Progress In Electromagnetics Research C, Vol. 123, 1–16, 2022

A Normal-Vector-Field-Based Preconditioner for a Spatial Spectral Domain-Integral Equation Method for Multi-Layered Electromagnetic Scattering Problems

Ligang Sun*, Roeland J. Dilz, and Martijn C. van Beurden

Abstract—A normal-vector-field-based block diagonal-preconditioner for the spatial spectral integral method is proposed for an electromagnetic scattering problem with multi-layered medium. This preconditioner has a block-diagonal matrix structure for both 2D TM polarization and 3D cases. Spectral analysis shows that the preconditioned system has a more clustered eigenvalue distribution, compared to the unpreconditioned system. For the cases with high contrast or negative permittivity, numerical experiments illustrate that the preconditioned system requires fewer iterations than the unpreconditioned system. The total computation time is reduced accordingly while the accuracy based on the normal-vector field formulation of the solution is preserved.

1. INTRODUCTION

In electrical engineering Maxwell solvers for electromagnetic scattering problems have wide and important applications, which range from semiconductor metrology in integrated circuits (ICs) production [1–3], to designing elements on nanophotonic chips [4, 5], and to analysing metamaterials [6, 7]. In these cases, it is required to have fast and accurate Maxwell solvers, especially for the cases where the number of unknowns is large.

Different types of Maxwell solvers have been developed in the past decades to solve electromagnetic scattering problems. When the incident fields and solutions are stationary or time-harmonic, one can solve the problem with a frequency-domain Maxwell solver. The frequency-domain solver can be more computationally efficient than a time-domain Maxwell solver, and it can be divided into two categories. The first kind relies on a differential form of Maxwell's equations, popular methods in this first category are the finite-difference (FD) [8] and finite element methods (FEM) [9]. The second category depends on an integral-equation formulation of Maxwell's equations, which incorporates the Green function and the volume is restricted to the support of the sources of the electromagnetic field. Both domain integral equations [10, 11] and surface integral equations [12] belong to the latter category.

In [13–15], a spatial spectral method is proposed to solve two-dimensional (2D) transverse electric (TE), 2D transverse magnetic (TM) and three-dimensional (3D) scattering problems in a layered medium, respectively. The main differences between this method and other volume integral equation solvers are: (1) a Gabor frame is used as a discretization in the transverse plane, which brings a fast and accurate Fourier transformation; and (2) a spectral integration path is chosen to avoid the singularities of the Green function in the spectral domain. The accuracy is improved by introducing an auxiliary field based on the local normal-vector field (NVF) formulation [16].

The above spatial spectral discretization approach leads to a high-dimensional linear system of equations. Usually iterative methods such as GMRES [17], BiCG-type methods [18–20], or IDR(s) [21]

Received 19 May 2022, Accepted 7 August 2022, Scheduled 14 August 2022
* Corresponding author: Ligang Sun (l.sun1@tue.nl).
The authors are with the Department of Electrical Engineering, Eindhoven University of Technology, the Netherlands.

are deployed to solve these large linear systems instead of a direct method [22]. For each iteration, this spatial spectral solver reaches a computational complexity of $O(N \log N)$ in terms of the matrix-vector product. However, convergence difficulties are observed in terms of a large number of iterations when the underlying physical problem has high-contrast or negative-permittivity scatterers embedded in the layered medium, or when the scatterer is large. Preconditioning is usually a vital component for high-dimensional linear systems with a poor convergence rate, to enable practical computations within a reasonable time [23]. A good preconditioner transforms the original system into a system that has the same solution, but exhibits better convergence performance. Furthermore, constructing and executing this preconditioner should be fast because it will be performed in every iteration as an extra matrix-vector product (MVP).

Optimal circulant preconditioners have been successfully used in domain integral equations in one-dimensional (1D) and 2D TE, or E-polarized cases to accelerate the iteration, see [24] and [25]. Circulant-type preconditioners have also proved effective to solve the system in the form of $I - GX$ with multi-level Toeplitz structure [26]. For scattering in periodic setups, the integral-equation formulation in the transverse directions exploits a continuous auxiliary field formulation together with a normal-vector field around object boundaries [16, 27]. In that case, the linear system corresponding to the integral equation can be written in the form $(C - GM)\mathbf{u} = \mathbf{f}$, where the matrices C and M are block-Toeplitz-Toeplitz-block (BTTB) matrices, and the matrix G represents the Green operator. In [28], the matrix C^{-1} and its approximations have been proposed as preconditioners and promising improvements were obtained after deploying these preconditioners. For the nonperiodic case, the spatial spectral method based on Gabor frames and an auxiliary field in combination with a normal-vector field formulation [14, 15] bears a close resemblance to the case of fully spectral methods for periodic structures. Therefore, it is a natural idea to extend the application of the C^{-1} preconditioner in [28] to the Gabor-frame based spatial spectral solver, which is the main objective of this paper. To be specific, we show that this NVF-based preconditioner has a block-diagonal structure and we illustrate that this preconditioner can reduce the number of iterations, while preserving the accuracy of the solution for high contrasts or negative permittivities.

This paper is organized as follows. In Section 2 we recall the most important details of the 2D TM and 3D spatial spectral Maxwell solver and we establish the NVF-based preconditioner. In Section 3 we discuss the effects of this NVF-based preconditioner based on spectral analysis. Numerical experiments are discussed in Section 4, which contains three experiments for which we show the reduction in the number of iterations, an accuracy validation, and a comparison in computation time. Section 5 contains the conclusions.

2. FORMULATION

Consider the following 2D or 3D scattering problem in Fig. 1. A multi-layered dielectric medium is placed in between two dielectric half-spaces. We define a Cartesian coordinate system such that all layers are stacked along the z direction as background materials, and each layer i ($1 \leq i \leq N$) has a constant relative permittivity ε_{rbi}. A scattering object, which is made of a different material, is located in a finite domain $D \in \mathbb{R}^3$ and is completely embedded within layer i. The relative permittivity of the scatterer is ε_{rs} and one can define a global relative permittivity function $\varepsilon_r(\mathbf{x})$ to distinguish all materials, where $\mathbf{x} = (x, y, z)$ denotes the spatial coordinates. In the absence of the scatterer, the incident electric field $\mathbf{E}^i(\mathbf{x})$ can be calculated as in [29].

2.1. Summary of the Spatial Spectral Method

The spatial spectral method [30] is developed based on the following domain integral representation:

$$\begin{aligned}\mathbf{E}^i(\mathbf{x}_T, z) =& \mathbf{E}(\mathbf{x}_T, z) \\ & -\mathcal{F}_T^{-1}\left\{\int_{\mathbb{R}} G(z'|\mathbf{k}_T, z) \cdot \mathcal{F}_T[\mathbf{J}(\mathbf{x}_T, z')]dz'\right\}\end{aligned} \tag{1}$$

where $\mathbf{x}_T$ denotes the spatial Cartesian coordinates in the transverse plane (i.e., $\mathbf{x}_T = (x, y)$ in the 3D case and $\mathbf{x}_T = x$ in the 2D case), and similarly, $\mathbf{k}_T$ denotes the spatial Fourier transform variables

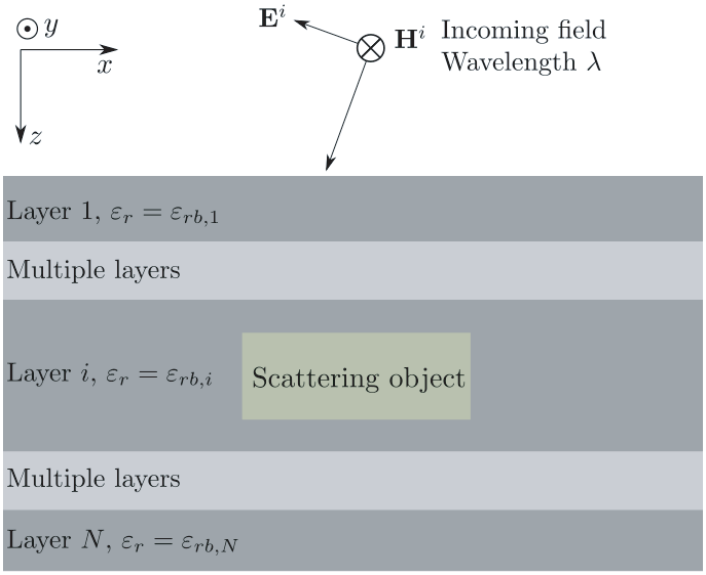

Figure 1. Geometric setting for a multi-layer medium with embedded scattering object.

