Received: 15 February 2019 | Revised: 17 June 2019 | Accepted: 2 July 2019

DOI: 10.1002/nla.2265

RESEARCH ARTICLE

WILEY

Uzawa methods for a class of block three-by-three saddle-point problems

Na Huang[1] | **Yu-Hong Dai**[2,3] | **QiYa Hu**[2,3]

[1]Department of Applied Mathematics, College of Science, China Agricultural University, Beijing, China

[2]LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

[3]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Correspondence

Yu-Hong Dai, LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 100190 Beijing, China; or School of Mathematical Sciences, University of Chinese Academy of Sciences, 100049 Beijing, China.
Email: dyh@lsec.cc.ac.cn

Funding information

National Postdoctoral Program for Innovative Talents, Grant/Award Number: BX201600182; China Postdoctoral Science Foundation, Grant/Award Number: 2016M600141; Chinese NSF, Grant/Award Number: 11631013, 11331012, 71331001 and 11971372; Key Project of Chinese National Programs for Fundamental Research and Development, Grant/Award Number: 2015CB856002; National Natural Science Foundation of China, Grant/Award Number: G11571352

Summary

In this work, we consider numerical methods for solving a class of block three-by-three saddle-point problems, which arise from finite element methods for solving time-dependent Maxwell equations and some other applications. The direct extension of the Uzawa method for solving this block three-by-three saddle-point problem requires the exact solution of a symmetric indefinite system of linear equations at each step. To avoid heavy computations at each step, we propose an inexact Uzawa method, which solves the symmetric indefinite linear system in some inexact way. Under suitable assumptions, we show that the inexact Uzawa method converges to the unique solution of the saddle-point problem within the approximation level. Two special algorithms are customized for the inexact Uzawa method combining the splitting iteration method and a preconditioning technique, respectively. Numerical experiments are presented, which demonstrated the usefulness of the inexact Uzawa method and the two customized algorithms.

KEYWORDS

inexact Uzawa method, saddle-point problem, Uzawa method

AMS CLASSIFICATION

65F10; 65F50

1 | INTRODUCTION

In this work, we consider the following block three-by-three saddle-point problems:

$$\mathcal{A}\ell := \begin{pmatrix} A & B^T & 0 \\ B & 0 & C^T \\ 0 & C & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f \\ g \\ h \end{pmatrix}, \tag{1}$$

where $A \in R^{n\times n}$ is a symmetric positive definite matrix, and $B \in R^{m\times n}$ and $C \in R^{l\times m}$ have full row rank. It is not difficult to verify that the coefficient matrix $\mathcal{A}$ is nonsingular, which means that the system (1) has a unique solution. The linear system of form (1) arises from the finite element methods for solving time-dependent Maxwell equations

https://doi.org/10.1002/nla.2265

with discontinuous coefficients in general three-dimensional Lipschitz polyhedral domains,[1] and the following quadratic program[2]:

$$\min\left\{\frac{1}{2}x^TAx+r^Tx+q^Ty\right\},\quad \text{s.t.}Bx+C^Ty=b,\quad x\in R^n,\quad y\in R^l,$$

where $r\in R^n$ and $q\in R^l$ are given vectors.

Numerical solutions for system of linear equations play an important role in many applications, such as computational fluid dynamics, diffuse optical tomography, molecular scattering, lattice quantum chromodynamics, optimal control, electronic networks, and computer graphics; see other works for example.[3–8] Linear systems of block two-by-two form,

$$\begin{pmatrix} H & E^T \\ E & -D \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} f \\ g \end{pmatrix} \tag{2}$$

known as classical saddle-point problems, have widely been studied in decades, where H is positive, D is positive semidefinite, and E is full row rank. There is a great variety of iteration methods for solving the classical saddle-point problem (2), such as Krylov subspace methods,[9,10] Uzawa schemes,[11–22] HSS-type methods,[23–28] iterative null space methods,[29–31] and iterative projection methods.[32] Although the block three-by-three linear system (1) can be seen as a special case of the problem (2), methods for the latter cannot be applied to (1) directly. This is because either the matrix H is symmetric indefinite or the matrix E is rank deficient if we partition $\mathcal{A}$ into the form of the coefficient matrix in (2) by

$$\left(\begin{array}{cc|c} A & B^T & 0 \\ B & 0 & C^T \\ \hline 0 & C & 0 \end{array}\right) \quad\text{or}\quad \left(\begin{array}{c|cc} A & B^T & 0 \\ \hline B & 0 & C^T \\ 0 & C & 0 \end{array}\right). \tag{3}$$

In addition, we emphasize that the saddle-point problem (1), in essence, is quite different from the block three-by-three linear systems considered in other works.[33–36] Actually, the latter one possesses the form of

$$\begin{pmatrix} A & B^T & -I \\ B & 0 & 0 \\ -Z & 0 & -X \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} f \\ g \\ h \end{pmatrix},$$

where X and Z are diagonal matrices and positive definite, and I is the identity matrix. Therefore, it is necessary to study the numerical methods for the system (1) based on its specific structure.

In this work, we extend the Uzawa method studied in other works[11–13,15,16,37,38] to solve the saddle-point problem (1). The direct extension of the Uzawa method requires the exact solution of a symmetric indefinite system of linear equations at each step. To avoid heavy computations at each step, we propose an inexact Uzawa method by an approximate solution to replace the computation of solving a linear system. Theoretical analyses show that the inexact Uzawa method converges to the unique solution of the system (1) under suitable assumptions on the approximation calculation level. Numerical results also show that the inexact Uzawa method is more effective than the minimal residual (MINRES) method.[10,39]

The organization of this work is as follows. We study the direct extension of the Uzawa method in Section 2. An inexact Uzawa method and its convergence properties are provided in Section 3. Two special cases of the inexact Uzawa method are proposed in Section 4. Several numerical experiments are presented in Section 5.

We end this section with an introduction of some notation that will be used in the subsequent analysis. For any matrix $H\in R^{l\times l}$, we shall often write its spectral radius, spectral set, inverse, and transpose as $\rho(H)$, $\text{sp}(H)$, H^{-1} and H^T, respectively. λ_H and Λ_H denote the minimum and maximum eigenvalues of a symmetric matrix H, respectively. $\|\cdot\|$ means the usual 2-norm. For an $l\times l$ symmetric positive definite matrix G, $\|x\|_G=\sqrt{\langle Gx,x\rangle}=\|G^{\frac{1}{2}}x\|$ for all $x\in R^l$ and $\|H\|_G=\sup_{x\neq 0}\frac{\|Hx\|_G}{\|x\|_G}=\|G^{\frac{1}{2}}HG^{-\frac{1}{2}}\|$ for all $H\in R^{l\times l}$. For each element $v=(v_1^T,v_2^T)^T\in R^{n+m+l}$ with $v_1\in R^{n+m}$ and $v_2\in R^l$, the energy-norm $|||\cdot|||$ is defined as follows.[18–20]

$$|||v|||^2=\|v_1\|_{K^{-1}}^2+\|v_2\|_{\tilde{S}}^2, \tag{4}$$

where $\widetilde{S} = C(BA^{-1}B^T)^{-1}C^T = CS^{-1}C^T$ and $K \in R^{(n+m)\times(n+m)}$ has the form of

$$K = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}. \tag{5}$$

2 | DIRECT EXTENSION OF THE UZAWA METHOD

The Uzawa method is an effective method for solving the saddle-point problem (2), which has the following scheme[37,38]:

$$\begin{cases} x_{k+1} = H^{-1}(f - E^T y_k), \\ y_{k+1} = y_k + \tau Q^{-1}(g - Ex_{k+1} + Dy_k), \end{cases} \tag{6}$$

where τ is a given parameter and Q is a symmetric positive definite matrix. Recently, there is a rapidly increasing literature, which is concerned with the inexact Uzawa methods because of the minimal memory requirements and easiness to be implemented; see other works for example.[11–18,38,40–42]

In this section, we will employ the iteration scheme (6) to solve the block three-by-three saddle-point problem (1) by repartitioning its coefficient matrix $\mathcal{A}$ into the form of (2). By the notation

$$H = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}, \quad E = (\, 0 \;\; C\,), \quad w = \begin{pmatrix} x \\ y \end{pmatrix}, \quad q = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{7}$$

the linear system (1) can be rewritten as

$$\begin{pmatrix} H & E^T \\ E & 0 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} q \\ h \end{pmatrix}.$$

Because A is symmetric positive definite and B has full row rank, the matrix H is nonsingular. Then, the Uzawa method (6) can be applied to the above repartition system directly, which possesses the following scheme:

$$\begin{cases} w_{k+1} = H^{-1}(q - E^T z_k), \\ z_{k+1} = z_k + \tau Q^{-1}(h - Ew_{k+1}), \end{cases} \tag{8}$$

where $w_{k+1} = (x_{k+1}^T, y_{k+1}^T)^T$.

We now present an instance for which (8) will be failed if we apply it to solve the linear system (1) directly.

Example 1. Consider the saddle-point problem (1), where

$$\mathcal{A} = \begin{pmatrix} 1e-15 & -1e-15 & 0 \\ -1e-15 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{9}$$

We choose the right-hand-side vector of (1) so that its exact solution is given by the vector $(1,1,1)^T$. We set $Q = 1$, $\tau = 1e-15$ and the initial vector $(w_0^T, z_0)^T = 0$. By Matlab, we get $(w_1^T, z_1)^T = (-10^{15}, -10^{15}, 1)^T$ and $(w_2^T, z_2)^T = (0.9992, 0.9992, 1)^T$. The later iterations keep the same, that is, $(w_k^T, z_k)^T = (0.9992, 0.9992, 1)^T, \quad \forall \quad k \geq 2$.

Therefore, we have to consider some modification of the iterative scheme (8). Noticing that the inverse of the matrix H is of the form

$$H^{-1} = \begin{pmatrix} A^{-1} - A^{-1}B^TS^{-1}BA^{-1} & A^{-1}B^TS^{-1} \\ S^{-1}BA^{-1} & -S^{-1} \end{pmatrix}, \tag{10}$$

where the Schur complement matrix $S = BA^{-1}B^T$, the first formula in (8) can be expressed as

$$\begin{aligned}\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} &= \begin{pmatrix} A^{-1} - A^{-1}B^TS^{-1}BA^{-1} & A^{-1}B^TS^{-1} \\ S^{-1}BA^{-1} & -S^{-1} \end{pmatrix} \left[\begin{pmatrix} f \\ g \end{pmatrix} - \begin{pmatrix} 0 \\ C^T \end{pmatrix} z_k \right] \\ &= \begin{pmatrix} A^{-1} - A^{-1}B^TS^{-1}BA^{-1} & A^{-1}B^TS^{-1} \\ S^{-1}BA^{-1} & -S^{-1} \end{pmatrix} \begin{pmatrix} f \\ g - C^Tz_k \end{pmatrix} \\ &= \begin{pmatrix} A^{-1}[f - B^TS^{-1}(BA^{-1}f + C^Tz_k - g)] \\ S^{-1}(BA^{-1}f + C^Tz_k - g) \end{pmatrix}. \end{aligned} \tag{11}$$

Then, we obtain

$$y_{k+1} = S^{-1}(BA^{-1}f + C^Tz_k - g)$$

and

$$x_{k+1} = A^{-1}[f - B^TS^{-1}(BA^{-1}f + C^Tz_k - g)] = A^{-1}(f - B^Ty_{k+1}).$$

These two schemes combined with

$$z_{k+1} = z_k + \tau Q^{-1} \left[h - (\,0 \;\; C\,) \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} \right] = z_k + \tau Q^{-1}(h - Cy_{k+1})$$

follow a new method, called direct Uzawa method, to solve the linear system (1).

Algorithm 1 (Direct Uzawa method). *Let $(x_0^T, y_0^T, z_0^T)^T \in R^{n+m+l}$ be an arbitrary initial vector. The iterates of the direct Uzawa method for solving the linear system (1) are computed according to the following procedure:*

$$\begin{cases} y_{k+1} = S^{-1}(BA^{-1}f + C^Tz_k - g), \\ x_{k+1} = A^{-1}(f - B^Ty_{k+1}), \\ z_{k+1} = z_k + \tau Q^{-1}(h - Cy_{k+1}), \end{cases} \tag{12}$$

where τ is a given positive constant and Q is a given symmetric positive definite matrix.

By the same settings as that in Example 1, Algorithm 1 for solving the saddle-point problem (1) converges to the exact solution within two steps.

Due to the indefiniteness of the matrix H, the convergence results in the work of Arrow et al.[37] are not appropriate for Algorithm 1. Therefore, it is necessary to analyze the convergence properties of Algorithm 1 based on the iteration scheme (12). By the definitions of the following three matrices:

$$\mathcal{M} = \begin{pmatrix} I & A^{-1}B^T & 0 \\ 0 & I & 0 \\ 0 & \tau Q^{-1}C & I \end{pmatrix}, \qquad \mathcal{N} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & S^{-1}C^T \\ 0 & 0 & I \end{pmatrix}$$

and

$$\mathcal{K} = \begin{pmatrix} A^{-1} & 0 & 0 \\ S^{-1}BA^{-1} & -S^{-1} & 0 \\ 0 & 0 & \tau Q^{-1} \end{pmatrix};$$

the scheme (12) can equivalently be reformulated as

$$\mathcal{M} \begin{pmatrix} x_{k+1} \\ y_{k+1} \\ z_{k+1} \end{pmatrix} = \mathcal{N} \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix} + \mathcal{K} \begin{pmatrix} f \\ g \\ h \end{pmatrix}.$$

This implies that the iterative matrix of (12) is

$$\mathcal{T} = \mathcal{M}^{-1}\mathcal{N} = \begin{pmatrix} I & -A^{-1}B^T & 0 \\ 0 & I & 0 \\ 0 & -\tau Q^{-1}C & I \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & S^{-1}C^T \\ 0 & 0 & I \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 0 & -A^{-1}B^TS^{-1}C^T \\ 0 & 0 & S^{-1}C^T \\ 0 & 0 & I - \tau\hat{Q} \end{pmatrix},$$

where $\hat{Q} = Q^{-1}CS^{-1}C^T$. Then, Algorithm 1 is convergent if and only if the spectral radius of the iterative matrix $\mathcal{T}$ is less than 1, that is, $\rho(\mathcal{T}) = \rho(I - \tau Q^{-1}CS^{-1}C^T) < 1$. From this point of view, we can establish the following convergence theorem.

