Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(8): 18392–18407.
DOI: 10.3934/math.2023935
Received: 15 February 2023
Revised: 21 April 2023
Accepted: 11 May 2023
Published: 30 May 2023

Research article

On the nonlinear matrix equation $X^s + A^H F(X)A = Q$

Yajun Xie[1], **Changfeng Ma**[1,*] **and Qingqing Zheng**[2]

[1] School of Big Data & Key Laboratory of Digital Technology and Intelligent Computing, Fuzhou University of International Studies and Trade, Fuzhou 350202, China

[2] School of Mathematics and Statistics, Fujian Normal University, Fuzhou 350007, China

* **Correspondence:** Email: mcf@fzfu.edu.cn; Tel: +8613763827962.

Abstract: Nonlinear matrix equation often arises in control theory, statistics, dynamic programming, ladder networks, and so on, so it has widely applied background. In this paper, the nonlinear matrix equation $X^s + A^H F(X)A = Q$ are discussed, where operator F are defined in the set of all $n \times n$ positive semi-definite matrices, and Q is a positive definite matrix. Sufficient conditions for the existence and uniqueness of a positive semi-definite solution are derived based on some fixed point theorems. It is shown that under suitable conditions an iteration method converges to a positive semi-definite solution. Moreover, we consider the perturbation analysis for the solution of this class of nonlinear matrix equations, and obtain a perturbation bound of the solution. Finally, we give several examples to show how this works in particular cases, and some numerical results to specify the rationality of the results we have obtain.

Keywords: nonlinear matrix equation; positive semi-definite solution; fixed point theorem; perturbation analysis
Mathematics Subject Classification: 15A24, 65H10

1. Introduction

Let $P(n)$ denote the set of $n \times n$ positive semi-definite matrices, and $M(n)$ denote the set of all $n \times n$ matrices. In this paper we consider the following class of nonlinear matrix equations

$$X^s + A^H F(X)A = Q, \tag{1.1}$$

where $F(\cdot) : P(n) \rightarrow M(n)$ is continuous, i.e., F transforms positive definite matrices into non-negative definite ones. $s \geq 1$, $A, Q \in M(n)$, and Q is a positive definite matrix. Note that $\overline{X}$ is a solution of Eq (1.1) if and only if it is a fixed point of the map

$$G(X) = (Q - A^H F(X)A)^{\frac{1}{s}}, \tag{1.2}$$

so the map $G(\cdot)$ plays an important role throughout this paper.

Several authors have considered such a nonlinear matrix equation problem. In [1], Qingchun Li and Panpan Liu considered the Eq (1.1), in the case that $s \geq 1$ and map $F(\cdot)$ is also defined in $P(n)$. The authors in [2] discussed an iteration method for the Eq (1.1) in the case $s = 1$ and with the condition $A^H F(Q)A < Q$, i.e., $G(Q) > O$. Moreover when $s = 1$, the perturbation theory for the Eq (1.1) can be found in [3]. Other authors discussed this equation for particular choices of the map $F(\cdot)$. For example, Beatrice Meini [4] established and proved theorems for the necessary and sufficient conditions of existence of a positive definite solution of the equation as $F(X) = \pm X^{-1}$, where Q is a positive definite matrix. In [5, 6], the case $F(X) = \pm X^{-2}$ is discussed. And the case $F(X) = \pm X^{-m}, m \in \{3, 4, \cdots\}$ is treated in [7, 8]. Xuefeng Duan and Anping Liao [9] considered the situation where $A^H F(X)A = -\sum_{i=1}^{m} A_i^H X^{\delta_i} A_i, (0 < |\delta_i| < 1)$ when the A_i $(i = 1, 2, 3, \cdots)$ denoted in [9] are equal with each other, they proved that the nonlinear matrix equation always has an unique positive definite solution, and multi-step stationary iterative method is proposed to work out the unique positive definite solution. In addition, Xuefeng Duan and Anping Liao solved a conjecture which is proposed in [10], and they obtained a conclusion that the nonlinear matrix equation $X^s - A^{\top} X^{-t} A = I$ does not always has an unique positive definite solution unless when some conditions are satisfied.

Our main contributions in this paper include the following aspects. Firstly, we give the existent interval of the solution of Eq (1.1) in the case that F maps into $P(n)$ or F maps into $-P(n)$, respectively. As compared to previous results, the results from which we have obtained have a better estimates for the solution $\overline{X}$ in some certain. Secondly, we also give the perturbation analysis of the solution of Eq (1.1) when $F : P(n) \to M(n)$, where the range of values of F is expanded compared with previous papers. Thirdly, we will promote and replenish the results in [1].

The paper is organized as follows. In Section 2, we shall derive some necessary and sufficient conditions for the existence of a solution for the nonlinear Eq (1.1) based on the the Schauder's fixed point theorem. And we also discuss the existence scope of the solution whenever the equation is solvable. Section 3 discusses the uniqueness of a solution. In Section 4, we analysis perturbation theory of the solution for Eq (1.1) and the perturbation bounds for the positive semi-definite solutions of these equations are given. Finally, Section 5 illustrates the correctness of the results which we have obtained with some numerical experiments and contains some examples on the matrix Eq (1.1) as well as the results in the preceding sections.

The following notations will be used throughout the rest of this paper. For $A \in M(n)$, $\lambda_1(A)$ and $\lambda_n(A)$ stand for the maximal and minimal eigenvalue of matrix A, respectively. A^H is the conjugate transpose of the matrix A, A^{-H} is the inversion of A^H. I is the identity matrix, and O is the 0-matrix. $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the l_2 norm and the Frobenius norm, respectively. With $A \geq O$ $(A > O)$ we denote that matrix A is positive semi-definite (positive definite). As a different notation for $A - B \geq O$ $(A - B > O)$, we will write $A \geq B (A > B)$, This induces a partial ordering on the Hermitian matrices. When we say that a Hermitian matrix is the smallest (largest) in some set, then this is always meant with respect to the partial ordering induced in this way. Further, the sets $[A, B]$ and (A, B) are defined by

$$[A, B] = \{X | A \leq X \leq B\}, \quad (A, B) = \{X | A < X < B\},$$

whereas $L_{A,B}$ denotes the line segment joining A and B, i.e., $L_{A,B} = \{tA + (1 - t)B | t \in [0, 1]\}$. $G(G(X))$ is denoted by $G^2(X)$, and the jth iterate of G on X is denoted by $G^j(X)$. For $B = (b_1, b_2, \cdots, b_n) = (b_{ij})$

and a matrix C, $B \otimes C = (b_{ij}C)$ is a Kronecker product and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A) = (a_1^\top, a_2^\top, \cdots, a_n^\top)^\top$. In order to develop the paper , we need that

$$\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}(X) \quad \text{and} \quad \|\mathrm{vec}(X)\|_2 = \|X\|_F,$$

where A, X and B are $n \times n$ complex matrix.

2. Existence and properties of solutions

In this section, we discuss the sufficient conditions for the existence of a positive definite or positive semi-definite solution of Eq (1.1) based on the Schauder's Fixed Point Theorem [11], and some properties of solutions are given.

Lemma 2.1. *(see Parodi [12]). If $A > B > O$(or $A \geq B > O$), then $A^\alpha > B^\alpha$ (or $A^\alpha \geq B^\alpha > O$) for all $\alpha \in (0, 1]$, and $A^\alpha < B^\alpha$(or $O < A^\alpha \leq B^\alpha$) for all $\alpha \in [-1, 0)$.*

Theorem 2.1. *Let $F : P(n) \to M(n)$ be continuous, If Eq (1.1) is solvable and $\overline{X}$ is a positive semi-definite solution of Eq (1.1), then the following results hold true.*
(i) If $F : P(n) \to P(n)$, then $O \leq A^H F(\overline{X})A \leq Q$, and $O \leq \overline{X} \leq Q^{\frac{1}{s}}$,
(ii) If $F : P(n) \to -P(n)$, then $\overline{X} \geq Q^{\frac{1}{s}}$.

