# ACCELERATED SOLUTION OF THE FREQUENCY-DOMAIN MAXWELL'S EQUATIONS BY ENGINEERING THE SPECTRUM OF THE OPERATOR USING THE CONTINUITY EQUATION

WONSEOK SHIN* AND SHANHUI FAN†

**Abstract.** We introduce a simple method to accelerate the convergence of iterative solvers of the frequency-domain Maxwell's equations. The method modifies the original Maxwell's equations to eliminate the high multiplicity of near-zero eigenvalues that exists in the original operator. The impact of the modified eigenvalue spectrum on the accelerated convergence is explained by visualizing the components of the residual vector with respect to the eigenvector basis at each iteration step. A comparison with the previous approach that is similar to our method is also presented.

**Key words.** Maxwell equations, iterative methods, eigenvalues and eigenfunctions, finite difference methods

**AMS subject classifications.** 35Q61, 78M20, 65F10, 65F15

**1. Introduction.** To understand electromagnetic (EM) and optical phenomena, it is essential to solve Maxwell's equations efficiently. In the frequency domain, assuming a time dependence $e^{+i\,\omega\,t}$ and nonmagnetic materials, the equations are

$$\nabla \times \mathbf{E} = -i\,\omega\,\mu_0\,\mathbf{H}, \tag{1.1}$$

$$\nabla \times \mathbf{H} = i\,\omega\,\varepsilon\,\mathbf{E} + \mathbf{J}, \tag{1.2}$$

where $\varepsilon$ is the electric permittivity; $\mu_0$ is the magnetic permeability of a vacuum; $\omega$ is the angular frequency; $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{J}$ are the electric field, magnetic field, and the electric current source density, respectively. To obtain the solutions $\mathbf{E}$ and $\mathbf{H}$ of the equations for a given $\mathbf{J}$, the main equation to solve is

$$\nabla \times \nabla \times \mathbf{E} - \omega^2\,\mu_0\,\varepsilon\,\mathbf{E} = -i\,\omega\,\mu_0\,\mathbf{J}, \tag{1.3}$$

which is formulated by eliminating $\mathbf{H}$ from (1.1) and (1.2); once $\mathbf{E}$ is obtained from (1.3), it is substituted into (1.1) to recover $\mathbf{H}$.

To solve (1.3) numerically, a large system of linear equations

$$A\,x = b, \tag{1.4}$$

where $A$ is a coefficient matrix, $x$ is an unknown column vector representing $\mathbf{E}$, and $b$ is a column vector representing $-i\,\omega\,\mathbf{J}$, is constructed by a method such as the finite-difference frequency-domain (FDFD) method or finite-element method (FEM). The coefficient matrix $A$ is sparse (with only 13 elements per row) and typically very large (often with more than 10 million rows and columns), especially for 3-dimensional (3D) problems. To solve a system with such a large and sparse coefficient matrix, iterative methods are usually preferred to direct methods [8].

However, it is well-known that convergence is quite slow when the iterative methods are directly applied to (1.4) constructed from (1.3) in the "low-frequency" regime; the low-frequency regime arises in geophysics and deep-subwavelength nanophotonics, for example, where the wavelengths used are much greater than the geometric features in problem domains. The huge null space of the operator $\nabla \times \nabla \times$ was proved to be the origin of the slow convergence, and several techniques to improve the convergence speed have been suggested.

Most of the techniques are based on the Helmholtz decomposition, which decomposes the $E$-field as $\mathbf{E} = \mathbf{\Psi} + \nabla\phi$ for a divergence-free vector field $\mathbf{\Psi}$ and a scalar function $\phi$. Thanks to the divergence-free nature of $\mathbf{\Psi}$, (1.3) is written as

$$-\nabla^2\mathbf{\Psi} - \omega^2\,\mu_0\,\varepsilon\,(\mathbf{\Psi} + \nabla\phi) = -i\,\omega\,\mu_0\,\mathbf{J}, \tag{1.5}$$

where the operator $\nabla \times \nabla \times$ with a huge null space is replaced with the negative Laplacian $-\nabla^2$. However, these techniques require solving an extra equation for the extra unknown $\phi$, which can be time-consuming.

*Department of Electrical Engineering, Stanford University, Stanford, CA, 94305 USA (wsshin@stanford.edu).
†Department of Electrical Engineering, Stanford University, Stanford, CA, 94305 USA (shanhui@stanford.edu).

A less explored technique is to modify (1.3) using the continuity equation

$$\nabla \cdot (\varepsilon \, \mathbf{E}) = \frac{i}{\omega} \nabla \cdot \mathbf{J}. \tag{1.6}$$

By dividing (1.6) by $\varepsilon$ and then taking the gradient, we obtain

$$\nabla \left[ \varepsilon^{-1} \nabla \cdot (\varepsilon \, \mathbf{E}) \right] = \frac{i}{\omega} \nabla \left[ \varepsilon^{-1} \nabla \cdot \mathbf{J} \right]. \tag{1.7}$$

Because (1.6) is derived from (1.3) by taking the divergence, the solution of (1.3) always satisfies (1.7). Moreover, (1.7) has the same physical unit as (1.3). Therefore, we can freely add any multiple of (1.7) to (1.3) to obtain a new equation with the same solution $\mathbf{E}$. The new equation may conform to iterative methods better than (1.3), without the penalty of introducing an extra unknown like the techniques based on the Helmholtz decomposition.

When this technique was first explored in [1], the authors added (1.7) to (1.3), hoping the operator of the resulting equation to be positive-definite, because a positive-definite operator is well-suited to iterative methods in general. Even though they were not able to prove whether the operator was positive-definite or not, they observed much faster convergence of iterative methods for the new equation than for (1.3).

In this paper, we re-examine the technique that uses (1.7) to accelerate the convergence of iterative methods in the low-frequency regime. We show that adding (1.7) to (1.3) as described in [1] accelerates the convergence only during the initial stage of iteration, but eventually stagnates the convergence. On the contrary, we find that *subtracting* (1.7) from (1.3) maintains accelerated convergence speeds throughout the entire iteration.

