Applied Mathematics and Computation 215 (2010) 3547–3562

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Least squares solution with the minimum-norm to general matrix equations via iteration ☆

Zhao-Yan Li [a], Yong Wang [a], Bin Zhou [b,*], Guang-Ren Duan [b]

[a] The Department of Mathematics, Harbin Institute of Technology, Harbin 150001, PR China
[b] The Center for Control Theory and Guidance Technology, Harbin Institute of Technology, Harbin 150001, PR China

ARTICLE INFO

Keywords:
Iterative algorithm
Gradient
General linear matrix equations
Minimal norm least squares
Optimal step size
Convergence rate

ABSTRACT

Two iterative algorithms are presented in this paper to solve the minimal norm least squares solution to a general linear matrix equations including the well-known Sylvester matrix equation and Lyapunov matrix equation as special cases. The first algorithm is based on the gradient based searching principle and the other one can be viewed as its dual form. Necessary and sufficient conditions for the step sizes in these two algorithms are proposed to guarantee the convergence of the algorithms for arbitrary initial conditions. Sufficient condition that is easy to compute is also given. Moreover, two methods are proposed to choose the optimal step sizes such that the convergence speeds of the algorithms are maximized. Between these two methods, the first one is to minimize the spectral radius of the iteration matrix and explicit expression for the optimal step size is obtained. The second method is to minimize the square sum of the F-norm of the error matrices produced by the algorithm and it is shown that the optimal step size exits uniquely and lies in an interval. Several numerical examples are given to illustrate the efficiency of the proposed approach.

1. Introduction

Linear matrix equations play an important role in linear systems theory. For example, the Sylvester matrix equation $AX + XB = C$ can be used to solve many control problems such as pole assignment [11], robust pole assignment [1], eigenstructure assignment [12] and fault detection. Its special forms include the well-known Lyapunov matrix equation $AX + XA^{T} = -C$ which has very important applications in stability analysis of linear systems [23].

Due to the important applications of this class of linear equations, many methods have been developed in the literature to provide both analytical and numerical solutions. For analytical solutions, Desouza et al. [5] have used the controllability and observability matrices to construct solutions to this class of equations. Besides analytical solutions, numerical methods for solving linear matrix equations are also well investigated in the literature. For example, by using the hierarchical

☆ The work of Bin Zhou and Zhao-Yan Li was partially supported by the National Natural Science Foundation of China under Grant Number 60904007; the work of Bin Zhou was partially supported by the Development Program for Outstanding Young Teachers at Harbin Institute of Technology under Grant Number HITQNJS.2009.054; the work of Guang-Ren Duan and Bin Zhou was partially supported by the Major Program of National Natural Science Foundation of China under Grant Number 60710002; and the work of Yong Wang and Zhao-Yan Li was partially supported by the National Natural Science Foundations of China under Grant Number 10771044 and the Natural Science Foundation of Heilongjiang Province under Grant Number 200605.

* Corresponding author. Address: Center for Control Theory and Guidance Technology, Harbin Institute of Technology, P.O. Box 416, Harbin, 150001 Heilongjiang, PR China.
E-mail addresses: zhaoyanlee@gmail.com (Z.-Y. Li), binzhoulee@163.com, binzhou@hit.edu.cn (B. Zhou).

0096-3003/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2009.10.052

identification principle, iterative algorithms are proposed in [6,8,9] to solve general linear matrix equations and coupled Sylvester matrix equations. For more references on this topic, see [13,16,18–20] and the references therein.

However, to the best of our knowledge, few results can be found in the literature for the following general linear matrix equation

$$\sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j = E, \tag{1}$$

where $A_i \in \mathbf{R}^{p\times m}, B_i \in \mathbf{R}^{n\times q}, C_j \in \mathbf{R}^{p\times n}, D_j \in \mathbf{R}^{m\times q}, i = 1,2,\ldots,r, j = 1,2,\ldots,s$, and $E \in \mathbf{R}^{p\times q}$ are known matrices and $X \in \mathbf{R}^{m\times n}$ is a matrix to be determined. Only some special cases of (1) were considered very recently. In [21], the following linear equation

$$AXB + CX^{\mathrm{T}}D = E, \tag{2}$$

where A, B, C and D are some known constant matrices of appropriate dimensions and X is a matrix to be determined, was considered. A more special case of (2), namely, the matrix equation $AX + X^{\mathrm{T}}C = B$, was investigated by Piao et al. [17]. The Moore–Penrose generalized inverse was used in [17] to find explicit solutions to this matrix equation. These results, however, are difficult to be extended to the more general case (1).

In this paper, we consider least squares solutions with minimal norm to the general linear matrix Eq. (1) by using iterative methods. In detail, we will solve the following least squares problem

$$\min_{X\in\mathbf{R}^{m\times n}} \left\| \sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j - E \right\|_{\mathrm{F}}.$$

Generally, solution to the above problem is not unique (see [4] for vector case). Therefore, we would like to search for the solution among them having the minimal norm, which is known as *minimal norm least squares solution*.

Using iterations to approximate exact solution to matrix equation has been well studied in the literature. For instance, the matrix sign function method was used in [3] to provide iterative solutions to the algebraic Riccati equations, cyclic Schur and Hessenberg-Schur methods were used in [2] to solve the periodic Lyapunov and Sylvester equations, the hierarchical identification principle, Hadmad product and star product were used in [6] to construct iterative algorithm for solving general Sylvester matrix equations and coupled Sylvester matrix equations, and an iterative algorithm was also proposed in [22] to solve the coupled discrete-time Markovian jump Lyapunov matrix equations. All these mentioned iteration based methods, however, are not directly applicable to obtaining the minimal norm least squares solution to the general linear matrix Eqs. (2) and (1).

In this paper, we also search for numerical solutions to the problem stated above by using iterations. Two iterative algorithms are proposed. Necessary and sufficient conditions that the step sizes in the algorithms should be satisfied to guarantee the convergence of the algorithms are presented. Moreover, we also provide two methods to choose the optimal step sizes in the algorithms such that the convergence rate, which is properly defined in this paper, is maximized. Some numerical examples are given to show the effectiveness of this method. Our results generalize our early results [15]. The merits of the proposed algorithms include: (1). They can be easily constructed without any factorizations on the coefficient matrices; (2). Only matrix multiplication is required during the iteration; (3). Convergence of the algorithms can be guaranteed provided the step sizes in the algorithms are small and (4). The optimal step sizes in the algorithm such that the convergence rates are maximized are given explicitly.

The remainder of this paper is organized as follows. Problem formulation is given in Section 2. The iterative algorithms to the problem and their convergence properties are proposed in Section 3. In Section 4, which contains two subsections, namely, Sections 4.1 and 4.2, we respectively give two methods to find the optimal step size such that the convergence rate of the algorithm is maximized. Some numerical examples are worked out in Section 5 to illustrate the effectiveness of the proposed algorithms and Section 6 concludes the paper.

Notations: Throughout this paper, we use $A^{\mathrm{T}}, \mathrm{tr}(A), \rho(A), \lambda(A), \|A\|_{\mathrm{F}}, \|A\|_2, \mathrm{Null}(A), \mathrm{Image}(A), \sigma_{\max}(A)$ and $\sigma_{\min}(A)$ to denote the transpose, trace, the spectral radius, the eigenvalue set, the Frobenius norm, the 2-norm, the null space, the image space, the maximal singular value and the minimal singular value of matrix A, respectively. The notation $P > 0$ means that P is symmetric and positive definite. I_n denotes an identity matrix of dimension n. If the subscript is omitted, the dimensions are consistent with the context. The Kronecker product of two matrices A and B is denoted by $A \otimes B$. The stretching function $\mathrm{vec}(A)$ where $A = [a_1, a_2, \ldots, a_m]$ is defined as $\mathrm{vec}(A) = \left[a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, \ldots, a_m^{\mathrm{T}}\right]^{\mathrm{T}}$. Let A be a nonsingular matrix. Then the condition number of A is defined as $\mathrm{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$.

2. Problem formulation

Consider the following general linear matrix equation

$$\sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j = E, \tag{3}$$

where $A_i \in \mathbf{R}^{p\times m}$, $B_i \in \mathbf{R}^{n\times q}$, $C_j \in \mathbf{R}^{p\times n}, D_j \in \mathbf{R}^{m\times q}$, $i = 1,2,\ldots,r, j = 1,2,\ldots,s, E \in \mathbf{R}^{p\times q}$ are known matrices and $X \in \mathbf{R}^{m\times n}$ is a matrix to be determined. The problem we are interested is stated as follows.

Problem 1. Let

$$\varpi_{\min} = \min_{X \in \mathbf{R}^{m\times n}} \left\{ \left\| \sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j - E \right\|_{\mathrm{F}} \right\}. \tag{4}$$

Find a matrix $X \in \mathbf{R}^{m\times n}$ such that $\|X\|_{\mathrm{F}}$ is minimized and

$$\left\| \sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j - E \right\|_{\mathrm{F}} = \varpi_{\min}. \tag{5}$$

Before giving solutions to Problem 1, we first introduce the following lemma.

Lemma 1 [4]. *Let $X \in \mathbf{R}^{m\times n}$ be any matrix. Then*

$$\mathrm{vec}(X^{\mathrm{T}}) = P(m,n)\mathrm{vec}(X),$$

where $P(m,n)$ is uniquely determined by the integers m and n. Moreover, the matrix $P(m,n)$ has the following properties.