in the transverse direction, i.e., with respect to $\mathbf{x}_T$. Note that $\mathcal{F}_T$ and $\mathcal{F}_T^{-1}$ denote a pair of Fourier transformations in the transverse plane between $\mathbf{x}_T$ and $\mathbf{k}_T$. G is the spectral-domain Green operator in the multi-layered medium. $\mathbf{E}(\mathbf{x}_T, z)$ represents the unknown total electric field, $\mathbf{J}(\mathbf{x}_T, z)$ is the contrast current density given by the spatial field-material interaction:

$$\mathbf{J}(\mathbf{x}_T, z) = j\omega\varepsilon_0\varepsilon_{rbi}\chi(\mathbf{x}_T, z)\mathbf{E}(\mathbf{x}_T, z), \tag{2}$$

where ε_0 is the permittivity of free space, and ε_{rbi} is the relative permittivity of the ith homogeneous layer in the background medium that contains the scatterers. Given the relative permittivity function $\varepsilon_r(\mathbf{x}_T, z)$, the contrast function $\chi(\mathbf{x}_T, z)$, bound to layer i, is defined as

$$\chi(\mathbf{x}_T, z) = \frac{\varepsilon_r(\mathbf{x}_T, z)}{\varepsilon_{rbi}(z)} - 1, \tag{3}$$

which is only supported on the domain of the scatterer.

One important feature of this spatial spectral method is that a Gabor frame is used to perform discretization in the transverse plane in both spatial and spectral domains. A Gabor-frame expansion for any $f(\mathbf{x}_T) \in \mathrm{L}^2(\mathbb{R}^2)$ in the spatial domain is

$$f(\mathbf{x}_T) = \sum_{\mathbf{m},\mathbf{n}} f_{\mathbf{m},\mathbf{n}} g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T), \tag{4}$$

in which $\mathbf{m}$ and $\mathbf{n}$ represent the spatial and the spectral shift number, respectively, $g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ is a Gabor frame function and $f_{\mathbf{m},\mathbf{n}}$ is a Gabor coefficient. Gabor coefficients are computed via the Gabor transformation:

$$f_{\mathbf{m},\mathbf{n}} = \int f(\mathbf{x}_T)\eta^*_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)d\mathbf{x}_T, \tag{5}$$

where $\eta_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ is the dual frame function and is computed via the Moore-Penrose inverse [31]. Full representations of the Gabor frame function and its dual frame function can be found in [14, 15]. The main advantage of this Gabor-frame-based discretization is that it establishes a fast relation between the spatial domain and the spectral domain. The Fourier transform of a spatial Gabor frame function $g_{\mathbf{m},\mathbf{n}}(\mathbf{x}_T)$ yields a Gabor frame in the spectral domain, and the Gabor coefficients of the spectral function $\hat{f}(\mathbf{k}_T)$ can be readily obtained via simple operations on the spatial Gabor coefficients $f_{\mathbf{m},\mathbf{n}}$ [32]. This property guarantees fast transformations between the spatial and the spectral domains and eventually contribute to the $O(N \log N)$ computational complexity for the matrix-vector product of the spatial

spectral method, where N represents the total number of unknowns after discretization. In [33], a set of basis functions is calculated based on equidistant Dirac delta test functions and an approximation of the exact Gabor-based discretization is introduced in [15] for 3D scattering problems. These new basis functions yield faster operations like multiplication and FFT-based Fourier transformation, which reduces the computation time and preserves accuracy.

In the z direction, the integral in Eq. (1) is discretized in terms of piecewise-linear (PWL) expansion functions:

$$\Lambda(z) = \begin{cases} 1 - \dfrac{|z - p\Delta_z|}{\Delta_z} & \text{if } |z - p\Delta_z| < \Delta_z \\ 0 & \text{if } |z - p\Delta_z| > \Delta_z \end{cases}, \tag{6}$$

where Δ_z is the discretization step in the z direction, and $1 \leq p \leq N_z$ denotes the index of the sample points along the z direction. N_z denotes the total number of sample points in the z direction. Another feature of the spatial spectral method is that a deformed integration path on the complex plane is chosen as alternative to an integration path on the real axis, to properly handle the branch cuts of the dielectric half-spaces and the poles that represent guided waves of the layered medium. Based on the reflection interfaces within a multi-layered medium [29], effective reflection coefficients are defined in [13–15]. Representing the Green function in the spectral domain along this integration path avoids the tedious calculation of Sommerfeld integrals.

To improve the accuracy and efficiency of the Gabor expansion in the presence of discontinuous permittivities in the transverse plane, a local normal-vector field formulation is used in this spatial spectral method. Based on the Li rules [34], which provide a framework to assess whether functions with discontinuities can be multiplied or not, the normal-vector field formulation was introduced by Popov and Nevière [27] to improve the convergence in Fourier analysis. The main idea of the normal-vector field formulation is to perform spatial multiplications on the continuous components of the electric field $\mathbf{E}$ and the electric flux density $\mathbf{D}$, which together constitute the auxiliary field $\mathbf{F}$, and then derive their discontinuous components from the multiplication by the field-material interactions. The normal-vector field $\mathbf{F}$ can then be transformed to the total electric field $\mathbf{E}$ and the contrast current function $\mathbf{J}$ through

$$\begin{aligned} \mathbf{E} &= C\mathbf{F}, \\ \mathbf{J} &= M\mathbf{F}. \end{aligned} \tag{7}$$

Explicit expressions of components of matrices C and M expressed in Cartesian coordinates are given in [14, 15] and [16].

2.2. The NVF-Based Block-Diagonal Preconditioner

Based on the domain integral representation (1) and the normal-vector field formulation (7), the spatial spectral method can be represented by the following linear system:

$$L\mathbf{u} = \mathbf{f}, \tag{8}$$

where $L \in \mathbb{C}^{N\times N}$ is the system matrix; the inhomogeneous term $\mathbf{f} \in \mathbb{C}^N$ represents the incident field $\mathbf{E}^i$; $\mathbf{u} \in \mathbb{C}^N$ contains the expansion coefficients of the auxiliary field $\mathbf{F}$ to be determined; and N represents the number of unknowns. The system matrix A can be decomposed as

$$L = C - G \cdot M, \tag{9}$$

where C and M transform the normal-vector field $\mathbf{F}$ into the total electric field $\mathbf{E}$ and the contrast current $\mathbf{J}$ through Eq. (7), and G denotes the Green tensor operation in combination with a pair of Fourier transformations. In the spatial spectral solver [30], the matrix L is implemented implicitly to avoid storing a full system matrix.

The structures of matrices L, C, G, M depend on the order of the discretization indexes associated with either the transverse plane or the z direction. When choosing the index associated to z-samples as the outermost one, i.e., the slowest changing index when moving row-wise or column-wise, matrices C and M have a block-diagonal structure with each block containing the Gabor coefficients of the operators related to the contrast χ (defined in [14] and [15]). The block-diagonal structure essentially comes from the direct (spatial) multiplication between the χ-related operators and the auxiliary field

F per z sample. The Green matrix G contains the Gabor transformation of the homogeneous-medium Green tensor and the reflected waves from the layer interfaces [13, 15] and therefore it has a denser structure at the block level. On the other hand, the fact that the Gabor frames have effectively a finite support in the spectral domain yields some sparsity per block of the matrix G. Simplified structures of matrices L, C, G and M are given in Fig. 2.

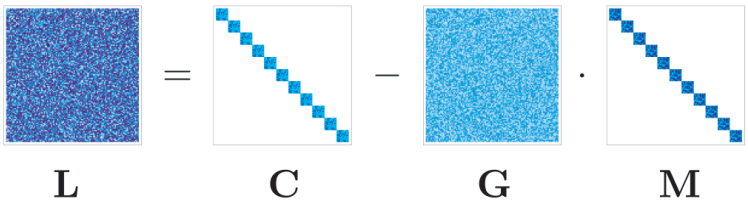

Figure 2. Sparsity patterns of the matrices L, C, G, and M.