Theorem 1. *Suppose that $A \in R^{n\times n}$ is symmetric positive definite, and $B \in R^{m\times n}$ and $C \in R^{l\times m}$ have full row rank. Let $Q \in R^{l\times l}$ be a symmetric positive definite matrix. If the parameter τ satisfies*

$$0 < \tau < \frac{2}{\Lambda_{\hat{Q}}},$$

then the sequence $\{x_k, y_k, z_k\}$ produced by Algorithm 1 converges to the unique solution of the saddle-point problem (1).

Proof. Let λ be an eigenvalue of $\hat{Q}$. It is easy to see that $1-\tau\lambda$ is an eigenvalue of the matrix $I-\tau\hat{Q}$. Hence, $\rho(I-\tau\hat{Q}) < 1$ holds if and only if $|1-\tau\lambda| < 1$ holds for any λ. Noting that both of matrices Q and $S = BA^{-1}B^T$ are symmetric positive definite, so all the eigenvalues of $\hat{Q}$ are positive. Then, it can be shown that the parameter τ should satisfy

$$0 < \tau < \min_{\lambda\in \mathrm{sp}(\hat{Q})} \frac{2}{\lambda} = \frac{2}{\Lambda_{\hat{Q}}},$$

which follows the results. □

In the following, we shall derive the optimal parameter τ_{opt} of Algorithm 1, that is, the value of τ, which minimizes the spectral radius of the iterative matrix $\mathcal{T}$.

Theorem 2. *Under the same conditions as in Theorem 1, the optimal parameter and optimal spectral radius are*

$$\tau_{\text{opt}} = \frac{2}{\lambda_{\hat{Q}} + \Lambda_{\hat{Q}}}, \qquad \rho(\mathcal{T}) = \frac{\Lambda_{\hat{Q}} - \lambda_{\hat{Q}}}{\Lambda_{\hat{Q}} + \lambda_{\hat{Q}}},$$

respectively.

Proof. It can be shown that

$$\rho(\mathcal{T}) = \rho(I - \tau\hat{Q}) = \max\{|1 - \tau\lambda_{\hat{Q}}|, |1 - \tau\Lambda_{\hat{Q}}|\}.$$

Therefore, $\rho(\mathcal{T})$ reaches the minimum if $|1 - \tau\lambda_{\hat{Q}}| = |1 - \tau\Lambda_{\hat{Q}}|$. After some algebraic manipulation, we complete the proof. □

3 | INEXACT UZAWA METHODS

Although Algorithm 1 is convergent for some proper parameter τ, at each step, it requires finding solutions of linear systems with the coefficient matrices $S \in R^{m\times m}$ and $A \in R^{n\times n}$, which is very costly and impractical in actual implementations. The main reason for this high computation is that we substitute the exact expression (10) into the Uzawa scheme (8) directly during deducing Algorithm 1. To overcome this disadvantage and improve the efficiency of Algorithm 1, we can solve the subproblems iteratively. More specifically, the first iteration scheme in (8) can be rewritten as $w_{k+1} = w_k + H^{-1}q_k$ with $q_k = q - Hw_k - E^Tz_k$. Then, we may employ some iterative methods, like preconditioned MINRES method in conjunction with a suitable block diagonal preconditioners, to solve the system $H\Delta w = q_k$ and get an approximation solution $\Phi_H(q_k)$ within proper margin of error. Here, we assume

$$\|\Phi_H(q_k) - H^{-1}q_k\|_K \leq \delta \|q_k\|_{K^{-1}} \tag{13}$$

for some $\delta \in [0, 1)$, where the matrix K is defined as in (5). Comparing the existing assumption given in the works of Bramble et al.[14] and Hu et al.,[19] it seems more rational to replace the assumption (13) by

$$\|\Phi_H(q_k) - H^{-1}q_k\|_K \leq \delta \|H^{-1}q_k\|_K.$$

However, this change makes the analysis more difficult. This means that the considered 3×3 saddle-point problem is essentially different from the 2×2 saddle-point problems studied in the most existing works.[14,16,18–20] Besides, if H is symmetric positive definite, we can set $K = H$. Then, the assumption (13) can be equivalent to

$$\|\Phi_H(q_k) - H^{-1}q_k\|_H \leq \delta \|q_k\|_{H^{-1}} = \delta \|H^{-1}q_k\|_H,$$

which has been made in other works[14,18–20] for the 2×2 saddle-point problems.

The inexact Uzawa method is defined as follows.

Algorithm 2. (Inexact Uzawa method). *Given an arbitrary initial vector $(x_0^T, y_0^T, z_0^T)^T \in R^{n+m+l}$. The iterates of the inexact Uzawa method for solving the linear system (1) are computed according to the following procedure, where $w_k = (x_k^T, y_k^T)^T, k = 0, 1, 2, \dots$.*

(1) Compute

$$q_k = q - Hw_k - E^T z_k = \begin{pmatrix} f - Ax_k - B^T y_k \\ g - Bx_k - C^T z_k \end{pmatrix}$$

and

$$w_{k+1} = w_k + \Phi_H(q_k). \tag{14}$$

(2) Update

$$z_{k+1} = z_k + \tau Q^{-1}(h - Ew_{k+1}),$$

where τ is a given positive constant and Q is a given symmetric positive definite matrix.

Let $((w^*)^T, (z^*)^T)^T := ((x^*)^T, (y^*)^T, (z^*)^T)^T$ be the exact solution of the saddle-point problem (1). Define the following three error vectors:

$$e_k^w = w^* - w_k, \quad e_k^z = z^* - z_k, \quad E_k = \begin{pmatrix} \sqrt{\delta} q_k \\ e_k^z \end{pmatrix}. \tag{15}$$

By the definition of q_k, we know that

$$q_k = q - Hw_k - E^T z_k = Hw^* + E^T z^* - Hw_k - E^T z_k = He_k^w + E^T e_k^z, \tag{16}$$

which follows $e_k^w = H^{-1}(q_k - E^T e_k^z)$. This together with (14) yields that

$$\begin{aligned} e_{k+1}^w &= w^* - w_{k+1} = w^* - w_k - \Phi_H(q_k) = e_k^w - \Phi_H(q_k) \\ &= H^{-1}(q_k - E^T e_k^z) - \Phi_H(q_k) \\ &= H^{-1}q_k - \Phi_H(q_k) - H^{-1}E^T e_k^z. \end{aligned} \tag{17}$$

Then, we obtain

$$\begin{aligned} e_{k+1}^z &= z^* - z_{k+1} = z^* - z_k - \tau Q^{-1}(Ew^* - Ew_{k+1}) \\ &= e_k^z - \tau Q^{-1}Ee_{k+1}^w = e_k^z - \tau Q^{-1}E\left(H^{-1}q_k - \Phi_H(q_k) - H^{-1}E^T e_k^z\right) \\ &= -\tau Q^{-1}E(H^{-1}q_k - \Phi_H(q_k)) + (I + \tau Q^{-1}EH^{-1}E^T)e_k^z. \end{aligned} \tag{18}$$

By (16) and (17), we have

$$\begin{aligned} q_{k+1} &= He^{w}_{k+1} + E^{T}e^{z}_{k+1} \\ &= (H - \tau E^{T}Q^{-1}E)\left(H^{-1}q_k - \Phi_H(q_k)\right) + \tau E^{T}Q^{-1}EH^{-1}E^{T}e^{z}_{k}. \end{aligned} \tag{19}$$

Noting that

$$\begin{aligned} EH^{-1}E^{T} &= (\,0\ \ C\,)\begin{pmatrix} A^{-1} - A^{-1}B^{T}S^{-1}BA^{-1} & A^{-1}B^{T}S^{-1} \\ S^{-1}BA^{-1} & -S^{-1} \end{pmatrix}\begin{pmatrix} 0 \\ C^{T} \end{pmatrix} \\ &= -CS^{-1}C^{T} = -\widetilde{S}, \end{aligned}$$

it follows from (18) and (19) that

$$q_{k+1} = (H - \tau E^{T}Q^{-1}E)\left(H^{-1}q_k - \Phi_H(q_k)\right) - \tau E^{T}Q^{-1}\widetilde{S}e^{z}_{k}$$

and

$$e^{z}_{k+1} = -\tau Q^{-1}E(H^{-1}q_k - \Phi_H(q_k)) + (I - \tau Q^{-1}\widetilde{S})e^{z}_{k}.$$

As the matrices K and $\widetilde{S}$ are symmetric positive definite, combining with the above equalities leads to

$$\begin{pmatrix} \sqrt{\delta}K^{-\frac{1}{2}}q_{k+1} \\ \widetilde{S}^{\frac{1}{2}}e^{z}_{k+1} \end{pmatrix} = \mathcal{Q}\begin{pmatrix} \frac{1}{\sqrt{\delta}}K^{\frac{1}{2}}\left(H^{-1}q_k - \Phi_H(q_k)\right) \\ \widetilde{S}^{\frac{1}{2}}e^{z}_{k} \end{pmatrix},$$

where

$$\mathcal{Q} = \begin{pmatrix} \delta K^{-\frac{1}{2}}(H - \tau E^{T}Q^{-1}E)K^{-\frac{1}{2}} & -\tau\sqrt{\delta}K^{-\frac{1}{2}}E^{T}Q^{-1}\widetilde{S}^{\frac{1}{2}} \\ -\tau\sqrt{\delta}\widetilde{S}^{\frac{1}{2}}Q^{-1}EK^{-\frac{1}{2}} & I - \tau\widetilde{S}^{\frac{1}{2}}Q^{-1}\widetilde{S}^{\frac{1}{2}} \end{pmatrix}. \tag{20}$$

Then, from (4), (13), and (15), we have

$$\begin{aligned} |||E_{k+1}||| &= \sqrt{\|\sqrt{\delta}q_{k+1}\|^2_{K^{-1}} + \|e^{z}_{k+1}\|^2_{\widetilde{S}}} \\ &= \left\|\begin{pmatrix} \sqrt{\delta}K^{-\frac{1}{2}}q_{k+1} \\ \widetilde{S}^{\frac{1}{2}}e^{z}_{k+1} \end{pmatrix}\right\| \le \rho(\mathcal{Q})\left\|\begin{pmatrix} \frac{1}{\sqrt{\delta}}K^{\frac{1}{2}}(H^{-1}q_k - \Phi_H(q_k)) \\ \widetilde{S}^{\frac{1}{2}}e^{z}_{k} \end{pmatrix}\right\| \\ &= \rho(\mathcal{Q})\sqrt{\frac{1}{\delta}\|H^{-1}q_k - \Phi_H(q_k)\|^2_{K} + \|\widetilde{S}^{\frac{1}{2}}e^{z}_{k}\|^2} \\ &\le \rho(\mathcal{Q})\sqrt{\|\sqrt{\delta}q_k\|^2_{K^{-1}} + \|e^{z}_{k}\|^2_{\widetilde{S}}} = \rho(\mathcal{Q})|||E_k|||. \end{aligned}$$

Therefore, in order to show the convergence of Algorithm 2, we just need to derive the conditions that guarantee $\rho(\mathcal{Q}) < 1$.

Substituting (5) and (7) into (20), we can rewrite the matrix $\mathcal{Q}$ as

$$\mathcal{Q} = \begin{pmatrix} \delta I & \delta A^{-\frac{1}{2}}B^{T}S^{-\frac{1}{2}} & 0 \\ \delta S^{-\frac{1}{2}}BA^{-\frac{1}{2}} & -\tau\delta S^{-\frac{1}{2}}C^{T}Q^{-1}CS^{-\frac{1}{2}} & -\tau\sqrt{\delta}S^{-\frac{1}{2}}C^{T}Q^{-1}\widetilde{S}^{\frac{1}{2}} \\ 0 & -\tau\sqrt{\delta}\widetilde{S}^{\frac{1}{2}}Q^{-1}CS^{-\frac{1}{2}} & I - \tau\widetilde{S}^{\frac{1}{2}}Q^{-1}\widetilde{S}^{\frac{1}{2}} \end{pmatrix}.$$

This implies that

$$\rho(\mathcal{Q}) \le \delta\rho(\mathcal{Q}_1) + \rho(\mathcal{Q}_2), \tag{21}$$

where

$$\mathcal{Q}_1 = \begin{pmatrix} I & A^{-\frac{1}{2}}B^{T}S^{-\frac{1}{2}} \\ S^{-\frac{1}{2}}BA^{-\frac{1}{2}} & 0 \end{pmatrix} \tag{22}$$

and

$$\mathcal{Q}_2 = \begin{pmatrix} -\tau\delta S^{-\frac{1}{2}}C^TQ^{-1}CS^{-\frac{1}{2}} & -\tau\sqrt{\delta}S^{-\frac{1}{2}}C^TQ^{-1}\widetilde{S}^{\frac{1}{2}} \\ -\tau\sqrt{\delta}\widetilde{S}^{\frac{1}{2}}Q^{-1}CS^{-\frac{1}{2}} & I-\tau\widetilde{S}^{\frac{1}{2}}Q^{-1}\widetilde{S}^{\frac{1}{2}} \end{pmatrix}. \tag{23}$$

Then, we will derive the upper bounds of $\rho(\mathcal{Q}_1)$ and $\rho(\mathcal{Q}_2)$, respectively.

