Preconditioners and Their Analyses for Edge Element Saddle-point Systems Arising from Time-harmonic Maxwell Equations

Hua Xiang * Shiyang Zhang † Jun Zou ‡

Abstract

We propose and analyze new preconditioners for the saddle-point systems arising from the edge element discretization of the time-harmonic Maxwell equations. The preconditioners come from a formula giving the inverse of the coefficient matrix of the saddle-point system with vanishing and non-vanishing wave numbers, and are generalizations of the preconditioner in [6]. we show theoretically and numerically that the conjugate gradient method (CG) with these new preconditioners can be applied efficiently when the wave number (k) is not too large (roughly $k \leq 4$ numerically). The spectral behaviors of the resulting preconditioned systems for the new and some existing preconditioners are analyzed and compared, and numerical experiments are presented to demonstrate and compare the efficiencies of these preconditioners.

Keywords: Time-harmonic Maxwell equations, saddle-point system, preconditioners.
AMS subject classifications: 65F10, 65N22, 65N30

1 Introduction

In this work we investigate and compare some effective preconditioning solvers for the following saddle-point system:

$$\mathcal{K}\begin{pmatrix} u \\ p \end{pmatrix} \equiv \begin{pmatrix} A - k^2 M & B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{1.1}$$

where $u \in \mathbb{R}^n$, $p \in \mathbb{R}^m$, A, $M \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times n}$, with $m \leq n$. We assume that $\mathcal{K}$ is nonsingular, so B must be of full row rank. We are particularly interested in the case where A is symmetric semi-positive definite, and $\dim(ker(A)) = m$, that is, A is maximally rank deficient [6, 7]. The matrix M is assumed to be symmetric positive definite, and to satisfy some extra basic constraint conditions given by (ii) of Proposition 2.1, and k is a given real number.

*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P. R. China. The research of this author was supported by the National Natural Science Foundation of China under grants 11571265 and 11471253. (hxiang@whu.edu.cn).

†School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P. R. China. (hydzhang@whu.edu.cn).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong. The work of this author was substantially supported by Hong Kong RGC General Research Fund (projects 14322516 and 14306814). (zou@math.cuhk.edu.hk)

The saddle-point system of form (1.1) with a maximal rank deficient A arises from many applications, including the numerical solution of time-harmonic Maxwell's equations [7, 8, 17] where k represents the wave number, the underdetermined norm-minimization problems and geophysical inverse problems; see more details in the recent paper [6]. It was a very inspiring and innovative work and developed a class of indefinite block preconditioners for the use with CG. CG may converge rapidly under certain conditions when it is applied for solving the general saddle-point system of form (1.1) with a maximal rank deficient A and $k = 0$. It was also pointed out in [6] that the saddle-point system under the above particular setting has not received as much attention as other situations, for example, the case of a symmetric positive definite A. But as it was demonstrated in [6] for the special case $k = 0$, when A is maximally rank deficient, some nice mathematical structures may be revealed and adopted to help construct efficient solution methods. This work is initiated and motivated by [6] and develop further in this direction, and show that new efficient numerical methods can be equally constructed for more general and difficult case $k \neq 0$.

Though most results of this work apply also to the general saddle-point system of form (1.1) with a maximal rank deficient A and proper constraints for M given by (ii) of Proposition 2.1, we mainly focus on the saddle-point system (1.1) that arises from time-harmonic Maxwell equations [3, 5, 10, 11, 17]:

$$\begin{cases} \nabla \times \nabla \times u - k^2 u + \nabla p = J & \text{in} \quad \Omega, \\ \nabla \cdot u = \rho & \text{in} \quad \Omega, \\ u \times n = 0 & \text{on} \quad \partial\Omega, \\ p = 0 & \text{on} \quad \partial\Omega \end{cases} \tag{1.2}$$

where u is a vector field, p is the scalar multiplier, J is the given external source and ρ is density of charge. Ω is a simply connected domain in $\mathbb{R}^3$ with a connected boundary $\partial\Omega$, with n being its outward unit normal. The wave number k is given by $k^2 = \omega^2 \varepsilon \mu$, where ω, ε and μ are positive frequency, permittivity and permeability of the medium, respectively. We assume that k^2 is not an interior Maxwell eigenvalue, but is allowed to be zero, and know the cases with appropriately small and large frequencies are physically relevant in magnetostatics, wave propagation and other applications [7]. We refer to [2, chapter 11] for a survey on this topic. Generalization of our preconditioners to cases with particular non-constant electromagnetic parameters is possible by checking the relations given in Proposition 2.1, which can be found to be true in [13]. The introduction of the Lagrange multiplier p in (1.2) may not be absolutely necessary for the general case $k \neq 0$, for which the divergence constraint does not need to be enforced directly and explicitly. namely it is possible to solve directly for u using the first equation in (1.2) with $p = 0$ mathematically [8], although it is still challenging to design an efficient numerical solver for this indefinite system. The saddle-point formulation (1.2) with the Lagrange multiplier p is stable and well-posed [5], especially it ensures the stability and Gauss's law directly when k is small and may better handle the singularity of the solution at the boundary of the domain [3, 5, 17]. More importantly, the mixed form (1.2) provides some extra flexibility on the computational aspect [13] and leads to better numerical stability and more efficient numerical solvers [6, 7]. And this is also the main motivation and focus and of the current work.

After discretizing (1.2) by using the Nédélec elements of the first kind [14, 15] for the approximation of the vector field u and the standard nodal elements for the multiplier p, we derive the saddle-point system (1.1) of our interest. We assume that the coefficient matrix $\mathcal{K}$ in (1.1) and its leading block $A - k^2 M$ ($k \neq 0$) are both nonsingular, which is true when the mesh size is sufficiently small [7].

Some very efficient preconditioners were proposed and analyzed recently in [6] for the special case of the saddle-point system (1.1), i.e., the wave number $k = 0$, and (1.1) reduces to

$$\mathcal{A}\begin{pmatrix} u \\ p \end{pmatrix} \equiv \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{1.3}$$

The following preconditioner $\mathcal{P}_0$ was proposed in [6] for solving the saddle-point system (1.3):

$$\mathcal{P}_0^{-1} = \begin{pmatrix} (A+M)^{-1}(I - B^T L^{-1} C^T) & CL^{-1} \\ L^{-1}C^T & 0 \end{pmatrix}, \tag{1.4}$$

where the matrix $L \in \mathbb{R}^{m\times m}$ is the discrete Laplacian, while $C \in \mathbb{R}^{n\times m}$ is the discrete gradient interpolator, whose columns span $ker(A)$ and can be formed easily and explicitly using the gradients of the standard nodal bases [6,7]. It was proved that the preconditioned system $\mathcal{P}_0^{-1}\mathcal{A}$ is simply diagonal, given by

$$\mathcal{P}_0^{-1}\mathcal{A} = \begin{pmatrix} (A+M)^{-1}(A + B^T L^{-1} B) & 0 \\ 0 & I \end{pmatrix}.$$

Note that both $A+M$ and $A+B^TL^{-1}B$ are symmetric positive definite, and then we can apply a CG-like method for the preconditioned system $\mathcal{P}_0^{-1}A$ in a non-standard inner product, even though both $\mathcal{A}$ and $\mathcal{P}_0$ are indefinite.

We should note that it is always possible to solve decomposed system instead of saddle point system [7]. For simpler cases with vanishing wave number, it is well-known that the saddle point system is equivalent to the following decomposed system:

- Pre-treatment: computation of Lagrange multiplier p by $p = L^{-1}C^T f$. One exact L-solver is needed here. In cases load vector f is discrete divergence-free, we have $p = 0$ and this step can be ignored.
- Solving the singular H(curl) system: $Au = f - B^T p$. This can be done for example, by the nodal auxiliary space preconditioning technique [9, 12]. In Hypre [22], instead of solving $Ax = b$, where $b \in \mathcal{R}(A)$, one can also choose to solve an equivalent system $(A + \alpha CC')x = b$, where α is a small positive number.
- Post-process: null space correction. Another exact L-solver is needed here. In this step we make sure that the solution satisfies the constraint condition $Bu = g$.