Proof. (i) Because of $\overline{X} \in P(n)$, we obtain that $F(\overline{X}) \geq O$. This implies that $A^H F(\overline{X})A \geq O$. Note that $\overline{X} \geq O, \overline{X}^s + A^H F(\overline{X})A = Q$, we have $A^H F(\overline{X})A \leq Q$, and $\overline{X}^s \leq Q$. From Lemma 2.1, we have $\overline{X} \leq Q^{\frac{1}{s}}$. This proves statement (i).

(ii) Because F maps into $-P(n)$, we know that $F(\overline{X}) \leq O$, which implies that $A^H F(\overline{X})A \leq O$. So we have

$$\overline{X}^s = Q - A^H F(\overline{X})A \geq Q.$$

According to Lemma 2.1, we obtain $\overline{X} \geq Q^{\frac{1}{s}}$, and this proves the second part of the theorem. □

Remark 2.1. *The positive semi-definite solution of the nonlinear matrix Eq (1.1) are not always exist, if the solution is existent, according to Theorem 2.1, we can obtain the existent interval of the solution in the case that F maps into $P(n)$ or F maps into $-P(n)$, respectively.*

Sufficient conditions for the existence of a positive semi-definite solution are derived in the following discussion.

Theorem 2.2. *Let $F : P(n) \to P(n)$ be continuous on $[O, Q^{\frac{1}{s}}]$, if inequality $A^H F(X)A \leq Q$ are satisfied for all $X \in [O, Q^{\frac{1}{s}}]$, then Eq (1.1) has a positive semi-definite solution in $[O, Q^{\frac{1}{s}}]$.*

Proof. From $F : P(n) \to P(n)$, we obtain $F(X) \geq O$, which implies that for all $X \in [O, Q^{\frac{1}{s}}]$ must have

$$O \leq G(X) = (Q - A^H F(X)A)^{\frac{1}{s}} \leq Q^{\frac{1}{s}}.$$

So we know that G maps $[O, Q^{\frac{1}{s}}]$ into $[O, Q^{\frac{1}{s}}]$. Moreover, because of the continuity of map F, we obtain that G is continuous. Obviously, interval $[O, Q^{\frac{1}{s}}]$ is a compact convex set, so according to the Schauder's Fixed Point Theorem, Eq (1.1) has a positive semi-definite solution in $[O, Q^{\frac{1}{s}}]$. □

Theorem 2.3. *Let $F : P(n) \to -P(n)$ be continuous on $[Q^{\frac{1}{s}}, +\infty)$, if there exists $B \geq Q$ such that*

$$Q - B \leq A^H F(X)A \leq O, \tag{2.1}$$

for all $X \in [Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$, then Eq (1.1) has a positive definite solution in $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$. Moreover, if the Eq (2.1) is satisfied for every $X \geq Q^{\frac{1}{s}}$, then all solutions of Eq (1.1) are in $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$.

Proof. From $F : P(n) \to -P(n)$, we obtain $F(X) \leq O$ for all $X \in [Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$. It follows from that $G(X) = (Q - A^H F(X)A)^{\frac{1}{s}} \geq Q^{\frac{1}{s}}$. Note that $Q - B \leq A^H F(X)A \leq O$, we obtain $Q - A^H F(X)A \leq B$, i.e., $G(X) \leq B^{\frac{1}{s}}$. So $G(X) \in [Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$, which implies G maps $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$ into $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$. Also we can prove that $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$ is a compact convex set, so according to the Schauder's Fixed Point Theorem, Eq (1.1) has a fixed point in $[Q^{\frac{1}{s}}, B^{\frac{1}{s}}]$. This fixed point is a positive definite solution solution of Eq (1.1).

Moreover, if the Eq (2.1) is satisfied for every $X \geq Q^{\frac{1}{s}}$, and let $\widehat{X}$ be a arbitrary solution of Eq (1.1), then

$$Q - A^H F(\widehat{X})A \leq B.$$

From Lemma 2.1, we have

$$Q^{\frac{1}{s}} \leq \widehat{X} = (Q - A^H F(\widehat{X})A)^{\frac{1}{s}} \leq B^{\frac{1}{s}}.$$

This completes the proof. □

An operator F is monotone if and only if $F(X) \geq F(Y)$ for all $X \geq Y$, and the operator F is anti-monotone if and only if $F(X) \leq F(Y)$ for all $X \geq Y$. More specific characters for these function can be seen in [13]. For the rest parts of this section we will discuss the map F which is either monotone or anti-monotone.

Theorem 2.4. *Let $F : P(n) \to P(n)$ be continuous, monotone and invertible on $[O, Q^{\frac{1}{s}}]$, and matrix A is invertible. If Eq (1.1) has a positive semi-definite solution $\overline{X}$, then $\overline{X} \in [max\{O, G(Q^{\frac{1}{s}})\}, min\{Q^{\frac{1}{s}}, F^{-1}(A^{-*}QA^{-1})\}]$.*

Proof. From Theory 2.1, we obtain $\overline{X} \in [O, Q^{\frac{1}{s}}]$ and $A^H F(\overline{X})A \in [O, Q]$, i.e., $A^H F(\overline{X})A \leq Q$. So we have $F(\overline{X}) \leq A^{-*}QA^{-1}$. Then we obtain $\overline{X} \leq F^{-1}(A^{-*}QA^{-1})$, which implies $\overline{X} \leq \min\{Q^{\frac{1}{s}}, F^{-1}(A^{-*}QA^{-1})\}$. Moreover, combine $\overline{X} \leq Q^{\frac{1}{s}}$ and the monotonicity of F, we obtain $F(\overline{X}) \leq F(Q^{\frac{1}{s}})$, so $A^H F(\overline{X})A \leq A^H F(Q^{\frac{1}{s}})A$, which implies that $\overline{X} = (Q - A^H F(\overline{X})A)^{\frac{1}{s}} \geq (Q - A^H F(Q^{\frac{1}{s}})A)^{\frac{1}{s}}$, then we have $\overline{X} \geq \max\{O, (Q - A^H F(Q^{\frac{1}{s}})A)^{\frac{1}{s}}\}$.

This complete the proof. □

Remark 2.2. *From Theorem 2.4, we obtain that if $Q > A^H F(Q^{\frac{1}{s}})A$, i.e., $G(Q^{\frac{1}{s}}) > O$, then the solution $\overline{X}$ of Eq (1.1) must be in $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$. On the contrary, if $G(Q^{\frac{1}{s}}) < O$, then the solution $\overline{X}$ must be in $[O, F^{-1}(A^{-*}QA^{-1})]$. Compared with Theorem 2.2, we can narrow the scope of the solution when $G(Q^{\frac{1}{s}}) \neq O$ are satisfied, but for the case that $G(Q) = O$, then the solution can't be narrowed.*

Next we will give the sufficient condition of the existence of the positive definite solution for the Eq (1.1) when $F : P(n) \to P(n)$ be continuous, monotone and invertible on $[O, Q^{\frac{1}{s}}]$.