It turns out that our method accelerates the convergence by engineering the eigenvalue spectrum of the operator into a form that is more preferable to iterative methods. We provide a detailed analysis to establish the relation between the convergence behavior and eigenvalue spectrum. There have been already many studies pointing out the dependence of the convergence behavior on the eigenvalue spectrum, but they mostly consist of mathematical proofs. In contrast, by visualizing the residual vector and residual polynomial at each iteration step, we develop a graphical mean to understand the impact of the eigenvalue spectrum on the convergence behavior. Such a graphical analysis provides a more intuitive understanding on the deterioration of the convergence speed caused by an eigenvalue cluster near zero and indefinite coefficient matrix.

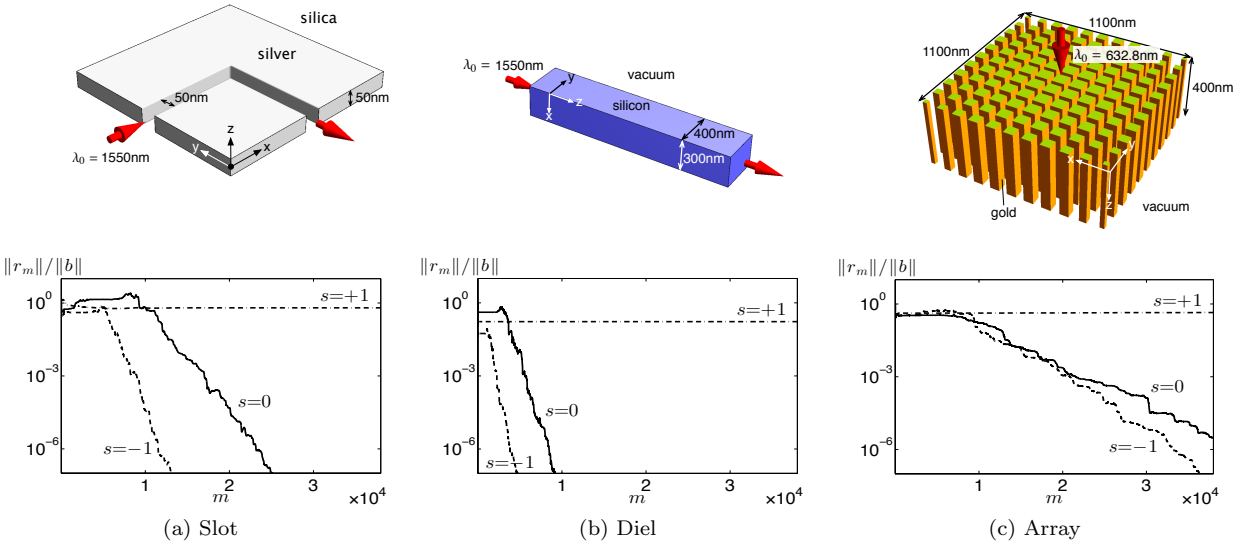
The rest of this paper is organized as follows. In Sec. 2, we formally introduce our equation and the equation solved in [1]. In Sec. 3, we solve a wide range of realistic 3D problems to compare the convergence behavior of an iterative method for the two equations as well as for the original equation (1.3). To understand the different convergence behavior for the different equations, in Sec. 4 we investigate the eigenvalue spectra of the operators used in the three equations for a simple homogeneous system. In Sec. 5, we relate the eigenvalue spectra to the convergence behavior of an iterative method, and we conclude in Sec. 6.

**2. Modified equations.** In this section, we formally describe the modified equations we investigate. When (1.7) is scaled by an arbitrary scale factor $s$ and added to (1.3), we obtain

$$\nabla \times \nabla \times \mathbf{E} + s \, \nabla \left[ \varepsilon^{-1} \nabla \cdot (\varepsilon \, \mathbf{E}) \right] - \omega^2 \, \mu_0 \, \varepsilon \, \mathbf{E} = -i \, \omega \, \mu_0 \, \mathbf{J} + s \, \frac{i}{\omega} \nabla \left[ \varepsilon^{-1} \nabla \cdot \mathbf{J} \right]. \tag{2.1}$$

As mentioned in Sec. 1, (2.1) has the same solution $\mathbf{E}$ as (1.3). However, because the operator on the left-hand side and the term on the right-hand side change with $s$, (2.1) produces a different system of linear equations of the form (1.4) with different $A$ and $b$ depending on $s$. Since the convergence behavior of iterative methods strongly depends on the properties of the coefficient matrix $A$, we can expect a different performance of iterative methods to solve (1.4) for different $s$.

In the following sections, we investigate three cases: $s = 0, -1, +1$; note that $s = 0$ corresponds to the original equation (1.3), and $s = +1$ corresponds to the equation solved in [1]. In the next section, we demonstrate the different convergence behavior of an iterative method for different $s$ by solving (2.1) for 3D systems.




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

FIG. 3.1. *Three inhomogeneous systems for which* (2.1) *is solver for* $s = 0, -1, +1$ *by QMR: (a) a slot waveguide bend formed in a thin silver film (Slot), (b) a straight silicon waveguide (Diel), and (c) an array of gold pillars (Array). In the figures in the first row, the dimensions and materials of the structures are specified. The directions of wave propagation are shown by red arrows, beside which the vacuum wavelengths used are indicated. In the figures in the second row, the convergence behavior of QMR is plotted. Note that for all three systems QMR converges fastest for* $s = -1$*, whereas it barely converges for* $s = +1$*.*

| | Slot | Diel | Array |
|---|---|---|---|
| $N_x \times N_y \times N_z$ | $192 \times 192 \times 240$ | $220 \times 220 \times 320$ | $220 \times 220 \times 130$ |
| $\Delta_x, \Delta_y, \Delta_z$ (nm) | $2 \sim 20$ | 10 | $5, 5, 20$ |

TABLE 3.1

*The specification of the finite-difference grids used for the three simulated systems described in Figures 3.1. The number of grid cells in each system is* $N_x N_y N_z$*, which results in* $3N_x N_y N_z$ *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.*

**3. Convergence behavior of QMR for 3D inhomogeneous systems.** In this section, we solve (2.1) for 3D inhomogeneous systems by an iterative method, and compare the convergence behavior for $s = 0, -1, +1$.