1. For two arbitrary integers m and n, $P(m,n)$ has the following explicit form

$$P(m,n) = \begin{bmatrix} E_{11}^{\mathrm{T}} & E_{12}^{\mathrm{T}} & \cdots & E_{1n}^{\mathrm{T}} \\ E_{21}^{\mathrm{T}} & E_{22}^{\mathrm{T}} & \cdots & E_{2n}^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ E_{m1}^{\mathrm{T}} & E_{m2}^{\mathrm{T}} & \cdots & E_{mn}^{\mathrm{T}} \end{bmatrix}_{mn\times mn},$$

   where $E_{ij}, i = 1,2,\ldots,m, j = 1,2,\ldots,n$ is an $m \times n$ matrix with the element at position (i,j) being 1 and the others being 0.
2. For two arbitrary integers m and n, $P(m,n)$ is a unitary matrix, i.e.,

$$P(m,n)P^{\mathrm{T}}(m,n) = P^{\mathrm{T}}(m,n)P(m,n) = I_{mn}.$$

3. For two arbitrary integers m and n, there holds $P(m,n) = P^{\mathrm{T}}(n,m)$.
4. Let m,n,p and q be four integers and $A \in \mathbf{R}^{m\times n}, B \in \mathbf{R}^{p\times q}$. Then

$$P(m,p)(B \otimes A) = (A \otimes B)P(n,q).$$

By using the Kronecker product, Lemma 1 and the well-known formulation

$$\mathrm{vec}(AXB) = (B^{\mathrm{T}} \otimes A)\mathrm{vec}(X), \tag{6}$$

the linear matrix Eq. (3) can be converted to the vector form $\Upsilon x = e$ with $x = \mathrm{vec}(X), e = \mathrm{vec}(E)$ and

$$\Upsilon = \sum_{i=1}^{r} \left(B_i^{\mathrm{T}} \otimes A_i\right) + \sum_{j=1}^{s} \left(D_j^{\mathrm{T}} \otimes C_j\right) P(m,n) \in \mathbf{R}^{pq\times mn}. \tag{7}$$

With these notations, the expression in (4) can be rewritten as

$$\varpi_{\min} = \min_{x \in \mathbf{R}^{mn}} \|\Upsilon x - \mathrm{vec}(E)\|_{\mathrm{F}}. \tag{8}$$

As a result, Problem 1 can be restated as the following new problem.

Problem 2. Let $\varpi_{\min}$ be defined as (8). Find a vector $x \in \mathbf{R}^{mn}$ such that $\|x\|_{\mathrm{F}}$ is minimized and

$$\|\Upsilon x - \mathrm{vec}(E)\|_{\mathrm{F}} = \varpi_{\min}. \tag{9}$$

Regarding solution to Problem 2, we have the following simple result whose proof is omitted (see, for example, [4]).

Lemma 2. *Problem 2 has a unique solution x_∞ given by $x_\infty = \Upsilon^{+}\mathrm{vec}(E)$ where Υ^{+} is the unique Moore–Penrose inverse of Υ. Especially, if*

$$\mathrm{rank}(\Upsilon) = mn, \tag{10}$$

then the unique solution is given by

$$x_\infty = (\Upsilon^{\mathrm{T}}\Upsilon)^{-1}\Upsilon^{\mathrm{T}}\mathrm{vec}(E), \tag{11}$$

and if

$$\mathrm{rank}(\Upsilon) = pq, \tag{12}$$

then the unique solution is given by

$$x_\infty = \Upsilon^{\mathrm{T}}(\Upsilon\Upsilon^{\mathrm{T}})^{-1}\mathrm{vec}(E). \tag{13}$$

In the remaining of this paper, the unique solution to Problem 1 is denoted by X_∞. Clearly, we have

$$x_\infty = \mathrm{vec}(X_\infty). \tag{14}$$

3. Iterative solution to Problem 1

The basic idea of our method is to use the gradient based iteration to approximate the exact solution to Problem 1. Denote

$$J(X) = \frac{1}{2}\left\|\sum_{i=1}^{r} A_i X B_i + \sum_{j=1}^{s} C_j X^{\mathrm{T}} D_j - E\right\|_{\mathrm{F}}^2. \tag{15}$$

Then the gradient of $J(X)$ can be easily computed. The result is given as the following lemma whose proof is omitted for simplicity.

Lemma 3. *The gradient $\frac{\partial J(X)}{\partial X}$ where $J(X)$ is defined as (15), is given by*

$$\frac{\partial J(X)}{\partial X} = \sum_{i=1}^{r} A_i^{\mathrm{T}}\Delta(X)B_i^{\mathrm{T}} + \sum_{j=1}^{s} D_j\Delta^{\mathrm{T}}(X)C_j, \tag{16}$$

where

$$\Delta(X) = \sum_{v=1}^{r} A_v X B_v + \sum_{l=1}^{s} C_l X^{\mathrm{T}} D_l - E. \tag{17}$$

Then our gradient based iterative algorithm can be constructed as follows:

$$X(k) = X(k-1) - \mu\left(\sum_{i=1}^{r} A_i^{\mathrm{T}}\Delta(X(k-1))B_i^{\mathrm{T}} + \sum_{j=1}^{s} D_j\Delta^{\mathrm{T}}(X(k-1))C_j\right), \tag{18}$$

where μ is the step size to be specified later and

$$\Delta(k) = \sum_{v=1}^{r} A_v X(k) B_v + \sum_{l=1}^{s} C_l X^{\mathrm{T}}(k) D_l - E. \tag{19}$$

We then have the following result regarding the convergence of iteration (18).

Theorem 1. *Assume that (10) is satisfied. Let $X(k), k = 1, 2, \ldots$ be iteratively given by (18) with initial condition $X(0)$. Then $X(k)$ converges to X_∞, i.e., $\lim_{k\to\infty} X(k) = X_\infty$, for arbitrary initial condition $X(0)$ if and only if*

$$0 < \mu < \mu_{\max} = \frac{2}{\sigma_{\max}^2(\Upsilon)}, \tag{20}$$

where Υ is defined as (7).

Proof. Taking vec on both sides of (19) and using Lemma 1 gives

$$\mathrm{vec}(\Delta(X(k))) = \Upsilon\mathrm{vec}(X(k)) - \mathrm{vec}(E). \tag{21}$$

Similarly, it follows from (18) that

$$\mathrm{vec}(X(k)) = \mathrm{vec}(X(k-1)) - \mu\left(\sum_{i=1}^{r}\left(B_i \otimes A_i^{\mathrm{T}}\right) + \sum_{j=1}^{s}\left(C_j^{\mathrm{T}} \otimes D_j\right)P(p,q)\right)\mathrm{vec}(\Delta(X(k-1))). \tag{22}$$

By using Lemma 1, we have

$$\sum_{j=1}^{s}\left(C_j^{\mathrm{T}} \otimes D_j\right)P(p,q) = P(n,m)\sum_{j=1}^{s}\left(D_j \otimes C_j^{\mathrm{T}}\right). \tag{23}$$

Therefore, (22) can be written as

$$\begin{aligned}\mathrm{vec}(X(k)) &= \mathrm{vec}(X(k-1)) - \mu\left(\sum_{i=1}^{r}\left(B_i\otimes A_i^{\mathrm{T}}\right) + P(n,m)\sum_{j=1}^{s}\left(D_j\otimes C_j^{\mathrm{T}}\right)\right)\mathrm{vec}(\Delta(X(k-1)))\\ &= \mathrm{vec}(X(k-1)) - \mu\left(\sum_{i=1}^{r}\left(B_i^{\mathrm{T}}\otimes A_i\right) + \sum_{j=1}^{s}\left(D_j^{\mathrm{T}}\otimes C_j\right)P^{\mathrm{T}}(n,m)\right)^{\mathrm{T}}\mathrm{vec}(\Delta(X(k-1)))\\ &= \mathrm{vec}(X(k-1)) - \mu\left(\sum_{i=1}^{r}\left(B_i^{\mathrm{T}}\otimes A_i\right) + \sum_{j=1}^{s}\left(D_j^{\mathrm{T}}\otimes C_j\right)P(m,n)\right)^{\mathrm{T}}\mathrm{vec}(\Delta(X(k-1)))\\ &= \mathrm{vec}(X(k-1)) - \mu\Upsilon^{\mathrm{T}}\mathrm{vec}(\Delta(X(k-1))).\end{aligned} \tag{24}$$

Inserting (21) into (24) gives

$$\mathrm{vec}(X(k)) = (I-\mu\Upsilon^{\mathrm{T}}\Upsilon)\mathrm{vec}(X(k-1)) + \mu\Upsilon^{\mathrm{T}}\mathrm{vec}(E). \tag{25}$$

Then it follows from (11) that

$$\Upsilon^{\mathrm{T}}\Upsilon\mathrm{vec}(X_\infty) = \Upsilon^{\mathrm{T}}\mathrm{vec}(E). \tag{26}$$

Hence, by substituting (26) into (25), we obtain

$$\mathrm{vec}(X(k)) - \mathrm{vec}(X_\infty) = (I-\mu\Upsilon^{\mathrm{T}}\Upsilon)\mathrm{vec}(X(k-1)) + \mu\Upsilon^{\mathrm{T}}\Upsilon\mathrm{vec}(X_\infty) - \mathrm{vec}(X_\infty) = (I-\mu\Upsilon^{\mathrm{T}}\Upsilon)(\mathrm{vec}(X(k-1)) - \mathrm{vec}(X_\infty)),$$

which, by denoting $\mathscr{X} = X - X_\infty$, can be written as

$$\mathrm{vec}(\mathscr{X}(k)) = (I-\mu\Upsilon^{\mathrm{T}}\Upsilon)\mathrm{vec}(\mathscr{X}(k-1)). \tag{27}$$

