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

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 $[ab0]^{\mathrm{T}}$ are

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

where

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

Between the two singular values, we are only interested in $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$, the smaller of the two. By straightforward algebra, we can show that the first derivative of $f_1 - f_2$ with respect to k_x is nonnegative. Hence, $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ 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. Thus, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}) = \min\{\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}, \tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}\}$ is also an increasing function of k_x. Since we are considering $k_x \geqslant 0$, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0})$ is minimized at $k_x = 0$.

Next, we consider $k_y > \omega/c$. In this case, $\tilde{\sigma}_{\mathbf{k},3}^{\mathrm{u}_0}$ of (A.2) is minimized at $k_x = k_{x0} \equiv s_x''[k_y^2 - \omega^2/c^2]^{1/2}$. In addition, since $\partial(f_1 - f_2)/\partial k_x$ is negative for $k_x < k_{x0}$ and positive for $k_x > k_{x0}$, $\tilde{\sigma}_{\mathbf{k},1}^{\mathrm{u}_0}$ is minimized at $k_x = k_{x0}$. Therefore, $\sigma_{\min}(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0})$ is minimized at $k_x = k_{x0}$.

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

$$\min_{k_x \geqslant 0} \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0}) \simeq \min_{k_x \geqslant 0} \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}\right) = \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=0} \simeq \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})_{k_x=0} \quad \text{for } k_y < \frac{\omega}{c}, \tag{A.5}$$

$$\min_{k_x \geqslant 0} \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0}) \simeq \min_{k_x \geqslant 0} \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}\right) = \sigma_{\min}\left(\widetilde{T}_{\mathbf{k}}^{\mathrm{u}_0}\right)_{k_x=k_{x0}} \simeq \sigma_{\min}(T_{\mathbf{k}}^{\mathrm{u}_0})_{k_x=k_{x0}} \quad \text{for } k_y > \frac{\omega}{c}, \tag{A.6}$$

which are (4.40) and (4.43) for $T_{\mathbf{k}} = T_{\mathbf{k}}^{\mathrm{u}_0}$, respectively.

Appendix B. First-order perturbation method for the nondegenerate singular values of symmetric matrices

In Appendix C, the singular values of symmetric matrices are calculated by a perturbation method, which we describe in this section. 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 [39].

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

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

where V^* is the complex conjugate of V. In other words, $U = V^*$ in (4.3) and $u_i = v_i^*$ in (4.4). The decomposition (B.1) is called Takagi's factorization or the symmetric SVD [40–42].

Suppose that $A^{(0)} \in \mathbb{C}^{n\times n}$ is a symmetric matrix whose SVD in the form (4.4) is

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

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

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

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

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

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 [41], because v_r is the unit eigenvector corresponding to a distinct eigenvalue σ_r^2 of the Hermitian eigenvalue problems (4.8)[6]; the same is true for $v_r^{(0)}$ corresponding to $\sigma_r^{(0)}$ of $A^{(0)}$. As a result,

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

because $A \to A^{(0)}$ as $\delta \to 0$. The nondegeneracy constraint is important in obtaining (B.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 $\mathrm{span}\{v_q^{(0)}, v_r^{(0)}\}$ instead.

[6] The phase factor $e^{i\theta_r}$ is arbitrary for the general SVD, but in fact it is not for Takagi's factorization [40]; the equality in (B.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, so we allow the freedom to vary the phase factor of v_r.