In [28], the system of a light scattering problem for a 2D-periodic structure was also represented in the form $C - GM$, but then for an expansion in terms of (discrete) Fourier modes. For that case, the matrix C is a block-diagonal matrix and each diagonal block of the matrix C is a so-called BTTB-block matrix. Preliminary investigations have shown that the number of iterations can be reduced significantly by taking the full inverse matrix C^{-1} as a preconditioner. Since the Gabor-frame transformation is a unitary transformation [35], it is natural to transfer the idea of this C^{-1} preconditioner to the spatial spectral method, where the Gabor representation is used in the transverse plane.

The N_z block matrices of C in Fig. 2 come from the N_z sampling points in z direction. Each block corresponds to the Gabor transformation of the function $\chi(\mathbf{x}_T, z_p)$ and the normal-vector field in the transverse plane with some fixed $z = z_p$ $(1 \leq z_p \leq N_z)$. Therefore, for a dielectric scatterer that has a uniform cross section in the z direction, the block submatrices of C are identical to each other. Together with the sparsity, owing to the block-diagonal structure, one can readily see that the matrix C^{-1} can be constructed by inverting one block submatrix of C. This simplifies the computational procedure in practice and makes C^{-1} a good candidate to precondition the original system $L\mathbf{u} = \mathbf{f}$. Hence we refer to the matrix C^{-1} as the normal-vector-field-based block-diagonal (NVF-BD) preconditioner for the spatial spectral method.

3. AN INDICATION OF A CLUSTERED SPECTRUM

It is well known that the convergence rate of an iterative method highly depends on the distribution of the eigenvalues of the system matrix: the more clustered the spectrum is, the faster the convergence rate will be, see [36, Chapter 1]. Hence, a good preconditioner should yield a clustered spectrum for the preconditioned system matrix and result in an increased convergence rate. The clustering effect has been studied in detail for various types of preconditioners, e.g., circulant preconditioners [37–39], Toeplitz preconditioners [40–42], and block Toeplitz preconditioners [43]. Following the analysis for the above preconditioners, we compare the eigenvalue distributions of the systems with and without applying the NVF-BD preconditioner.

To this end, we consider a 2D scattering problem as presented in Fig. 3, where one rectangular scatterer is embedded in a layer of SiO_2 enclosed by two vacuum half-spaces. The incident field is a plane wave with wavelength 425 nm that is normally incident with respect to the xy plane and the incident electric field $\mathbf{E}^i$ is polarized along the x direction. The substrate medium SiO_2 has a relative permittivity $\varepsilon_{rb} = 2.16$, and the scatterer has a relative permittivity $\varepsilon_r = 54$. Therefore the scatterer has contrast $\chi = 24$ and its length in the x direction is 200 nm. With such a high-contrast case we expect a better conditioned system after applying the NVF-BD preconditioner. In this example there are 15990 unknowns.

In Fig. 4 we compare the absolute values and real parts of the eigenvalues. Both the original system and the preconditioned system are indefinite but not strongly: among the 15990 eigenvalues only 14 of them have negative real parts. Throughout the rest of this article, 'org' represents the original system

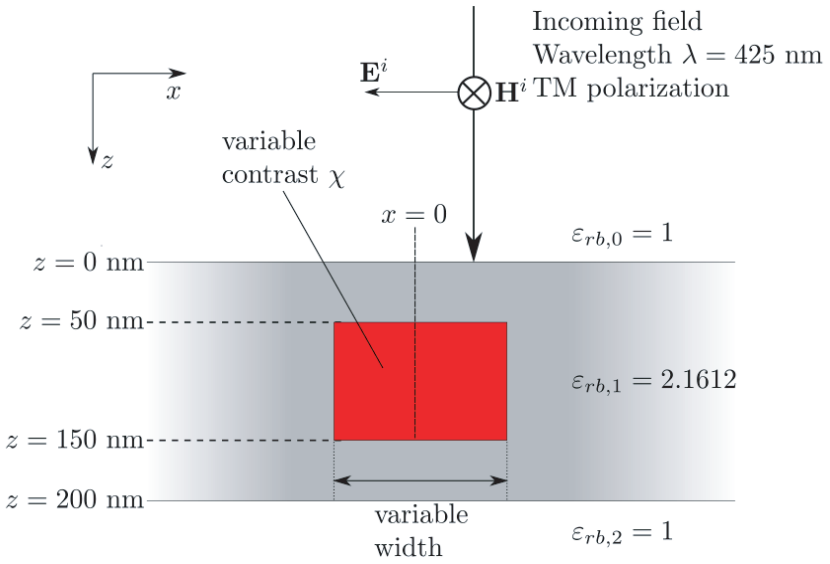

Figure 3. Scattering setup: a 2D TM polarized field is incident on a dielectric object (in red) embedded in a layered medium composed of SiO_2 and vacuum. $\varepsilon_{rb,i}$, for $0 \leq i \leq 2$, denotes the relative permittivities of these layers.

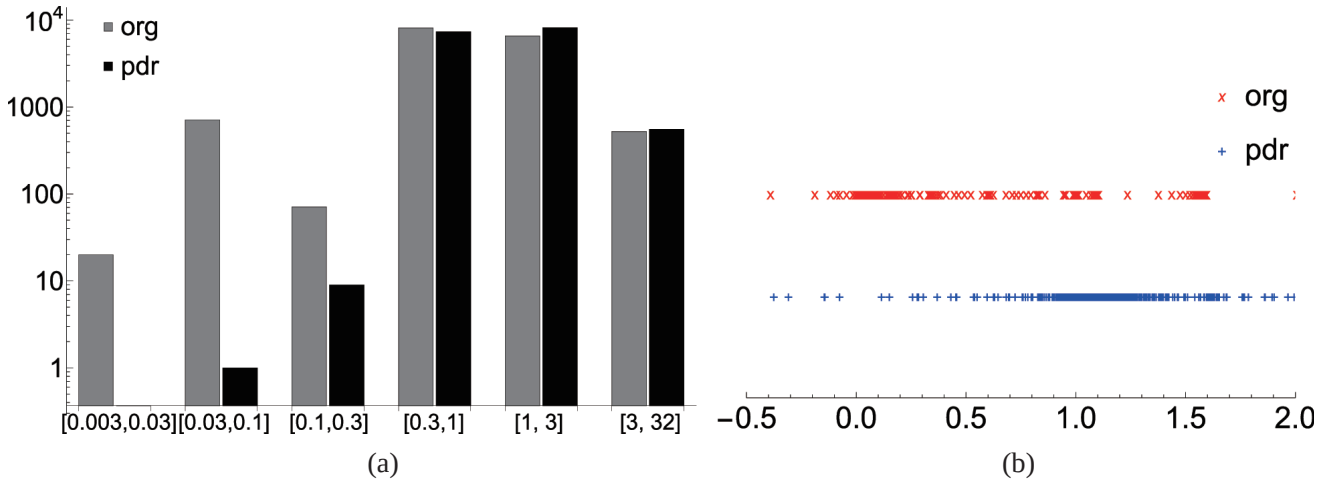

Figure 4. Comparison of eigenvalue distributions: (a) number of eigenvalues with absolute value located in per indicated interval, for the original system (org) and the preconditioned system (pdr). Note that the minimum and the maximum absolute eigenvalues of the original system are 0.0038 and 25, and the counterparts for the preconditioned system are 0.076 and 25. (b) Real parts of the eigenvalues $\mathrm{Re}(\lambda_i)$ in the range $[-0.5, 2]$.

and 'pdr' represents the system after applying the NVF-BD preconditioner. In Fig. 4(a), the horizontal axis denotes six intervals ranging from 0.003 to 32, and the vertical axis represents the number of the eigenvalues, which absolute values belong to each of the corresponding intervals, on a log scale. A significant difference is observed when comparing their minimum absolute eigenvalues: the minimum absolute eigenvalue is shifted away from the origin from 3.8×10^{-3} to 7.6×10^{-2}. In Fig. 4(b), we see the real part of those eigenvalues that satisfy $-0.5 \leq \mathrm{Re}(\lambda_i) \leq 2$. It is clear that without the NVF-BD preconditioner, the original system has much more eigenvalues close to 0, while after applying the NVF-BD preconditioner, only a few eigenvalues around 0 remain and there are much more eigenvalues clustered around 1. The distribution of eigenvalues plays a crucial role in a system's conditioning,

especially when the maximum eigenvalue does not change dramatically.