Lemma 1. *Let the matrix $\mathcal{Q}_1$ be defined as in (22). Then, the eigenvalues of $\mathcal{Q}_1$ are $(1-\sqrt{5})/2$, 1 and $(1+\sqrt{5})/2$. Therefore, the spectral radius of $\mathcal{Q}_1$ is $(1+\sqrt{5})/2$.*

Lemma 2. *Let the matrix $\mathcal{Q}_2$ be defined as in (23) and $\widehat{Q} = Q^{-1}CS^{-1}C^T$. Then, the upper bound of $\rho(\mathcal{Q}_2)$ meets the following estimate:*

$$\rho(\mathcal{Q}_2) = \frac{1}{2}\max\left\{ f(\lambda_{\widehat{Q}}) + \sqrt{f(\lambda_{\widehat{Q}})^2 + 4\delta\tau\lambda_{\widehat{Q}}}, \sqrt{f(\Lambda_{\widehat{Q}})^2 + 4\delta\tau\Lambda_{\widehat{Q}}} - f(\Lambda_{\widehat{Q}}) \right\},$$

where $f(\mu) = 1-(1+\delta)\tau\mu$.

Proof. Let $CS^{-\frac{1}{2}} = U(\,\Sigma\ 0\,)W^T$ be the singular value decomposition, where $\Sigma = \mathrm{diag}\{\sigma_1, \sigma_2, \dots, \sigma_l\}$, $\sigma_j > 0\,(j = 1, 2, \dots, l)$ is the singular value of $CS^{-\frac{1}{2}}$; $U \in R^{l\times l}$ and $W \in R^{m\times m}$ are two orthonormal matrices. Noting that

$$\widetilde{S} = CS^{-1}C^T = U(\,\Sigma\ 0\,)W^TW\begin{pmatrix}\Sigma\\0\end{pmatrix}U^T = U\Sigma^2U^T, \tag{24}$$

we have $\widetilde{S}^{\frac{1}{2}} = U\Sigma U^T$. Substituting this into (23), it follows

$$\mathcal{Q}_2 = \begin{pmatrix} -\tau\delta W\begin{pmatrix}\Sigma U^TQ^{-1}U\Sigma & 0\\ 0 & 0\end{pmatrix}W^T & -\tau\sqrt{\delta}W\begin{pmatrix}\Sigma U^TQ^{-1}U\Sigma\\ 0\end{pmatrix}U^T \\ -\tau\sqrt{\delta}U\left(\,\Sigma U^TQ^{-1}U\Sigma\ 0\,\right)W^T & I-\tau U\Sigma U^TQ^{-1}U\Sigma U^T \end{pmatrix}.$$

This implies that the matrix $\mathcal{Q}_2$ is similar to

$$\begin{pmatrix} -\tau\delta\Sigma U^TQ^{-1}U\Sigma & -\tau\sqrt{\delta}\Sigma U^TQ^{-1}U\Sigma & 0 \\ -\tau\sqrt{\delta}\Sigma U^TQ^{-1}U\Sigma & I-\tau\Sigma U^TQ^{-1}U\Sigma & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Assume that the eigenvalues of $\Sigma U^TQ^{-1}U\Sigma$ are $\mu_1, \mu_2, \dots, \mu_l$. Then, it can be shown that the matrix $\mathcal{Q}_2$ can further be similar to

$$\mathrm{diag}\left\{0, \begin{pmatrix} -\tau\delta\mu_1 & -\tau\sqrt{\delta}\mu_1 \\ -\tau\sqrt{\delta}\mu_1 & 1-\tau\mu_1 \end{pmatrix}, \dots, \begin{pmatrix} -\tau\delta\mu_l & -\tau\sqrt{\delta}\mu_l \\ -\tau\sqrt{\delta}\mu_l & 1-\tau\mu_l \end{pmatrix}\right\},$$

whose eigenvalues can be given explicitly by 0 and

$$\frac{1-(1+\delta)\tau\mu_i \pm \sqrt{[1-(1+\delta)\tau\mu_i]^2 + 4\delta\tau\mu_i}}{2}.$$

It follows from (24) that $\Sigma U^TQ^{-1}U\Sigma$ is similar to $Q^{-1}U\Sigma^2U^T = Q^{-1}\widetilde{S} = \widehat{Q}$. Therefore, $\mu_j(j = 1, 2, \dots, l)$ is also the eigenvalue of $\widehat{Q}$. Noticing that

$$\begin{aligned}\left(f(\mu) + \sqrt{f(\mu)^2 + 4\delta\tau\mu}\right)' &= \frac{f'(\mu)\sqrt{f(\mu)^2 + 4\delta\tau\mu} + f'(\mu)f(\mu) + 2\delta\tau}{\sqrt{f(\mu)^2 + 4\delta\tau\mu}} \\ &= -\frac{\tau\left[(1+\delta)\sqrt{f(\mu)^2 + 4\delta\tau\mu} + (1+\delta)f(\mu) - 2\delta\right]}{\sqrt{f(\mu)^2 + 4\delta\tau\mu}}\end{aligned}$$

and

$$f(\mu)^2 + 4\delta\tau\mu - \left[\frac{2\delta}{1+\delta} - f(\mu)\right]^2 = 4\delta\tau\mu - \frac{4\delta^2}{(1+\delta)^2} + \frac{4\delta}{1+\delta} f(\mu)$$
$$= -\frac{4\delta^2}{(1+\delta)^2} + \frac{4\delta}{1+\delta} > 0,$$

we know that the function $f(\mu) + \sqrt{f(\mu)^2 + 4\delta\tau\mu}$ is monotonically decreasing. This, together with the fact that $\sqrt{f(\mu)^2 + 4\delta\tau\mu} - f(\mu)$ is monotonically increasing, yields

$$\rho(Q_2) = \frac{1}{2}\max\left\{ f(\lambda_{\hat{Q}}) + \sqrt{f(\lambda_{\hat{Q}})^2 + 4\delta\tau\lambda_{\hat{Q}}}, \sqrt{f(\Lambda_{\hat{Q}})^2 + 4\delta\tau\Lambda_{\hat{Q}}} - f(\Lambda_{\hat{Q}}) \right\}.$$

This completes the proof. □

We are going to derive the convergence theorem of Algorithm 2. To this end, we introduce some notation, which will also be used frequently in subsequent sections. Let $a = (1+\sqrt{5})/2$ and

$$v_1 = \frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}},$$
$$v_2 = \frac{2\tau\Lambda_{\hat{Q}} - a\tau\Lambda_{\hat{Q}} + 3a - \sqrt{(2\tau\Lambda_{\hat{Q}} + a)^2 - a^2\tau(2-\tau\Lambda_{\hat{Q}})\Lambda_{\hat{Q}}}}{2a(a+\tau\Lambda_{\hat{Q}})}.$$

Then, we can establish the following theorem.

Theorem 3. *Suppose that $A \in R^{n\times n}$ is symmetric positive definite, and $B \in R^{m\times n}$ and $C \in R^{l\times m}$ have full row rank. Let $Q \in R^{l\times l}$ be a symmetric positive definite matrix. If the parameters τ and δ satisfy*

$$0 < \tau < \frac{2}{(1+\delta)\Lambda_{\hat{Q}}}, \qquad 0 \le \delta < \min\{v_1, v_2\},$$

then the sequence $\{x_k, y_k, z_k\}$ produced by Algorithm 2 converges to the unique solution of the saddle-point problem (1).

Proof. From (21) and Lemmas 1 and 2, we know that Algorithm 2 is convergent if

$$2a\delta + f(\lambda_{\hat{Q}}) + \sqrt{f(\lambda_{\hat{Q}})^2 + 4\delta\tau\lambda_{\hat{Q}}} < 2 \tag{25}$$

and

$$2a\delta + \sqrt{f(\Lambda_{\hat{Q}})^2 + 4\delta\tau\Lambda_{\hat{Q}}} - f(\Lambda_{\hat{Q}}) < 2. \tag{26}$$

It follows from (25) that

$$2 - 2a\delta - f(\lambda_{\hat{Q}}) > 0 \tag{27}$$

and

$$a(a - \tau\lambda_{\hat{Q}})\delta^2 - a(1+\tau\lambda_{\hat{Q}})\delta + \tau\lambda_{\hat{Q}} > 0. \tag{28}$$

From (27), we can derive that

$$\delta < \begin{cases} \frac{1+\tau\lambda_{\hat{Q}}}{2a-\tau\lambda_{\hat{Q}}}, & \text{if} \quad 0 < \tau\lambda_{\hat{Q}} < 2a; \\ 1, & \text{otherwise.} \end{cases} \tag{29}$$

If $\tau = \frac{a}{\lambda_{\hat{Q}}}$, from (28), we directly get

$$\delta < \frac{\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}})} < \frac{1+\tau\lambda_{\hat{Q}}}{2a-\tau\lambda_{\hat{Q}}}. \tag{30}$$

If $\tau < \frac{a}{\lambda_{\hat{Q}}}$, solving the inequality (28), we obtain

$$\delta < \frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}} \tag{31}$$

or

$$\delta > \frac{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}}{2a(a-\tau\lambda_{\hat{Q}})}. \tag{32}$$

It can be shown that

$$\frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}} < \frac{1+\tau\lambda_{\hat{Q}}}{2a-\tau\lambda_{\hat{Q}}}. \tag{33}$$

This, along with (29)–(32), yields

$$\delta < \frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}}. \tag{34}$$

If $\tau > \frac{a}{\lambda_{\hat{Q}}}$, as $\delta \geq 0$, from (28), we have

$$0 \leq \delta < \frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}}. \tag{35}$$

Similarly, we can show that (33) also holds if $\frac{a}{\lambda_{\hat{Q}}} < \tau < \frac{2a}{\lambda_{\hat{Q}}}$. Noticing that $\delta < 1$ and

$$\frac{2\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}}) + \sqrt{a^2(1-\tau\lambda_{\hat{Q}})^2 + 4a\tau^2\lambda_{\hat{Q}}^2}} = \frac{\tau\lambda_{\hat{Q}}}{a(1+\tau\lambda_{\hat{Q}})}$$

holds if $\tau = \frac{a}{\lambda_{\hat{Q}}}$, by (30), (34), and (35), we can derive $0 \leq \delta < \min\{1, \nu_1\}$.

On the other hand, from (26), we can get

$$2a\delta + \sqrt{f(\Lambda_{\hat{Q}})^2 + 4\delta\tau\Lambda_{\hat{Q}}} - f(\Lambda_{\hat{Q}}) < 2.$$

This shows that

$$|f(\Lambda_{\hat{Q}})| - f(\Lambda_{\hat{Q}}) < 2, \qquad 2 + f(\Lambda_{\hat{Q}}) - 2a\delta > 0 \tag{36}$$

and

$$a(a+\tau\Lambda_{\hat{Q}})\delta^2 - (2\tau\Lambda_{\hat{Q}} - a\tau\Lambda_{\hat{Q}} + 3a)\delta + 2 - \tau\Lambda_{\hat{Q}} > 0. \tag{37}$$

By the definition of $f(\mu)$ in Lemma 2 and (36), we get

$$\tau < \frac{2}{(1+\delta)\Lambda_{\hat{Q}}}, \qquad \delta < \frac{3-\tau\Lambda_{\hat{Q}}}{2a+\tau\Lambda_{\hat{Q}}}. \tag{38}$$

Solving the inequality (37) for δ, we obtain

$$\delta < \frac{2\tau\Lambda_{\hat{Q}} - a\tau\Lambda_{\hat{Q}} + 3a - \sqrt{(2\tau\Lambda_{\hat{Q}} + a)^2 - a^2\tau(2-\tau\Lambda_{\hat{Q}})\Lambda_{\hat{Q}}}}{2a(a+\tau\Lambda_{\hat{Q}})} \tag{39}$$

or

$$\delta > \frac{2\tau\Lambda_{\widehat{Q}} - a\tau\Lambda_{\widehat{Q}} + 3a + \sqrt{(2\tau\Lambda_{\widehat{Q}} + a)^2 - a^2\tau(2 - \tau\Lambda_{\widehat{Q}})\Lambda_{\widehat{Q}}}}{2a(a + \tau\Lambda_{\widehat{Q}})}. \tag{40}$$

We can directly check that

$$\frac{2\tau\Lambda_{\widehat{Q}} - a\tau\Lambda_{\widehat{Q}} + 3a - \sqrt{(2\tau\Lambda_{\widehat{Q}} + a)^2 - a^2\tau(2 - \tau\Lambda_{\widehat{Q}})\Lambda_{\widehat{Q}}}}{2a(a + \tau\Lambda_{\widehat{Q}})} \leq \frac{3 - \tau\Lambda_{\widehat{Q}}}{2a + \tau\Lambda_{\widehat{Q}}}.$$

This, together with (38)–(40) and $0 \leq \delta < 1$, yields $0 \leq \delta < \nu_2 < 1$. This combined with $0 \leq \delta < \min\{1, \nu_1\}$ completes the proof. □

Remark 1. As $0 < \tau < \frac{2}{(1+\delta)\Lambda_{\widehat{Q}}} < \frac{2}{\Lambda_{\widehat{Q}}}$, it can directly be shown that

$$\nu_1 \geq \frac{2\tau\lambda_{\widehat{Q}}}{a(1 + \tau\lambda_{\widehat{Q}}) + a|1 - \tau\lambda_{\widehat{Q}}| + 2\sqrt{a}\tau\lambda_{\widehat{Q}}} = \begin{cases} \frac{\tau\lambda_{\widehat{Q}}}{a + \sqrt{a}\tau\lambda_{\widehat{Q}}}, & \text{if}\tau\lambda_{\widehat{Q}} < 1; \\ \frac{1}{a + \sqrt{a}}, & \text{if}\tau\lambda_{\widehat{Q}} \geq 1 \end{cases}$$

and

$$\nu_2 \geq \frac{2\tau\Lambda_{\widehat{Q}} - a\tau\Lambda_{\widehat{Q}} + 3a - \sqrt{(2\tau\Lambda_{\widehat{Q}} + a)^2}}{2a(a + \tau\Lambda_{\widehat{Q}})} = \frac{2 - \tau\Lambda_{\widehat{Q}}}{2(a + \tau\Lambda_{\widehat{Q}})}.$$

Therefore, if $\tau = \frac{1}{\Lambda_{\widehat{Q}}} < \frac{2}{(1+\delta)\Lambda_{\widehat{Q}}}$, as $a = (1 + \sqrt{5})/2$, we can see that

$$\nu_1 \geq \frac{1}{a\kappa + \sqrt{a}} \approx \frac{1}{1.618\kappa + 1.272}$$

and

$$\nu_2 \geq \frac{1}{2(a + 1)} \approx 0.1910,$$

where $\kappa = \Lambda_{\widehat{Q}}/\lambda_{\widehat{Q}}$ is the condition number of $\widehat{Q}$. Therefore, from Theorem 3, we know that Algorithm 2 is convergent if

$$\tau = \frac{1}{\Lambda_{\widehat{Q}}}, \qquad 0 \leq \delta \leq \min\left\{0.1910, \frac{1}{1.618\kappa + 1.272}\right\}.$$

This illustrates that the convergence conditions of Algorithm 2 are reasonable if the condition number of $\widehat{Q}$ is not too large.