For the more general case $k \neq 0$, the block triangular preconditioners

$$\mathcal{M}_{\eta,\varepsilon} = \begin{bmatrix} A + (\eta - k^2)M & (1 - \eta\varepsilon)B^T \\ 0 & \varepsilon L \end{bmatrix} \tag{1.5}$$

with double variable relaxation parameters $\eta > k^2$ and $\varepsilon \neq 0$ were studied in [4, 7, 19, 20].

In this work, we construct some new preconditioners for (1.1). As it was shown in [6] that the aforementioned efficient preconditioners $\mathcal{P}_0^{-1}$ in (1.4) work very effectively for the special and simple case with vanishing wave number ($k = 0$), we demonstrate that similar preconditioners can be constructed and generalized also for the saddle-point linear system (1.1) with more general and difficult cases, i.e., $k \neq 0$, including high frequency waves (roughly $k \leq 4$). And we will see

analytically the spectral distributions of these new preconditioners are quite similar to the ones of the existing effective preconditioners (1.5). But the new preconditioner can be applied with CG iteration under a non-standard inner product although both the coefficient matrix $\mathcal{K}$ and the new preconditioner are indefinite, and numerically they perform mostly better and stabler than the existing preconditioners (1.5).

The rest of the paper is arranged as follows. We develop in Section 2 an important formula for computing the inverse of $\mathcal{K}$, based on which we propose in Section 3 a new preconditioner and compare its performance with existing preconditioners for the saddle-point system (1.1) with general wave numbers, and study and compare the spectral properties of the preconditioned matrices afterwards. Numerical experiments are presented in Section 4. Finally, in Section 5 after some discussions about some related issues we make a conclusion.

2 Computing the inverse of $\mathcal{K}$

We derive in this section some formulas for computing the inverse of the matrix $\mathcal{K}$ in (1.1). To do so, we first recall some useful properties of the matrices A, B, M, L and C, which are introduced in the Introduction.

Proposition 2.1. *The matrices A, B, M, L and C have the following properties [6, 7]:*

(i) $\mathbb{R}^n = ker(A) \oplus ker(B)$.

(ii) It holds that $C = M^{-1}B^T$, *and there exists a constant* $\bar{\alpha} > 0$ *independent of mesh size such that* $u^T A u \geq \bar{\alpha}\, u^T M u, \quad \forall\, u \in ker(B)$.

(iii) $L = BM^{-1}B^T$, *or* $L = BC$.

(iv) The inverse of $\mathcal{A}$ can be represented by

$$\mathcal{A}^{-1} = \begin{pmatrix} V & CL^{-1} \\ L^{-1}C^T & 0 \end{pmatrix}, \tag{2.1}$$

where the diagonal block V is given by

$$V = (A + B^T L^{-1} B)^{-1}(I - B^T L^{-1} C^T) = (A + B^T L^{-1} B)^{-1} - CL^{-1}C^T\,. \tag{2.2}$$

Remark. *The formula in* (iv) *can also be treated as a corollary to [21, Corollary 3.5] or [23] and we can generalize the formula for computing the inverse of the symmetric saddle-point matrix $\mathcal{A}$ in* (2.1) *to the more general case with following non-symmetric generalized saddle-point system (denotations in this remark should be treated separately):*

$$\mathcal{K}\begin{pmatrix} u \\ p \end{pmatrix} \equiv \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{2.3}$$

where all block matrices A, B, C and D are allowed to be non-square, with $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times k}$, $C \in \mathbb{R}^{l\times n}$, $D \in \mathbb{R}^{l\times k}$. But the entire matrix $\mathcal{K}$ is square, i.e., $m + l = n + k = t$.

We further assume that the rank of A is $m + n - t$. Then we write $C_r \in \mathbb{R}^{n\times l}$ as the matrix of full column rank whose columns span $ker(A)$, and $C_l \in \mathbb{R}^{k\times m}$ as the matrix of full row rank whose rows span the left null space of A, namely $C_l A = 0$, and $AC_r = 0$. We write $L_l = C_l B$ and $L_r = CC_r$, and the range space of A by $\mathcal{R}(A)$. Then it is easy to check the following statement from results in [21, Corollary 3.5], or [23]:

Corollary 2.1. *Assume that*

$$\mathcal{R}(A) \cap \mathcal{R}(B) = 0\,, \quad \mathcal{R}(A^T) \cap \mathcal{R}(C^T) = 0\,, \tag{2.4}$$

$$rank(A) = m + n - t\,, \quad rank(B) = k\,, \quad rank(C) = l\,, \tag{2.5}$$

then the matrix $\mathcal{K}$ is non-singular and its inverse can be represented by

$$\mathcal{K}^{-1} = \begin{pmatrix} N & C_r L_r^{-1} \\ L_l^{-1} C_l & 0 \end{pmatrix}, \tag{2.6}$$

where N satisfies

$$NA = I - C_r L_r^{-1} C\,, \quad AN = I - B L_l^{-1} C_l\,, \tag{2.7}$$

$$NB = -C_r L_r^{-1} D\,, \quad CN = -D L_l^{-1} C_l. \tag{2.8}$$

If $m = n$, it holds for any $X \in \mathbb{R}^{m\times l}$ such that $A + XC$ is non-singular that

$$N = (A + XC)^{-1}(I - B L_l^{-1} C_l - X D L_l^{-1} C_l). \tag{2.9}$$

We can easily see that the coefficient matrix in (1.3) *satisfies* (2.4) *and* (2.5).

Now we are ready to derive a formula for computing the inverse of the matrix $\mathcal{K}$ in (1.1). Recall that $\mathcal{K}$ and $A - k^2 M$ are invertible. We write the (1,1) block of the inverse $\mathcal{K}^{-1}$ as T, then we have the following representation of the inverse of the saddle-point matrix $\mathcal{K}$.

Theorem 2.2. *The inverse of $\mathcal{K}$ is given by*

$$\mathcal{K}^{-1} = \begin{pmatrix} T & C L^{-1} \\ L^{-1} C^T & k^2 L^{-1} \end{pmatrix}, \tag{2.10}$$

where T satisfies

$$(A - k^2 M) T = I - B^T L^{-1} C^T\,, \quad BT = 0\,. \tag{2.11}$$

Proof. We write $\mathcal{K}^{-1}$ as a perturbation of $\mathcal{A}^{-1}$ in the form

$$\mathcal{K}^{-1} = \mathcal{A}^{-1} + \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}, \tag{2.12}$$

then using the fact that $\mathcal{K}\mathcal{K}^{-1} = I$, namely

$$\left[\mathcal{A} + \begin{pmatrix} -k^2 M & 0 \\ 0 & 0 \end{pmatrix}\right] \cdot \left[\mathcal{A}^{-1} + \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}\right] = I\,,$$

we obtain by a direct computation that

$$-k^2 M(V + X_1) + A X_1 + B^T X_3 = 0, \tag{2.13}$$

$$-k^2 (B^T L^{-1} + M X_2) + A X_2 + B^T X_4 = 0, \tag{2.14}$$

$$B X_1 = 0\,, \quad B X_2 = 0. \tag{2.15}$$

By direct computation we have

$$AV = I - B^T L^{-1} C^T\,, \quad BV = 0\,. \tag{2.16}$$

From (2.12) we know that $V + X_1$ is the (1,1) block of $\mathcal{K}^{-1}$, so it follows from (2.15) and (2.16) that

$$BT = B(V + X_1) = 0.$$

Multiplying (2.13) by C^T we derive

$$-k^2 B(V + X_1) + LX_3 = 0,$$

which gives

$$X_3 = k^2 L^{-1} B(V + X_1) = 0. \tag{2.17}$$

Similarly, multiplying (2.14) by C^T we obtain

$$-k^2(I + BX_2) + LX_4 = 0.$$

By combining this equality with the second relation in (2.15), we come to

$$X_4 = k^2 L^{-1}. \tag{2.18}$$

Then we may substitute (2.18) into (2.14) to get

$$(A - k^2 M)X_2 = 0\,, \tag{2.19}$$

which proves $X_2 = 0$.

Noting that we have proved $X_3 = 0$, then (2.13) reduces to $-k^2 M(V + X_1) + AX_1 = 0$, or $(A - k^2 M)(V + X_1) = AV$, which completes the desired proof. □

The following result is important to help us understand the leading block T of the inverse of $\mathcal{K}$ in (2.10).