Figure 4 also shows that the preconditioned system's spectrum is more clustered around 1. To verify this observation, we counted how many eigenvalues are located within the interval $[1-\delta, 1+\delta]$ for a given $\delta > 0$. We obtain the percentages by dividing by the total number of eigenvalues and compare them in Table 1. Note that in the original system only 36.5% of the eigenvalues were located within the disc centered at 1 with radius 10^{-8} in the complex plane, while this number becomes 71.5% after applying the NVF-BD preconditioner. Clearly, there is a stronger clustering of the eigenvalues around 1 in the preconditioned system. Analogous to the clustering effects studied in other preconditioners [37–43], we expect this promising indicator of the NVF-BD preconditioner can reduce the number of iterations as well.

Table 1. Percentage comparison for eigenvalues located within the interval $[1-\delta, 1+\delta]$, given δ as a parameter.

δ	% of org.	% of pdr.
10^{-2}	89.5	93.0
10^{-4}	85.4	91.6
10^{-6}	67.0	81.4
10^{-8}	36.5	71.5

4. NUMERICAL RESULTS

To show the effectiveness of the NVF-BD preconditioner, we have tested the preconditioned system on the following three scattering problems: (A) a 2D TM rectangular object with high contrast, (B) a 2D TM metal grating problem with negative permittivity and (C) a 3D bar-shaped object with high contrast. In all cases we mainly focus on the reduction in the number of iterations after applying the NVF-BD preconditioner. We also show the reduction in computation time in case (A), and compare the solution in case (B) with an independent reference. The iterative method used in all three cases is the BiCGstab(2) algorithm [20], with the maximum number of iterations set to 1250. In Table 2 we summarize all Gabor parameters used in these three problems. Note that case (A-1), case (A-2) and case (A-3) are three variants of case (A), which are used to demonstrate the NVF-BD preconditioner's effects on larger-scatter cases and computation time. Following the notations in [13–15], we use X, m and n to denote the Gabor window length, the spatial shift number, and the frequency modulation number, respectively, and we use N_z to represent how many PWL functions are used in the z direction. Further, $p = 3$ and $q = 2$ are the oversampling parameters for the Gabor frames. Note that case (C) is a 3D problem and we use the same discretization parameters in both x and y directions.

Table 2. Discretization parameters used in simulation cases (A), (B) and (C).

case	X [nm]	m	n	N_z
(A)	100	$-7:7$	$-40:40$	41
(A-1)	100	varying	$-40:40$	41
(A-2)	100	$-5:5$	varying	101
(A-3)	100	$-5:5$	$-100:100$	varying
(B)	500	$-12:12$	$-40:40$	29
(C)	100	$-4:4$	$-10:10$	21

We define the relative error in step k, with corresponding solution vector $\mathbf{u}_k$, as $e_k = \frac{\|L\mathbf{u}_k - \mathbf{f}\|}{\|\mathbf{f}\|}$, with the system matrix L and the inhomogeneous term $\mathbf{f}$ introduced in Section 2, and $\|\cdot\|$ denotes the ℓ^2 norm of a vector. The iterative procedure is terminated once a relative error of 10^{-5} or less is reached.

It is known that different iterative methods yield differences in convergence behaviour, especially for high-contrast cases. However, comparing the difference in convergence of the various iterative methods is not the aim of this paper.

4.1. Case (A): a 2D TM High-Contrast Problem

In the first case we consider the 2D scattering problem in Fig. 3 again and keep the geometry parameters as introduced in Section 3. Table 2 displays the discretization parameters we used. Note that there are 81 frame functions used in each Gabor window length X, which yields a resolution of 1 nm in the x direction. In the z direction the PWL functions are employed with sample distance $\Delta_z = 2.5\,\text{nm}$. Discretization parameters are given under case (A) in Table 2.

To see the effect of the NVF-BD preconditioner on the number of iterations, we fix the object's size by taking its width $w = 200\,\text{nm}$ and change the value of the contrast χ. The contrast ranges from 2 to 64 and we are more interested in the high-contrast cases, since they are more challenging. We compare the number of iterations for the original solver and the preconditioned solver in Table 3. Due to the nature of the BiCGstab(2) algorithm, one iteration represents four matrix-vector products (MVPs). Note that in the low-contrast cases such as $\chi \leq 4$ the NVF-BD preconditioner saves about 50% of the iterations. For the cases where $8 \leq \chi \leq 24$ the total number of iterations is reduced by up to 90%, when $\chi \geq 32$ the unpreconditioned system fails to converge within 1250 iterations, whereas the NVF-BD preconditioner makes the solver converge within an acceptable number of iterations.

Table 3. Total number of iterations for Simulation case (A) for a scatterer with different contrast χ but the same geometric size. Note that "1250+" means the iterative solver fails to reach the desired relative error within 1250 iterations.

χ	org	pdr
2	7	4
4	17	9
8	50	22
16	245	66
24	1227	112
32	1250+	247
48	1250+	446
64	1250+	743

Figure 5 shows the evolution of the relative error versus the iteration count for the original system and the preconditioned system for the specific case $\chi = 24$, which corresponds to a relative permittivity $\varepsilon_r = 51.87$ for the rectangular scatterer. The horizontal axis denotes the number of iterations within the iterative solver, and the vertical axis denotes the relative error of the approximated solution at each iteration. Clearly, the preconditioned system significantly outperforms the original system in this high-contrast case. One possible reason for this significant reduction in number of iterations is that the NVF formulation plays a dominant role in the behavior of the iterative solver acting on the original system. The NVF-BD preconditioner improves the distribution of eigenvalues, as observed in Fig. 4, and also yields a much better conditioned system. The reduction in the number of iterations also saves a significant amount of computation time. We recorded the total computation times for this case on a single-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM. The original system takes 11,829.5 seconds, while the preconditioned system only needs 1,151.5 seconds to reach the desired relative error of $1 \cdot 10^{-5}$.

To explore the performance of the NVF-BD preconditioner on a larger scattering object, we change the scatterer's size and keep its contrast constant. The scatterer's width is changed from 200 nm to 1100 nm, which implies that the range of the x coordinate of the scatterer changes from $[-100, 100]\,\text{nm}$ to $[-550, 550]\,\text{nm}$. We set the spatial shift index m of the Gabor frame in Table 2 from $m = -8 : 8$ to

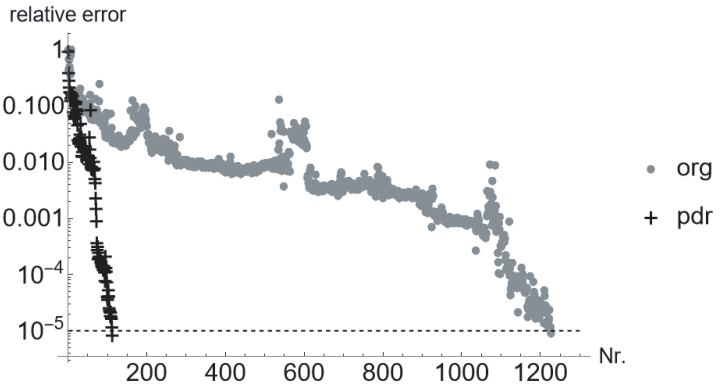

Figure 5. Convergence of the iterative solver for the high contrast case in Fig. 3 with $\chi = 24$. The dashed line denotes the desired relative error $1 \cdot 10^{-5}$.

$m = -12 : 12$, and therefore the corresponding computation domain is increased from $[-500, 500]$ nm to $[-850, 850]$ nm in the x direction, which covers the scatterer's domain and a part of its near field. Other discretization parameters are given under case (A-1) in Table 2. For all cases, the scatterer's contrast is kept at $\chi = 16$, which corresponds to a relative permittivity $\varepsilon_r = 36.74$. Table 4 presents the number of iterations for the original system and the preconditioned system. We observe that the preconditioned system can handle a significantly larger range of widths of the scatterer. Hence, the NVF-BD preconditioner reduces the number of iterations not only in cases of high contrast but also in cases where the scatterer has a larger width.