Clearly, it follows from the above relation that $\lim_{k\to\infty}\mathscr{X}(k) = 0$ for arbitrary initial condition $\mathscr{X}(0)$ if and only if $I-\mu\Upsilon^{\mathrm{T}}\Upsilon$ is Schur stable, i.e., $\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon) < 1$. We note that $I-\mu\Upsilon^{\mathrm{T}}\Upsilon$ is a symmetric matrix. Therefore, we have $\lambda(I-\mu\Upsilon^{\mathrm{T}}\Upsilon) = \{1-\mu\sigma_i^2(\Upsilon)\}_{i=1}^{mn}$ and

$$\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon) = \max_{1\leqslant i\leqslant mn}\{|1-\mu\sigma_i^2(\Upsilon)|\}.$$

Hence $\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon) < 1$ if and only if $|1-\mu\sigma_{\max}^2(\Upsilon)| < 1$ which is equivalent to (20). The proof is completed. □

Remark 1. If $\mu = \mu_{\max}$, then the algorithm in (18) does not converge to X_∞ for arbitrary initial condition according to Theorem 1. Nevertheless, it converges to a matrix that is bounded in norm and dependent on the initial condition $X(0)$. To see this, we note that

$$\lambda(I-\mu_{\max}\Upsilon^{\mathrm{T}}\Upsilon) = \{-1,\lambda_1,\lambda_2,\ldots,\lambda_p\},$$

where $|\lambda_i| < 1, i = 1,2,\ldots,p$, and moreover, the algebraic multiplicity and geometric multiplicity for the eigenvalue -1 are the same. Therefore, it follows from the discrete-time linear system theory (see, for instance, [14]) that the solution to the difference Eq. (27) is bounded in norm for arbitrary bounded initial condition. This fact can also be observed in the examples given later.

The following proposition can be obtained immediately.

Proposition 1. *Assume that (10) is satisfied. Let μ satisfy inequality (20) and $X(k), k = 1,2,\ldots$ be iteratively produced by the iteration (18). Then*

$$\|X(k) - X_\infty\|_{\mathrm{F}} < \|X(k-1) - X_\infty\|_{\mathrm{F}}, \quad \forall k \geqslant 1. \tag{28}$$

Proof. Denote

$$x(k) = \mathrm{vec}(\mathscr{X}(k)). \tag{29}$$

Then it follows from (27) that

$$x^{\mathrm{T}}(k)x(k) = x^{\mathrm{T}}(k-1)(I-\mu\Upsilon^{\mathrm{T}}\Upsilon)^2x(k-1),$$

which can be equivalently written as

$$x^{\mathrm{T}}(k)x(k) - x^{\mathrm{T}}(k-1)x(k-1) = x^{\mathrm{T}}(k-1)F(\mu)x(k-1),$$

with

$$F(\mu) = -\mu\Upsilon^{\mathrm{T}}\Upsilon(2I-\mu\Upsilon^{\mathrm{T}}\Upsilon).$$

Note that μ satisfies (20), then $2I-\mu\Upsilon^{\mathrm{T}}\Upsilon$ is positive definite. Since $\mu\Upsilon^{\mathrm{T}}\Upsilon$ and $2I-\mu\Upsilon^{\mathrm{T}}\Upsilon$ are commutable, the matrix $F(\mu)$ is symmetric and negative definite. As a result, we get

$$\|x(k)\|_2^2 < \|x(k-1)\|_2^2,$$

which is equivalent to (28) as

$$\|X\|_{\mathrm{F}}^2 = \|\mathrm{vec}(X)\|_2^2. \qquad \square \tag{30}$$

If we use $\|X(k)-X_\infty\|_{\mathrm{F}}$ to denote the distance between $X(k)$ and X_∞, then Proposition 1 indicates that the distance between $X(k)$ and X_∞ decreases monotonously as k increases. Therefore, the closer the initial condition to the exact solution X_∞, the fewer the iteration steps the algorithm will need. But it is difficult to guess an initial condition that is sufficiently close to X_∞. In practice, we can simply take $X(0)=0$.

Though Theorem 1 provides a necessary and sufficient condition to guarantee the convergence of the algorithm (18), the right hand side of (20) is difficult to calculate as the matrix Υ may have very high dimensions. Therefore, we next provide a sufficient condition that is easy to compute.

Corollary 1. *Assume that (10) is satisfied. Let $X(k), k=1,2,\ldots$ be iteratively produced by the iteration in (18). Then $\lim_{k\to\infty}X(k)=X_\infty$ holds true for arbitrary initial condition $X(0)$ if*

$$0<\mu<\frac{2}{v_1} \text{ or } 0<\mu<\frac{2}{v_2^2}, \tag{31}$$

where v_1 and v_2 are, respectively, given by

$$\begin{aligned} v_1 &= (r+s)\left(\sum_{i=1}^{r}\|B_i\|_2^2\|A_i\|_2^2+\sum_{j=1}^{s}\|D_j\|_2^2\|C_j\|_2^2\right),\\ v_2 &= \sum_{i=1}^{r}\|B_i\|_2\|A_i\|_2+\sum_{j=1}^{s}\|D_j\|_2\|C_j\|_2, \end{aligned} \tag{32}$$

and satisfy $v_2^2\leqslant v_1$

Proof. It follows from Lemma 1 that $P(m,n)$ is a unitary matrix. Therefore,

$$\begin{aligned} \sigma_{\max}(\Upsilon) &= \left\|\sum_{i=1}^{r}\left(B_i^{\mathrm{T}}\otimes A_i\right)+\sum_{j=1}^{s}\left(D_j^{\mathrm{T}}\otimes C_j\right)P(m,n)\right\|_2\\ &\leqslant \left\|\sum_{i=1}^{r}\left(B_i^{\mathrm{T}}\otimes A_i\right)\right\|_2+\left\|\sum_{j=1}^{s}\left(D_j^{\mathrm{T}}\otimes C_j\right)P(m,n)\right\|_2\\ &= \left\|\sum_{i=1}^{r}\left(B_i^{\mathrm{T}}\otimes A_i\right)\right\|_2+\left\|\sum_{j=1}^{s}\left(D_j^{\mathrm{T}}\otimes C_j\right)\right\|_2\\ &\leqslant \sum_{i=1}^{r}\left\|B_i^{\mathrm{T}}\otimes A_i\right\|_2+\sum_{j=1}^{s}\left\|D_j^{\mathrm{T}}\otimes C_j\right\|_2. \end{aligned}$$

Since $\|A\otimes B\|_2=\|A\|_2\|B\|_2$, we can further obtain

$$\sigma_{\max}(\Upsilon)\leqslant\sum_{i=1}^{r}\|B_i\|_2\|A_i\|_2+\sum_{j=1}^{s}\|D_j\|_2\|C_j\|_2, \tag{33}$$

which in turn implies that

$$\sigma_{\max}^2(\Upsilon)\leqslant\left(\sum_{i=1}^{r}\|B_i\|_2\|A_i\|_2+\sum_{j=1}^{s}\|D_j\|_2\|C_j\|_2\right)^2\leqslant(r+s)\left(\sum_{i=1}^{r}\|B_i\|_2^2\|A_i\|_2^2+\sum_{j=1}^{s}\|D_j\|_2^2\|C_j\|_2^2\right), \tag{34}$$

where we have used the following well-known inequality

$$\left(\sum_{i=1}^{p}a_i\right)^2\leqslant p\sum_{i=1}^{p}a_i^2.$$

Substituting (33) and (34) into (20) gives (31). The proof is completed. $\square$

Remark 2. The matrix Υ will have large dimensions if the matrices in the Eq. (3) have large dimensions. In this case, it is difficult to compute the singular value of Υ directly. This implies that the proposed iteration (18) may be not suitable for large matrices by using Theorem 1 where the singular value of Υ is required. From the point of selecting step size, Corollary 1 is more useful than Theorem 1.

Clearly, if the relation in (10) does not hold, then the iteration in (18) will not converge. In the following, we will present another iterative algorithm to solve Problem 1 when the condition in (12) is met. Our new iteration is constructed as follows:

$$Y(k) = Y(k-1) - \mu\left(\sum_{i=1}^{r} A_i \Delta(Y(k-1)) B_i + \sum_{j=1}^{s} C_j \Delta^{\mathrm{T}}(Y(k-1)) D_j - E\right), \tag{35}$$

where

$$\Delta(Y(k)) = \sum_{v=1}^{r} A_v^{\mathrm{T}} Y(k) B_v^{\mathrm{T}} + \sum_{l=1}^{s} D_l Y^{\mathrm{T}}(k) C_l, \tag{36}$$

with initial condition $Y(0)$ and step size μ that is to be determined later. Regarding the convergence of iteration (35), we can present the following result.