Theorem 2.3. *The matrix $A + \eta B^T L^{-1} B - k^2 M$ is non-singular for any $\eta \neq k^2$, and its null space is exactly the same as that of A for $\eta = k^2$.*

Proof. By means of (i) of Proportion 2.1, we can write for any $u \in \mathbb{R}^n$ that $u = u_A + u_B$ with $u_A \in ker(A)$ and $u_B \in ker(B)$. If $(A + \eta B^T L^{-1} B - k^2 M)u = 0$, then $(A - k^2 M)u_B + \eta B^T L^{-1} B u_A - k^2 M u_A = 0$. As the columns of C span the null space of A, there exists $p \in \mathbb{R}^m$ such that $u_A = Cp$. So we see $(A - k^2 M)u_B + (\eta - k^2)B^T p = 0$. Multiplying its both sides by C^T, we derive $p = 0$, hence $(A - k^2 M)u_B = 0$, yielding that $u_B = 0$. Hence we have proved $u = 0$, and also the non-singularity of the desired matrix.

Next, we consider the case with $\eta = k^2$. We show the two matrices $A + k^2 B^T L^{-1} B - k^2 M$ and A have the same null space. First, note that $(A + k^2 B^T L^{-1} B - k^2 M)C = 0$. Now we assume u is in the null space of $A + k^2 B^T L^{-1} B - k^2 M$. We still write $u = u_A + u_B$, and follow the earlier proof of the non-singularity of the matrix, but with $\eta = k^2$ now. Then we deduce $(A - k^2 M)u_B = 0$, which implies $u_B = 0$, hence we know $Au = 0$. □

The following result comes directly from (2.11) and Theorem 2.3. And it introduces a very crucial parameter η to the expression of the leading block T of the inverse of $\mathcal{K}$ in (2.10), and it can take an arbitrary value except for $\eta \neq k^2$.

Corollary 2.2. *For any $\eta \neq k^2$, it holds that*

$$\begin{aligned} T &= (A+\eta B^T L^{-1} B - k^2 M)^{-1}(I - B^T L^{-1} C^T) \\ &= (A+\eta B^T L^{-1} B - k^2 M)^{-1} - \frac{1}{\eta - k^2} C L^{-1} C^T. \end{aligned} \tag{2.20}$$

In conclusion, we can easily see from Theorem 2.2 and Corollary 2.2 the following formula with $\eta \neq k^2$ for computing the inverse of the matrix $\mathcal{K}$ in (1.1), which forms the basis in our construction of some new preconditioners that are discussed in the next section:

$$\mathcal{K}^{-1} = \begin{pmatrix} (A+\eta B^T L^{-1} B - k^2 M)^{-1}(I - B^T L^{-1} C^T) & C L^{-1} \\ L^{-1} C^T & k^2 L^{-1} \end{pmatrix}. \tag{2.21}$$

3 New preconditioners and their spectral properties

The formula (2.21) suggests us some natural preconditioners for the saddle-point matrix $\mathcal{K}$ in (1.1). Noting that the matrix $B^T L^{-1} B$ is a dense matrix, the action of the (1,1) block of (2.21) is very expensive to compute. To overcome the difficulty, we approximate the dense matrix $A + \eta B^T L^{-1} B - k^2 M$ by the spectrally equivalent [7] sparse matrix $A + \eta M - k^2 M$ to get the following simplified preconditioner for the matrix $\mathcal{K}$:

$$\mathcal{P}^{-1} \equiv \begin{pmatrix} (A+\eta M - k^2 M)^{-1}(I - B^T L^{-1} C^T) & C L^{-1} \\ L^{-1} C^T & k^2 L^{-1} \end{pmatrix}. \tag{3.1}$$

For the simple case with vanishing wave number ($k = 0$) and $\eta = 1$, the preconditioner (3.1) reduces to the existing one $\mathcal{P}_0^{-1}$ in (1.4). To ensure the non-singularity of the matrix $A + \eta M - k^2 M$ involved in (3.1), we can simply set the parameter $\eta > k^2$ so that it becomes symmetric positive definite. Moreover, this choice also ensures the non-singularity of the matrix on the right-hand side of (3.1), as discussed below.

Theorem 3.1. *For any $\eta > k^2$, the matrix on the right-hand side of* (3.1) *is nonsingular.*

Proof. It is direct to check that the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ is given by

$$\mathcal{P}^{-1}\mathcal{K} = \begin{pmatrix} (A+\eta M - k^2 M)^{-1}(A + k^2 B^T L^{-1} B - k^2 M) + C L^{-1} B & 0 \\ 0 & I \end{pmatrix}.$$

Using Proposition 2.1 (i), we can further write the (1, 1) block of the above matrix as follows:

$$\begin{aligned} &(A+\eta M - k^2 M)^{-1}(A + k^2 B^T L^{-1} B - k^2 M) + C L^{-1} B \\ =&(A+\eta M - k^2 M)^{-1}(A + k^2 B^T L^{-1} B - k^2 M) \\ &+ (A+\eta M - k^2 M)^{-1}(\eta M C L^{-1} B - k^2 M C L^{-1} B) \\ =&(A+\eta M - k^2 M)^{-1}(A + \eta B^T L^{-1} B - k^2 M), \end{aligned}$$

so the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ reads also as

$$\mathcal{P}^{-1}\mathcal{K} = \begin{pmatrix} (A+\eta M - k^2 M)^{-1}(A + \eta B^T L^{-1} B - k^2 M) & 0 \\ 0 & I \end{pmatrix}. \tag{3.2}$$

We know that the leading block of $\mathcal{P}^{-1}\mathcal{K}$ in (3.2) is non-singular by Theorem 2.3, hence the desired conclusion follows. □

Note that $A+\eta M-k^2M$ and its inverse are always symmetric positive definite for $\eta > k^2$. Actually, the original matrix $A+\eta B^TL^{-1}B-k^2M$ can be also symmetric positive definite as shown below.

Theorem 3.2. *If any $\eta > k^2$ and $k^2 < \bar{\alpha}$, the matrix $A+\eta B^TL^{-1}B-k^2M$ is symmetric positive definite.*

Proof. For any $u \in \mathbb{R}^n$, we can write $u = u_A + u_B$ with $u_A \in ker(A)$ and $u_B \in ker(B)$. By Proposition 2.1 we know $u_A^TMu_B = 0$ and $u_A^TB^TL^{-1}Bu_A = u_A^TMu_A$. Therefore,

$$\begin{aligned} & u^T(A+\eta B^TL^{-1}B-k^2M)u \\ =& u_A^T(A+\eta B^TL^{-1}B-k^2M)u_A + u_B^T(A+\eta B^TL^{-1}B-k^2M)u_B \\ =& u_A^T(\eta B^TL^{-1}B-k^2M)u_A + u_B^T(A-k^2M)u_B \\ =& u_B^T(A-k^2M)u_B + (\eta-k^2)u_A^TMu_A\,. \end{aligned} \tag{3.3}$$

But we have $u_B^TAu_B \geq \bar{\alpha}u_B^TMu_B$ by (ii) in Proposition 2.1, and this implies

$$u^T(A+\eta B^TL^{-1}B-k^2M)u \geq (\bar{\alpha}-k^2)u_B^TMu_B + (\eta-k^2)u_A^TMu_A > 0\,,$$

hence proves the desired result. □

For $\eta > k^2$ and $k^2 < \bar{\alpha}$, the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ is self-adjoint and positive definite with respect to the inner product

$$\langle x, y\rangle = x^T \begin{pmatrix} (A+\eta M-k^2M) & 0 \\ 0 & I \end{pmatrix} y. \tag{3.4}$$

So we can apply the CG iteration [1] in this special inner product for solving the preconditioned system associated with $\mathcal{P}^{-1}\mathcal{K}$. But Theorem 3.2 does not ensure the positive definiteness of the preconditioned system for $k^2 \geq \bar{\alpha}$, so the CG iteration may fail theoretically. We will see in Section 4 that numerically this is not a problem (see Figure 3).

We know the convergence rates of the CG and MINRES can be reflected often by the spectrum of the preconditioned system. For this purpose, we now study the spectral properties of the preconditioned system $\mathcal{P}^{-1}\mathcal{K}$. First, we present an interesting observation that the symmetric positive definiteness of the matrix $A+\eta B^TL^{-1}B-k^2M$ is determined by the parameter k.