Table 4. Total number of iterations recorded for the scatterers with different width but with a constant contrast $\chi = 16$.

width [nm]	org	pdr
200	232	66
300	432	91
400	1102	229
500	970	258
600	1250+	333
700	1250+	527
900	1250+	930
1100	1250+	1030

Next, we investigate the NVF-BD preconditioner's effect on computation time. The total computation time equals to the initialization time and solution time. During the initialization of the NVF-BD preconditioner, the matrix C^{-1} is computed based on Doolittle LU factorization, and one only has to compute the inverse of a single block matrix of C since the contrast χ is a constant along its height. The total solution time is equal to the product of the average solution time per iteration and the total number of iterations. One may not obtain a significant reduction in the total computation time for the preconditioned system if the time per iteration increases a lot due to the extra four MVPs induced by the preconditioner per iteration. We compare the computation time per iteration for the original solver with that for the preconditioned system by considering the single scatterer case in Fig. 3 with $\chi = 8$ and width $w = 200$ nm. Note that each block matrix C_i $(1 \leq i \leq N_z)$ has dimension N_x, and the MVP of matrix block C_i has a quadratic complexity. Hence we expect the MVP with matrix C^{-1} should have a complexity of $\mathcal{O}(N_z N_x^2)$.

In Fig. 6, we compare the solution time per iteration and the extra computation time per MVP of

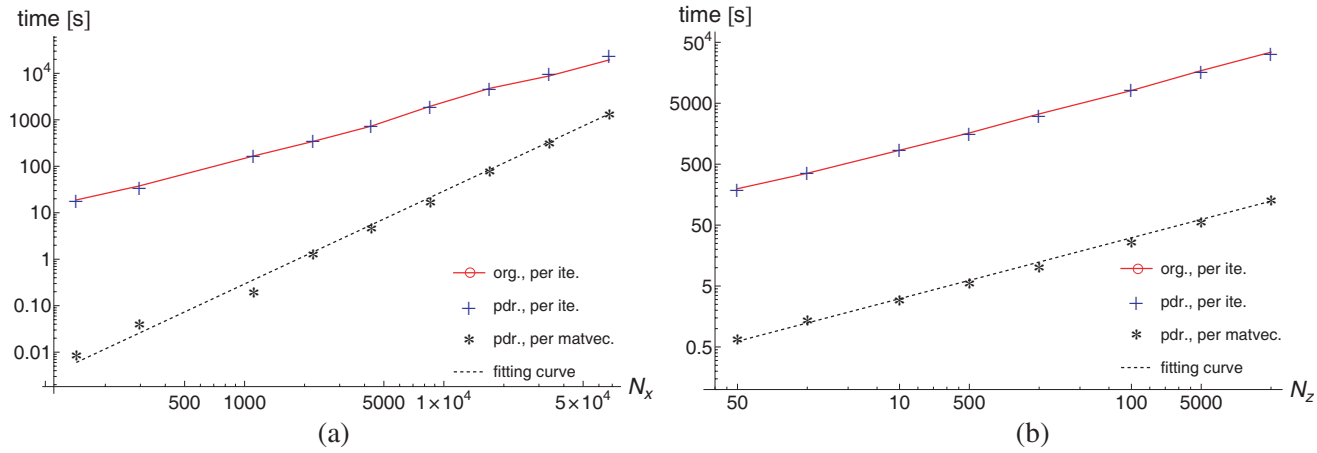

Figure 6. Average computation time per iteration for (a) N_x and (b) N_z unknowns. Note that "org. per ite." means the computation time per iteration for the unpreconditioned system, "pdr. per ite." means the computation time per iteration for the system after applying the NVF-BD preconditioner, "pdr. per matvec." means the time penalty due to the extra MVP in the preconditioned system, and "fitting curve" means the best fit based on the data points in the N_x and N_z cases, respectively.

the preconditioner by changing the number of unknowns in both x and z directions, respectively. The vertical axes denote the average solution time per iteration in seconds. The horizontal axis in Fig. 6(a) shows the number of unknowns N_x in the x direction. Note that $N_x = (2m+1)\cdot(2n+1)$, where the spatial shift index m satisfies $-5 \le m \le 5$, the frequency modulation index n satisfies $n \in \{-N, \ldots, N\}$ and N ranges from 6 to 3080. This corresponds to a resolution in the x direction that ranges from 6.231 nm to 0.013 nm. In the z direction, a total of 101 PWL functions are used with sample distance $\Delta_z = 1$ nm. All discretization parameters used in Fig. 6(a) are summarized under case (A-2) in Table 2. The horizontal axis in Fig. 6(b) denotes that N_z PWL functions are used in the calculation. N_z ranges from 10 to 2000, which corresponds to a resolution Δ_z in the z direction from 10 nm to 0.05 nm. In this case we have $-5 \le m \le 5$ and $-100 \le n \le 100$. All discretization parameters used in Fig. 6(b) are summarized under case (A-3) in Table 2. In both Figs. 6(a) and (b), the red dots and the blue crosses are computed based on the total solution time divided by the total number of iterations, and the gray stars are the computation time per MVP due to the preconditioner only. All the simulations were performed on a single-core Intel(R) Xeon(R) Gold 6148 CPU at 2.40 GHz with 755 GB RAM.

The average computation time per iteration for the original system and the preconditioned system, and the average computation time per MVP of the preconditioned system are displayed in Figs. 6(a) and (b). It is clear that the average computation time increases when a finer discretization is taken in either the x or z direction. The analytical representation of the fitting curve in Fig. 6(a) is $T(N_x) = 2.94 \cdot 10^{-7} N_x^2$ and in (b) it is $T(N_z) = 0.012 N_z$, where T is the average computation time per iteration. The recorded data points of the MVP time coincide with the fitting curve well. Therefore we confirm our prediction that the extra MVP operation in the preconditioned system has a complexity $\mathcal{O}(N_z N_x^2)$. Furthermore, we observe that both original system and preconditioned system have similar average computation time per iteration for almost the entire range of N_x and N_z cases, except for the last two data points in Fig. 6(a), where the time of the extra MVP due to the preconditioner is non-negligible compared with MVP of the original system and the other operations in the BiCGstab(2) iterative solver. In the z direction, a much larger N_z would be required to observe a similar effect, owing to the $\mathcal{O}(N_z N_x^2)$ complexity for the MVP of the preconditioner. In Fig. 7 we compare the total solution time for the original system and the preconditioned system. The vertical axes denote the total solution time in seconds. The horizontal axes represent the discretization parameters N_x and N_z in Figs. 7(a) and (b), respectively. Both figures suggest that in most cases (except for the cases with extremely large N_x) the total solution time can be reduced by a factor larger than 2, which is corresponding to the gained reduction factor in terms of the number of iterations for the $\chi = 8$ case in Table 3. For other cases in Table 3, the reduction factor in computation time is expected to be comparable to the reduction

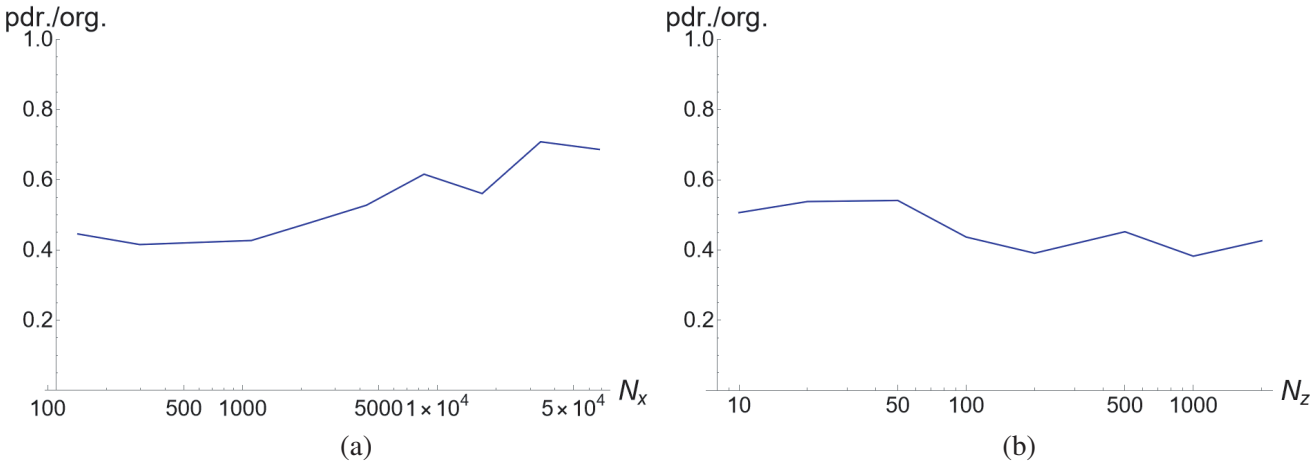

Figure 7. Comparison of computation time between systems "org" and "pdr". (a) Different discretization in the x direction. (b) Different discretization in the z direction.

factor in terms of the number of iterations, since the computation time per iteration in Figs. 6(a) and (b) is independent of the contrast χ.