4 | TWO SPECIAL CASES OF THE INEXACT UZAWA METHOD

The key step in Algorithm 2 is to choose an approach to produce the vector $\Phi_H(q_k)$. There have been many researches working on fast solvers for the system of linear equations with coefficient matrix H; see other works for example.[14,17,21,28,31,38,40,41] In what follows, we shall directly employ a splitting iteration method to solve the linear system $Hw = q - E^T z_k$ and get the next iteration w_{k+1}, which can be concluded in the following algorithm.

Algorithm 3. *Let $H = M - N$ be a splitting of H and $(x_0^T, y_0^T, z_0^T)^T \in R^{n+m+l}$ be an arbitrary initial vector. The iterates for solving the linear system (1) are computed according to the following procedure, where $w_k = (x_k^T, y_k^T)^T, k = 0, 1, 2, \ldots$.*

(1) Set $w_{(k,0)} = w_k$.

(2) For $j = 0, 1, \ldots, r - 1$, compute

$$w_{(k,j+1)} = M^{-1}Nw_{(k,j)} + M^{-1}(q - E^T z_k) \tag{41}$$

and set $w_{k+1} = w_{(k,r)}$.

(3) Update

$$z_{k+1} = z_k + \tau Q^{-1}(h - Ew_{k+1}),$$

where τ is a given positive constant and Q is a given symmetric positive definite matrix.

As Algorithm 3 is a special case of Algorithm 2, we will use Theorem 3 to derive the convergence conditions of Algorithm 3. Let $w^*_{k+1} = H^{-1}(q - E^T z_k)$. Then, we have

$$\begin{aligned} w_{k+1} = w_{(k,r)} &= M^{-1}Nw_{(k,r-1)} + M^{-1}(q - E^T z_k) \\ &= (M^{-1}N)^2 w_{(k,r-2)} + (I + M^{-1}N)M^{-1}(q - E^T z_k) \\ &= (M^{-1}N)^r w_k + \sum_{i=0}^{r-1}(M^{-1}N)^i M^{-1}(q - E^T z_k) \end{aligned}$$

and

$$w^*_{k+1} = (M^{-1}N)^r w^*_{k+1} + \sum_{i=0}^{r-1}(M^{-1}N)^i M^{-1}(q - E^T z_k).$$

This implies that

$$\begin{aligned} \Phi_H(q_k) - H^{-1}q_k &= w_{k+1} - w_k - \left(w^*_{k+1} - w_k\right) = w_{k+1} - w^*_{k+1} \\ &= (M^{-1}N)^r \left(w_k - w^*_{k+1}\right) = (M^{-1}N)^r \left[w_k - H^{-1}(q - E^T z_k)\right] \\ &= -(M^{-1}N)^r H^{-1} q_k, \end{aligned}$$

which follows

$$\|\Phi_H(q_k) - H^{-1}q_k\|_K \le \|(M^{-1}N)^r H^{-1}q_k\|_K \le \|K^{\frac{1}{2}}(M^{-1}N)^r H^{-1}K^{\frac{1}{2}}\| \|q_k\|_{K^{-1}}.$$

Setting $\delta = \|K^{\frac{1}{2}}(M^{-1}N)^r H^{-1}K^{\frac{1}{2}}\|$ in (13), we can see that

$$\begin{aligned} \delta &= \left\|K^{\frac{1}{2}}(M^{-1}N)^r K^{-\frac{1}{2}}K^{\frac{1}{2}}H^{-1}K^{\frac{1}{2}}\right\| \le \left\|K^{\frac{1}{2}}(M^{-1}N)^r K^{-\frac{1}{2}}\right\| \left\|K^{\frac{1}{2}}H^{-1}K^{\frac{1}{2}}\right\| \\ &= \rho\left(K^{\frac{1}{2}}H^{-1}K^{\frac{1}{2}}\right)\|(M^{-1}N)^r\|_K = \rho\left(Q_1^{-1}\right)\|(M^{-1}N)^r\|_K \\ &= \left(\frac{1+\sqrt{5}}{2}\right)\|(M^{-1}N)^r\|_K. \end{aligned}$$

This, combined with Theorem 3, yields the following results immediately.

Theorem 4. *Let $H = M - N$ be a splitting of H. Let K be defined in (5). If the parameter τ and the number r of the inner iteration satisfy*

$$0 < \tau \le \frac{1}{\Lambda_{\hat{Q}}}, \qquad 0 \le \|(M^{-1}N)^r\|_K < \frac{2}{1+\sqrt{5}}\min\{v_1, v_2\}, \tag{42}$$

then the sequence $\{x_k, y_k, z_k\}$ produced by Algorithm 3 converges to the unique solution of the saddle-point problem (1).

Remark 2. It is easy to see that $(M^{-1}N)^r \to 0 (r \to \infty)$ if $\rho(M^{-1}N) < 1$. Thus, there exists a constant r such that the condition (42) holds. Noting that

$$\begin{aligned} \delta &= \left\|\left(K^{\frac{1}{2}}M^{-1}NK^{-\frac{1}{2}}\right)^r K^{\frac{1}{2}}H^{-1}K^{\frac{1}{2}}\right\| \le \left\|K^{\frac{1}{2}}M^{-1}NK^{-\frac{1}{2}}\right\|^r \left\|K^{\frac{1}{2}}H^{-1}K^{\frac{1}{2}}\right\| \\ &= \frac{1+\sqrt{5}}{2}\|M^{-1}N\|_K^r, \end{aligned}$$

from Theorem 4, we know that if $\|M^{-1}N\|_K < 1$, (42) can further be restricted as

$$\|M^{-1}N\|_K^r \le \frac{2}{1+\sqrt{5}}\min\{v_1, v_2\}.$$

This shows that the number r of the inner iteration should be larger than $\frac{\log(0.6\min\{\nu_1,\nu_2\})}{\log\|M^{-1}N\|_K}$.

In addition to the splitting iteration scheme (41) in Algorithm 3, we can also find a matrix $\widehat{H}$ to approximate H, which is available to reach $\widehat{H}^{-1}q_k \approx H^{-1}q_k$. Here, we take $\widehat{H}$ to be a block lower triangular matrix

$$\widehat{H} = \begin{pmatrix} P_A & 0 \\ B & -\frac{1}{\omega}P_S \end{pmatrix}, \tag{43}$$

where $\omega > 0$ is a constant, and $P_A \in R^{n\times n}$ and $P_S \in R^{m\times m}$ are two symmetric positive definite matrices. Then, we can replace the exact form $H^{-1}q_k$ by this inexact $\widehat{H}^{-1}q_k$. This shows that (14) in Algorithm 2 can be embodied as

$$w_{k+1} = w_k + \Phi_H(q_k) = w_k + \widehat{H}^{-1}q_k.$$

Substituting w_k, q_k, and $\widehat{H}$ into this formula, we have

$$\begin{cases} x_{k+1} = x_k + P_A^{-1}(f - Ax_k - B^T y_k), \\ y_{k+1} = y_k + \omega P_S^{-1}(Bx_{k+1} + C^T z_k - g). \end{cases}$$

This, together with Algorithm 2, yields another inexact Uzawa method.

Algorithm 4. *Let $(x_0^T, y_0^T, z_0^T)^T \in R^{n+m+l}$ be an arbitrary initial vector. The iterates for solving the linear system (1) are computed according to the following procedure:*

$$\begin{cases} x_{k+1} = x_k + P_A^{-1}(f - Ax_k - B^T y_k), \\ y_{k+1} = y_k + \omega P_S^{-1}(Bx_{k+1} + C^T z_k - g), \\ z_{k+1} = z_k + \tau Q^{-1}(h - Cy_{k+1}), \end{cases} \tag{44}$$

where $P_A \in R^{n\times n}$, $P_S \in R^{m\times m}$ and $Q \in R^{l\times l}$ are given symmetric positive definite matrices, and ω and τ are two given positive constants.

Remark 3. It is not difficult to verify that the iteration scheme (44) can also be deduced by the splitting

$$\mathcal{A} = \begin{pmatrix} P_A & 0 & 0 \\ B & -\frac{1}{\omega}P_S & 0 \\ 0 & C & \frac{1}{\tau}Q \end{pmatrix} - \begin{pmatrix} P_A - A & -B^T & 0 \\ 0 & -\frac{1}{\omega}P_S & -C^T \\ 0 & 0 & \frac{1}{\tau}Q \end{pmatrix}. \tag{45}$$

Therefore, Algorithm 4 is a splitting method.

We are now ready to study the convergence property of Algorithm 4. It follows from Remark 3 that the convergence of Algorithm 4 can be establish by analyzing the spectral radius of the iterative matrix. Using (45), we can derive that

$$\begin{aligned} \mathcal{R} &= \begin{pmatrix} P_A & 0 & 0 \\ B & -\frac{1}{\omega}P_S & 0 \\ 0 & C & \frac{1}{\tau}Q \end{pmatrix}^{-1} \begin{pmatrix} P_A - A & -B^T & 0 \\ 0 & -\frac{1}{\omega}P_S & -C^T \\ 0 & 0 & \frac{1}{\tau}Q \end{pmatrix} \\ &= \begin{pmatrix} P_A^{-1} & 0 & 0 \\ \omega P_S^{-1}BP_A^{-1} & -\omega P_S^{-1} & 0 \\ -\omega\tau Q^{-1}CP_S^{-1}BP_A^{-1} & \omega\tau Q^{-1}CP_S^{-1} & \tau Q^{-1} \end{pmatrix} \begin{pmatrix} P_A - A & -B^T & 0 \\ 0 & -\frac{1}{\omega}P_S & -C^T \\ 0 & 0 & \frac{1}{\tau}Q \end{pmatrix} \\ &= \begin{pmatrix} \widehat{A} & -P_A^{-1}B^T & 0 \\ \omega P_S^{-1}B\widehat{A} & I - \omega P_S^{-1}BP_A^{-1}B^T & \omega P_S^{-1}C^T \\ -\omega\tau Q^{-1}CP_S^{-1}B\widehat{A} & \omega\tau Q^{-1}CP_S^{-1}BP_A^{-1}B^T - \tau Q^{-1}C & I - \omega\tau Q^{-1}CP_S^{-1}C^T \end{pmatrix}, \end{aligned}$$

where $\widehat{A} = I - P_A^{-1}A$. In what follows, we just study the specific case $P_A = A$. In this case, the iterative matrix $\mathcal{R}$ can be simplified to

$$\begin{aligned}
\mathcal{R} &= \begin{pmatrix} 0 & -A^{-1}B^T & 0 \\ 0 & I - \omega P_S^{-1}S & \omega P_S^{-1}C^T \\ 0 & \omega\tau Q^{-1}CP_S^{-1}S - \tau Q^{-1}C & I - \omega\tau Q^{-1}CP_S^{-1}C^T \end{pmatrix} \\
&\sim \begin{pmatrix} 0 & -A^{-1}B^T & 0 \\ 0 & I - \omega P_S^{-\frac{1}{2}}SP_S^{-\frac{1}{2}} & \omega P_S^{-\frac{1}{2}}C^TQ^{-\frac{1}{2}} \\ 0 & \omega\tau Q^{-\frac{1}{2}}CP_S^{-1}SP_S^{-\frac{1}{2}} - \tau Q^{-\frac{1}{2}}CP_S^{-\frac{1}{2}} & I - \omega\tau Q^{-\frac{1}{2}}CP_S^{-1}C^TQ^{-\frac{1}{2}} \end{pmatrix} \\
&= \begin{pmatrix} 0 & -A^{-1}B^T & 0 \\ 0 & I - \omega\widehat{S} & \omega\widehat{C}^T \\ 0 & \omega\tau\widehat{C}\widehat{S} - \tau\widehat{C} & I - \omega\tau\widehat{C}\widehat{C}^T \end{pmatrix},
\end{aligned}$$

where $\widehat{S} = P_S^{-\frac{1}{2}}SP_S^{-\frac{1}{2}}$ and $\widehat{C} = Q^{-\frac{1}{2}}CP_S^{-\frac{1}{2}}$. This shows that $\rho(\mathcal{R}) = \rho(\widehat{\mathcal{R}})$, where

$$\widehat{\mathcal{R}} = \begin{pmatrix} I - \omega\widehat{S} & \omega\widehat{C}^T \\ \omega\tau\widehat{C}\widehat{S} - \tau\widehat{C} & I - \omega\tau\widehat{C}\widehat{C}^T \end{pmatrix}. \tag{46}$$

Thus, it is concluded that Algorithm 4 with $P_A = A$ is convergent if and only if $\rho(\widehat{\mathcal{R}}) < 1$. To derive some sufficient conditions for $\rho(\widehat{\mathcal{R}}) < 1$, we introduce an existing important result.