Theorem 3.3. *For any two numbers $\eta_1, \eta_2 > k^2$, $A+\eta_1 B^TL^{-1}B-k^2M$ is symmetric positive definite if and only if $A+\eta_2 B^TL^{-1}B-k^2M$ is symmetric positive definite.*

Proof. For any $\eta_1 > k^2$, suppose that $A+\eta_1 B^TL^{-1}B-k^2M$ is not symmetric positive definite. As this matrix is nonsingular by Theorem 2.3, hence it is not symmetric semi-positive definite. Therefore, there exists $u \in \mathbb{R}^n$ satisfying $u^T(A+\eta_1 B^TL^{-1}B-k^2M)u < 0$. But we can write $u = u_A+u_B$ with $u_A \in ker(A)$ and $u_B \in ker(B)$. Then we can see that $u_B \neq 0$ and $u_B^T(A-k^2M)u_B < 0$ from (3.3). Now for any $\eta_2 > k^2$, we can easily check $u_B^T(A+\eta_2 B^TL^{-1}B-k^2M)u_B = u_B^T(A-k^2M)u_B < 0$, hence $A+\eta_2 B^TL^{-1}B-k^2M$ is not symmetric positive definite. □

Next we present several results about the eigenvalues of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$.

Lemma 3.4. *For any $\eta > k^2$, $\lambda = 1$ is an eigenvalue of $(A+\eta M-k^2M)^{-1}(A+\eta B^TL^{-1}B-k^2M)$ with its algebraic multiplicity being m. The rest of the eigenvalues are bounded by*

$$\frac{\bar{\alpha}-k^2}{\bar{\alpha}+\eta-k^2} < \lambda < 1. \tag{3.5}$$

Proof. The result was proved in [7, Theorem 5.1] for $\eta = 1$ and $k^2 < 1$. But our desired results for an arbitrary positive η can be done similarly. □

The following result is a direct consequence of Lemma 3.4 by using the formula (3.2).

Theorem 3.5. *For any $\eta > k^2$, $\lambda = 1$ is an eigenvalue of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ with its algebraic multiplicity being $2m$. The rest of the eigenvalues are bounded as in* (3.5).

Now we like to make some spectral comparisons between the two preconditioned systems generated by our new preconditioner $\mathcal{P}$ and the existing block tridiagonal one $\mathcal{M}_{\eta,\varepsilon}$ in (1.5) for the saddle-point matrix $\mathcal{K}$. We first recall the following results from [20, Theorem 2.6].

Theorem 3.6. *For any $\eta > k^2$, both $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{\varepsilon(\eta-k^2)}$ are the eigenvalues of $\mathcal{M}_{\eta,\varepsilon}^{-1}\mathcal{K}$, each with its algebraic multiplicity m. The rest of the eigenvalues are bounded as in* (3.5).

We see from Theorems 3.5 and 3.6 that the spectra of $\mathcal{P}^{-1}\mathcal{K}$ and $\mathcal{M}_{\eta,\varepsilon}^{-1}\mathcal{K}$ are quite similar, except that the latter has an extra eigenvalue λ_2, with its algebraic multiplicity being m. This will be also confirmed numerically in the next section.

The block tridiagonal preconitioners $\mathcal{M}_{\eta,\varepsilon}$ reduce to symmetric if we set $\varepsilon = \frac{1}{\eta}$:

$$\mathcal{M}_{\eta,1/\eta} = \begin{bmatrix} A+(\eta-k^2)M & 0 \\ 0 & \frac{1}{\eta}L \end{bmatrix}. \tag{3.6}$$

This preconditioner was analyzed and applied in [7, 19] along with the minimal residual (MINRES) iteration. We may observe from Theorems 3.5 and 3.6 that the eigenvalues of our preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ are a little better clustered than those of $\mathcal{M}_{\eta,1/\eta}^{-1}\mathcal{K}$ as its eigenvalue λ_2 is smaller than $\frac{\bar{\alpha}-k^2}{\bar{\alpha}+\eta-k^2}$. But our new preconditioner $\mathcal{P}$ can be applied with CG for $k^2 < \bar{\alpha}$, and MINRES for $k^2 \geq \bar{\alpha}$. And more importantly, as we will see from our numerical experiments in next section, we can also apply the new preconditioner $\mathcal{P}$ with CG even for $k^2 \geq \bar{\alpha}$ and the convergence is still rather stable, while CG with preconditioner $\mathcal{M}_{\eta,\varepsilon}$ in (3.6) breaks down most of the time.

On the other hand, if we choose $\varepsilon \neq 1/\eta$, the preconditioner $\mathcal{M}_{\eta,\varepsilon}$ is non-symmetric, and the methods like the generalized minimal residual method should be used, which are less economical than methods like CG or MINRES. Note that for $\varepsilon = -\frac{1}{\eta-k^2}$, we have $\lambda_2 = \lambda_1$, so $\lambda = 1$ is an eigenvalue of $\mathcal{M}_{\eta,\varepsilon}^{-1}\mathcal{K}$ with its algebraic multiplicity being $2m$, the same as for $\mathcal{P}^{-1}\mathcal{K}$.

Now we consider the inner iterations associated with the new preconditioner $\mathcal{P}$. For any two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, we can write

$$\begin{aligned}
\mathcal{P}^{-1}\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} (A+\eta M-k^2M)^{-1}(x-B^TL^{-1}C^Tx)+CL^{-1}y \\ L^{-1}C^Tx+k^2L^{-1}y \end{pmatrix} \\
&= \begin{pmatrix} (A+\eta M-k^2M)^{-1}x-\frac{1}{\eta-k^2}CL^{-1}C^Tx+CL^{-1}y \\ L^{-1}C^Tx+k^2L^{-1}y \end{pmatrix} \\
&= \begin{pmatrix} (A+\eta M-k^2M)^{-1}x+CL^{-1}(y-\frac{1}{\eta-k^2}C^Tx) \\ L^{-1}(C^Tx+k^2y) \end{pmatrix}.
\end{aligned}$$

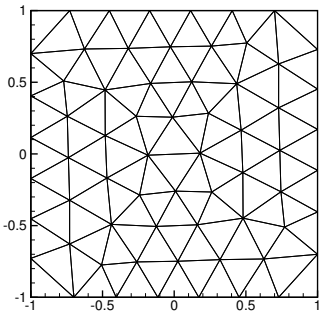


Figure 1: Mesh G1: n+m=187

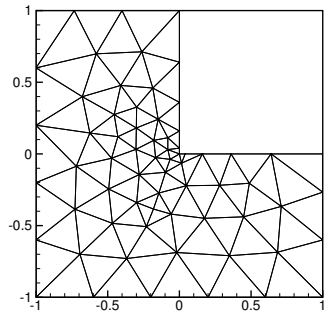


Figure 2: Mesh L1: n+m=185

In both forms we have to solve two linear systems associated with the discrete Laplacian L and one with $A+(\eta-k^2)M$ at each evaluation of the action of $\mathcal{P}^{-1}$. Many fast solvers are available for solving these two symmetric and positive definite systems [9, 13]. We use the first form to implement the action of $\mathcal{P}^{-1}$.

We know that the parameter $\bar{\alpha}$ depends only on the shape regularity of the mesh and the approximation order of the finite elements used, and is irrelevant to the size of the mesh [8, Theorem 4.7]. Numerically we may expect a upper bound for k^2 that ensures the positive definiteness of $A+\eta B^T L^{-1}B-k^2M$, and this bound should be independent of the mesh size. We test this numerically in the next section.

4 Numerical experiments

In this section, we present numerical experiments to demonstrate and compare the spectral distributions of the preconditioned systems of the saddle-point problem (1.1) with the existing preconditioner $\mathcal{M}_{\eta,1/\eta}$ in (1.5) and the new one $\mathcal{P}$ in (3.1), and the results of some Krylov subspace iterations. The edge elements of lowest order are used for the discretization of the system (1.2) in a square domain ($-1\le x\le 1, -1\le y\le 1$) or an L-shaped domain (see Figures 1 and 2), which is partitioned using unstructured simplicial meshes generated by EasyMesh [16]. For Meshes G1 through G5, the desired side lengths of the triangles that contain one of the vertices of the domain are set to be the same and meshes lead to linear systems of size $m+n=$ 187, 437, 1777, 7217, 23769 respectively. For Meshes L1 through L5, the desired side lengths of the triangles that contain the origin are one-tenth of the desired side lengths of the triangles that contain other vertices of the domain and $m+n=$ 185, 409, 1177, 5325, 29277 respectively.