Theorem 2. *Assume that (12) is satisfied. Let X_∞ be the unique solution to Problem 1. Then the iteration (35) converges to a finite matrix Y_∞ for arbitrary initial condition if and only if*

$$0 < \mu < \frac{2}{\sigma_{\max}^2(\Upsilon)}. \tag{37}$$

Furthermore, if (37) is satisfied and $\lim_{k\to\infty} Y(k) = Y_\infty$, then

$$X_\infty = \sum_{i=1}^{r} A_i^{\mathrm{T}} Y_\infty B_i^{\mathrm{T}} + \sum_{l=1}^{s} D_l Y_\infty^{\mathrm{T}} C_l. \tag{38}$$

Proof. Taking vec on both sides of (35) gives

$$\begin{aligned}
\mathrm{vec}(Y(k)) &= \mathrm{vec}(Y(k-1)) + \mu\mathrm{vec}(E) \\
&\quad - \mu\left(\sum_{i=1}^{r}\left(B_i^{\mathrm{T}} \otimes A_i\right) + \sum_{j=1}^{s}\left(D_j^{\mathrm{T}} \otimes C_j\right) P(m,n)\right)\mathrm{vec}(\Delta(Y(k-1))) \\
&= \mathrm{vec}(Y(k-1)) - \mu\Upsilon\mathrm{vec}(\Delta(Y(k-1))) + \mu\mathrm{vec}(E).
\end{aligned} \tag{39}$$

On the other hand, taking vec on both sides of (36) and using (23) gives

$$\begin{aligned}
\mathrm{vec}(\Delta(Y(k))) &= \left(\sum_{i=1}^{r}\left(B_i \otimes A_i^{\mathrm{T}}\right) + \sum_{j=1}^{s}\left(C_j^{\mathrm{T}} \otimes D_j\right) P(q,p)\right)\mathrm{vec}(Y(k)) \\
&= \left(\sum_{i=1}^{r}\left(B_i \otimes A_i^{\mathrm{T}}\right) + P(n,m)\sum_{j=1}^{s}\left(D_j \otimes C_j^{\mathrm{T}}\right)\right)\mathrm{vec}(Y(k)) \\
&= \left(\sum_{i=1}^{r}\left(B_i^{\mathrm{T}} \otimes A_i\right) + \sum_{j=1}^{s}\left(D_j \otimes C_j^{\mathrm{T}}\right) P^{\mathrm{T}}(n,m)\right)^{\mathrm{T}}\mathrm{vec}(Y(k)) \\
&= \left(\sum_{i=1}^{r}\left(B_i^{\mathrm{T}} \otimes A_i\right) + \sum_{j=1}^{s}\left(D_j \otimes C_j^{\mathrm{T}}\right) P(m,n)\right)^{\mathrm{T}}\mathrm{vec}(Y(k)) \\
&= \Upsilon^{\mathrm{T}}\mathrm{vec}(Y(k)).
\end{aligned} \tag{40}$$

Therefore, (39) can be written as

$$\mathrm{vec}(Y(k)) = (I - \mu\Upsilon\Upsilon^{\mathrm{T}})\mathrm{vec}(Y(k-1)) + \mu\mathrm{vec}(E).$$

Similar to the proof of Theorem 1, the above iteration converges if and only if (37) is satisfied. Moreover, when (37) is met, we have

$$\mathrm{vec}(Y_\infty) = (\Upsilon\Upsilon^{\mathrm{T}})^{-1}\mathrm{vec}(E). \tag{41}$$

On the other hand, similar to the procedure used in obtaining (40), we can deduce from (38) that

$$\mathrm{vec}(X_\infty) = \Upsilon^{\mathrm{T}}\mathrm{vec}(Y_\infty). \tag{42}$$

Consequently, it follows from (41) and (42) that

$$\mathrm{vec}(X_\infty) = \Upsilon^{\mathrm{T}}(\Upsilon\Upsilon^{\mathrm{T}})^{-1}\mathrm{vec}(E)$$

which clearly indicates that X_∞ is the unique solution to Problem 1 in view of Lemma 2 and (14). □

Remark 3. The iteration (35) can be regarded as the dual form of the iteration in (18). Similar results corresponding to Proposition 1 and Corollary 1 to the iteration in (35) can be obtained, which are omitted for brevity.

Although Theorem 1 and Corollary 1 give conditions to choose the step size μ to guarantee the convergence of the algorithm (18), they do not provide a way to choose the "optimal" step size μ such that the iteration in (18) converges fastest. In fact, convergence rate is an important index for measuring the ability of an iterative algorithm. In the next section, we will consider this problem in detail. As explained in Remark 3, iteration (35) can be regarded as the dual from of iteration (18), we will only consider the iteration in (18) since corresponding results to iteration (35) can be obtained in a very similar way.

4. Convergence rate analysis of the algorithms

4.1. Convergence rate analysis by using spectral radius

Consider a linear iteration

$$X(k) = AX(k-1) + B,\ X(k) \in \mathbf{R}^{n\times m}, \quad k \geqslant 1, \tag{43}$$

where A and B are constant matrices with appropriate dimensions. It is well known that the iteration in (43) converges to a finite matrix X_∞ for arbitrary initial condition $X(0)$ if and only if $\rho(A) < 1$ [4]. Moreover, the smaller the $\rho(A)$, the faster the iteration will converge. For this reason, the number $-\log(\rho(A))$ is usually used to denote the convergence rate of the iteration (43) [4]. For clarity, we firstly introduce the following definition for convergence rate of the iteration (43).

Definition 1 [15]. Assume that the iteration (43) converges to the unique matrix X_∞ for arbitrary initial condition $X(0)$. The α-convergence rate for the iteration (43) is a scalar $\gamma = -\log\beta$ with $0 < \beta < 1$ such that

$$\|X(k) - X_\infty\|_\alpha \leqslant K\beta^k\|X(0) - X_\infty\|_\alpha, \quad k \geqslant 0, \tag{44}$$

and there exists at least one $X(0)$ such that "=" hold in (44). In (44), K is a positive scalar independent of k and β, and α denotes a suitable matrix norm (e.g., $\alpha = 2$ or $\alpha = \mathrm{F}$).

Our next lemma shows that $-\log(\rho(A))$ can indeed be used to denote the 2-convergence rate of the iteration (43) in a special case. The proof is similar to the proof of Lemma 2 in [15] and is thus omitted.

Lemma 4. *Consider the iteration in (43) where $A \in \mathbf{R}^{n\times n}$ is a real symmetric matrix with $\rho(A) < 1$ and $X(k) = x(k) \in \mathbf{R}^n, \forall k \geqslant 0$, is a vector. Then the 2-convergence rate of the iteration (43) is $-\log(\rho(A))$ in the sense of Definition 1. Moreover, for arbitrary initial condition $x(0)$, there holds*

$$\|x(k) - x_\infty\|_2 \leqslant \rho^k(A)\|x(0) - x_\infty\|_2. \tag{45}$$

We recall another technical lemma that will be used later.

Lemma 5 [15]. *Assume that $m_i, i = 1, 2, \ldots, n$, are some given positive scalars. Denote $m_{\max} = \max_{1\leqslant i\leqslant n}\{m_i\}$ and $m_{\min} = \min_{1\leqslant i\leqslant n}\{m_i\}$. Then*

$$\min_{0<u<\frac{2}{m_{\max}}} \max_{1\leqslant i\leqslant n}\{|1 - um_i|\} = \frac{m_{\max} - m_{\min}}{m_{\max} + m_{\min}}. \tag{46}$$

Moreover, the unique u_{opt} such that the above relation holds is

$$u_{\mathrm{opt}} = \frac{2}{m_{\max} + m_{\min}}.$$

Then we can prove the following result.

Theorem 3. *Assume that (10) is satisfied. Let $X(k), k = 1, 2, \ldots$ be iteratively given by (18) with initial condition $X(0)$ and step size μ satisfying (20). Then the F-convergence rate of the algorithm (18) is maximized if*

$$\mu = \mu_{\mathrm{opt}}^{\mathrm{sr}} = \frac{2}{\sigma_{\max}^2(\Upsilon) + \sigma_{\min}^2(\Upsilon)}. \tag{47}$$

Moreover, if μ is chosen as (47), then

$$\|X(k) - X_\infty\|_{\mathrm{F}} \leqslant \left(\frac{\mathrm{cond}^2(\Upsilon) - 1}{\mathrm{cond}^2(\Upsilon) + 1}\right)^k \|X(0) - X_\infty\|_{\mathrm{F}}, \quad k \geqslant 0. \tag{48}$$

Proof. Note that $I - \mu\Upsilon^{\mathrm{T}}\Upsilon$ is a Schur stable and symmetric matrix. Therefore, it follows from Lemma 4 that $-\log(\rho(I - \mu\Upsilon^{\mathrm{T}}\Upsilon))$ is the 2-convergence rate of the iteration (27). Hence, the 2-convergence rate of the iteration (27) is

maximized if and only if $-\log(\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon))$ is maximized, or equivalently, $\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon)$ is minimized. That is to solve the following optimization problem

$$\min_{0<\mu<\mu_{\max}}\{\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon)\}=\min_{0<\mu<\mu_{\max}}\left\{\max_{1\leqslant i\leqslant mn}\{|1-\mu\sigma_i^2(\Upsilon)|\}\right\}. \tag{49}$$

We notice that (49) is in the form of (46). Therefore, according to Lemma 5, $\rho(I-\mu\Upsilon^{\mathrm{T}}\Upsilon)$ is minimized if μ is chosen as (47). Moreover,

$$\rho\left(I-\mu_{\mathrm{opt}}^{\mathrm{sr}}\Upsilon^{\mathrm{T}}\Upsilon\right)=\frac{\sigma_{\max}^2(\Upsilon)-\sigma_{\min}^2(\Upsilon)}{\sigma_{\max}^2(\Upsilon)+\sigma_{\min}^2(\Upsilon)}=\frac{\mathrm{cond}^2(\Upsilon)-1}{\mathrm{cond}^2(\Upsilon)+1}. \tag{50}$$