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

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

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

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

By applying A of (B.4) to v_p and substituting (B.3) and (B.6) in the result, we obtain

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

Subsequent application of $v_p^{(0)\mathsf{T}}$ to the right equation of (B.8) leads to

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

where (B.2) is used to obtain the first term of the right-hand side. Now, because of (B.7), all terms in the sum in (B.9) are in the order of δ^2 unless $r = p$. Hence,

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

or equivalently

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

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

$$-|O(\delta^2)| \leqslant \sigma_p - \left|\sigma_p^{(0)} + \delta\left(v_p^{(0)\mathsf{T}} A^{(1)} v_p^{(0)}\right)\right| \leqslant |O(\delta^2)|, \tag{B.12}$$

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

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

Appendix C. Estimation of the minimum of $\sigma_{\min}(T_{\mathbf{k}}^{\mathbf{u}_0})$ over k_x for a given $k_y > \omega/c$

In this section, we derive (4.44a) by examining $\sigma_{\min}(T_{\mathbf{k}}^{\mathbf{u}_0})$. Eq. (4.44b) can be similarly derived by examining $\sigma_{\min}(T_{\mathbf{k}}^{sc_0})$. The general assumptions $k_x \geqslant 0$, $k_y \geqslant 0$, $\varepsilon > 0$, $s_x'' \gg 1$ of Section 4.4 and the specific assumption $k_y > \omega/c$ apply here.

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

$$k_{x0} = s_x'' \sqrt{k_{y0}^2 - \frac{\omega^2}{c^2}} \tag{C.1}$$

and

$$\mathbf{k}_0 = \hat{\mathbf{x}} k_{x0} + \hat{\mathbf{y}} k_{y0}. \tag{C.2}$$

Then, the left-hand side of (4.44a) is $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathbf{u}_0})$, which we evaluate below.

We approximate $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathbf{u}_0})$ to first order in a small perturbation parameter δ. The perturbed quantity in $T_{\mathbf{k}_0}$ is the real part of s_x of (4.15), which is written as

$$s_x = -i s_x''(1+\delta), \tag{C.3}$$

where

$$\delta = \frac{i}{s_x''}. \tag{C.4}$$

Because $|\delta| \ll 1$ due to (4.16), the approximation of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathbf{u}_0})$ to first order in δ should be an accurate estimate of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathbf{u}_0})$.

To obtain the approximation of $T_{\mathbf{k}_0}^{\mathbf{u}_0}$, we approximate the three singular values of $\sigma_{\min}(T_{\mathbf{k}_0}^{\mathbf{u}_0})$ one by one. The singular value of $T_{\mathbf{k}_0}^{\mathbf{u}_0}$ corresponding to the singular vector $[001]^{\mathsf{T}}$ is $\sigma_{\mathbf{k}_0,3}^{\mathbf{u}_0}$, which is $\sigma_{\mathbf{k},3}^{\mathbf{u}_0}$ in (4.21) for $\mathbf{k} = \mathbf{k}_0$. Because (C.3) implies

$$\frac{1}{s_x^2} = -\frac{1}{s_x''^2(1+\delta)^2} = -\frac{1}{s_x''^2}(1-2\delta) + O(\delta^2), \tag{C.5}$$

we have

$$\sigma^{u_0}_{\mathbf{k}_0,3} = |s_x|\left|-\frac{k_{x0}^2}{s_x''^2\mu}(1-2\delta)+\frac{k_{y0}^2}{\mu}-\omega^2\varepsilon\right|+O(\delta^2) = 2|\delta||s_x|\left(\frac{k_{y0}^2}{\mu}-\omega^2\varepsilon\right)+O(\delta^2), \tag{C.6}$$

where k_{x0} is expressed in terms of k_{y0} using (C.1). Substituting (C.4) in (C.6) leads to

$$\sigma^{u_0}_{\mathbf{k}_0,3} = \frac{2\sqrt{s_x''^2+1}}{s_x''}\left(\frac{k_{y0}^2}{\mu}-\omega^2\varepsilon\right)+O(\delta^2). \tag{C.7}$$