Lemma 3 (See the work of Young[39]). *Both roots of the real quadratic equation $\lambda^2 - b\lambda + c = 0$ are less than one in modulus if and only if $|c| < 1$ and $|b| < 1 + c$.*

Theorem 5. *Suppose that $A \in R^{n\times n}$ is symmetric positive definite, and $B \in R^{m\times n}$ and $C \in R^{l\times m}$ have full row rank. If the parameters ω and τ satisfy*

$$0 < \omega < \frac{2}{\Lambda_{\widehat{S}}}, \qquad 0 < \tau < \frac{2(2 - \omega\Lambda_{\widehat{S}})}{\omega\Lambda_{\widehat{C}\widehat{C}^T}},$$

then the sequence $\{x_k, y_k, z_k\}$ produced by Algorithm 4 converges to the unique solution of the saddle-point problem (1).

Proof. Let λ and $(x^T, y^T)^T$ be the eigenvalue and eigenvector of the matrix $\widehat{\mathcal{R}}$, respectively. Then, by (46), we can get

$$\begin{cases} x - \omega\widehat{S}x + \omega\widehat{C}^Ty = \lambda x, \\ \omega\tau\widehat{C}\widehat{S}x - \tau\widehat{C}x + y - \omega\tau\widehat{C}\widehat{C}^Ty = \lambda y. \end{cases} \tag{47}$$

By the first equality in (47), it follows

$$\omega\tau\widehat{C}\widehat{S}x - \tau\widehat{C}x + y - \omega\tau\widehat{C}\widehat{C}^Ty = \tau\widehat{C}(\omega\widehat{S}x - x - \omega\widehat{C}^Ty) + y = -\lambda\tau\widehat{C}x + y.$$

Combined with the second equality in (47), we have

$$-\lambda\tau\widehat{C}x = (\lambda - 1)y. \tag{48}$$

If $\lambda = 1$, the first equality in (47) can be simplified to

$$\widehat{S}x - \widehat{C}^Ty = 0. \tag{49}$$

Because $\widehat{S}$ is symmetric positive definite, from (49), we have $x = \widehat{S}^{-1}\widehat{C}^Ty$. Substituting into (48) leads to $\widehat{C}\widehat{S}^{-1}\widehat{C}^Ty = 0$. Noticing that $\widehat{C}$ is row full rank, the matrix $\widehat{C}\widehat{S}^{-1}\widehat{C}^T$ is symmetric positive definite as well. This shows that $y = 0$.

Combining with (49) yields $x = 0$, which is contrary to the fact that $(x^T, y^T)^T$ is the eigenvector. Thereby, $\lambda \neq 1$ and (48) can equivalently be reformulated as

$$y = -\frac{\lambda\tau}{\lambda - 1}\widehat{C}x. \tag{50}$$

Substituting (50) into the first equality of (47), we obtain

$$x - \omega\widehat{S}x - \frac{\lambda\omega\tau}{\lambda - 1}\widehat{C}^T\widehat{C}x = \lambda x. \tag{51}$$

We assert that $x \neq 0$. Otherwise, it follows from (47) that $\widehat{C}^T y = 0$ and $(\lambda - 1)y + \omega\tau\widehat{C}\widehat{C}^T y = 0$, which leads to $y = 0$ immediately. This contradicts with $(x^T, y^T)^T \neq 0$. Then, multiplying both sides of (51) from the left with $x^*/(x^*x)$ yields

$$\lambda^2 + \lambda\left(\omega\theta(x) + \omega\tau\eta(x) - 2\right) + 1 - \omega\theta(x) = 0, \tag{52}$$

where $\theta(x)$ and $\eta(x)$ are defined by

$$\theta(x) = \frac{x^*\widehat{S}x}{x^*x}, \qquad \eta(x) = \frac{x^*\widehat{C}^T\widehat{C}x}{x^*x}.$$

If $x \in \text{null}(\widehat{C})$, from (51), we have $x - \omega\widehat{S}x = \lambda x$, which can be equivalent to $\lambda = 1 - \omega\theta(x)$. Then, $|\lambda| < 1$ holds if and only if $|1 - \omega\theta(x)| < 1$ holds for any $x \in \text{null}(\widehat{C})$.

If $x \notin \text{null}(\widehat{C})$, from Lemma 3 we know that $|\lambda| < 1$ holds if

$$|1 - \omega\theta(x)| < 1$$

and

$$|\omega\theta(x) + \omega\tau\eta(x) - 2| < 2 - \omega\theta(x). \tag{53}$$

Therefore, for any $x \neq 0$, it holds that $|1 - \omega\theta(x)| < 1$. By some simple calculations, we have $0 < \omega\theta(x) < 2$. As the matrix $\widehat{S}$ is symmetric positive definite and $x \neq 0$, we get $\theta(x) \neq 0$. This shows that

$$0 < \omega < \min_{x \neq 0} \frac{2}{\theta(x)} = \frac{2}{\Lambda_{\widehat{S}}}.$$

On the other hand, as $x \notin \text{null}(\widehat{C})$, we have $\eta(x) \neq 0$. Then, it follows from (53) that

$$0 < \tau < \frac{2(2 - \omega\theta(x))}{\omega\eta(x)}.$$

We can check that

$$\min_{x \notin \text{null}(\widehat{C})} \frac{2(2 - \omega\theta(x))}{\omega\eta(x)} \geq \frac{2(2 - \omega\lambda_{\max}(\widehat{S}))}{\omega\lambda_{\max}(\widehat{C}^T\widehat{C})} = \frac{2(2 - \omega\Lambda_{\widehat{S}})}{\omega\Lambda_{\widehat{C}\widehat{C}^T}}.$$

To ensure the convergence, τ should satisfy

$$0 < \tau < \frac{2(2 - \omega\Lambda_{\widehat{S}})}{\omega\Lambda_{\widehat{C}\widehat{C}^T}},$$

which follows the results. □

Using (52), it can be shown that the nonzero eigenvalues of iterative matrix $\mathcal{R}$ are

$$\lambda = \frac{-[\omega\theta(x) + \omega\tau\eta(x) - 2] \pm \sqrt{[\omega\theta(x) + \omega\tau\eta(x) - 2]^2 - 4[1 - \omega\theta(x)]}}{2}. \tag{54}$$

However, it is difficult to get the optimal parameters ω_{opt} and τ_{opt} by the above expressions as that in Theorem 2. In the following, we will give a reasonable way to choose the parameters in Algorithm 4 for the case $P_A = A$.

If $[\omega\theta(x)+\omega\tau\eta(x)-2]^2 < 4(1-\omega\theta(x))$, from (54) it is not difficult to verify that $|\lambda| = \sqrt{1-\omega\theta(x)}$. If $[\omega\theta(x)+\omega\tau\eta(x)-2]^2 \geq 4(1-\omega\theta(x))$, we have

$$|\lambda| = \frac{|\omega\theta(x)+\omega\tau\eta(x)-2| + \sqrt{[\omega\theta(x)+\omega\tau\eta(x)-2]^2 - 4[1-\omega\theta(x)]}}{2}.$$

Therefore, for any $\lambda \in \mathrm{sp}(\widehat{\mathcal{R}})$, it holds that

$$\begin{aligned}|\lambda| &\leq \frac{|\omega\theta(x)+\omega\tau\eta(x)-2| + \sqrt{[\omega\theta(x)+\omega\tau\eta(x)-2]^2 + 4|1-\omega\theta(x)|}}{2}\\ &\leq \frac{1}{2}\max\left\{ |\omega\Lambda_{\hat{S}} + \omega\tau\Lambda_{\hat{C}\hat{C}^T} - 2| + \sqrt{[\omega\Lambda_{\hat{S}} + \omega\tau\Lambda_{\hat{C}\hat{C}^T} - 2]^2 + 4|1-\omega\Lambda_{\hat{S}}|}\right.,\\ &\qquad \left. |\omega\lambda_{\hat{S}} + \omega\tau\lambda_{\hat{C}\hat{C}^T} - 2| + \sqrt{[\omega\lambda_{\hat{S}} + \omega\tau\lambda_{\hat{C}\hat{C}^T} - 2]^2 + 4|1-\omega\lambda_{\hat{S}}|}\right\}.\end{aligned}$$

Therefore, τ can be chosen such that

$$\max\{|\omega\lambda_{\hat{S}} + \omega\tau\lambda_{\hat{C}\hat{C}^T} - 2|, |\omega\Lambda_{\hat{S}} + \omega\tau\Lambda_{\hat{C}\hat{C}^T} - 2|\}$$

minimize. This implies that τ should satisfy

$$|\omega\Lambda_{\hat{S}} + \omega\tau\Lambda_{\hat{C}\hat{C}^T} - 2| = |\omega\lambda_{\hat{S}} + \omega\tau\lambda_{\hat{C}\hat{C}^T} - 2|,$$

which leads to

$$\tau = \frac{4-\omega(\lambda_{\hat{S}} + \Lambda_{\hat{S}})}{\omega(\lambda_{\hat{C}\hat{C}^T} + \Lambda_{\hat{C}\hat{C}^T})}. \tag{55}$$

In addition, we can take ω to satisfy $|1-\omega\Lambda_{\hat{S}}| = |1-\omega\lambda_{\hat{S}}|$, which minimizes the function $\max\{|1-\omega\lambda_{\hat{S}}|, |1-\omega\Lambda_{\hat{S}}|\}$. By some simple calculations, we have

$$\omega = \frac{2}{\lambda_{\hat{S}} + \Lambda_{\hat{S}}}.$$

This, together with (55), yields that

$$\tau = \frac{2}{\omega(\lambda_{\hat{C}\hat{C}^T} + \Lambda_{\hat{C}\hat{C}^T})} = \frac{\lambda_{\hat{S}} + \Lambda_{\hat{S}}}{\lambda_{\hat{C}\hat{C}^T} + \Lambda_{\hat{C}\hat{C}^T}}. \tag{56}$$

After some algebraic manipulation, it follows

$$|1-\omega\Lambda_{\hat{S}}| = \rho_1, \qquad |\omega\Lambda_{\hat{S}} + \omega\tau\Lambda_{\hat{C}\hat{C}^T} - 2| = \rho_1 + \rho_2,$$

where

$$\rho_1 = \frac{\Lambda_{\hat{S}} - \lambda_{\hat{S}}}{\Lambda_{\hat{S}} + \lambda_{\hat{S}}}, \qquad \rho_2 = \frac{\Lambda_{\hat{C}\hat{C}^T} - \lambda_{\hat{C}\hat{C}^T}}{\Lambda_{\hat{C}\hat{C}^T} + \lambda_{\hat{C}\hat{C}^T}}.$$

Thus, if $2\rho_1 + \rho_2 < 1$, the spectral radius of the iteration matrix $\mathcal{R}$ would be

$$\rho(\mathcal{R}) \leq \frac{\rho_1 + \rho_2 + \sqrt{(\rho_1+\rho_2)^2 + 4\rho_1}}{2} < 1.$$

5 NUMERICAL EXPERIMENTS

In this section, we present several numerical examples to test the performance of Algorithm 1 (denoted by "DUM"), Algorithm 2 (denoted by "IUM(δ)"), Algorithm 3 (denoted by "SIUM"), and Algorithm 4 (denoted by "SUM") for solving the three-by-three saddle-point problem (1). All experiments were performed on a Linux machine DELL Precision T7610 with 96 GB of RAM and 24-core processor Intel(R) Xeon(R) (2.60 GHz).

We also use the MINRES method (denoted by "MINRES") to solve the saddle-point problem (1). We compare the performance of these methods by reporting the number of iterations, the CPU time, and the relative residual, which are denoted by "Iter," "CPU," and "Res," respectively. Let $\ell_k = (x_k^T, y_k^T, z_k^T)^T$ be the k-th approximate solution. Then, the "Res" is defined by the relative error

$$\text{Res} := \frac{\|(f^T, g^T, h^T)^T - \mathcal{A}\ell_k\|_2}{\|(f^T, g^T, h^T)\|_2}.$$

In our implementation, we stop all considered algorithms when Res $\leq 10^{-6}$. All the initial guesses ℓ_0 are taken by zero vectors. We choose the right-hand-side vector of the saddle-point problem (1) so that its exact solution is given by the vector $(1, 1, \ldots, 1)^T \in R^{n+m+l}$. For the IUM($\delta$) method, we use MINRES method to derive $\Phi_H(q_k)$ such that the inequality

$$\frac{\|q_k - H\Phi_H(q_k)\|}{\|q_k\|} \leq \delta$$

holds for $\delta = 0.3, 0.5$. For the DUM method, we use the optimal parameter τ_{opt} derived in Theorem 2. For the SUM method, we take $P_A = A$, $\omega = 0.5$ and compute the parameter τ by the formula (56):

$$\tau = \frac{2}{\omega(\lambda_{\hat{C}\hat{C}^T} + \Lambda_{\hat{C}\hat{C}^T})}.$$

Example 2. Consider the saddle-point problem (1). The matrices A and B arise from the two dimensional "leaky" lid-driven cavity problem in a square domain $\Omega = (0 \leq x \leq 1, 0 \leq y \leq 1)$, that is, the following Stokes equations:

$$-\Delta\mathbf{u} + \nabla p = 0, \quad \text{in } \Omega,$$
$$\nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega. \tag{57}$$

A Dirichlet no-flow condition is applied on the side and bottom boundaries, and the nonzero horizontal velocity on the lid is $\{y = 1; -1 \leq x \leq 1 \,|\, \boldsymbol{u}_x = 1\}$. Here, $\boldsymbol{u}$ and p represent the velocity vector field and the pressure scalar field, Δ is the vector Laplacian in $\mathbb{R}^2$, ∇ denotes the gradient, and $\nabla\cdot$ is the divergence. To make the linear system (1) ill-conditioned and not too sparse, we consider the matrix C with the form of

$$C = (\text{diag}\{1, 3, 5, \ldots, 2l-1\} \quad \text{randn}(l, m-l)),$$

where randn$(l, m-l)$ is an l-by-$(m-l)$ matrix of normally distributed random numbers.