As a result, it follows from (45) in Lemma 4 that

$$\|x(k)\|_2\leqslant\rho^k\left(I-\mu_{\mathrm{opt}}^{\mathrm{sr}}\Upsilon^{\mathrm{T}}\Upsilon\right)\|x(0)\|_2, \tag{51}$$

which is equivalent to (48) in view of (30) and (50). At last, we show that $\mu=\mu_{\mathrm{opt}}^{\mathrm{sr}}$ satisfies the condition (20). Since Υ is of full column rank, we have $\sigma_{\min}(\Upsilon)\neq 0$. That is $\mu_{\mathrm{opt}}^{\mathrm{sr}}<\mu_{\max}$. The proof is completed. □

Remark 4. We recall the well-known Conjugate-Gradient (CG) method for solving the linear matrix equation

$$Ax=B, \tag{52}$$

where A is positive definite. Let x_∞ be the exact solution to (52) and $x(k)$ be iteratively given by the CG method with initial condition $x(0)$ (see [4]). Then the CG iteration satisfies the following relation

$$\|x(k)-x_\infty\|_A\leqslant 2\left(\frac{\sqrt{\mathrm{cond}(A)}-1}{\sqrt{\mathrm{cond}(A)}+1}\right)^k\|x(0)-x_\infty\|_A, \tag{53}$$

where $\|x\|_A^2=x^{\mathrm{T}}Ax$. It is very interesting to note that (48) is similar to (53). This similarity indicates that our method will suffer the same problem as the CG method that the convergence performance becomes poor if A is badly conditioned. To improve the convergence performance of the CG method when A is badly conditioned, the preconditioned CG method, which is still a studying subject in the literature (e.g., [4]), is used instead. Our further study should adequately take this problem into account.

For illustration of the proposed theory, see Example 1 in Section 5. However, $\mu_{\mathrm{opt}}^{\mathrm{sr}}$ is not always the optimal step size for the iteration in (18). This can be observed by another simple example – Example 2 in Section 5. This phenomenon can be explained as follows. We note that $\mu_{\mathrm{opt}}^{\mathrm{sr}}$ is the solution to the following min–max optimization problem

$$\min_{0<\mu<\mu_{\max}}\max_{1\leqslant i\leqslant N}\{|1-\mu\sigma_i^2(\Upsilon)|\}.$$

This problem is quite similar to the well-known $\mathscr{H}_\infty$ optimal control problem (which is also a min–max problem):

$$\min_{K(s)\ \text{stabilizing}\ G(s)}\sup_{\omega\in\mathbf{R}}\{\sigma_{\max}(T(\mathrm{j}\omega))\},$$

where $T(s)=G_{11}(s)+G_{12}(s)K(s)(I-G_{22}(s)K(s))^{-1}G_{21}(s)$ with $G(s)$ and $K(s)$ being two rational and proper transfer function matrices and $G(s)$ admitting the following partition

$$G(s)=\begin{bmatrix}G_{11}(s) & G_{12}(s)\\ G_{21}(s) & G_{22}(s)\end{bmatrix}.$$

Optimal solution to min–max solution is conservative in practice as it is optimal in the "worst case" which may not happen at all. This fact in $\mathscr{H}_\infty$ optimal control problem has been emphasized in many references (for example, [10,23]). In this paper, the "worst case" is that the initial condition $X(0)$ should be chosen such that

$$\left\|\left(I-\mu_{\mathrm{opt}}^{\mathrm{sr}}\Upsilon^{\mathrm{T}}\Upsilon\right)x(0)\right\|_2=\left\|\left(I-\mu_{\mathrm{opt}}^{\mathrm{sr}}\Upsilon^{\mathrm{T}}\Upsilon\right)\right\|_2\|x(0)\|_2,$$

which is generally not satisfied in practice.

Therefore, we will give in the next subsection another approach to select the optimal step size μ such that a more reasonable objective function for measuring the convergence performance of the iteration in (18) is minimized.

4.2. Convergence rate analysis by using error square sum

It is nature to choose the following index function

$$J_{\mathrm{e}}(\mu,X_0)=\sum_{k=0}^{\infty}\|X(k)-X_\infty\|_{\mathrm{F}}^2, \tag{54}$$

which can be understood as the square sum of the error of the iteration in (18), to measure the convergence rate of the algorithm in (18). Obviously, the smaller the $J_e(\mu, X_0)$, the better the convergence performance of the algorithm (18) is. In this subsection, we will study the property of $J_e(\mu, X_0)$ as a function of μ and X_0, and investigate how to choose the parameter μ such that $J_e(\mu, X_0)$ is minimized. First we need a compact expression for $J_e(\mu, X_0)$.

Lemma 6. *Assume that Υ is of full column rank and $X(k), k = 1, 2, \ldots$ are iteratively given by (18) with initial condition $X(0) = X_0$ and step size μ satisfying (20). Let $J_e(\mu, X_0)$ be defined as (54). Then*

$$J_e(\mu, X_0) = x_0^T Q(\mu) x_0, \tag{55}$$

where $x_0 = \mathrm{vec}(X_0 - X_\infty)$ and

$$Q(\mu) = \left(\mu \Upsilon^T \Upsilon \left(2I - \mu \Upsilon^T \Upsilon\right)\right)^{-1}. \tag{56}$$

Proof. Let $X(k)$ and $x(k)$ be related with (29). It follows from (27) that

$$x^T(k)x(k) = x_0^T((I - \mu \Upsilon^T \Upsilon)^2)^k x_0.$$

Therefore, in view of (30), we have

$$\begin{aligned} J_e(\mu, X_0) &= \sum_{k=0}^{\infty} \|\mathcal{X}(k)\|_F^2 \\ &= \sum_{k=0}^{\infty} \|x(k)\|_2^2 \\ &= \sum_{k=0}^{\infty} x^T(k)x(k) \\ &= \sum_{k=0}^{\infty} x_0^T((I - \mu \Upsilon^T \Upsilon)^2)^k x_0. \end{aligned} \tag{57}$$

Since μ satisfies (20), the matrix $(I - \mu \Upsilon^T \Upsilon)^2$ is Schur stable and

$$\sum_{k=0}^{\infty} ((I - \mu \Upsilon^T \Upsilon)^2)^k = (I - (I - \mu \Upsilon^T \Upsilon)^2)^{-1} = Q(\mu). \tag{58}$$

Substituting (58) into (57) gives (55). This completes the proof. □

We give another technical lemma whose proof is given in Appendix A.

Lemma 7. *Assume that Υ is of full column rank and μ satisfies (20). Let $Q(\mu)$ be defined as (56). Then*

$$\frac{dQ(\mu)}{d\mu} = -2(\mu \Upsilon^T \Upsilon (2I - \mu \Upsilon^T \Upsilon))^{-2} (I - \mu \Upsilon^T \Upsilon) \Upsilon^T \Upsilon, \tag{59}$$

$$\frac{d^2 Q(\mu)}{d\mu^2} = 8\mu^{-3} (\Upsilon^T \Upsilon)^{-1} (2I - \mu \Upsilon^T \Upsilon)^{-3} (I - \mu \Upsilon^T \Upsilon)^2 + 2(\mu(2I - \mu \Upsilon^T \Upsilon))^{-2}. \tag{60}$$

Moreover, $\frac{d^2}{d\mu^2} Q(\mu)$ is positive definite.

Let Υ admit the following singular value decomposition

$$\Upsilon = USV^T, \tag{61}$$

where U and V^T are two unitary matrices and S is a diagonal matrix with the following partitions

$$S = \begin{bmatrix} \sigma_1 I_{v_1} & 0 & \cdots & 0 \\ 0 & \sigma_2 I_{v_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_h I_{v_h} \end{bmatrix}, \quad V^T = \begin{bmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_h^T \end{bmatrix}, \tag{62}$$

in which $\sigma_1 > \sigma_2 > \cdots > \sigma_h > 0$ are the singular values of Υ and $v_i, i = 1, 2, \ldots, h$, satisfying $\Sigma_{i=1}^h v_i = mn$, are some positive scalars representing the multiplicities of the corresponding singular values $\sigma_i, i = 1, 2, \ldots, h$. Hence $\sigma_{\max}(\Upsilon) = \sigma_1$ and $\sigma_{\min}(\Upsilon) = \sigma_h$.