Theorem 2.5. *Let $F : P(n) \to P(n)$ be continuous, monotone and invertible on $[O, Q^{\frac{1}{s}}]$. If $G(Q^{\frac{1}{s}}) > O$, then Eq (1.1) has a positive definite solution in $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$. If $G(Q^{\frac{1}{s}}) < O, F(Q) < Q$, A is invertible, and $F^{-1}(A^{-*}QA^{-1}) \geq G(X)$ are satisfied for every $X \in [O, F^{-1}(A^{-*}QA^{-1})]$, then Eq (1.1) has a positive semi-definite definite solution in $[O, F^{-1}(A^{-*}QA^{-1})]$.*

Proof. Obviously, the map G is continuous and $G(X) \le Q^{\frac{1}{s}}$. Because F is monotone on $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, we can obtain $F(X) \le F(Q^{\frac{1}{s}})$ when $X \in [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$. This implies $Q - A^H F(X)A \ge Q - A^H F(Q^{\frac{1}{s}})A$, i,e., $G(X) \ge (Q - A^H F(Q^{\frac{1}{s}})A)^{\frac{1}{s}} = G(Q^{\frac{1}{s}})$, so we have G map $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$ into $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$. According to the Schauder's Fixed Point Theorem, G has a fixed point which is a positive definite solution of Eq (1.1) in $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$. This proved the first part of the theorem. One step closer, if $F^{-1}(A^{-*}QA^{-1}) \ge G(X)$ are satisfied for every $X \in [O, F^{-1}(A^{-*}QA^{-1})]$, then we have $A^H F(X)A \le F(Q)$ for $O \le X \le F^{-1}(A^{-*}QA^{-1})$ because of the monotonicity of F. This implies $Q - A^H F(X)A \ge Q - F(Q)$, so we can obtain $G(x) \ge (Q - F(Q))^{\frac{1}{s}} \ge O$, which indicate G maps $[O, F^{-1}(A^{-*}QA^{-1})]$ into $[O, F^{-1}(A^{-*}QA^{-1})]$. Then G has a fixed point which is a positive semi-definite definite solution of Eq (1.1) in $[O, F^{-1}(A^{-*}QA^{-1})]$. The proof is completed. □

Remark 2.3. *The assumption that F be continuous on* $[O, Q^{\frac{1}{s}}]$ *in the conditions of Theorem 2.5 is very important to guarantee existence of a solution of Eq (1.1). That is, if the map F is not continuous, then Eq (1.1) may have no solution in the interval. To show this, we give a simple example in the following.*

Example 2.1. *Consider the scalar case, take* $A = 1, Q = b > 1$, *and let F be a piecewise constant function as follows:*

$$F(X) = \begin{cases} c, \ X > \dfrac{b}{2}, \\ d, \ X \le \dfrac{b}{2}, \end{cases}$$

where $c \ge \frac{b}{2}$ and $d < \frac{b}{2}$, clearly there is no solution to Eq (1.1). So this account for the assumption that F be continuous play an important role to guarantee existence of a solution of Eq (1.1).

For the case that $F : P(n) \to P(n)$ be continuous, monotone, if the nonlinear matrix Eq (1.1) has a positive definite solution, then the existence scope of the solution had also been given in Theorem 2 of [1], where the conclusion is proved though mathematical induction. There the map F has some difference with here proposed in above Theorem, and need to solve two positive solutions of two equations, respectively. Also in [1] some conclusions had been given when the map F is anti-monotone, here we will give the existence scope of the solution when Eq (1.1) is solvable.

Theorem 2.6. *Let* $F : P(n) \to P(n)$ *be continuous, anti-monotone and invertible. If Eq (1.1) has a positive definite solution* $\overline{X}$, *then* $\overline{X} \in (F^{-1}(A^{-*}QA^{-1}), G(Q^{\frac{1}{s}})]$. *Moreover, if* $F^{-1}(A^{-*}QA^{-1}) \le G(X)$ *for all* $[F^{-1}(A^{-*}QA^{-1}), G(Q^{\frac{1}{s}})]$, *then Eq (1.1) has a positive definite solution.*

Proof. See Theorem 1 of [1]. □

3. Uniqueness of a solution and iterative method

In the previous section, some conditions were derived for the existence of a solution of Eq (1.1) and some properties about the solution. But nothing was said about uniqueness, here we will apply the Banach's Fixed Point Theorem [14] to deduce our conclusion.

Lemma 3.1. *For any positive integer s, and* $X, Y \in M(n)$, *we always have*

$$X^s - Y^s = \sum_{i=0}^{s-1} X^i (X - Y) Y^{s-1-i}.$$

Proof. Obviously by mathematical induction method we can proved the conclusion, here we omit the proof process. □

Lemma 3.2. *(cf. Theorem I.*1.8 *in [15]) Let $F : U \to M(n)$ ($U \subset M(n)$ open) be differentiable at any point of U, then*

$$\|F(X) - F(Y)\| \le \sup_{Z \in L_{X,Y}} \|D(F(Z))\| \|X - Y\|,$$

for all $X, Y \in U$.

Lemma 3.3. *(Theorem X.38 in [13]) If the operators X, Y satisfy $X \ge aI$ and $Y \ge aI$ for some positive number a, and $0 < r < 1$, then $\|X^r - Y^r\| \le ra^{r-1}\|X - Y\|$.*

In the following discussion $\phi(n)$ will denote a closed and bounded interval in $P(n)$, i.e., $\phi(n)$ will be of the form $[B; C] = \{X | B \le X \le C\}$, with $B, C \in P(n)$. Let $S_{\phi(n)}$ be the smallest positive value such that $\sup_{Z \in \phi(n)} \|D(F(Z))\| = \max_{Z \in \phi(n)} \|D(F(Z))\| \le S_{\phi(n)}$ holds.

Theorem 3.1. *When s is a positive integer, $G(Q^{\frac{1}{s}}) > O$, and $F : P(n) \to P(n)$ be continuous, monotone and invertible, $\|A^H\|_F \|A\|_F S_{\phi(n)} < s\widetilde{\lambda}^{s-1}$. If Eq (1.1) has a solution $\overline{X}$ in $\phi(n)$ then $\overline{X}$ is the unique solution in $\phi(n)$ which is positive definite. Here $\phi(n) = [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, $\widetilde{\lambda} = \lambda_n(G(Q^{\frac{1}{s}}))$.*

Proof. Assume that $\widetilde{X}$ be a solution of Eq (1.1) and $\widetilde{X} \ne \overline{X}$, from remark 2.1, we obtain that $\widetilde{X}, \overline{X} \in [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, so we have $\overline{X} \ge \widetilde{\lambda} I, \widetilde{X} \ge \widetilde{\lambda} I$. According to Lemma 3.1, we can obtain

$$\begin{aligned}
\|\overline{X}^s - \widetilde{X}^s\|_F &= \|A^H F(\widetilde{X})A - A^H F(\overline{X})A\|_F \\
&= \|A^H (F(\widetilde{X}) - F(\overline{X}))A\|_F \\
&\le \|A^H\|_F \|A\|_F \|F(\overline{X}) - F(\widetilde{X})\|_F \\
&\le \|A^H\|_F \|A\|_F S_{\phi(n)} \|\overline{X} - \widetilde{X}\|_F.
\end{aligned}$$

Also from Lemma 3.1, we have

$$\begin{aligned}
\|\overline{X}^s - \widetilde{X}^s\|_F &= \|\sum_{i=0}^{s-1} \overline{X}^i (\overline{X} - \widetilde{X}) \widetilde{X}^{s-1-i}\|_F \\
&= \|(\sum_{i=0}^{s-1} \widetilde{X}^{s-1-i} \otimes \overline{X}^i) \mathrm{vec}(\overline{X} - \widetilde{X})\|_2 \\
&\ge s\widetilde{\lambda}^{s-1} \|\overline{X} - \widetilde{X}\|_F.
\end{aligned}$$

Combine above two inequality, we have $s\widetilde{\lambda}^{s-1}\|\overline{X} - \widetilde{X}\|_F \le \|A^H\|_F \|A\|_F S_{\phi(n)} \|\overline{X} - \widetilde{X}\|_F$. Because

$G(Q^{\frac{1}{s}}) > O, \widetilde{\lambda} > 0$ is guarantee, so we have $\|\overline{X} - \widetilde{X}\|_F \le \dfrac{\|A^H\|_F \|A\|_F S_{\phi(n)}}{s\widetilde{\lambda}^{s-1}} \|\overline{X} - \widetilde{X}\|_F < \|\overline{X} - \widetilde{X}\|_F$.