The remaining two singular values of $T^{u_0}_{\mathbf{k}_0}$ correspond to the singular vectors of the form $[ab0]^T$. Therefore, we can derive the two singular values by applying the perturbation method established in Appendix B to the top-left 2×2 block of $T^{u_0}_{\mathbf{k}_0}$. Using (C.3) and

$$\frac{1}{s_x} = -\frac{1}{is_x''(1+\delta)} = -\frac{1}{is_x''}(1-\delta)+O(\delta^2), \tag{C.8}$$

we approximate the top-left 2×2 block of $T^{u_0}_{\mathbf{k}}$ of (4.20b) for $\mathbf{k}=\mathbf{k}_0$ as

$$A = \begin{bmatrix} \frac{k_{y0}^2}{s_x\mu}-\frac{\omega^2\varepsilon}{s_x} & -\frac{k_{x0}k_{y0}}{s_x\mu} \\ -\frac{k_{x0}k_{y0}}{s_x\mu} & \frac{k_{x0}^2}{s_x\mu}-s_x\omega^2\varepsilon \end{bmatrix} \simeq \begin{bmatrix} \left(-\frac{k_{y0}^2}{is_x''\mu}+\frac{\omega^2\varepsilon}{is_x''}\right)(1-\delta) & \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) \\ \frac{k_{x0}k_{y0}}{is_x''\mu}(1-\delta) & -\frac{k_{x0}^2}{is_x''\mu}(1-\delta)+is_x''(1+\delta)\omega^2\varepsilon \end{bmatrix}. \tag{C.9}$$

Following the notations in Appendix B, (C.9) is decomposed as

$$A \simeq A^{(0)}+\delta A^{(1)} = \begin{bmatrix} -\frac{k_{x0}^2}{is_x''^3\mu} & \frac{k_{x0}k_{y0}}{is_x''\mu} \\ \frac{k_{x0}k_{y0}}{is_x''\mu} & \frac{is_x''k_{y0}^2}{\mu} \end{bmatrix} + \delta\begin{bmatrix} \frac{k_{x0}^2}{is_x''^3\mu} & -\frac{k_{x0}k_{y0}}{is_x''\mu} \\ -\frac{k_{x0}k_{y0}}{is_x''\mu} & \frac{k_{x0}^2}{is_x''\mu}+is_x''\omega^2\varepsilon \end{bmatrix}, \tag{C.10}$$

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

We obtain the two singular values $\sigma^{u_0}_{\mathbf{k}_0,1}$ and $\sigma^{u_0}_{\mathbf{k}_0,2}$ of $T^{u_0}_{\mathbf{k}_0}$ from A. However, since eventually we are interested in $\sigma_{\min}\left(T^{u_0}_{\mathbf{k}_0}\right)$, we focus on the smaller of the two, which is denoted by $\sigma^{u_0}_{\mathbf{k}_0,1}$. 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 $\sigma^{u_0}_{\mathbf{k}_0,1}$ 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 (4.8). The result is

$$v_1^{(0)} = \frac{1}{\sqrt{k_{x0}^2/s_x''^2+s_x''^2k_{y0}^2}}\begin{bmatrix} -is_x''k_{y0}^2 \\ -ik_{x0}^2/s_x'' \end{bmatrix}. \tag{C.11}$$

Using (C.10) and (C.11) in (B.13), we obtain

$$\sigma^{u_0}_{\mathbf{k}_0,1} = \left|\sigma_1^{(0)}+\delta\left(v_1^{(0)T}A^{(1)}v_1^{(0)}\right)\right|+O(\delta^2) = 2\omega^2\varepsilon\frac{k_{y0}^2-\omega^2\mu\varepsilon}{(s_x''^2+1)k_{y0}^2-\omega^2\mu\varepsilon}+O(\delta^2), \tag{C.12}$$

where (C.1), (C.3), and (C.4) are used to simplify the result.

Taking the ratio between (C.7) and (C.12), we can easily see that $\sigma^{u_0}_{\mathbf{k}_0,1} < \sigma^{u_0}_{\mathbf{k}_0,3}$ in the leading order. Therefore, we conclude that

$$\sigma_{\min}\left(T^{u_0}_{\mathbf{k}_0}\right) = 2\omega^2\varepsilon\frac{k_{y0}^2-\omega^2\mu\varepsilon}{(s_x''^2+1)k_{y0}^2-\omega^2\mu\varepsilon}+O(\delta^2), \tag{C.13}$$

which is (4.44a).