The three inhomogeneous systems we consider are illustrated in the first row of Fig. 3.1. We enclose each system by the perfectly matched layer (PML) boundary condition appropriately to simulate an infinite space. For each system, we construct three systems of linear equations of the form (1.4) corresponding to $s = 0, -1, +1$ by the FDFD method. The number of the grid cells in the finite-difference grid used to discretize each system is shown in Table 3.1, together with the grid edge lengths in the $x$-, $y$-, $z$-directions. Note that the the grid edge lengths are much smaller than the vacuum wavelengths $\lambda_0$ indicated in the first row of Fig. 3.1 so that the problems are in the low-frequency regime.

The constructed systems of linear equations are solved by the quasi-minimal residual (QMR) iterative method [4]. At the $m$-th step of the QMR iteration, an approximate solution $x_m$ is generated. As $m$ increases, $x_m$ eventually converges to the exact solution of the system of linear equations $A\,x = b$. We assume that convergence is achieved when the residual vector

$$r_m = b - A\,x_m \tag{3.1}$$

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

The second row of Fig. 3.1 shows the convergence behavior of QMR for the three systems. Note that for all three systems $s = -1$ induces in the fastest convergence, and $s = +1$ barely results in convergence.

The three systems tested in this section are chosen deliberately to include geometries with different degrees of complexities, and different materials such as dielectrics and metals. Therefore, Fig. 3.1 suggests that $s = -1$ and $s = +1$ lead to faster and slower convergence speeds than $s = 0$, respectively, 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 [6]. Hence, we conclude that the different convergence speeds for $s = 0, -1, +1$ originate from the difference in the fundamental properties of the operator in (2.1) for different $s$.

In the next section, we consider a simple homogeneous system, and show that different $s$ induces a very different eigenvalue spectrum of the operator. Such a different eigenvalue spectrum is used to explain the different convergence speed for different $s$ later.

**4. Eigenvalue spectra of the operators for a homogeneous system.** In this section, we examine the operator in (2.1) for a homogeneous system to investigate the properties of the operator for different $s$.

For a homogeneous system where $\varepsilon$ is constant, the both sides of (1.7) are simplified as

$$\nabla \left[\varepsilon^{-1} \nabla \cdot (\varepsilon \, \mathbf{E})\right] = \nabla(\nabla \cdot \mathbf{E}) \,, \quad \frac{i}{\omega} \nabla \left[\varepsilon^{-1} \nabla \cdot \mathbf{J}\right] = \frac{i}{\omega \, \varepsilon} \nabla(\nabla \cdot \mathbf{J}). \tag{4.1}$$

Substituting (4.1), (2.1) reduces to

$$\nabla \times \nabla \times \mathbf{E} + s \, \nabla(\nabla \cdot \mathbf{E}) - \omega^2 \, \mu_0 \, \varepsilon \, \mathbf{E} = -i \, \omega \, \mu_0 \, \mathbf{J} + s \, \frac{i}{\omega \, \varepsilon} \nabla(\nabla \cdot \mathbf{J}). \tag{4.2}$$

The convergence behavior of iterative methods to solve (4.2) strongly depends on the eigenvalue spectrum of its operator

$$T = \nabla \times (\nabla \times \quad) + s \, \nabla(\nabla \cdot \quad) - \omega^2 \, \mu_0 \, \varepsilon. \tag{4.3}$$

Since we consider a constant $\varepsilon$ in this section, the term $-\omega^2 \, \mu_0 \, \varepsilon$ in (4.3) only shifts the eigenvalue spectrum of the operator

$$T_0 = \nabla \times (\nabla \times \quad) + s \, \nabla(\nabla \cdot \quad) \tag{4.4}$$

by $-\omega^2 \, \mu_0 \, \varepsilon$, which is a small amount in the low-frequency regime. Therefore, the eigenvalue spectrum of $T$ can be approximated very well by the eigenvalue spectrum of $T_0$.

In Appendix A, we show that

$$\lambda = 0 \,, \;\; |\mathbf{k}|^2 \,, \;\; |\mathbf{k}|^2 \tag{4.5}$$

and

$$\lambda = -|\mathbf{k}|^2 \,, \;\; 0 \,, \;\; 0 \tag{4.6}$$

are the eigenvalues of the operators $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$, respectively, where $\mathbf{k}$ is the wavevector of spatially harmonic eigenfunctions. Moreover, we also show in Appendix A that (4.5) and (4.6) correspond to the same set of eigenfunctions. Therefore, the operator $T_0$ has

$$\lambda = -s \, |\mathbf{k}|^2 \,, \;\; |\mathbf{k}|^2 \,, \;\; |\mathbf{k}|^2 \tag{4.7}$$

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

The eigenvalues (4.7) predict very different eigenvalue spectra of $T_0$'s for the different values of $s$ we consider. If $s = 0$, (4.7) reduces to (4.5). Because (4.5) has 0 as an eigenvalue for each $\mathbf{k}$, $T_0$ has a very high multiplicity of the eigenvalue 0. In addition, the three values in (4.7) are all nonnegative, so $T_0$ is positive-semidefinite for $s = 0$.

If $s = -1$, the high multiplicity of the eigenvalue 0 occurred for $s = 0$ is removed, but the multiplicity of the eigenvalue $|\mathbf{k}|^2$ increases by 1.5 times for each $\mathbf{k}$. Note that $T_0$ for $s = -1$ is also positive-semidefinite.