This is a contradiction, so $\widetilde{X} = \overline{X}$.

The proof is completed. □

When s is not a positive integer, we can also give a sufficient condition for the existence of the unique solution of Eq (1.1).

Theorem 3.2. *Let $H(Y) = Q - A^H F(Y^{\frac{1}{s}})A$, and $F : P(n) \to P(n)$ be continuous, monotone. $a = \frac{1}{s}S_{[H(Q),Q]}\|A\|^2\lambda_n^{1-s}(H(Q)) < 1$. Then the following results hold.*

(i) Equation (1.1) has and only has a positive semi-definite definite solution $\overline{X}$ in $[(H(Q))^{\frac{1}{s}}, Q^{\frac{1}{s}}] = [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$.

(ii) If we consider the following iterative method:

$$\forall Y_0 \in [H(Q), Q], Y_{k+1} = Q - A^H F(Y_k^{\frac{1}{s}})A = H(Y_k), k = 0, 1, 2, \cdots . \tag{3.1}$$

Then the sequence $\{Y_k\}_{k=0}^{\infty}$ in (3.1) converges to the unique solution $\overline{Y}$ of the equation $Y + A^H F(Y^{\frac{1}{s}})A = Q$, and $\overline{X} = \overline{Y}^{\frac{1}{s}}$, moreover, we can obtain $\|Y_{k+1} - Y_k\| < a\|Y_k - Y_{k-1}\|$.

Proof. (i) Because F maps into $P(n)$,we have $H(Y) \le Q$ for all $Y \in [H(Q), Q]$. Note that F is monotone, we have $F(Y^{\frac{1}{s}}) \le F(Q^{\frac{1}{s}})$, which implies $H(Y) \ge Q - A^H F(Q^{\frac{1}{s}})A = H(Q)$. So H maps $[H(Q), Q]$ into $[H(Q), Q]$. Moreover, $\forall Y_1, Y_2 \in [H(Q), Q]$, we have $Y_1 \ge \lambda_n(H(Q))I, Y_2 \ge \lambda_n(H(Q))I$, combine Lemma 3.1–3.3, we can obtain

$$\begin{aligned} \|H(Y_1) - H(Y_2)\| &= \|A^H F(Y_2^{\frac{1}{s}})A - A^H F(Y_1^{\frac{1}{s}})A\| \\ &= \|A^H(F(Y_2^{\frac{1}{s}}) - F(Y_1^{\frac{1}{s}}))A\| \\ &\le \|A\|^2 \max_{Z \in L_{Y_1^{\frac{1}{s}}, Y_2^{\frac{1}{s}}}} \|D(Z)\|\|Y_1^{\frac{1}{s}} - Y_2^{\frac{1}{s}}\| \\ &\le \frac{1}{s}S_{[H(Q),Q]}\|A\|^2\lambda_n^{1-s}(H(Q))\|Y_1 - Y_2\|. \end{aligned}$$

From $a < 1$, we obtain that H is a contraction map on $[H(Q), Q]$. Because $[H(Q), Q]$ is a closed subset of $P(n)$ which implies $[H(Q), Q]$ is a complete metric space. Then it follows from Banach's Fixed Point Theorem that the map H has a unique fixed point $\overline{Y}$ in $[H(Q), Q]$ i.e., $H(\overline{Y}) = \overline{Y}$. Let $\overline{X} = \overline{Y}^{\frac{1}{s}} \in [(H(Q))^{\frac{1}{s}}, Q^{\frac{1}{s}}] = [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, then the matrix $\overline{X}$ is the unique solution of Eq (1.1). This prove the first part of the Theorem.

(ii) From the proof of the (i), H is a contraction map which maps $[H(Q), Q]$ into $[H(Q), Q]$, and for all$Y_1, Y_2 \in [H(Q), Q]$, we have $\|H(Y_1) - H(Y_2)\| < a\|Y_1 - Y_2\|$. So the sequence $\{Y_k\}_{k=0}^{\infty}$ in (3.1) converges to the unique solution $\overline{Y}$ of the equation $Y + A^H F(Y^{\frac{1}{s}})A = Q$, and $\overline{X} = \overline{Y}^{\frac{1}{s}}$ is the unique solution of Eq (1.1), also we have $\|Y_{k+1} - Y_k\| < a\|Y_k - Y_{k-1}\|$. The proof is completed. □

Corollary 3.1. *If the conditions of the Theorem 3.2 are satisfied, then we have*

$$\|Y_k - Y\| \le \frac{a^k}{1-a}\|Y_1 - Y_0\|, \|Y_k - Y\| \le \frac{a}{1-a}\|Y_k - Y_{k-1}\|.$$

Proof. From the prove of Theorem 3.2, we obtain that $\|Y_{k+1} - Y_k\| < a\|Y_k - Y_{k-1}\|$. So we have

$$\begin{aligned} \|Y_{k+p} - Y_k\| &\le \sum_{i=1}^{p} \|Y^{k+i} - Y^{k+i-1}\| \\ &\le (a^{p-1} + \cdots + a + 1)\|Y_{k+1} - Y_k\| \end{aligned} \tag{3.2}$$

$$\leq \quad \frac{a^k}{1-a}||Y_1 - Y_0||.$$

Let $p \to \infty$, then we get $||Y_k - Y|| \leq \frac{a^k}{1-a}||Y_1 - Y_0||$, and from the second inequality, i.e.,(3.1) we can obtain that $||Y_k - Y|| \leq \frac{a}{1-a}||Y_k - Y_{k-1}||$. This implies that the error of approximate solution can be estimated by the scope of difference between the iteration sequences Y_k and Y_{k+1}. □

For the case that $F : P(n) \to P(n)$ is continuous, anti-monotone and invertible had been discussed in The Theorem 3 of [1], there the conclusion is similar with the above Theorem which we have obtain.

The next corollary describes the number of iterations to be taken to ensure that $||Y_k - Y|| \leq ||\varepsilon||$.

Corollary 3.2. *If $Y_0 = Q^{\frac{1}{s}}$ and ε is a convergence tolerance then the number k of iterations to be taken is at most*

$$k = \left[\frac{\ln \varepsilon + \ln ||A^H F(Q^{\frac{1}{s}})A||}{\ln a}\right] + 1.$$

Proof. Let $Y_0 = Q^{\frac{1}{s}}$, from Lemma 3.1, we have $||Y_k - Y|| \leq \frac{a^k}{1-a}||Y_1 - Y_0|| = \dfrac{a^k}{1-a}||A^H F(Q^{\frac{1}{s}})A||$. If want to ensure that $||Y_k - Y|| \leq ||\varepsilon||$, we just need make $\frac{a^k}{1-a}||A^H F(Q^{\frac{1}{s}})A|| \leq \varepsilon$, then we obtain

$$k \leq \frac{\ln \varepsilon + \ln ||A^H F(Q^{\frac{1}{s}})A||}{\ln a}.$$

This implies that the number k of iterations to be taken is at most

$$\left[\frac{\ln \varepsilon + \ln ||A^H F(Q^{\frac{1}{s}})A||}{\ln a}\right] + 1.$$

□

Theorem 3.3. *If $F : P(n) \to P(n)$ is monotone, continuous and $G(Q^{\frac{1}{s}}) > O$, then the following results hold true.*

(i) G is anti-monotone on $P(X)$ which maps $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$ into $[G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, and for any positive definite matrix X for which $G(X)$ is positive definite we have $G(Q^{\frac{1}{s}}) \leq G^2(X) \leq Q^{\frac{1}{s}}$.