To derive the matrices A and B, we discretize the Stokes Equations (57) by the Q2-P1 finite elements on the uniform grids with the grid parameters $h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}$. Then, we have $n = 2\left(1 + \frac{1}{h}\right)^2$ and $m = \frac{3}{4h^2}$. This discrete process can be accomplished by the IFISS software package developed by Elman et al.[7] In Example 2, we take $l = \frac{1}{4h^2}$, $P_S = \text{diag}(\bar{A}^{-1}B^T)$ and $Q = \text{diag}(CP_S^{-1}C^T)$, where $\bar{A} = \text{diag}(A)$ and diag(A) is a square diagonal matrix with the elements of the main diagonal elements of A. We set $\tau = 1$ for the IUM(δ) method and the SIUM method. For the SIUM method, we set $r = 3$ and

$$M = \begin{pmatrix} LL^T & 0 \\ B & P_S/2 \end{pmatrix}, \qquad N = \begin{pmatrix} LL^T - A & -B^T \\ 0 & P_S/2 \end{pmatrix},$$

where L produced by the incomplete Cholesky decomposition of A.

The resulting numerical results are listed in Tables 1 and 2 and Figure 1. To see the role of the parameters in the convergence behaviors of the test methods, we draw the characteristic curves of the number of iterations versus the parameters in Figure 2 and the region of the parameters that the SUM method satisfies Res $\leq 10^{-6}$ in 2,000 steps in Figure 3 for $h = \frac{1}{16}$.

In Examples 3 and 4, we take $Q = C\mathrm{diag}(P_S)^{-1}C^T$ for all tested methods, where $P_S = BL^{-T}L^{-1}B^T$ and L is produced by the incomplete Cholesky decomposition of A with the droptol being 0.01.

TABLE 1 The parameter $\tau(\tau_{\mathrm{opt}})$ of DUM and SUM for Example 2

	h	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	$\frac{1}{256}$
	DOF	834	3,202	12,546	49,666	197,634
DUM	τ_{opt}	0.4598	0.3462	0.2271	0.1340	-
	CPU	0.0049	0.0109	0.0997	2.1411	-
SUM	τ	1.6174	1.6290	1.6625	1.7429	1.8006
	CPU	0.0048	0.0069	0.1503	2.0295	57.2394

TABLE 2 Numerical results for Example 2

DOF	DUM			SUM		
	Iter	CPU	Res	Iter	CPU	Res
834	68	0.1065	7.8952e−07	103	0.1297	9.9156e−07
3,202	107	1.6510	8.5264e−07	129	0.9529	5.9680e−07
12,546	189	107.5597	9.3794e−07	156	8.1733	9.4665e−07
49,666	366	8,191.6794	9.6989e−07	188	56.8353	9.3230e−07
197,634	-	-	-	219	628.3964	9.4622e−07
	SIUM			MINRES		
834	50	0.0122	5.1118e−07	1,415	0.0676	9.9843e−07
3,202	85	0.0730	8.5038e−07	7,123	2.4616	9.9348e−07
12,546	207	2.5183	7.7242e−07	16,737	105.7489	9.9989e−07
49,666	576	78.3841	8.9125e−07	13,498	904.1798	9.9994e−07
197,634	3,203	5,571.1346	9.2067e−07	10,554	26,207.6693	9.9986e−07
	IUM(0.3)			IUM(0.5)		
834	38	0.0540	8.3234e−07	60	0.0579	8.9228e−07
3,202	42	0.5191	7.5624e−07	68	0.4334	8.8801e−07
12,546	48	3.8121	9.2203e−07	80	4.8170	8.6244e−07
49,666	103	54.2726	7.2330e−07	105	55.0844	8.1486e−07
197,634	457	2,302.0706	9.7774e−07	445	2,091.2695	9.5075e−07

Note. Iter =iterations; Res =relative residual; MINRES =minimal residual.

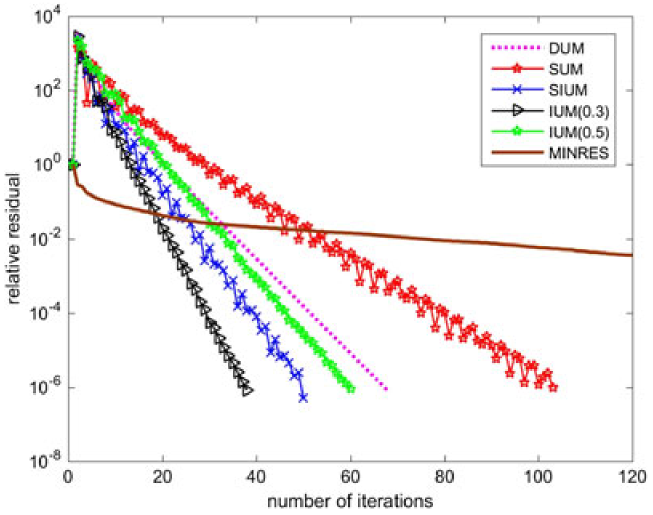

FIGURE 1 The iteration curves of the tested methods for Example 2 with DOF= 834. MINRES =minimal residual

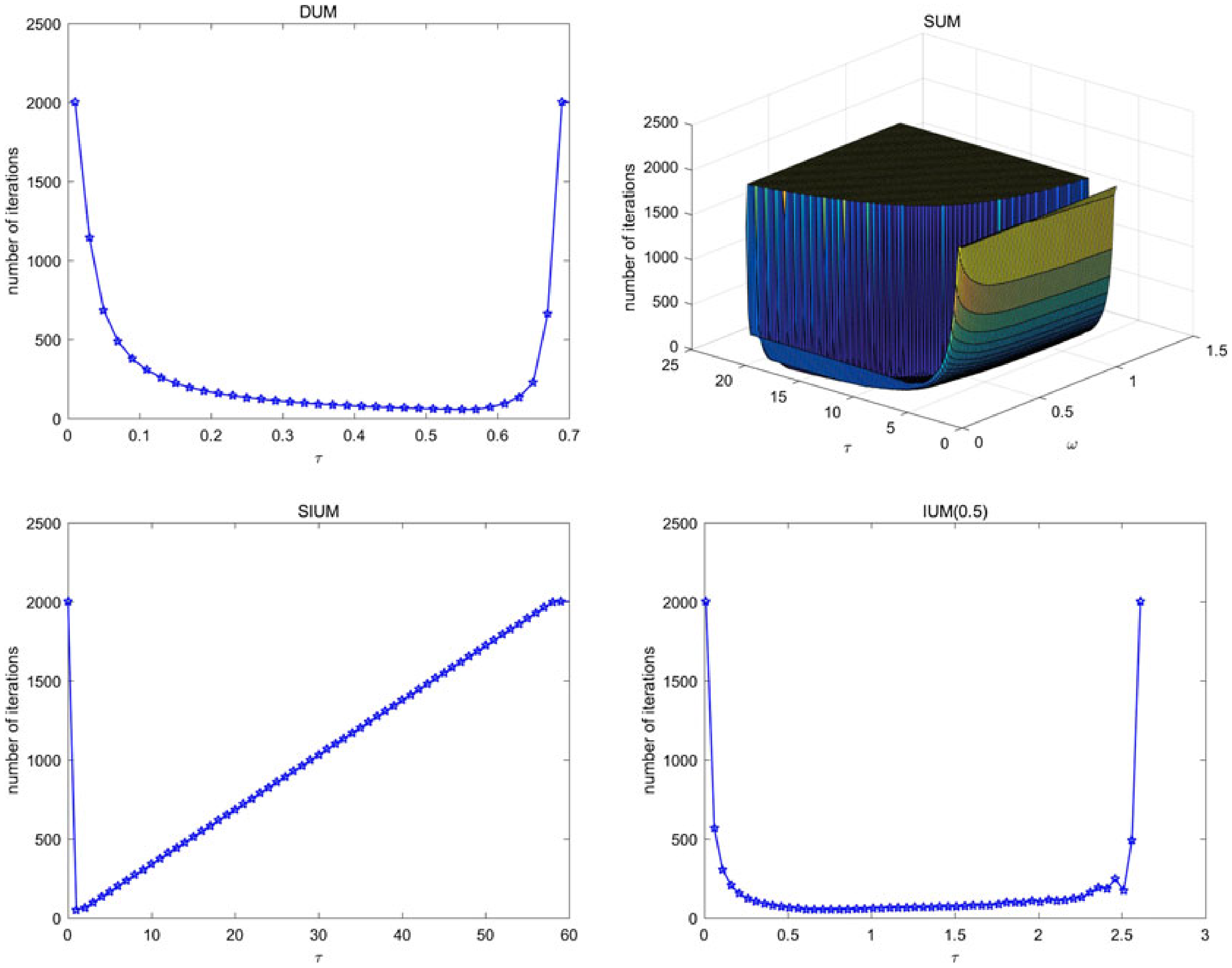

FIGURE 2 The characteristic curves of the number of iterations versus the parameter τ for Example 2 with DOF=834

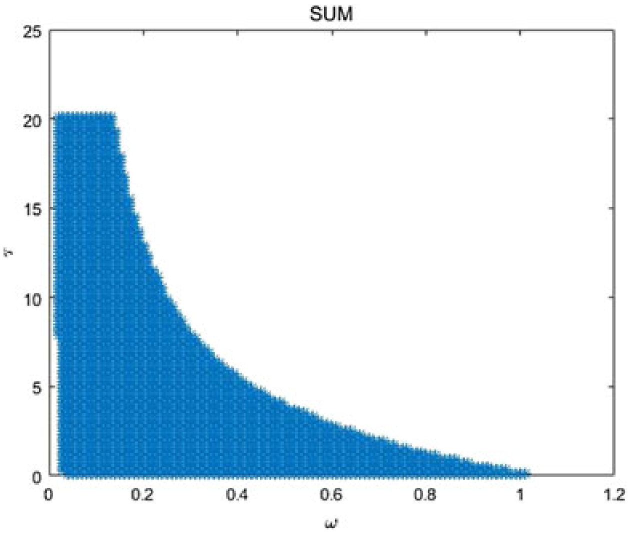

FIGURE 3 The region of the parameters that the SUM method satisfies Res $\leq 10^{-6}$ in 2,000 steps for Example 2 with DOF=834

Example 3. Consider the saddle-point problem (1), where

$$A = \mathrm{diag}(2W^TW + D_1, D_2, D_3) \in R^{n\times n}$$

is a block-diagonal matrix,

$$B = [E, -I_{2\tilde{p}}, I_{2\tilde{p}}] \in R^{m\times n} \quad \text{and} \quad C = F \in R^{l\times m}$$

are both full row-rank matrices, where $\tilde{p} = p^2, \hat{p} = p(p+1)$. Furthermore, $W = (w_{ij}) \in R^{\hat{p}\times\hat{p}}$ with $w_{ij} = e^{-2((i/3)^2+(j/3)^2)}$; $D_1 = I_{\hat{p}}$ is an identity matrix; $D_i = \mathrm{diag}(d_j^{(i)}) \in R^{2\tilde{p}\times 2\tilde{p}}$, $i = 2, 3$, are diagonal matrices, with

$$d_j^{(2)} = \begin{cases} 1, & \text{for} \quad 1 \le j \le \tilde{p}; \\ 10^{-5}(j-\tilde{p})^2, & \text{for} \quad \tilde{p}+1 \le j \le 2\tilde{p}, \end{cases}$$

$$d_j^{(3)} = 10^{-5}(j+\tilde{p})^2, \quad \text{for} \quad 1 \le j \le 2\tilde{p},$$

and

$$E = \begin{pmatrix} \hat{E} \otimes I_p \\ I_p \otimes \hat{E} \end{pmatrix}, \qquad \hat{E} = \begin{pmatrix} 2 & -1 & & & \\ & 2 & -1 & & \\ & & \ddots & \ddots & \\ & & & 2 & -1 \end{pmatrix} \in R^{p\times(p+1)},$$

and

$$F = \left(\hat{F} \otimes I_p \;\; I_p \otimes \hat{F} \right), \hat{F} = \begin{pmatrix} p+1 & -\frac{p-1}{p} & \frac{p-2}{p} & \dots & \frac{(-1)^{p-1}}{p} \\ -\frac{p-1}{p} & 2p+1 & \ddots & \ddots & \vdots \\ \frac{p-2}{p} & \ddots & \ddots & \ddots & \frac{p-2}{p} \\ \vdots & \ddots & \ddots & \ddots & -\frac{p-1}{p} \\ \frac{(-1)^{p-1}}{p} & \dots & \frac{p-2}{p} & -\frac{p-1}{p} & p^2+1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \in R^{(p+1)\times p}.$$

In Example 3, the matrices A and B arise in computing the descent directions in the Newton steps involved in the modified primal-dual interior point method used to solve the nonsmooth and nonconvex minimization problems from restorations of piecewise constant images.[43] In the computations, we take $\tau = 0.5$ for the IUM(δ) method and the SIUM method. For the SIUM method, we set $r = 3$ and

$$M = \begin{pmatrix} A_L & 0 \\ B & \mathrm{diag}(P_S) \end{pmatrix}, \qquad N = \begin{pmatrix} A_L - A & -B^T \\ 0 & \mathrm{diag}(P_S) \end{pmatrix},$$

where A_L is the lower triangular part of A. We test this problem by choosing different values of p. The resulting numerical results are listed in Tables 3 and 4 and Figure 4. To see the role of the parameters in the convergence behaviors of the test methods, we draw the characteristic curves of the number of iterations versus the parameters in Figure 5 and the region of the parameters that the SUM method satisfies Res $\le 10^{-6}$ in 2,000 steps in Figure 6 for $p = 16$.