We use MATLAB (inter(R) Core(TM) i7-4510U CPU @ 2.00 GHz 2.60 GHz, 4 GB RAM) to implement all numerical iterative solvers, and the $A+(\eta-k^2)M$-solvers are achieved by preconditioned CG method with the Hiptmair-Xu preconditioners [9], while the Laplacian-solvers achieved with an incomplete Cholesky factorization as an preconditioner. The right-hand side of (1.1), denoted by b, is set to be a vector with all components being ones, and the zero vector is used as the initial guess $x^{(0)}$ for all iterations. We run respectively the CG [18, a revised version of Algorithms 1] with the new preconditioner $\mathcal{P}$ for solving the saddle-point system (1.1), and the preconditoned MINRES with the block diagonal preconditioner $\mathcal{M}_{\eta,1/\eta}$, and denote these two methods by $\mathcal{P}$-CG and $\mathcal{M}_{\eta,1/\eta}$-MINRES respectively. The outer iterations are terminated

based on the criterion $\|b - \mathcal{K}x^{(k)}\|_2 \leq 10^{-6} \cdot \|b\|_2$, where $x^{(k)}$ is the kth iterate. We take the parameter $\eta = k^2 + 1$ and set the stopping criterion for all Laplacian-solvers (include both L-solvers and those Laplacian-solvers inside the Hiptmair-Xu preconditioners) to be a relative l_2-norm error of the residual less than a same tolerance, unless otherwise stated. The unit for iteration time is second and the time spent by incomplete Cholesky factorizations of two kinds of Laplacian matrices are added on for both methods.

Table 1: The rows $\mathcal{P}$-CG and $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES: iteration numbers (time) of the methods $\mathcal{P}$-CG and $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES respectively with various wave numbers k. The rows *Ratios*: ratios between the time spent by $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES and $\mathcal{P}$-CG. We take the parameter $\eta = k^2 + 1$ and the meshes G1 through G5.

k	0	1.0	1.55	1.6	2	4
Mesh G1						
$\mathcal{P}$-CG	5(0.6802)	6(0.7485)	11(1.2736)	11(1.268)	11(1.2619)	25(2.6965)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	7(0.8907)	9(1.1619)	15(1.6414)	15(1.6309)	14(1.5287)	31(3.2299)
Ratios	1.3095	1.5523	1.2888	1.2863	1.2114	1.1978
Mesh G2						
$\mathcal{P}$-CG	5(0.949)	7(1.1789)	12(1.9092)	12(1.8898)	11(1.7518)	28(4.1755)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	7(1.2051)	9(1.4634)	15(2.3171)	15(2.261)	15(2.2492)	31(4.4716)
Ratios	1.2698	1.2414	1.2137	1.1965	1.2839	1.0709
Mesh G3						
$\mathcal{P}$-CG	5(1.9158)	6(2.1352)	11(3.5818)	11(3.6182)	11(3.5805)	25(7.5431)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(2.7483)	9(2.9572)	15(4.6596)	15(4.6461)	14(4.337)	31(9.0741)
Ratios	1.4346	1.385	1.3009	1.2841	1.2113	1.203
Mesh G4						
$\mathcal{P}$-CG	5(6.2827)	6(7.2006)	9(10.3151)	9(10.2293)	11(12.1859)	24(25.2278)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	7(8.1305)	9(10.0955)	12(13.2987)	12(13.0848)	14(15.0183)	29(29.4057)
Ratios	1.2941	1.402	1.2892	1.2791	1.2324	1.1656
Mesh G5						
$\mathcal{P}$-CG	5(32.9871)	6(37.8760)	9(53.7789)	9(53.6557)	11(63.9640)	23(127.2520)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	7(43.0383)	8(48.4078)	12(69.4350)	12(69.3690)	14(79.6365)	29(156.8750)
Ratios	1.3047	1.2781	1.2911	1.2929	1.2450	1.2328

4.1 Iteration performance

In Tables 1 and 2 we show numbers and time of iteration for the methods $\mathcal{P}$-CG and $\mathcal{M}_{\eta,1/\eta}$-MINRES with different meshes and wave numbers, using the same tight inner tolerance 10^{-6} for all $A + (\eta - k^2)M$-solvers and Laplacian-solvers. The required number of iteration for the new method $\mathcal{P}$-CG is slightly smaller than that for the method $\mathcal{M}_{\eta,1/\eta}$-MINRES, which is consistent with our theoretical prediction in Section 3. The rows *Ratios* give the ratio between the time spent by $\mathcal{M}_{\eta,1/\eta}$-MINRES and $\mathcal{P}$-CG. The time spent by $\mathcal{M}_{\eta,1/\eta}$-MINRES is about $19\% \sim 55\%$ more than that by $\mathcal{P}$-CG for $k \leq 2$, while the difference becomes lightly smaller when we take the parameter $k = 4$. We can also observe that the required numbers of iteration are basically independent of mesh size. These experiments are just used to make a rough comparison between these two methods and in practice more efficient Laplacian-solvers should be used.

We also ran in Tables 3 and 4 the two methods with a variety of inner tolerances for $A + (\eta -$

Table 2: The rows $\mathcal{P}$-CG and $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES: iteration numbers (time) of $\mathcal{P}$-CG and $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES respectively with various wave numbers k. The rows *Ratios*: ratios between time spent by $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES and $\mathcal{P}$-CG. We take the parameter $\eta = k^2 + 1$ and the meshes L1 through L5.

k	0	1.0	1.2	1.25	2	4
Mesh L1						
$\mathcal{P}$-CG	5(0.8006)	7(1.0285)	9(1.2749)	8(1.1593)	10(1.3678)	25(3.3425)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(1.1821)	9(1.3191)	12(1.7786)	10(1.3660)	15(1.9633)	31(3.8042)
Ratios	1.4765	1.2826	1.3951	1.1783	1.4353	1.1382
Mesh L2						
$\mathcal{P}$-CG	6(1.3000)	7(1.3541)	9(1.7064)	8(1.5272)	12(2.2195)	28(4.8256)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(1.5563)	9(1.6863)	12(2.1860)	10(1.8770)	15(2.7282)	32(5.4724)
Ratios	1.1972	1.2454	1.2810	1.2290	1.2292	1.1340
Mesh L3						
$\mathcal{P}$-CG	5(1.8053)	7(2.1953)	9(2.7333)	8(2.5120)	12(3.5724)	25(7.0046)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(2.5005)	9(2.7071)	11(3.2302)	11(3.3899)	15(4.2672)	31(8.4044)
Ratios	1.3851	1.2331	1.1818	1.3495	1.1945	1.1998
Mesh L4						
$\mathcal{P}$-CG	5(5.3455)	7(6.9149)	8(7.7637)	8(7.7648)	12(11.0665)	24(21.0020)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(7.7230)	9(8.5107)	12(11.0335)	12(11.0477)	15(13.4594)	31(26.6709)
Ratios	1.4448	1.2308	1.4212	1.4228	1.2162	1.2699
Mesh L5						
$\mathcal{P}$-CG	5(53.0033)	7(69.6419)	8(78.3754)	8(78.5196)	10(95.8931)	26(231.7050)
$\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES	8(78.9650)	9(87.1453)	10(95.7683)	10(95.6296)	13(121.6840)	30(261.2080)
Ratios	1.4898	1.2513	1.2219	1.2179	1.2690	1.1273

$k^2)M$-solver (listed on the left side) and L-solver (listed on the top), while tolerances for the Laplacian-solvers inside the Hitmair-Xu preconditioners are specially fixed as 10^{-1}. First we can observe that times on the lower triangular part are smaller than these on the upper triangular part, which shows that a looser tolerance for $A + (\eta - k^2)M$-solver and a tighter tolerance for L-solver would be a good combination strategy. The $A + (\eta - k^2)M$-solver is relatively more expensive than L-solver. Having acknowledged this fact, it makes sense for us to adopt the above strategy. Table 5 gives ratios between the time listed in Table 4 and 3. Still we can observe that the new method $\mathcal{P}$-CG has a non-negligible advantage on $\mathcal{M}_{\eta,1/\eta}$-MINRES. We will have a more detailed discussion for the best inner tolerance(s) later.