Theorem 4. *Let Υ be of full column rank and $X(k), k = 1, 2, \ldots$ be iteratively given by (18) with initial condition $X(0)$ and step size μ satisfying (20). Assume that $X_0 \neq X_\infty$. Then there exists a unique optimal value*

$$\mu_{\mathrm{opt}}^{\mathrm{ess}} = \mu_{\mathrm{opt}}^{\mathrm{ess}}(X_0) \in \begin{cases} (0, \mu_{\max}) & V_1^T x_0 \neq 0 \\ (0, \mu_{\max}] & V_1^T x_0 = 0 \end{cases} \tag{63}$$

such that the index function $J_e(\mu, X_0)$ defined as (54) is minimized. Furthermore, the optimal value μ_{opt}^{ess} has the following estimation

$$\frac{1}{\sigma_{max}^2(\Upsilon)} \leqslant \mu_{opt}^{ess} \leqslant \frac{1}{\sigma_{min}^2(\Upsilon)}. \tag{64}$$

Proof. Denote $x_0 = \mathrm{vec}(X_0 - X_\infty)$. Since μ satisfies (20), it follows from Lemma 7 that $\frac{d^2}{d\mu^2}Q(\mu)$ is positive definite. As $X_0 \neq X_\infty \Rightarrow x_0 \neq 0$, we conclude that

$$\frac{d^2 J_e(\mu, X_0)}{d\mu^2} = x_0^T \frac{d^2}{d\mu^2} Q(\mu) x_0 > 0.$$

That is to say, for arbitrary initial condition $X_0 \neq X_\infty$, the index function $J_e(\mu, X_0)$ is a convex function of μ with $\mu \in (0, \mu_{max})$. Therefore, there exists one, and only one, optimal value $\mu_{opt}^{ess} = \mu_{opt}^{ess}(X_0) \in [0, \mu_{max}]$ such that $J_e(\mu, X_0)$ is minimized.

Since $x_0 \neq 0$. It follows from (57) that

$$\lim_{\mu \to 0^+} J_e(\mu, X_0) = \sum_{k=0}^{\infty} x_0^T x_0 = \infty. \tag{65}$$

On the other hand, by using (61) and (62), we have

$$\mu_{max} \Upsilon^T \Upsilon - I = V \widehat{S} V^T,$$

with

$$\widehat{S} = \begin{bmatrix} I_{v_1} & 0 & \cdots & 0 \\ 0 & \widehat{\sigma}_2 I_{v_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\sigma}_h I_{v_h} \end{bmatrix},$$

and $\widehat{\sigma}_i = \frac{2\sigma_i}{\sigma_1} - 1, i = 2, 3, \ldots, h$. It is easy to verify that $|\widehat{\sigma}_i| < 1, i = 2, 3, \ldots, h$. Therefore, we can obtain

$$J_e(\mu, X_0) = \sum_{k=0}^{\infty} x_0^T V \widehat{S}^{2k} V^T x_0 = \sum_{k=0}^{\infty} x_0^T V_1 V_1^T x_0 + \sum_{i=2}^{h} \frac{x_0^T V_i V_i^T x_0}{1 - \widehat{\sigma}_i^2}. \tag{66}$$

Now we consider the following two cases.

Case 1: $V_1^T x_0 \neq 0$. Then it follows from (66) and $V_1^T x_0 \neq 0$ that

$$\lim_{\mu \to \mu_{max}^-} J_e(\mu, X_0) = \infty. \tag{67}$$

As $J_e(\mu, X_0)$ is a convex function with respect to μ, Eqs. (65) and (67) clearly imply that $\mu_{opt}^{ess} \in (0, \mu_{max})$ which is the top expression in (63).

Case 2: $V_1^T x_0 = 0$. In this case, in view of (66), we have

$$\lim_{\mu \to \mu_{max}^-} J_e(\mu, X_0) = \sum_{i=2}^{h} \frac{x_0^T V_i V_i^T x_0}{1 - \widehat{\sigma}_i^2} < \infty. \tag{68}$$

The above equation and (54) imply that $\lim_{k \to \infty} \mathscr{X}(k) = 0$, i.e., the iteration in (18) converges to X_∞ Eq. (68) also indicates that $J_e(\mu, X_0)$ may be minimized when $\mu = \mu_{max}$. Therefore, we have the second expression in (63).

We next show (64). It follows from (59) that $\frac{d}{d\mu}Q(\mu) < 0$ if and only if $I - \mu \Upsilon^T \Upsilon > 0$ which holds if and only if $\mu < 1/\sigma_{max}^2(\Upsilon)$. That is to say, if $\mu < 1/\sigma_{max}^2(\Upsilon)$, then for arbitrary initial condition X_0, $J_e(\mu, X_0)$ decreases as μ increases. Therefore, the inequality in the left hand side of (64) should be satisfied. Similarly, $\frac{d}{d\mu}Q(\mu) > 0$ if and only if $I - \mu \Upsilon^T \Upsilon < 0$ which is equivalent to $\mu > 1/\sigma_{min}^2(\Upsilon)$. That is to say, if $\mu > 1/\sigma_{min}^2(\Upsilon)$, then for arbitrary initial condition $X_0, J_e(\mu, X_0)$ increases as μ increases. Therefore, we have $\mu_{opt}^{ess} < 1/\sigma_{min}^2(\Upsilon)$ which is just the inequality in the right hand side of (64). The proof is completed. □

Remark 5. Notice that $V_1^T x_0 = 0$ can be equivalently written as

$$x_0 \in \mathrm{Null}\left(V_1^T\right) = \mathrm{Image}\left(\begin{bmatrix} V_2^T \\ \vdots \\ V_h^T \end{bmatrix}\right),$$

where $V_i^T \in \mathbf{R}^{v_i \times mn}, i = 1, 2, \ldots, h$. Since v_1 is a small integer in general, it follows that $V_1^T x_0 \neq 0$ is satisfied for "almost all" initial condition X_0. Therefore $\mu_{opt}^{ess} \in (0, \mu_{max})$ for "almost all" initial condition X_0.

Remark 6. Though the existence of the optimal step size $\mu_{\text{opt}}^{\text{ess}}$ is guaranteed according to Theorem 4, it is hard to obtain such optimal value in practice as it is dependent on the initial condition X_0 and the exact solution X_∞. Note that we have $\sigma_{\max}^2(\Upsilon) \gg \sigma_{\min}^2(\Upsilon)$ in general, the inequality (64) can be simplified as

$$\frac{1}{\sigma_{\max}^2(\Upsilon)} \leqslant \mu_{\text{opt}}^{\text{ess}} < \frac{2}{\sigma_{\max}^2(\Upsilon)} = \mu_{\max}.$$

Our experiences from simulations show that $\mu_{\text{opt}}^{\text{ess}}$ chosen as $\frac{\alpha}{\sigma_{\max}^2(\Upsilon)}$, $1 \leqslant \alpha \leqslant 1.7$ can lead to good convergence performances (see one of the examples in Section 5).

Remark 7. As pointed out by the reviewer, since $\mu_{\text{opt}}^{\text{ess}}$ depends on the initial matrix X_0, its value cannot be used to measure the convergence of the algorithm generally. However, the convergence of $J_e(\mu, X_0)$ defined in (54) and the iteration (18) is consistent, namely, if $J_e(\mu_1, X_0) < J_e(\mu_2, X_0)$, then the iteration (18) with $\mu = \mu_1$ converges faster than that with $\mu = \mu_2$. Moreover, determining μ also involves the singular value of large matrix Υ. From this point of view, the result in Theorem 4 has more theoretical meaning than practical one.

An illustrative example for the developed theory is Example 3 given in Section 5.

5. Some numerical examples

In this section, we use several examples to validate the effectiveness of the developed results.

Example 1. Consider the following linear matrix equation

$$AX + X^{\mathrm{T}}B = C, \tag{69}$$

where A and B are, respectively, given by

$$A = \begin{bmatrix} 0.9268 & 0.3739 & 0.5080 \\ 0.3157 & 0.1542 & 0.4521 \\ 0.3271 & 0.3044 & 0.3816 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1834 & 0.5337 & 0.9326 \\ 0.1499 & 0.8615 & 0.0326 \\ 0.9278 & 0.1393 & 0.0036 \end{bmatrix}.$$

The matrix C is obtained by substituting a specified X_∞ into (69). In particular, we set

$$X_\infty = \begin{bmatrix} 1.000 & 1.000 & 1.000 \\ -1.000 & -1.000 & -1.000 \\ -1.000 & 1.000 & 1.000 \end{bmatrix}, \quad C = \begin{bmatrix} -0.8494 & 0.5938 & 2.7051 \\ 0.6707 & 0.4251 & 1.8256 \\ 0.9022 & 1.9388 & 1.9819 \end{bmatrix}.$$

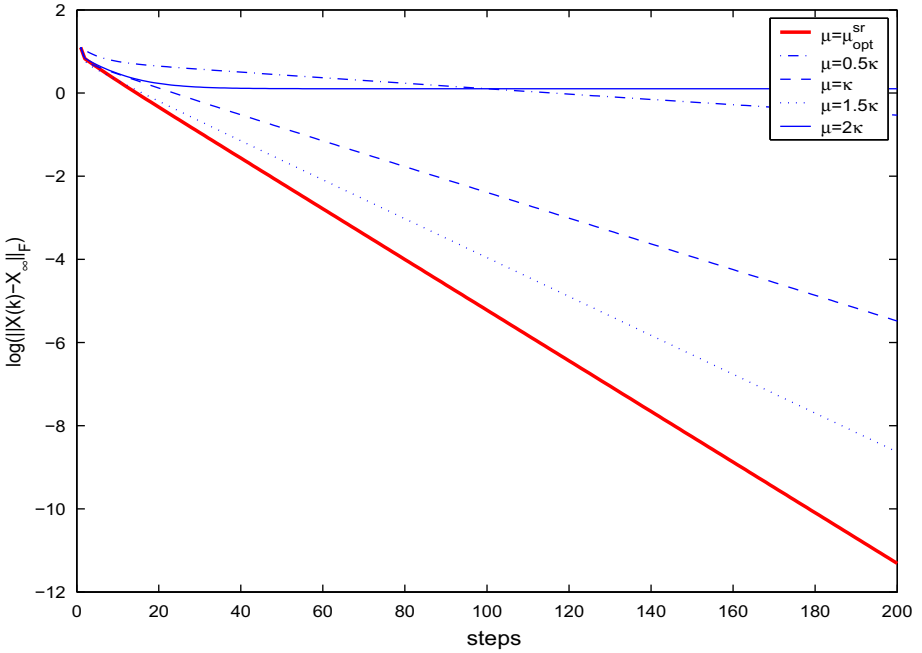

Fig. 1. Convergence performances of the algorithm (18) for Example 1 with different step size μ. In the figure, $\kappa = \frac{1}{\sigma_{\max}^2(\Upsilon)}$.