References

[1] J.-P. Bérenger, A perfectly matched layer for the absorption of electromagnetic waves, Journal of Computational Physics 114 (1994) 185–200.
[2] G. Veronis, S. Fan, Theoretical investigation of compact couplers between dielectric slab waveguides and two-dimensional metal-dielectric-metal plasmonic waveguides, Optics Express 15 (2007) 1211–1221.
[3] L. Verslegers, P. Catrysse, Z. Yu, W. Shin, Z. Ruan, S. Fan, Phase front design with metallic pillar arrays, Optics Letters 35 (2010) 844–846.
[4] A. Taflove, S.C. Hagness, Computational Electrodynamics: The Finite-Difference Time-Domain Method, third ed., Artech House Publishers, 2005.
[5] Z. Sacks, D. Kingsland, R. Lee, J.-F. Lee, A perfectly matched anisotropic absorber for use as an absorbing boundary condition, IEEE Transactions on Antennas and Propagation 43 (1995) 1460–1463.

[6] W.C. Chew, W.H. Weedon, A 3D perfectly matched medium from modified Maxwell's equations with stretched coordinates, Microwave and Optical Technology Letters 7 (1994) 599–604.
[7] C.M. Rappaport, Perfectly matched absorbing boundary conditions based on anisotropic lossy mapping of space, Microwave and Guided Wave Letters, IEEE 5 (1995) 90–92.
[8] R. Mittra, U. Pekel, A new look at the perfectly matched layer (PML) concept for the reflectionless absorption of electromagnetic waves, Microwave and Guided Wave Letters, IEEE 5 (1995) 84–86.
[9] J. Roden, S. Gedney, Convolution PML (CPML): An efficient FDTD implementation of the CFS-PML for arbitrary media, Microwave and Optical Technology Letters 27 (2000) 334–339.
[10] J.-Y. Wu, D. Kingsland, J.-F. Lee, R. Lee, A comparison of anisotropic PML to Berenger's PML and its application to the finite-element method for EM scattering, IEEE Transactions on Antennas and Propagation 45 (1997) 40–50.
[11] Y. Botros, J. Volakis, A robust iterative scheme for FEM applications terminated by the perfectly matched layer (PML) absorbers, Proceedings of the Fifteenth National Radio Science Conference, 1998, pp. D11/1–D11/8.
[12] B. Stupfel, A study of the condition number of various finite element matrices involved in the numerical solution of Maxwell's equations, IEEE Transactions on Antennas and Propagation 52 (2004) 3048–3059.
[13] P. Talukder, F.-J. Schmuckle, R. Schlundt, 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.
[14] Y. Botros, J. Volakis, Preconditioned generalized minimal residual iterative scheme for perfectly matched layer terminated applications, Microwave and Guided Wave Letters, IEEE 9 (1999) 45–47.
[15] J.-M. Jin, W. Chew, Combining PML and ABC for the finite-element analysis of scattering problems, Microwave and Optical Technology Letters 12 (1996) 192–197.
[16] K. Yee, Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media, IEEE Transactions on Antennas and Propagation 14 (1966) 302–307.
[17] J. Smith, Conservative modeling of 3-D electromagnetic fields, part I: Properties and error analysis, Geophysics 61 (1996) 1308–1318.
[18] N.J. Champagne II, J. Berryman, H. Buettner, FDFD: A 3D finite-difference frequency-domain code for electromagnetic induction tomography, Journal of Computational Physics 170 (2001) 830–848.
[19] K.S. Kunz, R.J. Luebbers, The Finite Difference Time Domain Method for Electromagnetics, CRC-Press, 1993. Section 3.2.