Table 3: Iteration numbers (time) for $\mathcal{P}$-CG on Grid L4 with different inner tolerances for $A + (\eta - k^2)M$-solvers (listed on the left side) and L-solvers (listed on the top). Tolerance for the Laplacian-solvers inside the Hitmair-Xu preconditioners is specially fixed as 1e-1. we take the parameters $k = 0$ and $\eta = 1$ and denote 10^{-1} by 1e-1 etc.

	1e-5	1e-4	1e-3	1e-2	1e-1
1e-5	5(2.0808)	5(2.1045)	5(2.1666)	6(2.3585)	13(4.5092)
1e-4	5(1.6821)	5(1.6615)	5(1.6706)	6(2.0957)	13(3.9118)
1e-3	5(1.3027)	5(1.3421)	5(1.3220)	6(1.5743)	13(3.0280)
1e-2	7(1.1731)	6(1.1331)	6(1.1461)	6(1.1881)	13(2.3217)
1e-1	8(0.7564)	9(0.7813)	8(0.7640)	7(0.8225)	13(1.5585)

Table 4: Iteration numbers (time) for $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES on Grid L4 with different inner tolerances for $A+(\eta-k^2)M$-solvers (listed on the left side) and L-solvers (listed on the top). Tolerance for the Laplacian-solvers inside the Hitmair-Xu preconditioners is specially fixed as 1e-1. we take the parameters $k=0$ and $\eta=1$.

	1e-5	1e-4	1e-3	1e-2	1e-1
1e-5	8(2.9245)	8(3.0198)	9(3.1409)	10(3.5018)	21(7.5691)
1e-4	8(2.4044)	8(2.4364)	9(2.6340)	10(2.9974)	21(5.8159)
1e-3	8(1.7549)	8(1.7433)	9(2.0404)	10(2.3071)	21(4.5918)
1e-2	10(1.4443)	10(1.4053)	9(1.3736)	10(1.6640)	21(3.5088)
1e-1	18(1.2268)	18(1.2463)	15(0.9492)	17(1.1446)	21(2.0830)

Table 5: Ratios between the time listed in Table 3 and Table 4. the new method $\mathcal{P}$-CG has a non-negligible advantage on $\mathcal{M}_{\eta,1/\eta}$-MINRES in terms of time.

	1e-5	1e-4	1e-3	1e-2	1e-1
1e-5	1.4054	1.4349	1.4497	1.4848	1.6786
1e-4	1.4294	1.4664	1.5767	1.4303	1.4868
1e-3	1.3471	1.299	1.5435	1.4655	1.5164
1e-2	1.2311	1.2402	1.1985	1.4005	1.5113
1e-1	1.6219	1.5952	1.2425	1.3915	1.3365

In Tables 6 we report required numbers of iteration for $\mathcal{P}$-CG ($\mathcal{M}_{\eta,1/\eta}$-MINRES) with different values of $\eta-k^2$ for the given less accurate tolerance of the inner iterations associated with L and $A+(\eta-k^2)M$. If we take the parameter $k=0$ or 1, it is good enough to take the parameter $\eta-k^2=1$; However, this is not true for the cases when $k=2$ or $k=4$. In these cases, each iteration method with sightly larger values of $\eta-k^2$, for example, $\eta-k^2=1/2k^2$ perform better.

Table 6: The time spent by $\mathcal{P}$-CG ($\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES) on Grid L4 with different values of k and $\eta-k^2$. Inner tolerances for $A+(\eta-k^2)M$-solver and L-solver are set to be 1e-2 and 1e-4 respectively.

η	k^2+1	k^2+2	k^2+4	k^2+8	k^2+16
$k=0$	1.5594(1.9346)	1.5704(2.1003)	1.7223(2.2196)	2.0911(2.7166)	3.1828(3.7524)
$k=1$	2.2820(2.5947)	2.2033(2.6732)	2.2419(2.7203)	2.3651(3.7658)	3.4197(4.2749)
$k=2$	3.8534(4.5152)	3.4527(3.7501)	3.1498(3.5301)	3.8618(5.0507)	4.3371(6.1675)
$k=4$	9.3866(11.9675)	8.5725(8.8261)	6.4138(7.3868)	7.3863(8.9812)	7.9024(9.1783)

From Tables 3 and 4 we may also observe the $\mathcal{P}$-CG and $\mathcal{M}_{\eta,1/\eta}$-MINRES methods with the tolerance pair (1e-1, 1e-5) (for $A+(\eta-k^2)M$-solver and L-solver respectively) performs best when we take the parameter $k=0,\eta=1$. In Table 7 we make some more extra experiments to further investigate this phenomenon. It shows that when $k=0,1,1.2$ or 1.25, the method with tolerance pair (1e-1,1e-5) performs best. However, while we take the parameter $k=4$, one may need a relatively tighter inner solver. Also in Table 7 we list the ratios between iteration time for $\mathcal{M}_{\eta,1/\eta}$-MINRES and $\mathcal{P}$-CG.

Table 7: The first column: tolerance pair for $A + (\eta - k^2)M$-solver/L-solver; The first line for each tolerance pair: time spent by $\mathcal{P}$-CG ($\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES) on Grid L4 with different wave numbers k ($\eta = 1$ if $k^2 < 1$, otherwise $\eta = 1.5k^2$). The second line for each tolerance pair: ratios between the time spent by $\mathcal{M}_{\eta,\frac{1}{\eta}}$-MINRES and $\mathcal{P}$-CG.

k	0	1	1.2	1.25	2	4
1e-4/1e-5	3.4148(4.5445)	3.7516(5.6651)	4.6408(5.9792)	4.5615(5.9787)	5.9447(7.5844)	12.3785(13.2688)
	1.3308	1.5100	1.2884	1.3107	1.2758	1.0719
1e-3/1e-5	2.3986(3.1542)	3.2000(4.4446)	3.4226(4.7965)	3.7015(4.8802)	4.4020(5.4902)	10.1951(10.8177)
	1.3150	1.3889	1.4014	1.3185	1.2472	1.0611
1e-2/1e-5	1.5173(2.0689)	1.9599(2.6223)	2.6071(3.6012)	2.6189(3.2503)	3.0453(3.4254)	6.8830(8.1561)
	1.3635	1.3380	1.3813	1.2411	1.1248	1.1850
1e-1/1e-5	0.8272(1.3258)	1.2529(1.5368)	2.7211(4.9149)	1.8538(3.4134)	3.7716(8.3334)	7.5735(9.7269)
	1.6027	1.2266	1.8062	1.8413	2.2095	1.2843

Table 8: Smallest eigenvalue of the matrix A_η on different meshes. $\eta = k^2+1$. Eigenvalues above dotted lines are positive (then we have A_η is positive definite), while eigenvalues underneath dotted lines are negative (A_η is not positive definite).

k	G1	G2	G3	G4
0	0.4677	0.4738	0.4776	0.4769
1	0.4677	0.4738	0.4776	0.4769
1.55	0.0340	0.0360	0.0369	0.0373
1.6	-0.0544	-0.0543	-0.0536	-0.0533
2	-0.8719	-0.8988	-0.8875	-0.8823
4	-7.7031	-7.9434	-7.8425	-7.7907

k	L1	L2	L3	L4
0	0.4787	0.4582	0.4758	0.4654
1	0.2496	0.2575	0.2704	0.2753
1.2	0.0039	0.0128	0.0175	0.0200
1.25	-0.0646	-0.0556	-0.0530	-0.0512
2	-1.4349	-1.4249	-1.4580	-1.4674
4	-8.3127	-8.3664	-8.3974	-8.4450

4.2 Spectral analysis

In numerical experiments in Table 8 we compute the smallest eigenvalue of the matrix

$$A_\eta = \begin{pmatrix} A + \eta B^T L^{-1} B - k^2 M & 0 \\ 0 & I_m \end{pmatrix},$$

denoted by $\lambda_{\min}(A_\eta)$, to test definiteness of A_η. It is easy to see by (3.2) that we can apply CG with our new preconditioner, under the special inner product defined in (3.4) if A_η is symmetric positive definite. The first observation is that on Meshes G1 through G4, the matrices A_η are symmetric positive definite with $k = 0, 1$, or 1.55, and not positive definite anymore with $k = 1.6, 2$ or 4. Similarly, on Meshes L1 through L4, the matrices A_η are symmetric positive definite with $k = 0, 1$, or 1.2, and not positive definite with $k = 1.25, 2$, or 4. As predicted by Theorem 3.2, the definiteness of A_η is independent of mesh size. It is important to note that for smaller k, A_η is symmetric positive definite, thus CG can be used with the new preconditioner $\mathcal{P}$ instead of MINRES, though the original system $\mathcal{K}$ is indefinite. When k is not small enough, the corresponding preconditioned matrices are no longer positive definite even under the special non-standard inner product. Thus the corresponding minimum residual method should be used, theoretically; However, the numerical results in Subsection 4.1 indicate that CG still does not fail even when this violation occurs, and actually converge equally stably and fast. We would explain the reason for this important phenomenon in the next paragraph.