(ii) There always exists either a periodic orbit of period 2 of the map G or a fixed point of G. The sequence of matrices $G^{2j}(Q^{\frac{1}{s}})_{j=0}^{\infty}$ is a decreasing sequence of positive definite matrices converging to a positive definite matrix X_∞, and the sequence of matrices $G^{2j+1}(Q^{\frac{1}{s}})_{j=0}^{\infty}$ is an increasing sequence of positive definite matrices converging to a positive definite matrix $X_{-\infty}$, and the matrices X_∞, $X_{-\infty}$ form either a periodic orbit of G of period 2,or $X_\infty = X_{-\infty}$, in which case it is a fixed point of G, and hence solution of Eq (1.1).

(iii) Moreover, G maps the set $[X_{-\infty}, X_\infty]$ into itself, and any periodic orbit of G is contained in this set. In particular, any solution of Eq (1.1) is in between $X_{-\infty}$ and X_∞, and if $X_\infty = X_{-\infty}$, then there is an unique positive definite solution.

(iv) In the case where $X_\infty = X_{-\infty}$ this matrix is the global attractor for the map G in the following sense: For any positive definite X for which $G(X)$ is positive definite as well, we have $\lim j_{\to\infty} G^j(X) = X_\infty$.

(v) In the case where $X_\infty = X_{-\infty}$ the following holds: if $X \leq X_{-\infty}$, then the orbit of X under G converges to the periodic orbit $X_{-\infty}$, X_∞ in the sense that $\lim j_{\to\infty} G^{2j-1}(X) = X_\infty$, *and* $\lim j_{\to\infty} G^{2j}(X) = X_{-\infty}$. *If $X \geq X_\infty$ and $G(X)$ is positive definite, then the orbit of X under G converges to the periodic orbit $X_{-\infty}$, X_∞ in the sense that* $\lim j_{\to\infty} G^{2j-1}(X) = X_{-\infty}$ *and* $\lim j_{\to\infty} G^{2j}(X) = X_\infty$.

Proof. (i) Noting F is monotone on $P(n)$, for all $X, Y \in P(n)$ and $X \leq Y$, we have $F(X) \leq F(Y)$ which implies $G(X) = (Q - A^H F(X)A)^{\frac{1}{s}} \geq (Q - A^H F(Y)A)^{\frac{1}{s}} = G(Y)$, so G is anti-monotone. Because F maps into $P(n)$, we can obtain $G(X) = (Q - A^H F(X)A)^{\frac{1}{s}} \leq Q^{\frac{1}{s}}$, also from the monotonicity of F, we have $F(X) \leq F(Q^{\frac{1}{s}})$ for all $X \in [G(Q^{\frac{1}{s}}), Q^{\frac{1}{s}}]$, which implies that $Q - A^H F(X)A \geq Q - A^H F(Q^{\frac{1}{s}})A$, i.e., $G(X) \geq G(Q^{\frac{1}{s}})$. For any positive definite matrix X for which $G(X)$ is positive definite, we have $G(X) \leq Q^{\frac{1}{s}}$ and $G^2(X) \leq Q^{\frac{1}{s}}$. Combine $G(X) \leq Q^{\frac{1}{s}}$ and that G is anti-monotone, then we have $G^2(X) \geq G(Q^{\frac{1}{s}})$.

The proof for (ii)-(v) are similar to the interpretation of Theorem 2.2 in [2], which is omitted by this paper. □

Theorem 3.4. *If $F : P(n) \rightarrow -P(n)$ is monotone, continuous, Then the sequence of matrices $\{G^{2j}(Q^{\frac{1}{s}})\}_1^\infty$ is a increasing sequence of positive definite matrices converging to a positive definite matrix $\widetilde{X}_{-\infty}$, and the sequence of matrices $\{G^{2j+1}(Q^{\frac{1}{s}})\}_1^\infty$ is an decreasing sequence of positive definite matrices converging to a positive definite matrix $\widetilde{X}_\infty$. If $\widetilde{X}_\infty = \widetilde{X}_{-\infty} = \widetilde{X}$, then $\widetilde{X}$ is the unique definite solution of Eq (1.1).*

Proof. Because F maps into $-P(n)$, we have $G(X) = (Q - A^H F(X)A)^{\frac{1}{s}} \geq Q^{\frac{1}{s}}$, which implies $G(Q^{\frac{1}{s}}) \geq Q^{\frac{1}{s}} > O$ and $G^2(Q^{\frac{1}{s}}) \geq Q^{\frac{1}{s}}$. Similar with the proof of Theorem 3.2, we obtain that G is anti-monotone, and G^2 is monotone. So we have $G^3(Q^{\frac{1}{s}}) \leq G(Q^{\frac{1}{s}})$, then applying G^2 repeatedly, we see that the monotonicity of G^2 on this set implies that the sequence $\{G^{2j+1}(Q^{\frac{1}{s}})\}_1^\infty$ is a decreasing sequence of positive definite matrices that is bounded below by the positive definite matrix $Q^{\frac{1}{s}}$. Hence it converges to a positive definite matrix $\widetilde{X}_\infty$. Also apply the monotonicity of G^2 to inequality $G^2(Q^{\frac{1}{s}}) \geq Q^{\frac{1}{s}}$, we can obtain the sequence $\{G^{2j}(Q^{\frac{1}{s}})\}_1^\infty$ is a increasing sequence of positive definite matrices. Note G is anti-monotone and $G(Q^{\frac{1}{s}}) \geq Q^{\frac{1}{s}}$, we have $G^2(Q^{\frac{1}{s}}) \leq G(Q^{\frac{1}{s}})$. Combine the monotonicity of G^2 and $\{G^{2j+1}(Q^{\frac{1}{s}})\}_1^\infty$ is a decreasing sequence that we have proved, we can obtain $G^{2j}(Q^{\frac{1}{s}}) \leq G^{2j-1}(Q^{\frac{1}{s}}) \leq G^{2j-3}(Q^{\frac{1}{s}}) \leq \cdots \leq G(Q^{\frac{1}{s}})$, which implies the sequence $\{G^{2j}(Q^{\frac{1}{s}})\}_1^\infty$ is bounded above by the positive definite matrix $G(Q^{\frac{1}{s}})$. So it converges to a positive definite matrix $\widetilde{X}_{-\infty}$. Obviously, when $\widetilde{X}_\infty = \widetilde{X}_{-\infty} = \widetilde{X}$, then $\widetilde{X}$ is a definite solution of Eq (1.1). In the following we prove uniqueness.