If $s = +1$, the multiple eigenvalue 0 occurred for $s = 0$ spreads into different eigenvalues $-|\mathbf{k}|^2$ for different $\mathbf{k}$. As a result, (4.7) has both positive and negative values, so $T_0$ becomes an indefinite operator for $s = +1$.

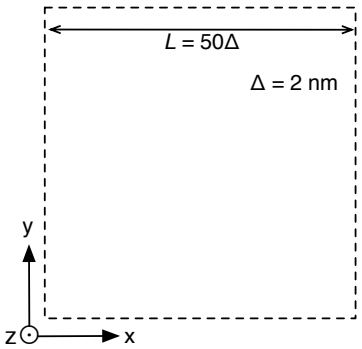
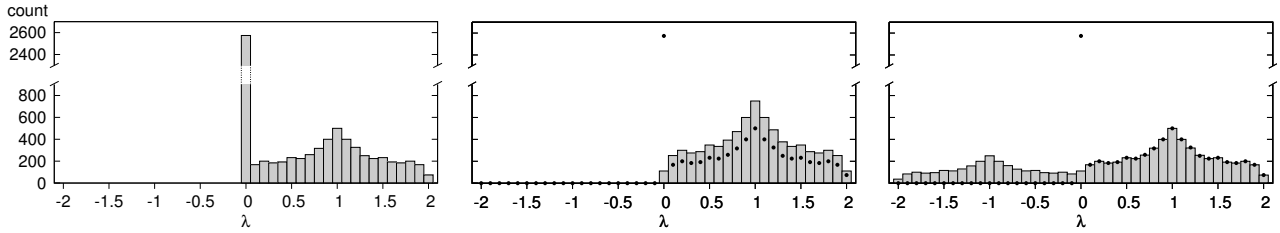



FIG. 4.1. *A 2D square domain filled with a vacuum for which the eigenvalue spectrum of $T_0$ is calculated numerically for $s = 0, -1, +1$. The domain is homogeneous in the z-direction, whereas its x- and y-boundaries are subject to the periodic boundary conditions. The square domain is discretized on a finite-difference grid with grid edge length $\Delta = 2\,\mathrm{nm}$. The domain is composed of $50 \times 50$ grid cells, which lead to 7500 eigenvalues in total.*




(a) Spectrum of $A_0$ for $s = 0$ (b) Spectrum of $A_0$ for $s = -1$ (c) Spectrum of $A_0$ for $s = +1$

FIG. 4.2. *The eigenvalue spectrum of $A_0$ discretized from $T_0$ for (a) $s = 0$, (b) $s = -1$, and (c) $s = +1$ for the vacuum-filled domain illustrated in Fig. 4.1. The height of each bar is equal to the number of eigenvalues contained in the bin indicated by the base of the bar. All 7500 eigenvalues $\lambda$ of $A_0$ are calculated for each $s$, and categorized into 41 bins in the horizontal axis. The unit of the horizontal axes is $\mathrm{nm}^{-2}$, and $-2\,\mathrm{nm}^{-2} \le \lambda \le 2\,\mathrm{nm}^{-2}$. In (b) and (c), black dots indicate the eigenvalue spectrum for $s = 0$ shown in (a). The vertical axes are broken to save space for the extremely high peak at $\lambda \simeq 0$ in the eigenvalue spectrum for $s = 0$. The three spectra are essentially the eigenvalue spectra of $A$ discretized from $T$.*

We verify the above predictions on the eigenvalue spectra by numerically calculating the spectra for a simple 2-dimensional (2D) domain filled with a vacuum (Fig. 4.1). The domain is discretized on a finite-difference grid with $N_x \times N_y = 50 \times 50$ grid cells. Therefore, the coefficient matrix $A_0$ discretized from $T_0$ has $3\,N_x\,N_y = 7500$ rows and columns, where the extra factor 3 accounts for the three Cartesian components of the $E$-field.

The spectra of the numerically calculated eigenvalues of the coefficient matrices $A_0$ for $s = 0, -1, +1$ are shown in Fig. 4.2. The spectra precisely agree with the predictions made above. In the spectrum for $s = 0$ (Fig. 4.2a), a very high peak exists at $\lambda \simeq 0$. Moreover, the peak is about 2500, which corresponds to a third of the total number of eigenvalues; this makes sense because one of the three eigenvalues (4.7) is 0 for each $\mathbf{k}$ when $s = 0$. In the spectrum for $s = -1$ (Fig. 4.2b), a peak does not exist at $\lambda \simeq 0$, but the heights of the remaining bars are increased by about 1.5 times. The spectrum for $s = +1$ (Fig. 4.2c) does not have a peak at $\lambda \simeq 0$, either, but it has negative eigenvalues. We also note that $A_0$ is positive-semidefinite for $s = 0, -1$, whereas it is indefinite for $s = +1$.

In the next section, we put an electric dipole current source in the homogeneous system illustrated in Fig. 4.1, and numerically solve (4.2) for the system by an iterative method to understand the impact of the different eigenvalue spectra studied in this section on the convergence behavior of the iterative method.

**5. Impact of the eigenvalue spectrum on the convergence behavior of GMRES.** In this section, we put an $x$-polarized electric dipole current source at the center of the system illustrated in Fig. 4.1, and solve the resulting equation (4.2) numerically for $s = 0, -1, +1$ by an iterative method. By closely investigating the evolution of the residual vector generated by the iterative method, we understand the impact of the eigenvalue spectrum of the coefficient matrix to the convergence behavior.