We make some more experiments to further compare the stability of the new and existing preconditioner $\mathcal{P}$ and $\mathcal{M}_{\eta,1/\eta}$. We can clearly see from Subsection 4.1 that CG can be always applied numerically with the new preconditioner $\mathcal{P}$ and it converges very well, though it may not guarantees to converge theoretically. But this is not the case for the preconditioner $\mathcal{M}_{\eta,1/\eta}$ when CG is used. To see this, we re-ran all the experiments in Table 4, but with CG iteration now instead MINRES. In each of the 30 numerical experiments, we have always experienced the case that one dividend becomes too small, which causes the break-down of the iterative process. The reasons behind are in fact very simple: we needs to divide by $p_k^T \mathcal{K} p_k$ (with p_k being the k-th search direction) at the k-th CG iteration with preconditioner $\mathcal{M}_{\eta,1/\eta}$, and to divide by $p_k^T A_\eta p_k$ at the k-th CG iteration with the new preconditioner $\mathcal{P}$, due to the existence of a special inner product (3.4). Figure 3 shows the distributions of the eigenvalues smaller than 0.3 of the two matrices $\mathcal{K}$ and A_η for $k = 4$, and these smaller and negative eigenvalues contribute mainly to the break-down of the iterations (most eigenvalues are larger than 0.3, but not shown in the figure). As one can see from the figure, there are many more eigenvalues in the red part for $\mathcal{K}$ than in the blue part for A_η, which explains clearly the highly instability of CG with the preconditioner $\mathcal{M}_{\eta,1/\eta}$ and good stability of CG with the new preconditioner $\mathcal{P}$.

Figure 3: Distributions of eigenvalues smaller than 0.3 of the coefficient matrix $\mathcal{K}$ (red part) and the matrix A_η (blue part) on Grid G3 for $k = 4$, and $\eta = k^2 + 1$. There are many more eigenvalues close to zero in the red part for $\mathcal{K}$ than in the blue part for A_η

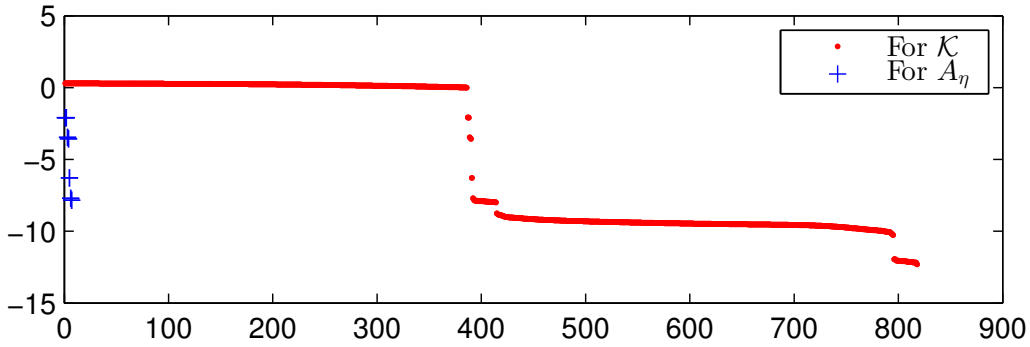


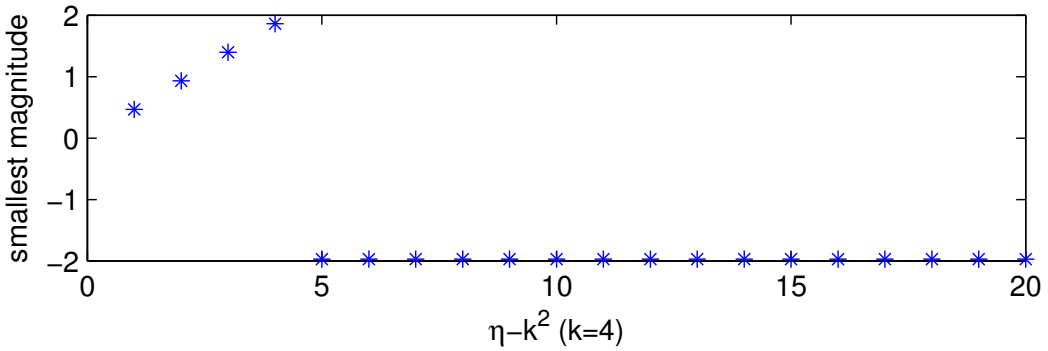
Figure 4 show influence of values of $\eta - k^2$ on the smallest magnitude eigenvalue of A_η. A larger $\eta - k^2$ makes the smallest magnitude eigenvalue of A_η bigger in terms of magnitude when $\eta - k^2$ is too small.

Table 9: Smallest magnitude eigenvalue of the matrix A_η on different meshes. These eigenvalues are well bounded from zero.

Mesh	G1	G2	G3	G4	L1	L2	L3	L4
$k = 4, \eta = 17$	0.4677	0.4738	0.4776	0.4769	0.4787	0.4582	0.4758	0.4654
$k = 4, \eta = 24$	1.8963	-2.0601	-2.0846	-2.0966	-1.9134	-1.9650	-1.9311	-1.9705
$k = 2, \eta = 5$	0.4677	0.4738	0.4776	0.4769	-0.2541	-0.2640	-0.2705	-0.2692
$k = 2, \eta = 6$	0.5061	0.5310	0.5311	0.5338	-0.2540	-0.2640	-0.2705	-0.2692

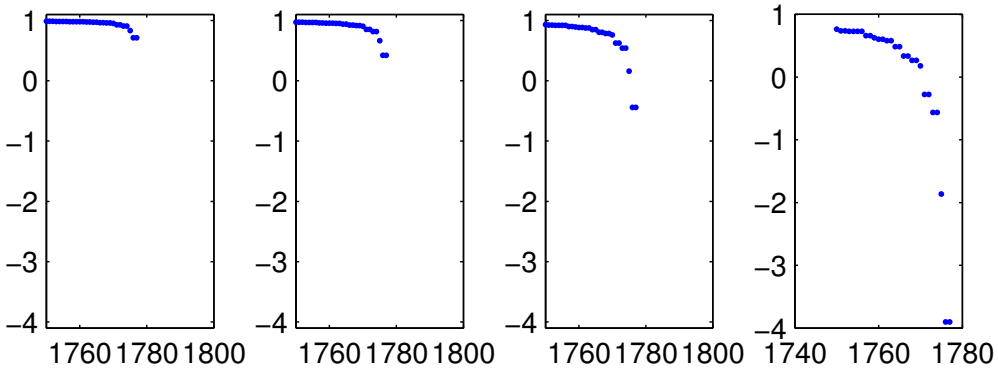
To check weather we can use CG with the new preconditioner $\mathcal{P}$ with proper parameter $k \leq 4$ on arbitrary large mesh, we in Table 9 investigate the influence of mesh size on the

Figure 4: smallest magnitude eigenvalue of the matrix A_η on Grid L4 for $k = 4$, and $\eta - k^2 = 1, 2 \cdots 20$. This figure shows that the parameter $\eta - k^2$ should not be two small.



smallest magnitude eigenvalue of A_η. We can observe that the smallest magnitude eigenvalue of A_η almost keeps consistent with the mesh size, in terms of magnitude. This phenomenon suggest that we can use CG with the new preconditioner even when $k^2 \geq \bar{\alpha}$. All in all, the numerical experiment shows very good stability and convergence of CG with the new preconditioner $\mathcal{P}$.