We conclude that the total solution time can be reduced by applying the NVF-BD preconditioner. We also observe that the reduction in computation time for the preconditioned system gets lowered for large N_x cases due to the computational complexity of the preconditioner.

4.2. Case (B): A 2D TM Metal Grating Problem

In the second case, a grating device made of aluminium is embedded in air and supported by an aluminium half-space, see Fig. 8. A plane wave with wavelength 700 nm is incident under an angle of 22.9° with respect to the z-axis and the incident electric field $\mathbf{E}^i$ is polarized in the xz plane. The relative permittivity of aluminum is $-63.6 - 31.29j$. The air in the six grooves are considered as the scatterers, which have negative contrast $\chi = -1.0108 + 0.0064j$. Table 2 case (B) displays the discretization parameters we used in this simulation. Notice that in total 81 Gabor frame functions are used for each Gabor window length X, which yields a resolution of 5 nm in the x direction. In the z

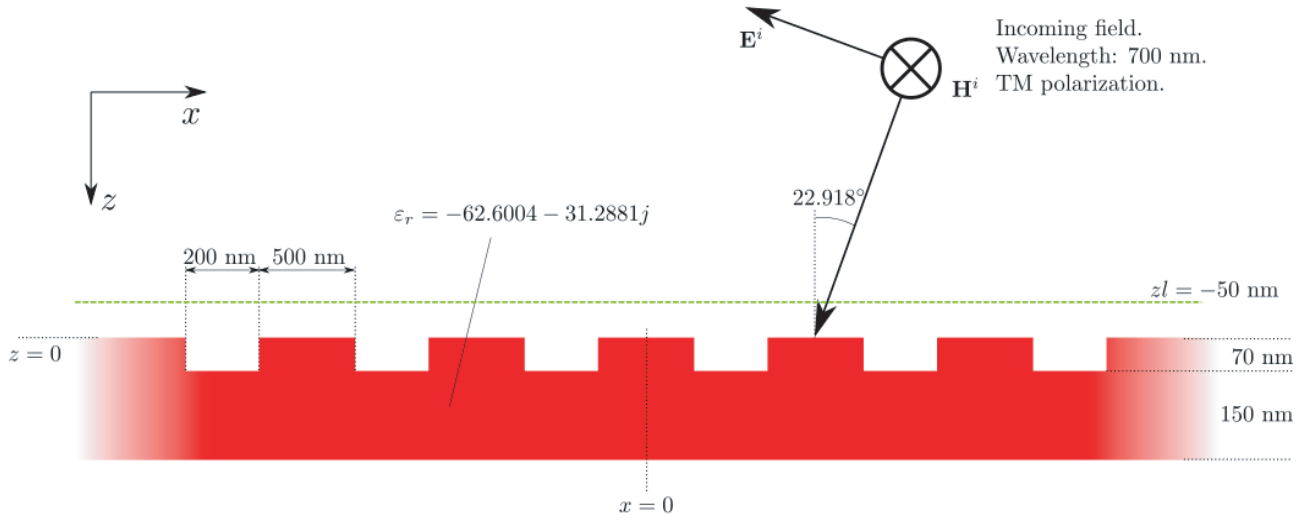

Figure 8. Geometry setting of a 2D TM metal grating problem.

direction PWL functions are employed with sampling distance $\Delta = 2.5\,\text{nm}$. Our goal in this example is to demonstrate the effectiveness of the NVF-BD preconditioner as compared with the original system in terms of convergence with an acceptable relative error. Also, in this 2D TM case, computing the NVF-BD preconditioner requires only moderate memory requirements due to the relatively low-dimensional block matrix C_i and we compute C^{-1} directly based upon the Doolittle LU factorization.

Figure 9 shows the convergence of the relative error versus the number of iterations for the original system and for the preconditioned system. It is clear that without applying the NVF-BD preconditioner, the system does not converge, while the preconditioned system reaches the desired relative error of 10^{-5} in 159 iterations.

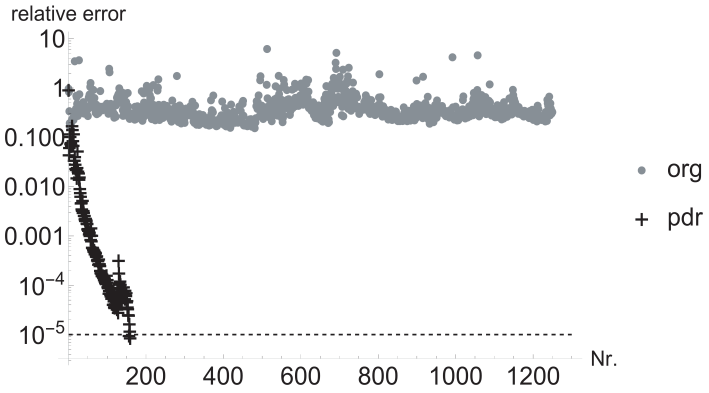

Figure 9. Iteration details for the negative permittivity case in Fig. 8 with $\chi = -1.01 + 0.0064j$. The horizontal dashed line denotes the desired accuracy goal $1 \cdot 10^{-5}$.

We have validated the preconditioned system's solution against the commercial FEM code JCMWave [44]. Fig. 10(a) presents the x-component of the total electric field E_x, where the red line denotes the JCMWave reference and the blue dashed lines denote the solution from the preconditioned system, so the solutions can be compared. Fig. 10(b) displays the absolute error between the solution from the preconditioned system in the near field for $z = -50\,\text{nm}$, just above the upper interface. One can observe that some high-frequency Gibbs ringings occur near the grooves' boundaries, where the contrast function is discontinuous. Gibbs phenomena can be the dominant contribution to the error

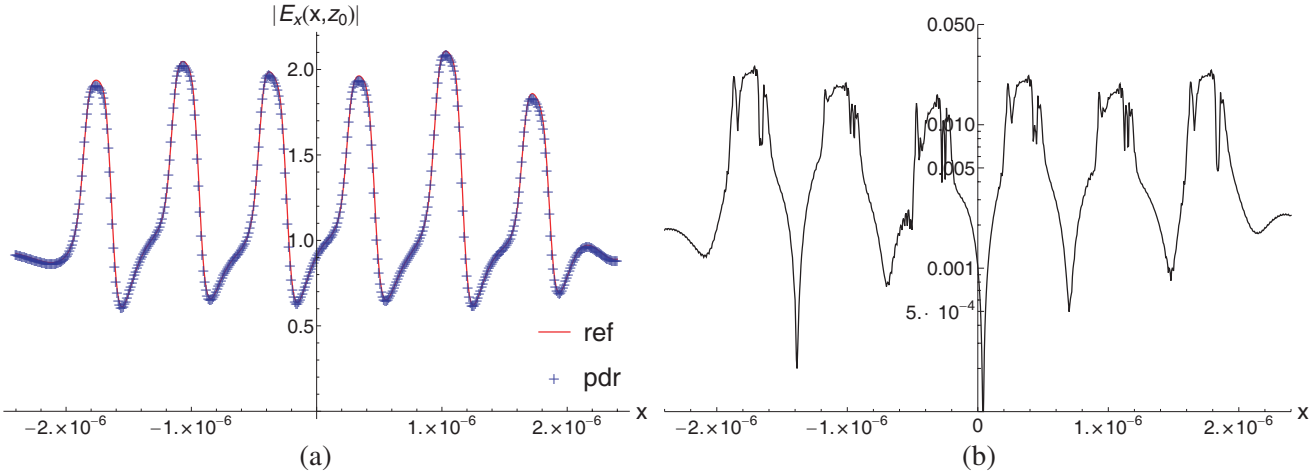

Figure 10. The electric fields for the case in Fig. 8. In (a) both reference and the solution $E_x(x, z_0)$ from the preconditioned system are displayed at $z_0 = -50\,\text{nm}$, indicated by the horizontal green line in Fig. 8. In (b) the absolute error in $E_x(x, z_0)$ between the solutions from the preconditioned system and the JCMWave reference is displayed on a log scale.

in the near field, just as observed in the original system [14, 15], but it does not propagate over a long distance. Figs. 10(a) and (b) together suggest that the solution obtained from the preconditioned system matches the reference well.