TABLE 3 The parameter $\tau(\tau_{\mathrm{opt}})$ of DUM and SUM for Example 3

	p	16	32	64	96	160	192
	DOF	2,080	8,256	32,896	73,920	205,120	295,296
DUM	τ_{opt}	0.3445	0.1416	0.0496	0.0331	0.0254	-
	CPU	0.0764	0.7257	12.4311	68.9215	761.6523	-
SUM	τ	0.7966	0.3815	0.2042	0.1623	0.1401	0.1360
	CPU	0.0695	0.4141	7.5794	44.4337	542.8541	1,539.3429

TABLE 4 Numerical results for Example 3

DOF	DUM			SUM		
	Iter	CPU	Res	Iter	CPU	Res
2,080	79	0.3353	9.1290e−07	53	0.0407	8.0054e−07
8,256	250	2.9031	9.5301e−07	107	0.2092	9.2266e−07
32,896	820	44.7342	9.9537e−07	176	1.2331	9.4136e−07
73,920	1,302	170.4552	9.9227e−07	237	5.1914	9.8501e−07
205,120	1,768	792.1864	9.9322e−07	291	28.0130	9.9133e−07
295,296	-	-	-	303	47.2449	9.5101e−07
	SIUM			MINRES		
2,080	52	0.0281	9.5861e−07	4,131	0.8477	9.9935e−07
8,256	113	0.1543	9.1551e−07	15,852	11.1118	9.9996e−07
32,896	211	1.3817	9.6552e−07	23,600	82.4627	9.9998e−07
73,920	247	5.0324	9.8457e−07	26,040	288.2650	9.9990e−07
205,120	275	23.9967	9.9842e−07	32,237	1,841.8354	9.9976e−07
295,296	283	44.8465	9.3441e−07	39,419	3,867.4327	9.9992e−07
	IUM(0.3)			IUM(0.5)		
2,080	67	0.0836	6.7912e−07	85	0.0754	8.5905e−07
8,256	96	0.6466	8.3181e−07	107	0.4636	9.3049e−07
32,896	120	3.4918	9.0591e−07	190	3.7424	7.9134e−07
73,920	438	27.9282	9.5505e−07	278	13.8619	9.9461e−07
205,120	572	120.1403	9.1910e−07	446	81.4302	9.3584e−07
295,296	-	-	-	636	173.5964	7.6594e−07

Note. Iter =iterations; Res =relative residual; MINRES =minimal residual.

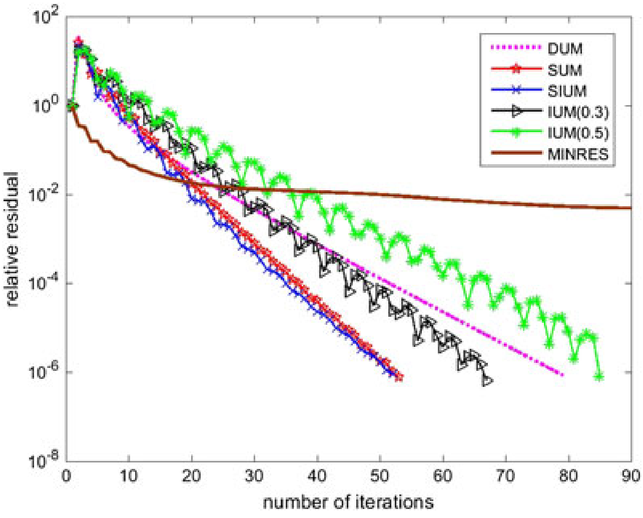

FIGURE 4 The iteration curves of the tested methods for Example 3 with DOF=2,080. MINRES = minimal residual

Example 4. Consider the saddle-point problem (1), where

$$A = \frac{1}{10}\text{sprandsym}(n, 0.01) + W_n, \quad B = [W_m, \text{sprand}(m, n-m, 0.05)],$$

$$C = [W_l, \text{sprand}(l, m-l, 0.05)],$$

where $W_i = \text{diag}\{1, 2, \ldots, i\}$ is a diagonal matrix, sprandsym$(n, 0.01)$ is a symmetric random, n-by-n, sparse matrix with approximately $0.01n^2$ nonzeros and each entry being the sum of one or more normally distributed random samples, and sprand$(m, l, 0.05)$ is a random, m-by-l, sparse matrix with approximately $0.05ml$ uniformly distributed nonzero entries.

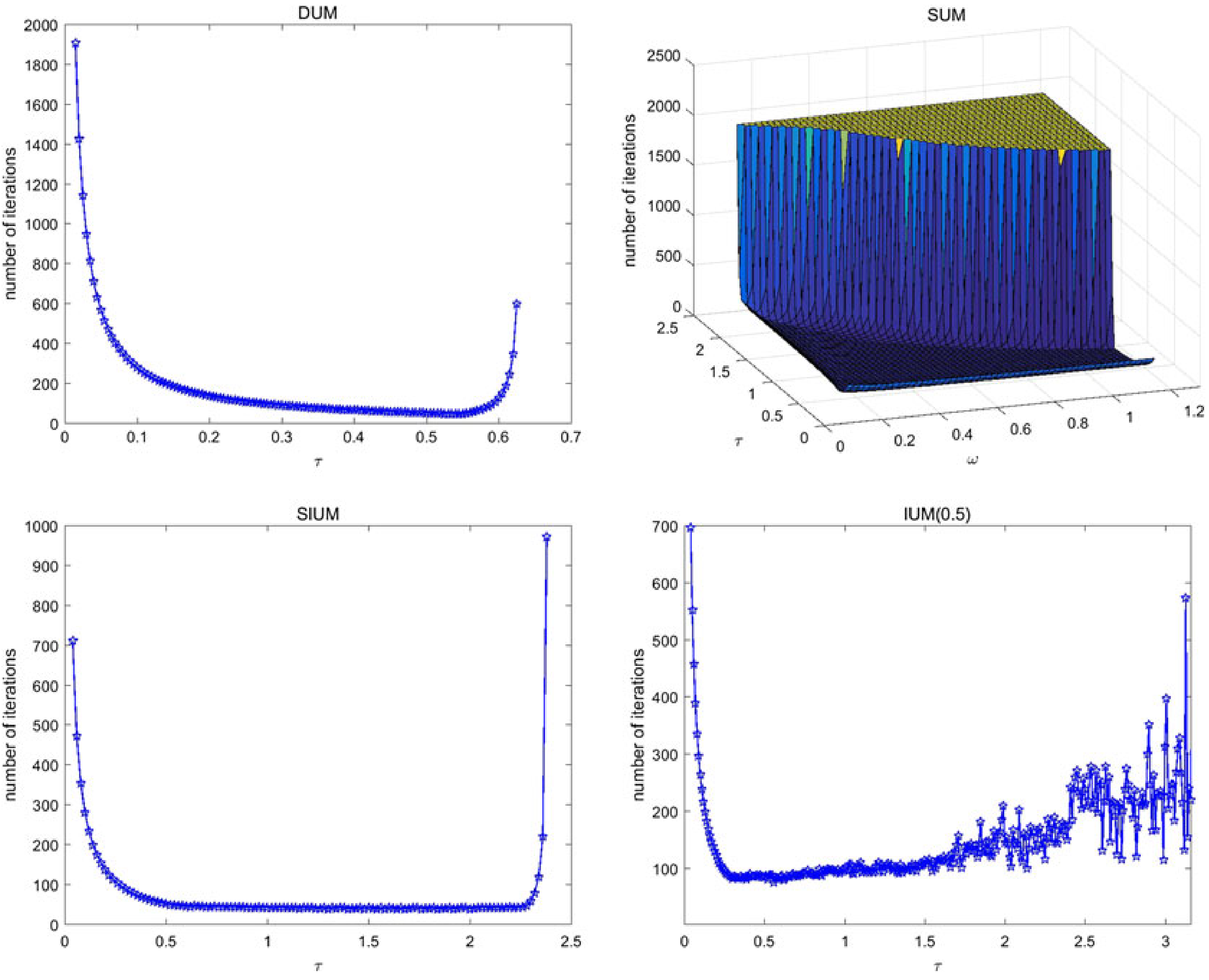

FIGURE 5 The characteristic curves of the number of iterations versus the parameters for Example 3 with DOF=2,080

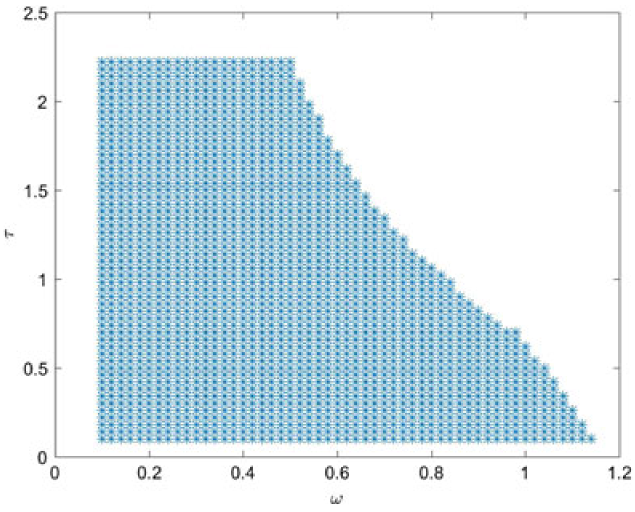

FIGURE 6 The region of the parameters that the SUM method satisfies Res $\leq 10^{-6}$ in 2,000 steps for Example 3 with DOF=2,080

In Example 4, we take $m = 0.8n$ and $l = 0.6n$. We choose $\omega = 1$ and $\tau = 1$ for the IUM(δ) method, the SIUM method and the SUM method. In addition, for the SIUM method, we set $r = 3$ and

$$M = \begin{pmatrix} LL^T & 0 \\ B & P_S \end{pmatrix}, \qquad N = \begin{pmatrix} LL^T - A & -B^T \\ 0 & P_S \end{pmatrix},$$

where L is produced by the incomplete Cholesky decomposition of A with the droptol being 0.01. We test this problem by choosing different values of n. The resulting numerical results are listed in Tables 5 and 6 and Figure 7. To see the role of the parameters in the convergence behaviors of the test methods, the characteristic curves of the number of iterations versus the parameters and the region of the parameters that the SUM method satisfies Res $\leq 10^{-6}$ in 2,000 steps are drawn out in Figure 8 and in Figure 6 for $n = 1,000$, respectively.

TABLE 5 The optimal parameter τ_{opt} of DUM for Example 4

	n	**1,000**	**5,000**	**10,000**	**30,000**	**50,000**
	DOF	2,400	12,000	24,000	72,000	120,000
DUM	τ_{opt}	0.9858	0.9827	0.9744	1.0001	-
	CPU	0.3962	11.0720	42.5889	57,216.2057	-

TABLE 6 Numerical results for Example 4

DOF		**DUM**			**SUM**	
	Iter	**CPU**	**Res**	**Iter**	**CPU**	**Res**
2,400	6	0.1450	2.1778e−08	7	0.0819	9.5071e−07
12,000	6	14.1205	6.9141e−08	7	8.1557	5.3868e−07
24,000	6	112.1068	1.9155e−07	7	62.2347	7.4345e−07
72,000	4	406.0063	7.6398e−07	6	286.3246	5.9487e−07
120,000	-	-	-	6	1,172.3025	4.8371e−07
		SIUM			MINRES	
2,400	14	0.1069	5.4387e−07	2,670	0.4529	9.9544e−07
12,000	14	8.6994	5.4336e−07	9,749	21.6911	9.9981e−07
24,000	14	47.1220	5.4391e−07	12,340	107.1463	9.9997e−07
72,000	14	436.2883	5.5322e−07	12,862	1,526.1595	9.9979e−07
120,000	14	1,430.8008	5.7400e−07	12,874	4,477.8089	9.9985e−07
		IUM(0.3)			IUM(0.5)	
2,400	7	0.0455	7.1540e−07	11	0.0636	1.4226e−07
12,000	8	2.0729	4.4867e−07	11	2.8031	4.9508e−07
24,000	9	12.4735	3.6772e−07	11	14.9362	9.6348e−07
72,000	9	114.0061	6.3576e−07	10	115.0651	9.1804e−07
120,000	9	315.2170	7.8701e−07	9	282.2954	8.1965e−07

Note. Iter =iterations; Res =relative residual; MINRES =minimal residual.