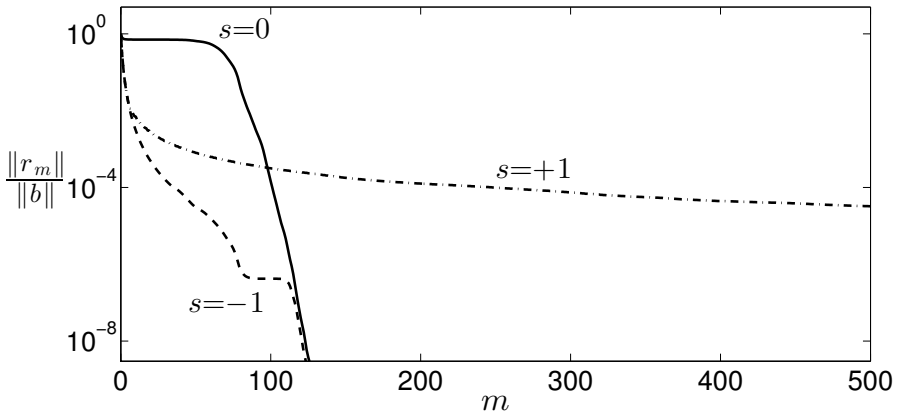



FIG. 5.1. *Convergence behavior of GMRES for the vacuum-filled domain illustrated in Fig. 4.1. Three systems of linear equations discretized from* (4.2) *for* $s = 0, -1, +1$ *are solved by GMRES, while monitoring the relative residual norms* $\|r_m\|/\|b\|$ *at each iteration step* $m$*. Note the the plot for* $s = 0$ *stagnates initially. Also note that the plot for* $s = +1$ *does not stagnate initially, but stagnates later on for a long iteration period.*

The iterative method of choice in this section is the generalized minimal residual (GMRES) method [7]. Fig. 5.1 shows $\|r_m\|/\|b\|$ versus the iteration step $m$ for the three values of $s$, where $r_m$ is the residual vector (3.1) generated at each iteration step. As can be see in the figure, the convergence behavior is quite different for different $s$.

The overall trend of the convergence behavior is consistent with the well-known theories of the Krylov subspace methods [8]. More specifically, it is known that GMRES stagnates initially if the coefficient matrix $A$ has many eigenvalues $\lambda$ close to 0 [3]. In addition, the convergence of the Krylov subspace methods is known to be generally much slower for indefinite $A$ than for definite $A$ [2]. The initial stagnation for $s = 0$ and slow convergence for $s = -1$ in Fig. 5.1 are consistent with these theories, because as shown in Fig. 4.2, our $A$ has a very high multiplicity near $\lambda = 0$ for $s = 0$, and is indefinite for $s = -1$. However, a more concrete and intuitive analysis of the convergence behavior can be given by using the residual polynomial as follows.

We first review the residual polynomial of GMRES briefly. For any Krylov subspace method, the residual vector $r_m$ generated by the method is expressed as [8]

$$r_m = p_m(A)\, r_0, \tag{5.1}$$

where $p_m$ is called the residual polynomial of degree at most $m$ that satisfies

$$p_m(0) = 1. \tag{5.2}$$

Different Krylov subspace methods construct $p_m$ differently. GMRES, as can be inferred from its name, construct $p_m$ so that it minimizes $\|r_m\|$ while satisfying (5.2), where $\|\cdot\|$ is the 2-norm of a column vector.

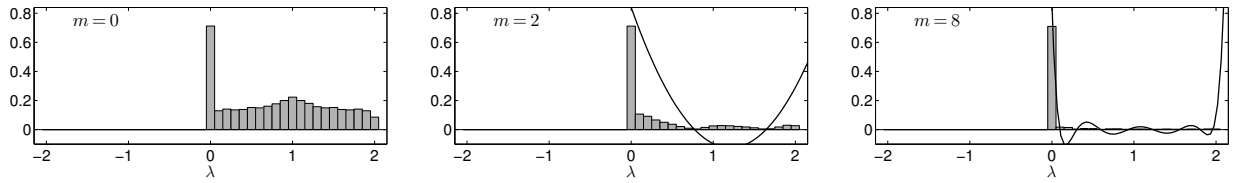
Now, consider the coefficient matrix $A \in \mathbb{C}^{n\times n}$ discretized from $T$ of (4.2). Because $T$ is a Hermitian operator, $A$ is a Hermitian matrix. Thus the eigenvalues $\lambda_i$ of $A$ are all real, and we can find an orthonormal set of eigenvectors $\{v_1, \ldots, v_n\}$. Hence, $A$ can be decomposed as

$$A = V\, \Lambda\, V^\dagger, \tag{5.3}$$

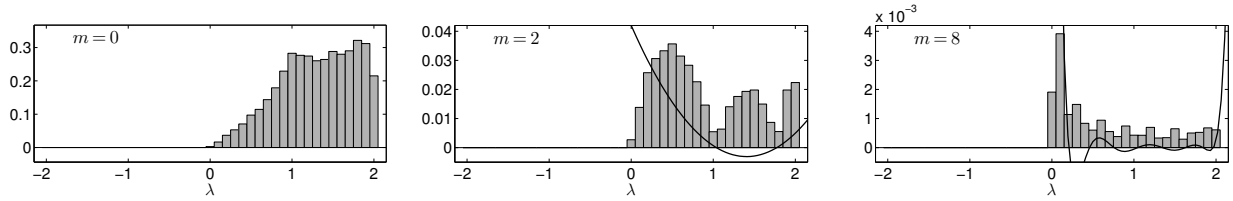
where

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \tag{5.4}$$

and $V^\dagger$ is the conjugate transpose of $V$; note that $V$ is unitary, i.e., $V\, V^\dagger = I$.