Assume that $\overline{X}$ is a definite solution of Eq (1.1), then we have $\overline{X} = G(\overline{X}) \geq Q^{\frac{1}{s}}$, so we obtain $G^{2j+1}(Q^{\frac{1}{s}}) \geq G^{2j}(Q^{\frac{1}{s}})$. Hence, letting $j \rightarrow \infty$, we see that $\overline{X} \geq \widetilde{X}_{-\infty} = \widetilde{X}$. Moreover, from $\overline{X} = G(\overline{X}) \geq Q^{\frac{1}{s}}$, we have $\overline{X} = G^2(\overline{X}) \geq G(Q^{\frac{1}{s}})$. Then applying G^2 repeatedly, we see that the monotonicity of G^2 implies that $G^{2j}(Q^{\frac{1}{s}}) \leq G^{2j-1}(Q^{\frac{1}{s}})$. So we obtain $\overline{X} \leq \widetilde{X}_\infty = \widetilde{X}$. Hence, we obtain $\widetilde{X} = \overline{X}$, i.e., $\widetilde{X}$ is the unique definite solution of Eq (1.1). This complete the proof. □

4. Perturbation analysis

In this section we will consider the perturbation analysis of the solution of Eq (1.1). The perturbed equation will be as follows:

$$X^s + \widetilde{A}^H F(X)\widetilde{A} = \widetilde{Q}, \tag{4.1}$$

where $\widetilde{Q}$ be positive definite, $\widetilde{A}$ and $\widetilde{Q}$ is a small perturbation of A and Q, respectively, and s is a positive integer. Denote $\triangle A = \widetilde{A} - A, \triangle Q = \widetilde{Q} - Q, \triangle X = \widetilde{X} - \overline{X}$. Then the following theorems give us upper bounds for $\|\overline{X} - \widetilde{X}\|_F$, here $\overline{X}$ is a solution of Eq (1.1), and $\widetilde{X}$ is a solution of Eq (4.1).

Theorem 4.1. *If $F : P(n) \to M(n)$ is differentiable, $\|F(\cdot)\|_F \le c$, $\overline{X}, \widetilde{X}$ be the positive definite semi-definite solutions of Eq (1.1) and its perturbation Eq (4.1). $\overline{X} \in [B_1, C_1]$, $\widetilde{X} \in [B_2, C_2]$. We have $\|\triangle X\|_F \le \dfrac{d}{s\lambda^{s-1} - l}$, and $\dfrac{\|\triangle X\|_F}{\|X\|_F} \le \dfrac{d}{\lambda(s\lambda^{s-1} - l)}$ when $l < s\lambda^{s-1}$ are satisfied. Here $\widehat{\lambda} = max\{\|X\|_F | X \in \Omega\}$, $\lambda = min\{\lambda_n(B_1), \lambda_n(B_2)\}$, $d = c\widehat{\lambda}(\|\widetilde{A}^H\|_F + \|A\|_F)\|\triangle A\|_F + \|\triangle Q\|_F, l = r_\Omega \|A\|_F \|\widetilde{A}^H\|_F, r_\Omega = max_{Z\in\Omega}\|DF(Z)\|_F, \Omega = \{\alpha X + (1-\alpha)Y | \alpha \in [0,1], X \in [B_1, C_1], Y \in [B_2, C_2]\}$.*

Proof. Because $\overline{X}, \widetilde{X}$ be the positive definite semi-definite solutions of Eq (1.1) and its perturbation Eq (4.1), we have

$$\begin{aligned} \overline{X}^s - \widetilde{X}^s &= \widetilde{A}^H F(\widetilde{X})\widetilde{A} - \widetilde{Q} - A^H F(\overline{X})A + Q \\ &= \widetilde{A}^H F(\widetilde{X})(\widetilde{A} - A) + \widetilde{A}^H (F(\widetilde{X}) - F(X))A + (\widetilde{A}^H - A^H)F(X)A + \triangle Q. \end{aligned} \tag{4.2}$$

Noting $\overline{X} \in [B_1, C_1]$, $\widetilde{X} \in [B_2, C_2]$, $\widehat{\lambda} = \max\{\|X\|_F | X \in \Omega\}$, $\lambda = \min\{\lambda_n(B_1), \lambda_n(B_2)\}$, we can obtain $\overline{X} \ge B_1 \ge \lambda_n(B)_1 I \ge \lambda I$, $\widetilde{X} \ge B_2 \ge \lambda_n(B_2)I \ge \lambda I$, and $\|\overline{X}\|_F \le \widehat{\lambda}, \|\widetilde{X}\|_F \le \widehat{\lambda}$. Combining Lemma 3.2 and substituting $r_\Omega = \max_{Z\in\Omega}\|DF(Z)\|_F, d = c\widehat{\lambda}(\|\widetilde{A}^H\|_F + \|A\|_F)\|\triangle A\|_F + \|\triangle Q\|_F$ for the Eq (4.2) we get

$$\begin{aligned} \|\overline{X}^s - \widetilde{X}^s\|_F &= \|\widetilde{A}^H F(\widetilde{X})(\widetilde{A} - A) + \widetilde{A}^H (F(\widetilde{X}) - F(X))A + (\widetilde{A}^H - A^H)F(X)A + \triangle Q\|_F \\ &\le \|\widetilde{A}^H\|_F \|F(\cdot)\|_F \|\widetilde{X}\|_F \|\triangle A\|_F + \|\widetilde{A}^H\|_F \max_{Z\in L_{\overline{X},\widetilde{X}}} \|DF(Z)\|_F \|\widetilde{X} - \overline{X}\|_F \|A\|_F \\ &\quad + \|\triangle A\|_F \|F(\cdot)\|_F \|\overline{X}\|_F \|A\|_F + \|\triangle Q\|_F \\ &= c\widehat{\lambda}(\|\widetilde{A}^H\|_F + \|A\|_F)\|\triangle A\|_F + \|\triangle Q\|_F + r_\Omega \|A\|_F \|\widetilde{A}^H\|_F \|\overline{X} - \widetilde{X}\|_F. \end{aligned} \tag{4.3}$$

From Lemma 3.2, we have

$$\begin{aligned} \|\overline{X}^s - \widetilde{X}^s\|_F &= \|\mathrm{vec}(\overline{X}^s - \widetilde{X}^s)\|_2 \\ &= \Big\|\mathrm{vec}\Big(\sum_{i=0}^{s-1} \overline{X}^i (\overline{X} - \widetilde{X})\widetilde{X}^{s-1-i}\Big)\Big\|_2 \\ &= \Big\|\Big(\sum_{i=0}^{s-1} \widetilde{X}^{s-1-i} \otimes \overline{X}^i\Big)\mathrm{vec}(\overline{X} - \widetilde{X})\Big\|_2 \\ &\ge s\lambda^{s-1}\big\|\overline{X} - \widetilde{X}\big\|_F. \end{aligned} \tag{4.4}$$

Combining (4.3) and (4.4), we can obtain

$$s\lambda^{s-1}\|\overline{X} - \widetilde{X}\|_F \le c\widehat{\lambda}(\|\widetilde{A}^H\|_F + \|A\|_F)\|\triangle A\|_F + \|\triangle Q\|_F + r_\Omega \|A\|_F \|\widetilde{A}^H\|_F \|\overline{X} - \widetilde{X}\|_F. \tag{4.5}$$

Substituting $d = c\widehat{\lambda}(\|\widetilde{A}^H\|_F + \|A\|_F)\|\triangle A\|_F + \|\triangle Q\|_F, l = r_\Omega \|A\|_F \|\widetilde{A}^H\|_F$ for (4.5), we can obtain

$$\|\triangle X\|_F \le \frac{d}{s\lambda^{s-1} - l}.$$

Moreover, we can get

$$\frac{\|\triangle X\|_F}{\|X\|_F} \le \frac{d}{\lambda(s\lambda^{s-1} - l)}.$$

This complete the proof. □

In fact, if we don't consider the high order quantity, i.e., $\|\triangle A\|^2$, then we can obtain a simpler error bound for $\|\overline{X} - \widetilde{X}\|_F$, which do not need calculate the value of $\|\widetilde{A}\|_F$.