[20] G. Veronis, S. Fan, Modes of subwavelength plasmonic slot waveguides, Journal of Lightwave Technology 25 (2007) 2511–2521. In the private communication with the authors, the use of the 1 nm grid edge length in this paper was confirmed.
[21] P.B. Johnson, R.W. Christy, Optical constants of the noble metals, Physical Review B 6 (1972) 4370–4379.
[22] E.D. Palik (Ed.), Handbook of Optical Constants of Solids, Academic Press, 1985.
[23] D.R. Lide (Ed.), CRC Handbook of Chemistry and Physics, 88th ed., CRC Press, 2007.
[24] R. Freund, N. Nachtigal, QMR: a quasi-minimal residual method for non-Hermitian linear systems, Numerische Mathematik 60 (1991) 315–339.
[25] S. Balay, J. Brown, K. Buschelman, W.D. Gropp, D. Kaushik, M.G. Knepley, L.C. McInnes, B.F. Smith, H. Zhang, PETSc Web page, 2011. Available from: <http://www.mcs.anl.gov/petsc>.
[26] D.A.H. Jacobs, A generalization of the conjugate-gradient method to solve complex systems, IMA Journal of Numerical Analysis 6 (1986) 447–452.
[27] M. Benzi, G.H. Golub, J. Liesen, Numerical solution of saddle point problems, Acta Numerica 14 (2005) 1–137. Section 9.2.
[28] B.N. Datta, Numerical Linear Algebra and Applications, 2nd ed., SIAM, 2010. Section 6.8.
[29] G.H. Golub, C.F. Van Loan, Matrix Computations, 3rd ed., The Johns Hopkins University Press, 1996. Section 2.5.6; 2.3.1; 2.3.3.
[30] R.B. Lehoucq, K. Maschhoff, D.C. Sorensen, C. Yang, ARPACK Web page, 2011. Available from: <http://www.caam.rice.edu/software/ARPACK>.
[31] MATLAB Web page, 2011. Available from: <http://www.mathworks.com/products/matlab>.
[32] R.B. Lehoucq, D.C. Sorensen, C. Yang, ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, SIAM, 1998.
[33] J.W. Goodman, Introduction to Fourier Optics, 3rd ed., Roberts & Company Publishers, 2005. Section 2.3.2.
[34] A.V. Oppenheim, R.W. Schafer, J.R. Buck, Discrete-Time Signal Processing, 2nd ed., Prentice Hall, 1999. Section 2.6.1; 4.2.
[35] C. Wolfe, U. Navsariwala, S. Gedney, A parallel finite-element tearing and interconnecting algorithm for solution of the vector wave equation with PML absorbing medium, IEEE Transactions on Antennas and Propagation 48 (2000) 278–284.
[36] C. Kottke, A. Farjadpour, S. Johnson, Perturbation theory for anisotropic dielectric interfaces, and application to subpixel smoothing of discretized numerical methods, Physical Review E 77 (2008) 036611. Appendix.
[37] F. Teixeira, W. Chew, General closed-form PML constitutive tensors to match arbitrary bianisotropic and dispersive linear media, Microwave and Guided Wave Letters, IEEE 8 (1998) 223–225.
[38] S. Gedney, An anisotropic perfectly matched layer-absorbing medium for the truncation of FDTD lattices, IEEE Transactions on Antennas and Propagation 44 (1996) 1630–1639.
[39] L.D. Landau, E.M. Lifshitz, Quantum Mechanics: Non-relativistic Theory, Course of Theoretical Physics, 3rd ed., vol. 3, Butterworth-Heinemann, 1977.
[40] T. Takagi, On an algebraic problem related to an analytic theorem of Carathedory and Fejer and on an allied theorem of Landau, Japanese Journal of Mathematics 1 (1924) 82–93.
[41] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, 1985. Corollary 4.4.4; Theorem 7.3.5.
[42] A. Bunse-Gerstner, W. Gragg, Singular value decompositions of complex symmetric matrices, Journal of Computational and Applied Mathematics 21 (1988) 41–54.