4.3. Case (C): A 3D High Contrast Problem

In the third case we consider a bar-shaped scatterer with a high relative permittivity $\varepsilon_r = 17$ in free space. The scatterer's dimensions are $300 \times 200 \times 100$ nm. The incident plane wave is characterized by the Cartesian wavevector $\mathbf{k} = (0, 0, k_0)$, with the electric field polarized in the x direction and with unit amplitude. The plane wave has a wavelength of $\lambda = 425$ nm. The geometry setting is given in Fig. 11. Table 2, case (C) displays the discretization parameters that are used in this simulation. Note that the frequency modulation number $-10 \leq n_x, n_y \leq 10$ and therefore there are 21 frame functions used per Gabor window length X and Y, which yields a resolution of 3.86 nm in both x and y directions. In the z direction PWL functions are employed with sample distance $\Delta = 5$ nm.

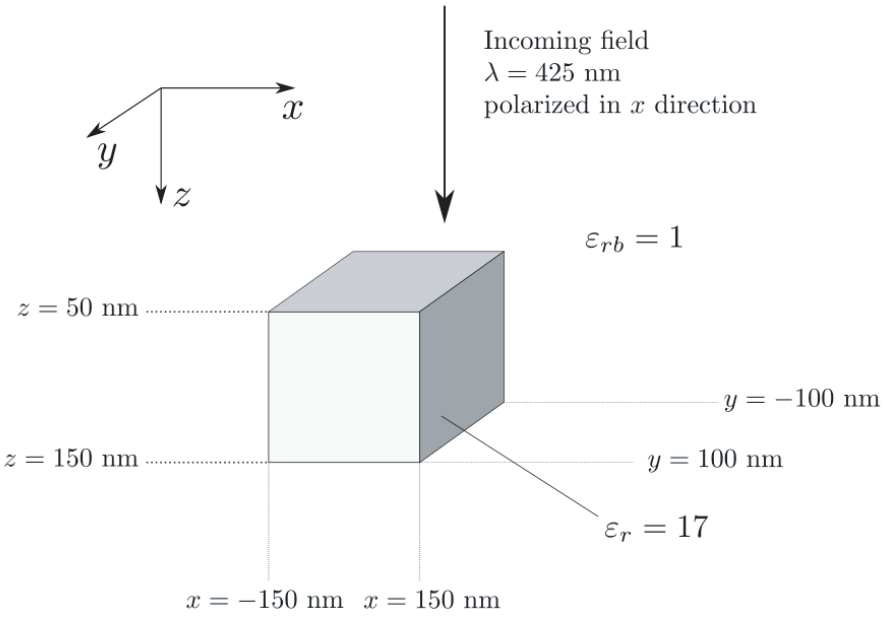

Figure 11. 3D scattering problem for a bar-shaped scatterer with relative permittivity $\varepsilon_r = 17$ embedded in air.

Usually the dimension of the system equation in 3D cases is huge. In this simulation there are $2.3 \cdot 10^6$ unknowns after performing the discretization based on Table 2 case (C), and the dimension of the block submatrices C_i in Fig. 2 is 1.0×10^5. Therefore it is unrealistic to store the full submatrix C_i due to its excessive memory requirement. As an alternative strategy we have implemented the preconditioned system such that the extra MVP for C^{-1} is executed based on an inner iterative solver. We also use the BiCGstab(2) algorithm in the inner iterative process and this inner iterative process is terminated once a relative error of less than or equal to 10^{-15} is reached. The inner iterative solver takes much fewer MVPs than the outer solver. However, this double-iterative method should be improved in future work, to make the entire solution process more efficient. Therefore, we focus on the effect of the NVF-BD preconditioner on the reduction in the number of iterations, instead of computation time, in this 3D case.

Figure 12 shows the evolution of the relative error versus the iteration count for the original system and the preconditioned system. It is clear that the preconditioned system outperforms the original system in this 3D high contrast problem with $\chi = 16$. The preconditioned system takes 454 iterations to reach the required relative error with a relatively fast rate of convergence. However, the original system failed to converge to the desired relative error within 1250 iterations. Notice that, from iteration

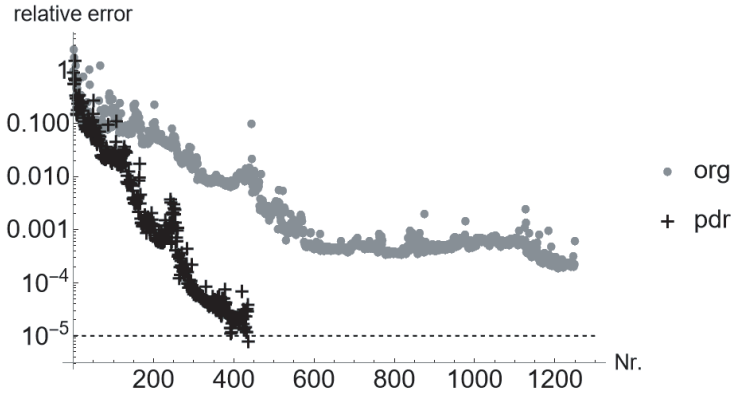

Figure 12. Iteration details for the high contrast case with $\chi = 16$. The dashed line denotes the default accuracy goal 1×10^{-5}.

550 to 1250, the residual vector of the original system gained less than 1-digit accuracy. This is a clear example that shows how the NVF-BD preconditioner can reduce the number of iterations in a 3D high-contrast case.

5. CONCLUSION

We proposed a normal-vector-field-based block-diagonal (NVF-BD) preconditioner for the original system of a spatial spectral solver with Gabor discretization for 2D TM polarization and 3D cases. The block-diagonal structure of the matrix that incorporates the normal-vector field formulation and previous work motivated us to apply this preconditioner to this spatial spectral Maxwell solver. We observed a more clustered eigenvalue distribution after applying this NVF-BD preconditioner, which is a good sign in the sense of expecting a reduction in the number of iterations. The NVF-BD preconditioner is either computed via a direct LU decomposition, in the 2D TM cases, or performed via an inner iterative procedure in the 3D problem.

We tested this NVF-BD preconditioner on three types of problems: (A) a 2D TM scattering problem with high contrast values and large geometry size, (B) a 2D TM metal grating problem, and (C) a 3D high contrast problem. The numerical experiments reveal that the number of iterations can be significantly reduced by applying the NVF-BD preconditioner, which therefore extends the capability of the original spatial spectral solver to cases with higher contrast, negative permittivity, or larger geometrical dimension. Computation-time analysis shows that the total solution time can also be reduced after applying the NVF-BD preconditioner, even though the reduction effect can be dampened when a large number of transverse basis functions N_x is used, due to the extra MVP for the preconditioner with $\mathcal{O}(N_z N_x^2)$ computational complexity. The proposed NVF-BD preconditioner itself can readily benefit from parallel computing, since the NVF-BD preconditioner has the same per-z-sample block-diagonal structure as the matrices C and M. However, a similar speed increase due to parallelization at the z-sampling level will not readily obtained for the original system due to the communication overhead associated with the Green function, for which many z-samples need to be combined.

ACKNOWLEDGMENT

This work was funded by NWO-TTW as part of the HTSM program under project number 16184. The authors thank Artur Palha from ASML Holding for providing the references based on JCMWave. The authors are grateful to the anonymous reviewers for their helpful comments and suggestions.