Shown in Fig. 1 are the convergence performances of the algorithm with different step size μ and the same initial condition $X(0) = 0$. It is clear to see that the convergence performance associated with $\mu = \mu_{\text{opt}}^{\text{sr}}$ is better than that with the other step sizes, which coincides with Theorem 3.

Example 2. Still consider matrix Eq. (69) with the following coefficient matrices A, B, C and the unique solution X_∞:

$$A = \begin{bmatrix} 0.1476 & 0.6364 & 0.2561 \\ 0.8492 & 0.5904 & 0.6943 \\ 0.9883 & 0.1258 & 0.9416 \end{bmatrix}, \quad B = \begin{bmatrix} 0.4434 & 0.2236 & 0.3336 \\ 0.4588 & 0.1729 & 0.0788 \\ 0.2192 & 0.8514 & 0.0130 \end{bmatrix},$$

$$X_\infty = \begin{bmatrix} 1.000 & 1.000 & 1.000 \\ -1.000 & -1.000 & -1.000 \\ -1.000 & 1.000 & 1.000 \end{bmatrix}, \quad C = \begin{bmatrix} -0.9795 & -1.0333 & 1.2819 \\ -0.2316 & 1.8553 & 2.4017 \\ 1.0424 & 3.0520 & 2.4810 \end{bmatrix}.$$

The computing results are given in Fig. 2 with different step sizes and the same initial condition $X(0) = 0$. We note that the convergence performance associated with $\mu_{\text{opt}}^{\text{sr}}$ is in fact not the best one. On the other hand, the optimal step size is about $\frac{1.7}{\sigma_{\max}^2(\Upsilon)}$ observed from the figure.

Example 3. Consider a linear matrix equation in the form of

$$AXB + CXD + EX^{\mathrm{T}}F = G. \tag{70}$$

The coefficient matrices A, B, C, D, E, F and G and the unique solution X_∞ are, respectively, given by

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

$$E = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad G = \begin{bmatrix} 9 & -5 \\ -2 & 12 \end{bmatrix}, \quad X_\infty = \begin{bmatrix} 1 & 1 \\ -1 & 2 \end{bmatrix}.$$

In the above, the matrices A, B, C and D are borrowed from [9], X_∞ is specified and G is obtained by substituting A, B, C, D, E, F and X_∞ into Eq. (70). We also compare our method with the one proposed in [9].

Shown in Fig. 3 is the convergence performance comparison for different step size μ and the algorithm proposed in [9]. With the given data, $\mu_{\max} = 0.0539$, $\mu_{\text{opt}}^{\text{sr}} = 0.0499$ and $\mu_{\text{opt}}^{\text{ess}} = 0.0523$ which is obtained by running a searching program for the initial condition $X(0) = 0$. We note that $\mu_{\text{opt}}^{\text{ess}} = \frac{1.94}{\sigma_{\max}^2(\Upsilon)}$. It is clear to see that the convergence rate with $\mu = \mu_{\text{opt}}^{\text{ess}}$ is larger than that with $\mu = \mu_{\text{opt}}^{\text{sr}}$ and the algorithm proposed in [9]. Finally, the CPU time of our method with $\mu_{\text{opt}}^{\text{ess}} = \frac{1.94}{\sigma_{\max}^2(\Upsilon)}$ is 0.83 s

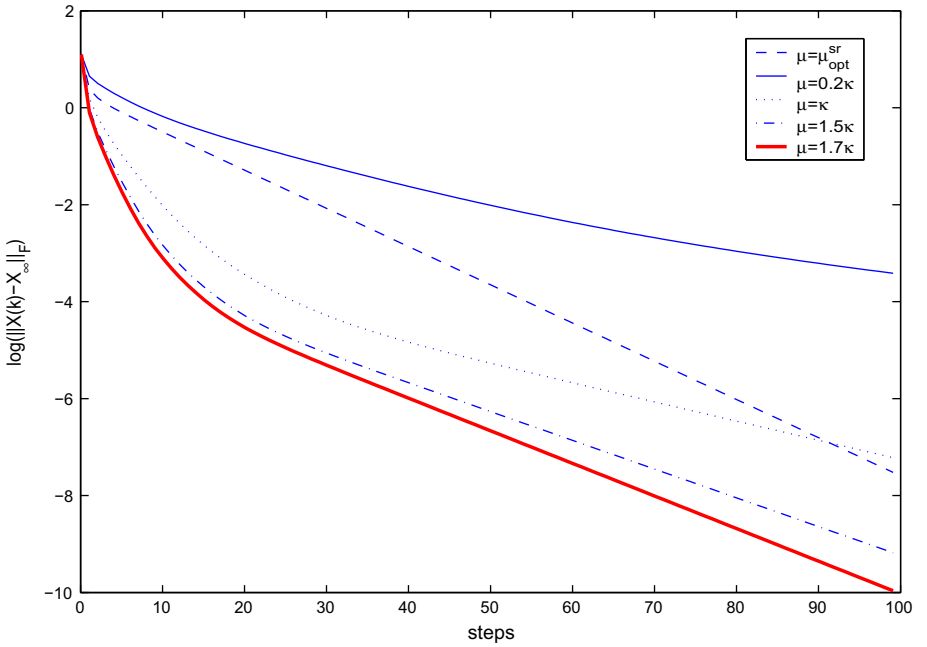

Fig. 2. Convergence performances of the algorithm (18) for Example 2 with different step size μ. In the figure, $\kappa = \frac{1}{\sigma_{\max}^2(\Upsilon)}$.

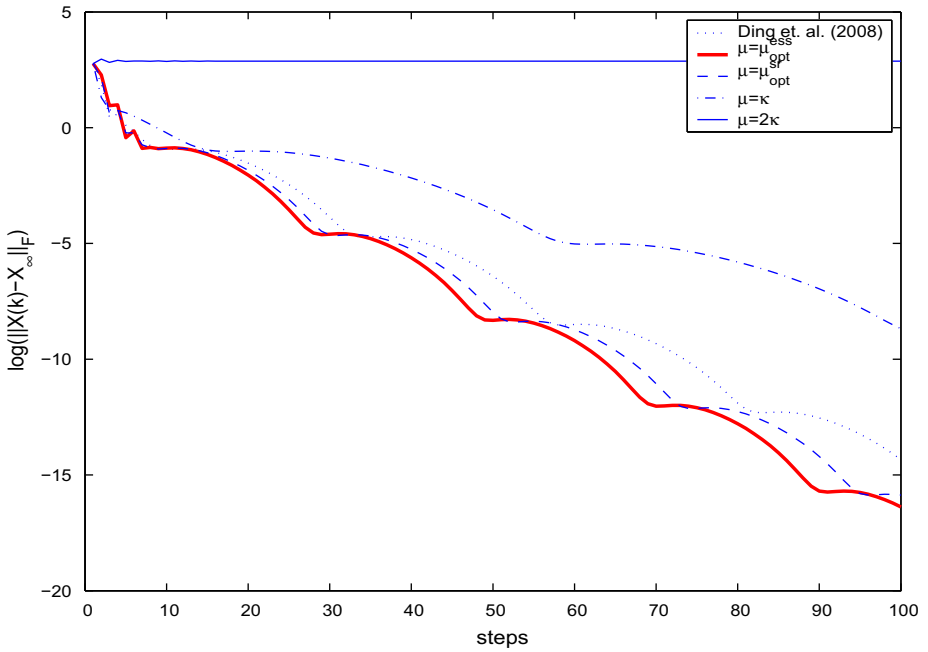

Fig. 3. Convergence performances of the algorithm (18) for Example 3 with different step size μ. In the figure, $\kappa = \frac{1}{\sigma^2_{\max}(\Upsilon)}$.

while the CPU time of the method in [9] is 0.92 s (both algorithms are run in an Intel 1.73-GHz computer and are stopped if $\log(\|X(k) - X_\infty\|_2) \leqslant -16$).

Example 4. Consider the following least squares problem

$$\min_X \|AXB + CX^{\mathrm{T}}D - E\|_{\mathrm{F}}, \tag{71}$$

where $A, C \in \mathbf{R}^{20\times 20}, B, D \in \mathbf{R}^{20\times 30}$ and $E \in \mathbf{R}^{20\times 30}$ are known and $X \in \mathbf{R}^{20\times 20}$ is to be determined. Similar to [7], the matrices A, B, C, D and E are generated in Matlab by using the following code:

```
rand('state', 0);
A = [triu(rand(20, 20), 1) + diag(10 + diag(rand(20)))];
B = [triu(rand(20, 20), 1) + diag(10 + diag(rand(20))), 0.1 * rand(20, 10)];
C = [triu(rand(20, 20), 1) + diag(10 + diag(rand(20)))];
D = [triu(rand(20, 20), 1) + diag(10 + diag(rand(20))), 0.1 * rand(20, 10)];
E = 0.1 * rand(20, 30);
```

Notice that $\Gamma \in \mathbf{R}^{600\times 400}$ is of full column rank. Therefore, we can use the iteration in (18) and (19) to produce the unique least squares solution to problem (71). Here the step size is chosen as $\mu = \frac{2}{\nu_2^2}$ where ν_2 is given by (32) and the initial condition is chosen as $X(0) = 0$. The exact solution to the least squares problem (71) is computed as $\mathrm{vec}(X_\infty) = (\Upsilon^{\mathrm{T}}\Upsilon)^{-1}\Upsilon^{\mathrm{T}}\mathrm{vec}(E)$. Using such direct Kronecker product approach to solve this high-dimensional equation, the computational time is 80.90 s on an Intel 1.73-GHz computer. The computational time is 2.6 s on the same computer by using our algorithm. This result clearly implies the effectiveness of the proposed approach.