(a) $\bar{r}_m$ for $s = 0$




(b) $\bar{r}_m$ for $s = -1$

FIG. 5.2. *Initial evolution of $\bar{r}_m$ for (a) $s = 0$ and (b) $s = -1$. Bars whose heights are $h_{mj}$ of (5.9) are plotted for the three iteration steps $m = 0, 2, 8$. In the figures for $m = 2$ and $m = 8$, the residual polynomials $p_m$ are also plotted as solid curves. The values of $p_m$ are scaled appropriately to show the oscillation and roots of the polynomials clearly. Note that in (b) the vertical axis is gradually zoomed in with increasing $m$. Also notice that the peak at $\lambda \simeq 0$ persists in (a).*

We define

$$\bar{r}_m \equiv V^\dagger \, r_m, \tag{5.5}$$

which is the representation of $r_m$ with respect to the orthonormal eigenvector basis. Because the unitary transformation $V$ preserves the 2-norm of a vector, we have

$$\|r_m\| = \|\bar{r}_m\|. \tag{5.6}$$

In addition, using (5.1), (5.3), and (5.5) we obtain

$$\bar{r}_m = V^\dagger \, r_m = V^\dagger \, p_m(A) \, r_0 = V^\dagger \, V \, p_m(\Lambda) \, V^\dagger \, r_0 = p_m(\Lambda) \, \bar{r}_0, \tag{5.7}$$

which relates $i$-th components of $\bar{r}_m$ and $\bar{r}_0$:

$$\bar{r}_{mi} = p_m(\lambda_i) \, \bar{r}_{0i}. \tag{5.8}$$

Eqs. (5.6) and (5.8) indicate that the evolution of $\|r_m\|$ in the iterative process can be monitored by observing $\bar{r}_m$, and $\bar{r}_m$ is directly related to the residual polynomial.

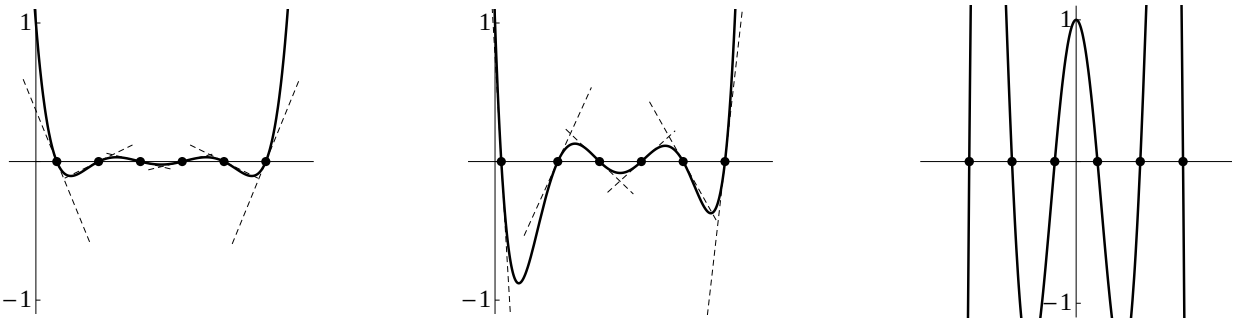
In Fig. 5.2, we visualize $\bar{r}_m$ for $s = 0$ and $s = -1$ at three initial iteration steps $m = 0, 2, 8$. The bins of each histogram is the same as those used in Fig. 4.2, but the height above the $j$-th bin $B_j$ at the iteration step $m$ is

$$h_{mj} = \frac{1}{\|b\|} \sqrt{\sum_{\lambda_i \in B_j} |\bar{r}_{mi}|^2} = \frac{1}{\|b\|} \sqrt{\sum_{\lambda_i \in B_j} |p_m(\lambda_i) \, \bar{r}_{0i}|^2}. \tag{5.9}$$

The quantity (5.9) is the norm of the projection of $r_m/\|b\|$ onto the space spanned by the eigenvectors $v_i$ corresponding to $\lambda_i \in B_j$. Therefore, $\|r_m\|/\|b\| = \sqrt{\sum_j |h_{mj}|^2}$. Note that if the bin $B_j$ contains a root of the residual polynomial $p_m$, the bar height (5.9) almost vanishes because $p_m(\lambda_i) \simeq 0$ for $\lambda_i \in B_j$.

Fig. 5.2 shows that $\bar{r}_0$'s for $s = 0$ and $s = -1$ have components only for $\lambda \gtrsim 0$, which is in agreement with the eigenvalue spectra for the two $s$ shown in Fig. 4.2. As the iteration step $m$ increases, GMRES constructs the residual polynomials $p_m$ (shown by the solid curves in the figures[1]) so that they have roots in the interval $[0, 2]$ where the components of $\bar{r}_0$ are nonzero. The second column in Fig. 5.2 clearly shows that

[1] The residual polynomials shown in Figs. 5.2 and 5.4 are calculated by solving optimization problems using CVX [5].




(a) Smallest root less close to zero (b) Smallest root closer to zero (c) Positive and negative roots

FIG. 5.3. *Residual polynomials $p_m$ of degree 6 with different locations of the roots. Solid lines represent residual polynomials. Black dots on the horizontal axis indicate the locations of the roots. The residual polynomials always satisfy $p_m(0) = 1$. In (a) and (b), the roots are all positive, but the smallest root is closer to zero in (b) than in (a); the locations of the other five roots are the same for the two cases. Note that the lines tangential to the polynomials at the roots become steeper as the smallest root moves toward zero, and thus the variation of the residual polynomial becomes more rapid. In (c), the residual polynomial with both positive and negative roots is illustrated. The amplitudes of the oscillation of the residual polynomial is much larger in (c) than in (a) and (b).*

the bars almost vanish at the roots of the residual polynomials. As the iteration step $m$ increases, GMRES puts more and more roots of $p_m$ in the interval $[0, 2]$, and the components of $\bar{r}_m$ decrease in magnitude overall.

However, it is noticeable in Fig. 5.2a that the peak at $\lambda \simeq 0$ does not decrease with increasing $m$ for $s = 0$. This is because the penalty for the residual polynomial's having a root very close to zero is too expensive as explained intuitively below. Because the residual polynomial $p_m$ satisfies the condition (5.2), it is factored as

$$p_m(x) = \prod_{i=1}^{d_m}(1 - x/\zeta_i), \tag{5.10}$$

where $d_m \leq m$ is the degree of $p_m$ and $\zeta_i$'s are the roots of $p_m$. The first derivative of $p_m(x)$ is

$$p'_m(x) = -\sum_{j=1}^{d_m} \tfrac{1}{\zeta_j} \prod_{i\neq j}(1 - x/\zeta_i), \tag{5.11}$$

and thus the slope of the tangential line to the polynomial at a root $\zeta_k$ is

$$p'_m(\zeta_k) = -\tfrac{1}{\zeta_k} \prod_{i\neq k}(1 - \zeta_k/\zeta_i). \tag{5.12}$$

Now, suppose that $0 < \zeta_1 < \cdots < \zeta_{d_m}$. We can easily show that $|p'_m(\zeta_k)|$ increases when $\zeta_1$ decreases toward zero. In other wards, for a given residual polynomial whose roots are all positive, if we make the smallest root even smaller, then the variation of the polynomial at the roots becomes more rapid. This situation is illustrated in Figs. 5.3a and 5.3b for a residual polynomial of degree 6.