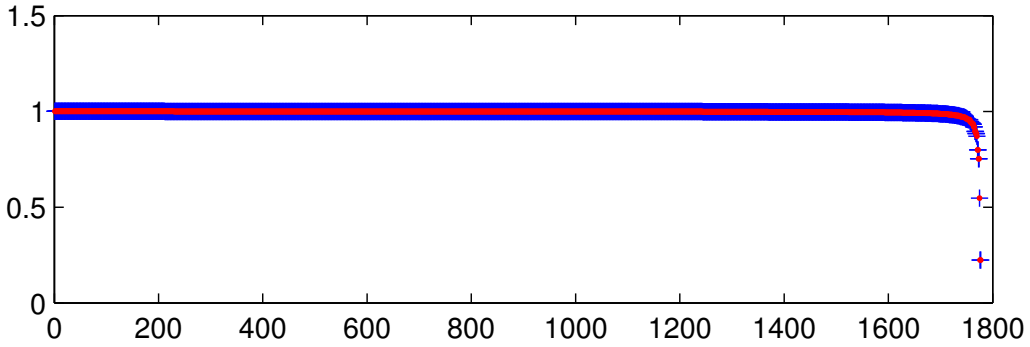
Figure 5: The distributions of the smallest 27 eigenvalues of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ on Grid G3 (from left to right: $k = 0, 1, 2, 4$.) The parameter η satisfies $\eta = k^2 + 1$. The distribution gets worse when k becomes lager.



Next we demonstrate the distributions of the smallest 27 eigenvalues of the preconditioned matrices $\mathcal{P}^{-1}\mathcal{K}$. Figure 5 plots the eigen-distribution of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ on Mesh G3 with different wave number k. We can see that the eigenvalues for $k = 0$ and $k = 1$ is well bounded, and there are only a few eigenvalues that lie between 0.22 and 0.8, while all the remaining eigenvalues stay in the range 0.8 and 1, with many of them being 1. These results are consistent with our theoretical prediction (Theorem 3.5). For $k = 2$ and $k = 4$, we see negative eigenvalues. The higher the wave number is, the worse eigenvalues distribute.

Figure 6 shows that the eigenvalues of the preconditioned matrices $\mathcal{P}^{-1}\mathcal{K}$ and $\mathcal{M}_{\eta,\varepsilon}^{-1}\mathcal{K}$ with $\varepsilon = -\frac{1}{\eta - k^2}$ are exactly the same for Mesh G3, with $k = 1.3$ and $\eta = k^2 + 1$.

Figure 6: The eigenvalue distributions of the preconditioned matrix $\mathcal{P}^{-1}\mathcal{K}$ (red part) and $\mathcal{M}_{\eta,\varepsilon}^{-1}\mathcal{K}$ (blue part) on Grid G3 for $k = 1.3$, $\varepsilon = -\frac{1}{\eta - k^2}$ and $\eta = k^2 + 1$. These two distributions are the same.



5 Conclusion Remarks

In this paper, based on the work in [6] and [7], we present generalized preconditioning techniques for saddle-point systems arising from the edge element discretization of stationary Maxwell equations and time-harmonic Maxwell equations. A parameter η is introduced to make the whole method more adaptable. Spectral properties of the preconditioners are briefly analyzed, which appear similar but slightly better compared with existing block diagonal precondtioners. Numerically we also find that our method has a considerable advantage in terms of iteration time, compared with MINRES with the block precondtioner $\mathcal{M}_{\eta,\frac{1}{\eta}}$. In numerical experiments we also test the influence of the introduced parameter η and inexact inner solvers.

More experiments are needed for three-dimensional Maxwell problems, where discontinuous coefficients with large jumps or more realistic domain should be considered. Efficient multigrid methods should be applied instead of incomplete Cholesky preconditioners for inner Laplacian solvers. Moreover we need also to test the parameter η and tolerances for inner solvers in these cases. There are several tips for possibly improving the performance of these methods. First, though a same tight tolerance for all L-solvers is good enough, different strategies for different L-solvers can be considered. Different ways to implement the action of $\mathcal{P}^{-1}$ can also be checked numerically for comparison, when different kind of right hand sides are given.

At last it should be noted that all our theoretical analyses base on the exact inner solvers. So in order to using CG, we may have to use exact preconditioned iterations for inner solvers. However, in our numerical experiments tolerance for $A + (\eta - k^2)M$-solvers can be quite loose and inexact inner solvers would enormously accelerate the whole iterations. Though it is a risk full of challenges to using inexact inner iterations, we think in future we should investigate stabilization of inexact inner solvers more carefully.

References

[1] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[2] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Vol. 44 of Springer Series in Computational Mathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[3] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[4] G.-H. Cheng, T.-Z. Huang, and S.-Q. Shen, *Block triangular preconditioners for the discretized time-harmonic Maxwell equations in mixed form*, Comput. Phys. Commun., 180 (2009), pp. 192–196.

[5] L. Demkowicz and L. Vardapetyan, *Modeling of electromagnetic absorption/scattering problems using hp-adaptive finite elements*, Comput. Methods Appl. Mech. Engrg., 152 (1998), pp. 103–124.

[6] R. Estrin and C. Greif, *On nonsingular saddle-point systems with a maximally rank deficient leading block*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 367–384.

[7] C. Greif and D. Schötzau, *Preconditioners for the discretized time-harmonic Maxwell equations in mixed form*, Numer. Lin. Algebra Appl., 14 (2007), pp. 281–297.

[8] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numerica, 11 (2002), pp. 237-339.

[9] R. Hiptmair and J. Xu, *Nodal Auxiliary Space Preconditioning in H(curl) and H(div) Spaces*, SIAM J. Numer. Anal., 45 (2007), pp. 2483–2509.

[10] P. Houston, I. Perugia, and D. Schötzau, *Mixed discontinuous Galerkin approximation of the Maxwell operator: non-stabilized formulation*, J. Sci. Comput., 22-23 (2005), pp. 315–346.

[11] Q. Hu and J. Zou, *Substructuring preconditioners for saddle-point problems arising from Maxwell's equations in three dimensions*, Math. Comput., 73 (2004), pp. 35–61.

[12] T. Kolev and P. Vassilevski, *Some experience with a H^1-based auxiliary space AMG for H(curl) Problems*, Report UCRL-TR-221841, LLNL, Livermore, CA, 2006.

[13] D. Li, C. Greif, and D. Schötzau, *Parallel numerical solution of the time-harmonic Maxwell equations in mixed form*, Numer. Lin. Algebra Appl., 19 (2012), pp. 525–539.

[14] P. Monk, *Analysis of a finite element method for Maxwell's equations*, SIAM J. Numer. Anal., 29 (1992), pp. 714–729.

[15] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[16] B. Niceno, *EasyMesh*. `http://web.mit.edu/easymesh_v1.4/www/easymesh.html`.

[17] I. Perugia, D. Schötzau, and P. Monk, *Stabilized interior penalty methods for the time-harmonic Maxwell equations*, Comput. Methods Appl. Mech. Eng., 191 (2002), pp. 4675–4697.

[18] J. Pestana and A. J. Wathen, *Combination preconditioning of saddle point systems for positive definiteness*, Numer. Lin. Algebra Appl., 20 (2013), pp. 785–808.

[19] S. L. Wu, T. Z. Huang, and C. X. Li, *Modified block preconditioners for the discretized time-harmonic Maxwell equations in mixed form*, J. Comput. Appl. Math., 237 (2013), pp. 419–431.

[20] Y. Zeng and C. Li, *New preconditioners with two variable relaxation parameters for the discretized time-harmonic Maxwell wquations in mixed form*, Math. Comput. Probl. Eng., 2012 (2012), pp. 1–13.

[21] Tian Y. and Takane Y., *The inverse of any two-by-two nonsingular partitioned matrix and three matrix inverse completion problems*, Comp. Math. Appl., 2009, 57(8), pp. 1294–1304.

[22] *hypre* : High performance preconditioners. `http://www.llnl.gov/CASC/hypre/`.

[23] J. M. Miao, *General expressions for the Moore-Penrose inverse of a* 2×2 *block matrix,* Linear Algebra Appl., 151 (1991), pp. 1-15.