6. Concluding remarks

This paper is concerned with numerical solutions to the minimal norm least squares solution to general linear matrix equations. Two iterative algorithms are proposed. The first one is established by using the gradient based optimization principle while the other one can be viewed as its dual form. Necessary and sufficient conditions are given for the step sizes such that the algorithms converge for arbitrary initial conditions. Also, a simple sufficient condition that is easy to test is also presented. To ensure a good convergence performance of the proposed algorithms, two methods are proposed to select the optimal step size such that the convergence rate of the algorithms is maximized. Between these two methods, the first one is established based on the criterion of minimizing the spectral radius of the iteration matrix while the second one is obtained

by minimizing the square sum of the F-norm of the error matrices associated with the algorithm. Several numerical examples are given to illustrate the effectiveness of the proposed method.

The proposed method is easily extended to solve coupled Sylvester matrix equations [6], periodic Lyapunov matrix equations and periodic Sylvester matrix equations [2]. Upon the present results, two interesting problems listed below are required to be considered in the future.

1. How to use the preconditioning method in CG algorithm to improve the convergence performance of the proposed algorithm when the condition number of Υ (see (7)) is large.
2. How to generate this technique to the field of solving nonlinear matrix equations such as algebraic Riccati equations.

Acknowledgment

The authors would like to thank the reviewer for many constructive suggestions. Especially, they would like to thank her/him for providing the minimal norm least squares Problem 1 and current simple proof of Theorem 1. The reviewer's comments also inspirit the authors to present Theorem 2 in this paper.

Appendix A. *Proof of Lemma 7*

Note that $\Upsilon^T\Upsilon, (\Upsilon^T\Upsilon)^{-1}, 2I-\mu\Upsilon^T\Upsilon, (2I-\mu\Upsilon^T\Upsilon)^{-1}, I-\mu\Upsilon^T\Upsilon$ and $(I-\mu\Upsilon^T\Upsilon)^{-1}$ are all commutable. Let $A(t)$ be a matrix function of the scalar t. Then

$$\frac{dA^{-1}(t)}{dt} = -A^{-1}(t)\frac{dA(t)}{dt}A^{-1}(t). \tag{72}$$

With the help of this formulation, we get

$$\begin{aligned}\frac{dQ(\mu)}{d\mu} &= -(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}\frac{d(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))}{d\mu}(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}\\ &= -(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}(\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon)+\mu\Upsilon^T\Upsilon(-\Upsilon^T\Upsilon))(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}\\ &= -2(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}X(I-\mu\Upsilon^T\Upsilon)(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-1}\\ &= -2(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}(I-\mu\Upsilon^T\Upsilon)\Upsilon^T\Upsilon,\end{aligned} \tag{73}$$

which is (59). Using formulation (72) again, we have

$$\begin{aligned}&\frac{d}{d\mu}(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}\\ &\quad= -(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}\frac{d(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^2}{d\mu}(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}\\ &\quad= -(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}(4\mu\Upsilon^T\Upsilon^2(I-\mu\Upsilon^T\Upsilon)(2I-\mu\Upsilon^T\Upsilon))(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}\\ &\quad= -4\mu^{-1}(I-\mu\Upsilon^T\Upsilon)(2I-\mu\Upsilon^T\Upsilon)^{-1}(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}.\end{aligned} \tag{74}$$

It follows from (73) that

$$\begin{aligned}\frac{d^2Q(\mu)}{d\mu^2} &= -2\frac{d(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}}{d\mu}(I-\mu\Upsilon^T\Upsilon)\Upsilon^T\Upsilon + 2(\mu\Upsilon^T\Upsilon(2I-\mu X\Upsilon^T\Upsilon))^{-2}(\Upsilon^T\Upsilon)^2\\ &= -2\frac{d(\mu\Upsilon^T\Upsilon(2I-\mu\Upsilon^T\Upsilon))^{-2}}{d\mu}(I-\mu\Upsilon^T\Upsilon)\Upsilon^T\Upsilon + 2(\mu(2I-\mu\Upsilon^T\Upsilon))^{-2}.\end{aligned}$$

Substituting (74) into the above equation and simplifying, gives (60). Obviously, $\frac{d^2}{d\mu^2}Q(\mu)$ is symmetric and positive definite. This completes the proof.

References

[1] S.P. Bhattacharyya, E. De Souza, Pole assignment via Sylvester's equation, Systems and Control Letters 1 (1972) 261–263.
[2] R. Byers, N. Rhee, Cyclic Schur and Hessenberg–Schur numerical methods for solving periodic Lyapunov and Sylvester equations. Technical Report, Dept. of Mathematics, Univ. of Missouri at Kansas City, 1995.
[3] R. Byers, Solving the algebraic Riccati equation with the matrix sign function, Linear Algebra and Its Applications 85 (1987) 267–279.
[4] J.L. Chen, X.H. Chen, Special Matrices, Tsinghua University Press, 2002 (in Chinese).
[5] E. Desouza, S.P. Bhattacharyya, Controllability, observability and the solution of $AX - XB = C$, Linear Algebra and Its Applications 39 (1981) 167–188.
[6] F. Ding, T. Chen, Iterative least squares solutions of coupled Sylvester matrix equations, Systems and Control Letters 54 (2) (2005) 95–107.
[7] F. Ding, T. Chen, Gradient based iterative algorithms for solving a class of matrix equations, IEEE Transactions on Automatic Control 50 (8) (2005) 1216–1221.
[8] F. Ding, T. Chen, On iterative solutions of general coupled matrix equations, SIAM Journal on Control and Optimization 44 (6) (2006) 2269–2284.

[9] F. Ding, P.X. Liu, J. Ding, Iterative solutions of the generalized Sylvester matrix equations by using the hierarchical identification principle, Applied Mathematics and Computation 197 (1) (2008) 41–50.
[10] J.C. Doyle, K. Glover, P.P. Khargonekar, B.A. Francis, State-space solutions to standard $\mathscr{H}_2$ and $\mathscr{H}_\infty$ control problems, IEEE Transactions on Automatic Control 34 (8) (1989) 831–847.
[11] G.R. Duan, Solutions to matrix equation $AV + BW = VF$ and their application to eigenstructure assignment in linear systems, IEEE Transactions on Automatic Control 38 (2) (1993) 276–280.
[12] G.R. Duan, On the solution to Sylvester matrix equation $AV + BW = EVF$, IEEE Transactions on Automatic Control 41 (4) (1996) 612–614.
[13] G.X. Huang, F. Yin, K. Guo, An iterative method for the skew-symmetric solution and the optimal approximate solution of the matrix equation $AXB = C$, Journal of Computational and Applied Mathematics 212 (2) (2008) 231–244.
[14] T. Kailath, Linear Systems, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[15] Z. Li, Y. Wang, Iterative algorithm for minimal norm least squares solution to general linear matrix equations, International Journal of Computer Mathematics, doi:10.1080/00207160802684459.
[16] Adem Kılı cman, Zeyad Abdel Aziz Al Zhour, Vector least-squares solutions for coupled singular matrix equations, Journal of Computational and Applied Mathematics 206 (2007) 1051–1069.
[17] F. Piao, Q. Zhang, Z. Wang, The solution to matrix equation $AX + X^{\mathrm{T}}C = B$, Journal of the Franklin Institute 344 (8) (2007) 1056–1062.
[18] X.Y. Peng, X.Y. Hu, L. Zhang, The reflexive and anti-reflexive solutions of the matrix equation $A^{\mathrm{H}}XB = C$, Journal of Computational and Applied Mathematics 200 (2) (2007) 749–760.
[19] Z.H. Peng, X.Y. Hu, L. Zhang, An effective algorithm for the least-squares reflexive solution of the matrix equation $A_1XB_1 = C_1, A_2XB_2 = C_2$, Applied Mathematics and Computation 181 (2006) 988–999.
[20] Y.Y. Qiu, Z.Y. Zhang, J.F. Lu, Matrix iterative solutions to the least squares problem of $BXA^{\mathrm{T}} = F$ with some linear constraints, Applied Mathematics and Computation 185 (2007) 284–300.
[21] M.H. Wang, X.H. Cheng, M.H. Wei, Iterative algorithms for solving the matrix equation $AXB + CX^{\mathrm{T}}D = E$, Applied Mathematics and Computation 187 (2) (2007) 622–629.
[22] Q. Wang, J. Lam, Y. Wei, T. Chen, Iterative solutions of coupled discrete Markovian jump Lyapunov equations, Computers and Mathematics with Applications 55 (4) (2008) 843–850.
[23] K. Zhou, J. Doyle, K. Glover, Robust and optimal control, Prentice-Hall, 1996.