The above discussion means that the peak at $\lambda \simeq 0$ in Fig. 5.2a can be resolved by the residual polynomials only at the price of making the other bars higher. Therefore, the peak remains unresolved until the other bars become sufficiently small. This makes $\|r_m\|/\|b\|$ for $s = 0$ stagnates during the initial GMRES iteration as shown in Fig. 5.1, because the peak at $\lambda \simeq 0$ constitutes a significant portion of the initial residual norm $\|r_0\|/\|b\|$ as indicated in the $m = 0$ figure in Fig. 5.2a. In fact, the bar at $\lambda \simeq 0$ remains unresolved also for $s = -1$, but this does not stagnates $\|r_m\|/\|b\|$, because the bar constitutes only a small portion of $\|r_0\|/\|b\|$ as can be seen in the $m = 0$ figure in Fig. 5.2b.

We next visualize $\bar{r}_m$ for $s = +1$ at several iteration steps $m$ in Fig. 5.4. It looks like $\bar{r}_0/\|b\|$ has components only for $\lambda \lesssim 0$ as shown in the $m = 0$ figure, which seems contradictory to the eigenvalue spectrum for $s = +1$ shown in Fig. 4.2c. In fact, $\bar{r}_0/\|b\|$ has components for $\lambda \gtrsim 0$, but they are much smaller than the components for $\lambda \lesssim 0$ and invisible in the $m = 0$ figure. Because most components of $\bar{r}_0$ correspond to $\lambda \lesssim 0$, initially GMRES resolves these components by placing the roots of the residual polynomials in the interval $[-2, 0]$ as shown in the $m = 2$ figure.

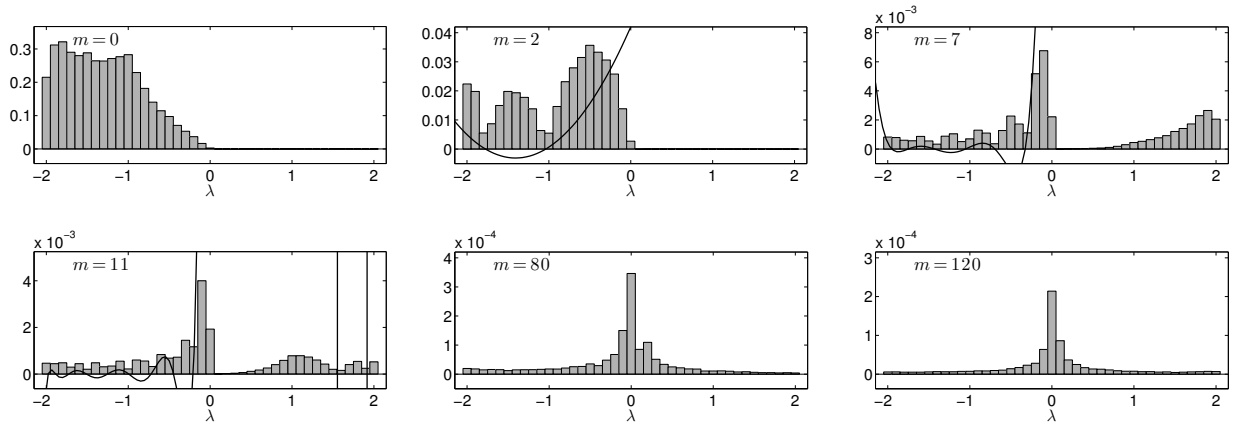



FIG. 5.4. *Evolution of $\bar{r}_m$ for $s = +1$. Bars whose heights are $h_{mj}$ of (5.9) are plotted for the six iteration steps $m = 0, 2, 7, 11, 80, 120$. In the figures for $m = 2, 7, 11$, the residual polynomials $p_m$ are also plotted as solid curves. The values of $p_m$ are scaled appropriately to show the oscillation and roots of the polynomials clearly. Notice that the peak at $\lambda \simeq 0$ in the $m = 80$ figure has barely decreased after 40 steps in the $m = 120$ figure.*

However, such polynomials have very large values in the $[0, 2]$ interval, so the components of $\bar{r}_m/\|b\|$ corresponding to $\lambda \gtrsim 0$ are gradually amplified as the iteration proceeds as indicated in the $m = 7$ figure. Eventually, the components of $\bar{r}_m/\|b\|$ for $\lambda \gtrsim 0$ become so amplified that they cannot be ignored, and GMRES starts to place the roots of the residual polynomials in the $[0, 2]$ interval as well.

As the iteration proceeds, GMRES places more and more roots on both the $[-2, 0]$ and $[0, 2]$ intervals. However, as can be seen in the $m = 80$ and $m = 120$ figures, the components of $\bar{r}_m/\|b\|$ corresponding to $\lambda \simeq 0$ persist even after long iteration. This persistent components has the same origin as the persisting peak at $\lambda \simeq 0$ we had in Fig. 5.2a; in other words, the components of $\bar{r}_m/\|b\|$ at $\lambda \simeq 0$ cannot be resolved easily because such resolution makes the other components of $\bar{r}_m/\|b\|$ larger.