Theorem 4.2. *The conditions are similar to the conditions of Theorem 4.1 which we have obtained, if we don't consider the high order quantity, i.e.,* $\|\triangle A\|_F^2$*, then we have the following absolute error*

$$\|\triangle X\|_F = \|\overline{X} - \widetilde{X}\|_F \leq \frac{2\widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F}{s\lambda^{s-1} - r_\Omega\|A\|_F^2}$$

and the relative error

$$\frac{\|\triangle X\|_F}{\|\overline{X}\|_F} \leq \frac{2\widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F}{\lambda(s\lambda^{s-1} - r_\Omega\|A\|_F^2)}.$$

Proof. Because $\overline{X}, \widetilde{X}$ be the positive definite semi-definite solutions of Eq (1.1) and its perturbation Eq (4.1), we have

$$\begin{aligned}
\overline{X}^s - \widetilde{X}^s &= \widetilde{A}^H F(\widetilde{X})\widetilde{A} - A^H F(\overline{X})A + Q - \widetilde{Q} \\
&= A^H(F(\widetilde{X}) - F(\overline{X}))A + A^H F(\widetilde{X})(\widetilde{A} - A) + (\widetilde{A}^H - A^H)F(\widetilde{X})A \\
&\quad +(\widetilde{A}^H - A^H)F(\widetilde{X})(\widetilde{A} - A) + Q - \widetilde{Q}. \qquad (4.6)
\end{aligned}$$

So we can obtain

$$\begin{aligned}
\|\overline{X}^s - \widetilde{X}^s\|_F &= \|A^H(F(\widetilde{X}) - F(\overline{X}))A + A^H F(\widetilde{X})(\widetilde{A} - A) + (\widetilde{A}^H - A^H)F(\widetilde{X})A \\
&\quad +(\widetilde{A}^H - A^H)F(\widetilde{X})(\widetilde{A} - A) + Q - \widetilde{Q}\|_F \\
&\leq \|A\|_F^2\|F(\widetilde{X}) - F(\overline{X})\|_F + \|A^H\|_F\|\triangle A\|_F\|F(\widetilde{X})\|_F + \|A\|_F\|\triangle A\|_F\|F(\widetilde{X})\|_F \\
&\quad +\|\triangle A\|_F^2\|F(\widetilde{X})\|_F + \|\triangle Q\|_F \\
&\leq r_\Omega\|A\|_F^2\|\overline{X} - \widetilde{X}\|_F + \widehat{c\lambda}\|A^H\|_F\|\triangle A\|_F + \widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F. \qquad (4.7)
\end{aligned}$$

Similar to Theorem 4.1, we have

$$\|\overline{X}^s - \widetilde{X}^s\|_F \geq s\lambda^{s-1}\|\overline{X} - \widetilde{X}\|_F. \qquad (4.8)$$

Combine (4.7) and (4.8), we obtain

$$s\lambda^{s-1}\|\overline{X} - \widetilde{X}\|_F \leq r_\Omega\|A\|_F^2\|\overline{X} - \widetilde{X}\|_F + \widehat{c\lambda}\|A^H\|_F\|\triangle A\|_F + \widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F.$$

So we have the absolute error

$$\|\triangle X\|_F = \|\overline{X} - \widetilde{X}\|_F \leq \frac{2\widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F}{s\lambda^{s-1} - r_\Omega\|A\|_F^2},$$

and we can also get the the relative error

$$\frac{\|\triangle X\|_F}{\|\overline{X}\|_F} \leq \frac{2\widehat{c\lambda}\|A\|_F\|\triangle A\|_F + \|\triangle Q\|_F}{\lambda(s\lambda^{s-1} - r_\Omega\|A\|_F^2)}.$$

The proof is completed. □

5. Examples and numerical results

So far we have considered the general nonlinear matrix Eq (1.1) and achieved general conditions for the existence of a positive definite solution or a positive definite semi-definite solution for this class of equations. Now we give several examples to show how this works in particular cases, and some numerical results to specify the rationality of the results we have obtain above.

Example 5.1. *Take $F(X) = X^{\alpha}$ $(\alpha \in (0,1])$, $Q = I$, and let A be an arbitrary square matrix with $\|A\|_2 < 1$. Then Eq (1.1) has a positive definite solution in $[(I - A^HA)^{\frac{1}{s}}, I]$.*

Proof. Because of $\|A\|_2 < 1$, so we have $\sqrt{\lambda_1(A^HA)} < 1$, that is, $\lambda_1(A^HA) < 1$, which implies that $A^HA < I$. So we have $A^HF(Q^{\frac{1}{s}})A = A^HIA = A^HA < I = Q$, i.e., $G(Q^{\frac{1}{s}}) = (Q - A^HF(Q^{\frac{1}{s}})A)^{\frac{1}{s}} > O$. Then all conditions of Theorem 2.5 are satisfied. And from the conclusion of Theorem 2.5, we obtain that equation $X^s + A^HX^{\alpha}A = I$ has a positive definite solution in $[(I - A^HA)^{\frac{1}{s}}, I]$. □

Example 5.2. *Take $F(X) = -\sum\limits_{i=1}^{m} X^{-\delta^i}$ $(X \geq I, m \geq 1, \delta^i \in (0,1])$, $Q = I$, and let A be an arbitrary square matrix . Then Eq (1.1) has a positive definite solution in interval $[I, (I + mA^HA)^{\frac{1}{s}}]$, in particularly, all solutions of equation $X^s - \sum\limits_{i=1}^{m} A^HX^{-\delta^i}A = I$ are in $[I, (I + mA^HA)^{\frac{1}{s}}]$.*

Proof. From Lemma 3.1, we obtain X^{δ^i} are monotonous for $i = 1, 2, \cdots, m$, so $X^{-\delta^i}$ are anti-monotonous for $i = 1, 2, \cdots, m$. This implies that $X^{-\delta^i} \leq I$ $(i = 1, 2, \cdots, m)$ when $X \geq I$, i.e., $F(X) \geq -mI$, then $A^HF(X)A \geq -mA^HA$. Let $B = I + mA^HA$, then Eq (2.1) of Theorem 2.3 is set up for all $X \in [Q^{\frac{1}{s}}, B^{\frac{1}{s}}] = [I, (I + mA^HA)^{\frac{1}{s}}]$. So all conditions of Theorem 2.3 are satisfied. According to Theorem 2.3, equation $X^s - \sum\limits_{i=1}^{m} A^HX^{-\delta^i}A = I$ has a positive definite solution in $[I, (I + mA^HA)^{\frac{1}{s}}]$. And more specifically, all solutions of the above equation are in $[I, (I + mA^HA)^{\frac{1}{s}}]$. □

Example 5.3. *Let $F(X) = X^{-t}, t \in (0,1]$, if Eq (1.1) has a positive definite solution $\overline{X}$, then from the conclusion of Theorem 2.6, we have $\overline{X} \in (F^{-1}(A^{-*}QA^{-1}), G(Q^{\frac{1}{s}}))$. Compared with the conclusion from Theorem 2.1 in [7] that $\overline{X} \in ((AQ^{-1}A^H)^{\frac{1}{t}}, Q^{\frac{1}{s}})$, the conclusion in our paper, i.e., the result from Theorem 2.6 which we have obtained have a better estimates for the solution $\overline{X}$.*

Proof. Note that F maps into $-P(n)$, we obtain $G(Q^{\frac{1}{s}}) \leq Q^{\frac{1}{s}}$. From $F(X) = X^{-t}$, we have $F^{-1}(A^{-*}QA^{-1}) = (AQ^{-1}A^H)^{\frac{1}{t}}$. So $\overline{X} \in (F^{-1}(A^{-*}QA^{-1}), G(Q^{\frac{1}{s}})) \in ((AQ^{-1}A^H)^{\frac{1}{t}}, Q^{\frac{1}{s}})$. The proof is completed. □

The above example (i.e., example 5.4) which we have given show that the conclusion in our paper may have a better estimates for the solution of Eq (1.1) to some extent.