REFERENCES

1. Domnenko, V., B. Küchler, W. Hoppe, J. Preuninger, U. Klostermann, W. Demmerle, M. Bohn, D. Krüger, R. R. H. Kim, and L. E. Tan, "Euv computational lithography using accelerated topographic mask simulation," *Design-Process-Technology Co-optimization for Manufacturability XIII*, Vol. 10962, 109620O, International Society for Optics and Photonics, 2019.
2. Ku, Y.-S., H.-L. Pang, W.-T. Hsu, and D.-M. Shyu, "Accuracy of diffraction-based overlay metrology using a single array target," *Optical Engineering*, Vol. 48, No. 12, 123601, 2009.
3. Diebold, A. C., *Handbook of Silicon Semiconductor Metrology*, CRC Press, 2001.
4. Wang, L., Y. Wang, and X. Zhang, "Embedded metallic focus grating for silicon nitride waveguide with enhanced coupling and directive radiation," *Optics Express*, Vol. 20, No. 16, 17509–17521, 2012.
5. Dzibrou, D. O., J. J. van der Tol, and M. K. Smit, "Tolerant polarization converter for InGaAsP-InP photonic integrated circuits," *Optics Letters*, Vol. 38, No. 18, 3482–3484, 2013.
6. Yu, N., P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, "Light propagation with phase discontinuities: Generalized laws of reflection and refraction," *Science*, Vol. 334, No. 6054, 333–337, 2011.
7. Jahani, S. and Z. Jacob, "All-dielectric metamaterials," *Nature Nanotechnology*, Vol. 11, No. 1, 23–36, 2016.
8. Shlager, K. L. and J. B. Schneider, "A selective survey of the finite-difference time-domain literature," *IEEE Antennas and Propagation Magazine*, Vol. 37, No. 4, 39–57, 1995.
9. Larson, M. G. and F. Bengzon, *The Finite Element Method: Theory, Implementation, and Applications*, Vol. 10, Springer Science & Business Media, 2013.
10. Sancer, M. I., K. Sertel, J. L. Volakis, and P. Van Alstine, "On volume integral equations," *IEEE Transactions on Antennas and Propagation*, Vol. 54, No. 5, 1488–1495, 2006.
11. Botha, M. M., "Solving the volume integral equations of electromagnetic scattering," *Journal of Computational Physics*, Vol. 218, No. 1, 141–158, 2006.
12. P. Ylä-Oijala, M. Taskinen, and S. Järvenpää, "Surface integral equation formulations for solving electromagnetic scattering problems with iterative methods," *Radio Science*, Vol. 40, No. 6, 2005.
13. Dilz, R. J. and M. C. van Beurden, "A domain integral equation approach for simulating two dimensional transverse electric scattering in a layered medium with a Gabor frame discretization," *Journal of Computational Physics*, Vol. 345, 528–542, 2017.
14. Dilz, R. J., M. G. van Kraaij, and M. C. van Beurden, "2D TM scattering problem for finite dielectric objects in a dielectric stratified medium employing Gabor frames in a domain integral equation," *JOSA A*, Vol. 34, No. 8, 1315–1321, 2017.
15. Dilz, R. J., M. G. van Kraaij, and M. C. van Beurden, "A 3D spatial spectral integral equation method for electromagnetic scattering from finite objects in a layered medium," *Optical and Quantum Electronics*, Vol. 50, No. 5, 1–22, 2018.
16. Van Beurden, M. C. and I. D. Setija, "Local normal vector field formulation for periodic scattering problems formulated in the spectral domain," *JOSA A*, Vol. 34, No. 2, 224–233, 2017.
17. Saad, Y. and M. H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on Scientific and Statistical Computing*, Vol. 7, No. 3, 856–869, 1986.
18. Fletcher, R., "Conjugate gradient methods for indefinite systems," *Numerical Analysis*, 73–89, Springer, 1976.
19. Van der Vorst, H., "A fast and smoothly convergent variant of BI-CG for the solution of nonsymmetrical linear systems," *SIAM Journal on Scientific and Statistical Computing*, Vol. 13, 631–644, 1992.
20. Sleijpen, G. L. and D. R. Fokkema, "BiCGstab (ell) for linear equations involving unsymmetric matrices with complex spectrum," *Electronic Transactions on Numerical Analysis*, Vol. 1, 11–32, 1993.

21. Sonneveld, P. and M. B. van Gijzen, "IDR (s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations," *SIAM Journal on Scientific Computing*, Vol. 31, No. 2, 1035–1062, 2009.
22. Saad, Y., *Iterative Methods for Sparse Linear Systems*, SIAM, 2003.
23. Wathen, A. J., "Preconditioning," *Acta Numerica*, Vol. 24, 2015.
24. Remis, R., "Circulant preconditioners for domain integral equations in electromagnetics," *2012 International Conference on Electromagnetics in Advanced Applications*, 337–340, IEEE, 2012.
25. Remis, R., "Preconditioning techniques for domain integral equations," *2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 235–238, IEEE, 2013.
26. Groth, S. P., A. G. Polimeridis, A. Tambova, and J. K. White, "Circulant preconditioning in the volume integral equation method for silicon photonics," *JOSA A*, Vol. 36, No. 6, 1079–1088, 2019.
27. Popov, E. and M. Nevière, "Maxwell equations in fourier space: fast-converging formulation for diffraction by arbitrary shaped, periodic, anisotropic media," *JOSA A*, Vol. 18, No. 11, 2886–2894, 2001.
28. Schneider, F., "Approximation of inverses of BTTB matrices," Master's thesis, Eindhoven University of Technology, 2016.
29. Van Kraaij, M. G. M. M., "Forward diffraction modelling: analysis and application to grating reconstruction," Ph.D. thesis, Eindhoven University of Technology, 2011.
30. Dilz, R. J., "A spatial spectral domain integral equation solver for electromagnetic scattering in dielectric layered media," Technische Universiteit Eindhoven, 2017.
31. Feichtinger, H. G. and T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications*, Springer Science & Business Media, 2012.
32. Dilz, R. J. and M. C. van Beurden, "The Gabor frame as a discretization for the 2D transverse electric scattering-problem domain integral equation," *Progress In Electromagnetics Research*, Vol. 69, 117–136, 2016.
33. Dilz, R. J. and M. C. van Beurden, "Fast operations for a Gabor-frame-based integral equation with equidistant sampling," *IEEE Antennas and Wireless Propagation Letters*, Vol. 17, No. 1, 82–85, 2017.
34. Li, L., "Use of Fourier series in the analysis of discontinuous periodic structures," *JOSA A*, Vol. 13, No. 9, 1870–1876, 1996.
35. Morgenshtern, V. I. and H. Bölcskei, "A short course on frame theory," arXiv preprint arXiv:1104.4300, 2011.
36. Axelsson, O. and V. A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, SIAM, 2001.
37. Chan, T. F., "An optimal circulant preconditioner for Toeplitz systems," *SIAM Journal on Scientific and Statistical Computing*, Vol. 9, No. 4, 766–771, 1988.
38. Chan, R. H., "Circulant preconditioners for Hermitian Toeplitz systems," *SIAM Journal on Matrix Analysis and Applications*, Vol. 10, No. 4, 542–550, 1989.
39. Tyrtyshnikov, E. E., "Optimal and superoptimal circulant preconditioners," *SIAM Journal on Matrix Analysis and Applications*, Vol. 13, No. 2, 459–473, 1992.
40. Chan, R. H., "Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions," *SIMA Journal of Numerical Analysis*, Vol. 11, No. 3, 333–345, 1991.
41. Chan, R. H. and K.-P. Ng, "Toeplitz preconditioners for Hermitian Toeplitz systems," *Linear Algebra and Its Applications*, Vol. 190, 181–208, 1993.
42. Noutsos, D. and P. Vassalos, "New band Toeplitz preconditioners for ill-conditioned symmetric positive definite Toeplitz systems," *SIAM Journal on Matrix Analysis and Applications*, Vol. 23, No. 3, 728–743, 2002.
43. Lin, F.-R., "Preconditioners for block Toeplitz systems based on circulant preconditioners," *Numerical Algorithms*, Vol. 26, No. 4, 365–379, 2001.
44. Burger, S., L. Zschiedrich, J. Pomplun, and F. Schmidt, "Finite-element based electromagnetic field simulations: Benchmark results for isolated structures," arXiv preprint arXiv:1310.2732, 2013.