Moreover, since the eigenvalue spectrum for $s = +1$ spans to both positive and negative values as shown in Fig. 4.2c, the roots of the residual polynomial need to be placed very close to zero on both positive and negative sides to fully resolve the components corresponding to $\lambda \simeq 0$. Therefore, the penalty for resolving the components of $\bar{r}_m/\|b\|$ corresponding to $\lambda \simeq 0$ is even more expensive for an indefinite operator than a definite operator (see Fig. 5.3c). Hence, the peak at $\lambda \simeq 0$ is harder to resolve for $s = +1$ than for $s = -1$. As a result, $\|r_m\|/\|b\|$ for $s = +1$ stagnates in the GMRES iteration process for a far extended period than it does for $s = 0$ as demonstrated in Fig. 5.1.

In summary of this section, we have shown that there are two kinds of stagnation of the GMRES convergence: one caused by a high multiplicity of eigenvalues near zero and the other caused by an indefinite coefficient matrix. We have also shown that the latter kind of stagnation takes much longer iteration than the first kind. The two kinds of stagnation explain the reason $s = -1$ induces faster convergence than $s = 0$ and $s = +1$.

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

The operator of the newly formulated equation does not have the high multiplicity of the near-zero eigenvalues that stagnates the convergence for the original operator. Especially, unlike the previously known approach that similarly removes the high multiplicity, our method leaves the operator nearly positive-semidefinite as the original operator, which results in a better performance than the previous approach.

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

**Appendix A.** Using the **k**-space representations of the operators, here we derive the eigenvalues (4.5) of $\nabla \times (\nabla \times \quad)$ and (4.6) of $\nabla(\nabla \cdot \quad)$ as well as their corresponding eigenfunctions.

Because both $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ are translationally invariant operators, their eigenfunctions vary

harmonically in space, and thus have the form

$$\mathbf{F}_{\mathbf{k}}\, e^{-i\,\mathbf{k}\cdot\mathbf{r}}, \tag{A.1}$$

where $\mathbf{r}$ is a position variable, $\mathbf{k} = \hat{\mathbf{x}}\,k_x + \hat{\mathbf{y}}\,k_y + \hat{\mathbf{z}}\,k_z$ is a real constant wavevector, and $\mathbf{F}_{\mathbf{k}} = \hat{\mathbf{x}}\,F^x_{\mathbf{k}} + \hat{\mathbf{y}}\,F^y_{\mathbf{k}} + \hat{\mathbf{z}}\,F^z_{\mathbf{k}}$ is a vector that is constant for a given $\mathbf{k}$.

Substituting (A.1) for $\mathbf{F}$ in the eigenvalue equations $\nabla\times(\nabla\times\mathbf{F}) = \lambda\,\mathbf{F}$ and $\nabla(\nabla\cdot\mathbf{F}) = \lambda\,\mathbf{F}$, we represent the equations in the $\mathbf{k}$-space. For $\nabla\times(\nabla\times\quad)$ we have

$$\begin{bmatrix} k_y^2+k_z^2 & -k_xk_y & -k_xk_z \\ -k_yk_x & k_z^2+k_x^2 & -k_yk_z \\ -k_zk_x & -k_zk_y & k_x^2+k_y^2 \end{bmatrix} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \lambda \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix}, \tag{A.2}$$

and for $\nabla(\nabla\cdot\quad)$ we have

$$-\begin{bmatrix} k_x^2 & k_xk_y & k_xk_z \\ k_yk_x & k_y^2 & k_yk_z \\ k_zk_x & k_zk_y & k_z^2 \end{bmatrix} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \lambda \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix}. \tag{A.3}$$

The matrices on the left-hand sides of (A.2) and (A.3) are the $\mathbf{k}$-space representations of the operators $\nabla\times(\nabla\times\quad)$ and $\nabla(\nabla\cdot\quad)$.

By solving (A.2) and (A.3) for a given $\mathbf{k}$, we obtain

$$\lambda = 0\,,\ \ |\mathbf{k}|^2\,,\ \ |\mathbf{k}|^2 \tag{A.4}$$

and

$$\lambda = -|\mathbf{k}|^2\,,\ \ 0\,,\ \ 0 \tag{A.5}$$

as the eigenvalues (4.5) of $\nabla\times(\nabla\times\quad)$ and (4.6) of $\nabla(\nabla\cdot\quad)$, respectively. It turns out that the two sets of the eigenvalues correspond to the same set of eigenfunctions in the form of (A.1) with

$$\mathbf{F}_{\mathbf{k}} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix},\ \ \begin{bmatrix} k_z \\ 0 \\ -k_x \end{bmatrix},\ \ \begin{bmatrix} -k_y \\ k_x \\ 0 \end{bmatrix}. \tag{A.6}$$

We note from (A.4) and (A.5) that $\nabla\times(\nabla\times\quad)$ and $\nabla(\nabla\cdot\quad)$ are positive-semidefinite and negative-semidefinite, respectively.

## REFERENCES


[1] K. Beilenhoff, W. Heinrich, and H. Hartnagel, *Improved finite-difference formulation in frequency domain for three-dimensional scattering problems*, Microwave Theory and Techniques, IEEE Transactions on, 40 (1992), pp. 540–546.

[2] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta numerica, 14 (2005), pp. 1–137. Section 9.2.

[3] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, *GMRES and the minimal polynomial*, BIT Numerical Mathematics, 36 (1996), pp. 664–675.

[4] Roland Freund and Noël Nachtigal, *Qmr: a quasi-minimal residual method for non-hermitian linear systems*, Numerische Mathematik, 60 (1991), pp. 315–339.

[5] Michael Grant and Stephen Boyd, *CVX Web page*, 2012. http://cvxr.com/cvx.

[6] D. A. H. Jacobs, *A generalization of the conjugate-gradient method to solve complex systems*, IMA Journal of Numerical Analysis, 6 (1986), pp. 447–452.

[7] Youcef Saad and Martin H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[8] Valeria Simoncini and Daniel B. Szyld, *Recent computational developments in krylov subspace methods for linear systems*, Numerical Linear Algebra with Applications, 14 (2007), pp. 1–59.