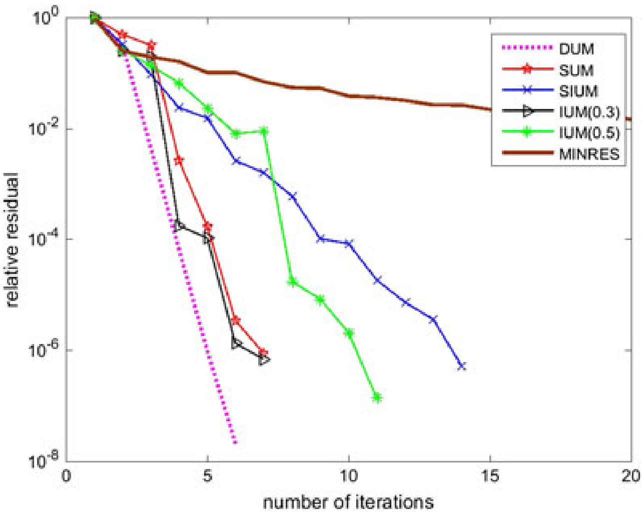

FIGURE 7 The iteration curves of the tested methods for Example 4 with DOF=2,400. MINRES = minimal residual

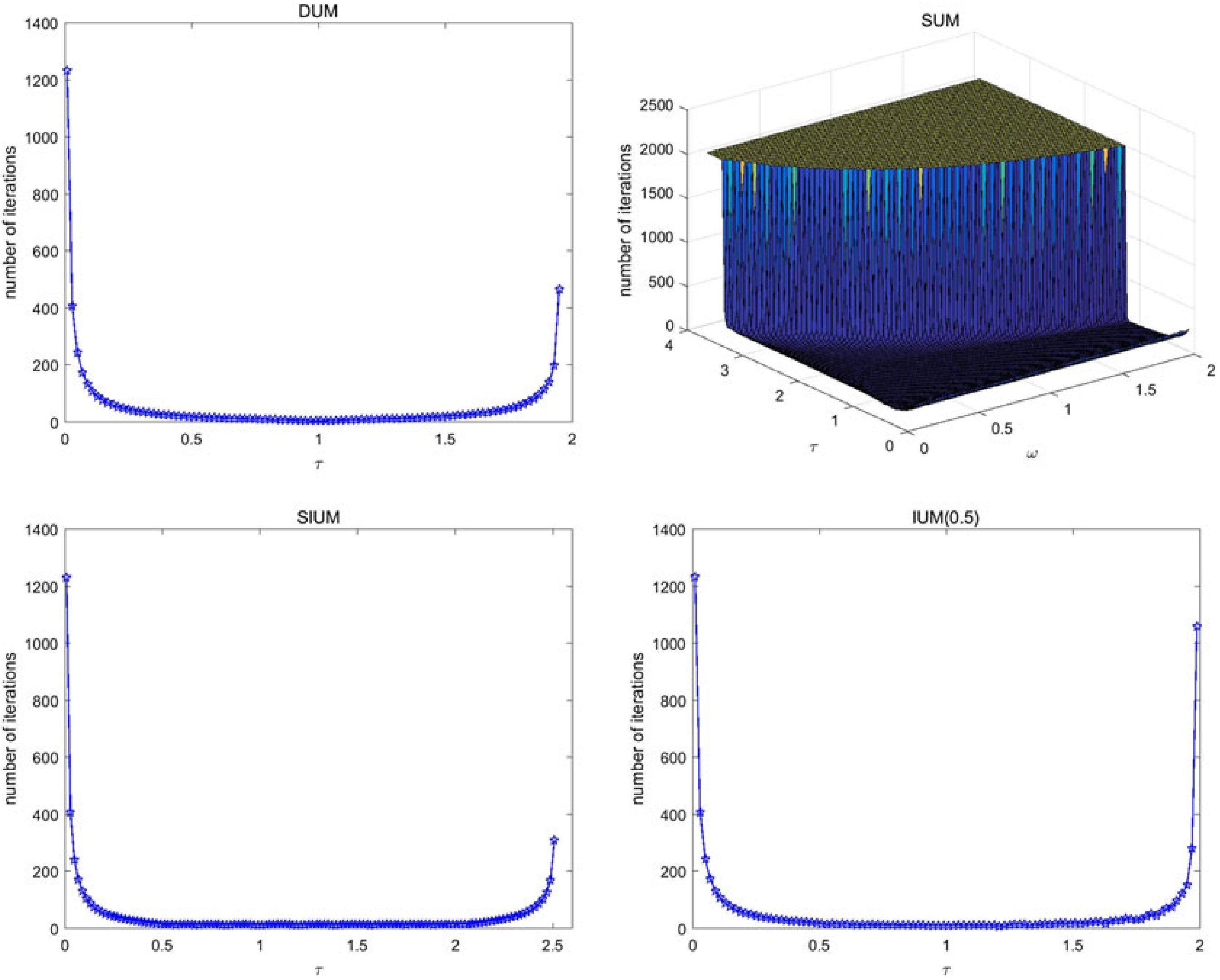

FIGURE 8 The characteristic curves of the number of iterations versus the parameter τ for Example 4 with DOF=2,400

Tables 1–6 and Figures 1–9 illustrate that Algorithms 2, 3, and 4 are feasible and efficient. From Tables 1–6, we can see that Algorithm 1 is quite impractical and expensive in terms of the CPU times. In Tables 3 and 5, the required CPU times of computing the optimal parameter τ_{opt} increase sharply with the dimension increased. This shows that deriving the

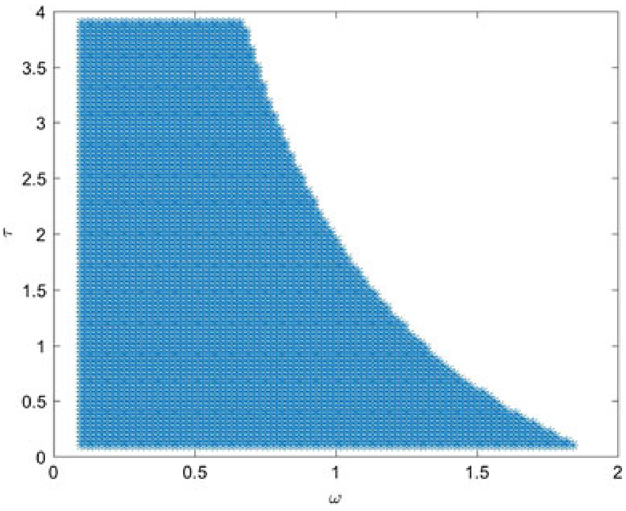

FIGURE 9 The region of the parameters that the SUM method satisfies Res $\leq 10^{-6}$ in 2,000 steps for Example 4 with DOF=2,400

optimal parameter $\tau_{\rm opt}$ is a time-consuming process. We can see from Tables 2, 4, and 6 that inexact Uzawa methods are evidently more efficient than exact methods. Figures 2, 5, and 8 indicate that the convergence rate of the tested methods depends strongly on the parameters.

6 | CONCLUSION

In this work, we have extended the Uzawa method (6) to solve the block three-by-three saddle-point problem (1). To avoid heavy computations at each step, we have also proposed an inexact Uzawa method (Algorithm 2), which solved the symmetric indefinite linear system $H(w_{k+1} - w_k) = q_k$ in some inexact way. Under suitable assumptions on the parameter τ and the approximation level δ, we have shown that the inexact Uzawa method converges to the unique solution of the saddle-point problem (1). Combining the splitting iteration method and a preconditioning technique, Algorithms 3 and 4 were customized for the inexact Uzawa method, respectively. Numerical results have demonstrated the effectiveness and robustness of Algorithms 2, 3, and 4.

The inexact Uzawa method still has a shortcoming, which needs to choose the parameter τ. Therefore, an interesting topic is how to compute the parameter τ. A future direction of research is introducing a variational parameter τ_k, which is computed by some inner products of vectors.

ACKNOWLEDGEMENTS

We would like to thank the referees for a number of very helpful comments and suggestions. This work does not have any conflicts of interest. The first author was supported by the National Postdoctoral Program for Innovative Talents (BX201600182) and the China Postdoctoral Science Foundation (2016M600141). The second author was supported by Chinese NSF grants (11631013, 11331012, 71331001 and 11971372) and the Key Project of Chinese National Programs for Fundamental Research and Development (2015CB856002). The third author was supported by the National Natural Science Foundation of China (G11571352).

ORCID

Yu-Hong Dai https://orcid.org/0000-0002-6932-9512

REFERENCES

1. Chen ZM, Du Q, Zou J. Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients. SIAM J Numer Anal. 2000;37:1542–1570.
2. Han DR, Yuan XM. Local linear convergence of the alternating direction method of multipliers for quadratic programs. SIAM J Numer Anal. 2013;51:3446–3457.
3. Arridge SR. Optical tomography in medical imaging. Inverse Problems. 1999;15:41–93.
4. Brezzi F, Fortin M. Mixed and hybrid finite element methods. New York, NY: Springer-Verlag; 1991.
5. Christiansen SH. Discrete Fredholm properties and convergence estimates for the electric field integral equation. Math Comput. 2004;73:143–167.
6. Day D, Heroux MA. Solving complex-valued linear systems via equivalent real formulations. SIAM J Sci Comput. 2001;23:480–498.
7. Elman HC, Ramage A, Silvester DJ. Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow. ACM Trans Math Softw. 2007;33:1–18.
8. Perugia I, Simoncini V. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. Numer Linear Algebra Appl. 2000;7:585–616.
9. Hestenes MR, Stiefel E. Methods of conjugate gradients for solving linear systems. J Res Nat Bur Stand. 1952;49:409–436.
10. Van der Vorst H. Iterative Krylov methods for large linear systems. Cambridge, UK: Cambridge University Press; 2003.
11. Bacuta C. A unified approach for Uzawa algorithms. SIAM J Numer Anal. 2006;44:2633–2649.
12. Bacuta C. Schur complements on Hilbert spaces and saddle point systems. J Comput Appl Math. 2009;225:581–593.
13. Bacuta C, Mccracken B, Shu L. Residual reduction algorithms for non-symmetric saddle point problems. J Comput Appl Math. 2011;235:1614–1628.
14. Bramble JH, Pasciak JE, Vassilev AT. Analysis of the inexact Uzawa algorithm for saddle point problems. SIAM J Numer Anal. 1997;34:1072–1092.
15. Bramble JH, Pasciak JE, Vassilev AT. Uzawa type algorithms for nonsymmetric saddle point problems. Math Comput. 2000;69:667–689.

16. Cheng X. On the nonlinear inexact Uzawa algorithm for saddle-point problems. SIAM J Numer Anal. 2000;37:1930–1934.
17. Elman HC, Golub GH. Inexact and preconditioned Uzawa algorithms for saddle point problems. SIAM J Numer Anal. 1994;31:1645–1661.
18. Hu Q, Zou J. An iterative method with variable relaxation parameters for saddle-point problems. SIAM J Matrix Anal Appl. 2001;23:317–338.
19. Hu Q, Zou J. Two new variants of nonlinear inexact Uzawa algorithms for saddle-point problems. Numerische Mathematik. 2002;93:333–359.
20. Hu Q, Zou J. Nonlinear inexact Uzawa algorithms for linear and nonlinear saddle-point problems. SIAM J Optim. 2006;16:798–825.
21. Zhang G-F, Yang J-L, Wang S-S. On generalized parameterized inexact Uzawa method for a block two-by-two linear system. J Comput Appl Math. 2014;255:193–207.
22. Zulehner W. Analysis of iterative methods for saddle point problems: a unified approach. Math Comput. 2002;71:479–505.
23. Bai Z-Z, Golub GH, Lu L-Z, Yin J-F. Block triangular and skew-Hermitian splitting method for positive-definite linear systems. SIAM J Sci Comput. 2005;26:844–863.
24. Bai Z-Z, Golub GH, Ng MK. Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems. SIAM J Matrix Anal Appl. 2003;24:603–626.
25. Bai Z-Z, Golub GH, Pan J-Y. Preconditioned Hermitian and skew-Hermitian splitting methods for non-Hermitian positive semidefinite linear systems. Numerische Mathematik. 2004;98:1–32.
26. Dyn N, Ferguson WE. The numerical solution of equality constrained quadratic programming problems. Math Comput. 1983;41:165–170.
27. Golub GH, Wathen AJ. An iteration for indefinite systems and its application to the Navier-Stokes equations. SIAM J Sci Comput. 1998;19:530–539.
28. Golub GH, Wu X, Yuan J-Y. SOR-like methods for augmented systems. BIT Numer Math. 2001;41:71–85.
29. Arioli M, Manzini G. A null space algorithm for mixed finite-element approximations of Darcy's equation. Commun Numer Meth Eng. 2002;18:645–657.
30. Gould NIM, Hribar ME, Nocedal J. On the solution of equality constrained quadratic programming problems arising in optimization. SIAM J Sci Comput. 2001;23:1376–1395.
31. Sarin V, Sameh A. An efficient iterative method for the generalized Stokes problem. SIAM J Sci Comput. 1998;19:206–226.
32. Benzi M. Solution of equality-constrained quadratic programming problems by a projection iterative method. Rend Mat Appl. 1993;13:275–296.
33. Forsgren A. Inertia-controlling factorizations for optimization algorithms. Appl Numer Math. 2002;43:91–107.
34. Greif C, Moulding E, Orban D. Bounds on eigenvalues of matrices arising from interior-point methods. SIAM J Optim. 2014;24:49–83.
35. Morini B, Simoncini V, Tani M. Spectral estimates for unreduced symmetric KKT systems arising from interior point methods. Numer Linear Algebra Appl. 2016;23:776–800.
36. Morini B, Simoncini V, Tani M. A comparison of reduced and unreduced KKT systems arising from interior point methods. Comput Optim Appl. 2017;68:1–27.
37. Arrow K, Hurwicz L, Uzawa H. Studies in nonlinear programming. Stanford, CA: Stanford University Press; 1958.
38. Bai Z-Z, Wang Z-Q. On parameterized inexact Uzawa methods for generalized saddle point problems. Linear Algebra Appl. 2008;428:2900–2932.
39. Young DM Jr. Iterative Solution for large linear systems. New York, NY: Academic Press; 1971.
40. Bank RE, Welfert BD, Yserentant H. A class of iterative methods for solving saddle point problems. Numerische Mathematik. 1990;56:645–666.
41. Chen X. On preconditioned Uzawa methods and SOR methods for saddle-point problems. J Comput Appl Math. 1998;100:207–224.
42. Queck W. The convergence factor of preconditioned algorithms of the Arrow-Hurwicz type. SIAM J Numer Anal. 1989;26:1016–1030.
43. Nikolova M, Ng MK, Zhang SQ, Ching W-K. Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. SIAM J Imaging Sci. 2008;1:2–25.

How to cite this article: Huang N, Dai Y-H, Hu Q. Uzawa methods for a class of block three-by-three saddle-point problems. *Numer Linear Algebra Appl*. 2019;e2265. https://doi.org/10.1002/nla.2265