An application of Theorem 2.4 is given to discuss the property of the positive definite solutions of Eq (1.1) in the following example.

Example 5.4. *For the matrix Eq (1.1), we choose $s = 5$, $F(X) = X^{0.5}$, then F is monotone by Lemma 3.1. Let*

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 7.15 & 3.02 & 0.11 \\ 3.02 & 6.20 & 2.01 \\ 0.11 & 2.01 & 6.50 \end{pmatrix}$$

and $Q = X^s + A^H F(X)A = X^5 + X^{0.5}$, *i.e.*,

$$Q = \begin{pmatrix} 5.1837 & 4.7204 & 2.1658 \\ 4.7204 & 4.7578 & 2.7761 \\ 2.1658 & 2.7761 & 2.4251 \end{pmatrix}$$

with every element multiplying 10^4.

Obviously, X is an positive matrix, that is, $X \geq O$. By using Matlab 7.0 we can obtain

$$\lambda_1(Q) = 1.1099 \times 10^5, \lambda_2(Q) = 0.1248 \times 10^5, \lambda_3(Q) = 0.0019 \times 10^5.$$

So we have $Q \geq O$. Now let us judge if $G(Q^{\frac{1}{5}}) = (Q - Q^{\frac{1}{10}})^{\frac{1}{5}} > O$ or not. By the calculation, we obtain that

$$G(Q^{\frac{1}{5}}) = \begin{pmatrix} 7.1500 & 3.0200 & 0.1100 \\ 3.0200 & 6.0200 & 2.0100 \\ 0.1100 & 2.0100 & 6.5000 \end{pmatrix}$$

and we also calculate the eigenvalue of $G(Q^{\frac{1}{5}})$, and obtain

$$\lambda_1 G(Q^{\frac{1}{5}}) = 6.5952, \lambda_2 G(Q^{\frac{1}{5}}) = 10.2107, \lambda_3 G(Q^{\frac{1}{5}}) = 2.8641.$$

This implies that $G(Q^{\frac{1}{5}}) > O$. Equation (1.1) with A and Q above has at least one positive definite solution as a result of $Q = X^s + A^H F(X)A = X^5 + X^{0.5}$. So according to the inherent meaning of Theorem 2.4, all positive semi-definite solutions of Eq (1.1) must be in $[G(Q^{\frac{1}{5}}), Q^{\frac{1}{5}}]$, that is all the positive semi-definite solutions $\overline{X}$ of equation $X^5 + X^{0.5} = Q$ satisfy

$$\begin{pmatrix} 7.1500 & 3.0200 & 0.1100 \\ 3.0200 & 6.0200 & 2.0100 \\ 0.1100 & 2.0100 & 6.5000 \end{pmatrix} \leq \overline{X} \leq \begin{pmatrix} 7.1515 & 3.0180 & 0.1109 \\ 3.0180 & 6.0229 & 2.0085 \\ 0.1109 & 2.0085 & 6.5010 \end{pmatrix}.$$

In the remainder of this section, we report some numerical results. These numerical results describe the correctness of Theorem 3.4 that we have obtained. The numerical experiments were carried out using MATLAB 2010a on ZWX-PC Intel i3 processor with 2.10 GHz and 4.0GB RAM computer with double precision.The rounding unit is approximately 1.11×10^{-16}. In the example we take $s = 1, Q = I, F(X) = -X^{-2}$, then the assumptions of Theorem 3.4 are satisfied. In Table 1, k denotes the number of iterations, ε_k denotes $\|G^{2k}(Q^{\frac{1}{s}}) - G^{2k+1}(Q^{\frac{1}{s}})\|_\infty$, X_k, Y_k, respectively, denote the iteration results of $G^{2k}(Q^{\frac{1}{s}})$, $G^{2k+1}(Q^{\frac{1}{s}})$ of the *kth* step. $\widetilde{X}_{-\infty}$, $\widetilde{X}_\infty$, respectively, is taken to be the final iterate of $G^{2k}(Q^{\frac{1}{s}})$, $G^{2k+1}(Q^{\frac{1}{s}})$ after $\varepsilon_k < 10^{-8}$ is satisfied. The algorithm to work out the $G^k(Q^{\frac{1}{s}})$ is shown in the appendix of section seven.

Table 1. Error analysis for $X_k - Y_k$.

k	0	1	2	3
ε_k	0.4098	0.0023	$1.8243e-006$	$8.9540e-011$

After calculating, we can also give the X_k, and Y_k, $k = 1, 2, 3$ as follows, where omit the precision of computer's mechanical error.

$$X_1 = \begin{pmatrix} 1.0533 & 0.0698 & -0.0318 & -0.0657 \\ 0.0698 & 1.1242 & -0.0536 & -0.1185 \\ -0.0318 & -0.0536 & 1.0269 & 0.0603 \\ -0.0657 & -0.1185 & 0.0603 & 1.1388 \end{pmatrix}, \ Y_1 = \begin{pmatrix} 1.0538 & 0.0704 & -0.0321 & -0.0662 \\ 0.0704 & 1.1249 & -0.0540 & -0.1191 \\ -0.0321 & -0.0540 & 1.0271 & 0.0606 \\ -0.0662 & -0.1191 & 0.0606 & 1.1393 \end{pmatrix},$$

$$X_2 = \begin{pmatrix} 1.0529 & 0.0704 & -0.0321 & -0.0659 \\ 0.0704 & 1.1249 & -0.0540 & -0.1191 \\ -0.0321 & -0.0540 & 1.0257 & 0.0606 \\ -0.0659 & -0.1191 & 0.0606 & 1.1390 \end{pmatrix}, \ Y_2 = \begin{pmatrix} 1.0531 & 0.0704 & -0.0321 & -0.0661 \\ 0.0704 & 1.1250 & -0.0540 & -0.1191 \\ -0.0321 & -0.0540 & 1.0269 & 0.0606 \\ -0.0661 & -0.1191 & 0.0606 & 1.1393 \end{pmatrix},$$

$$X_3 = \begin{pmatrix} 1.0538 & 0.0704 & -0.0321 & -0.0662 \\ 0.0704 & 1.1249 & -0.0540 & -0.1191 \\ -0.0321 & -0.0540 & 1.0271 & 0.0606 \\ -0.0662 & -0.1191 & 0.0606 & 1.1393 \end{pmatrix}, \ Y_3 = \begin{pmatrix} 1.0538 & 0.0704 & -0.0321 & -0.0662 \\ 0.0704 & 1.1249 & -0.0540 & -0.1191 \\ -0.0321 & -0.0540 & 1.0271 & 0.0606 \\ -0.0662 & -0.1191 & 0.0606 & 1.1393 \end{pmatrix}.$$

Obviously, $X_3 \geq X_2 \geq X_1 \geq X_0$, $Y_3 \leq Y_2 \leq Y_1 \leq Y_0$, and $X_3 \simeq Y_3$. In fact, from the numerical test process we can see that the sequence $\{X_k\}_1^\infty$ are increasing and converging to X_3, in the contrary, the sequence $\{Y_k\}_1^\infty$ are decreasing and converging to Y_3. Here the numerical results we obtained is consistent with the internal interpretation of Theorem 3.4. Moreover, $\widetilde{X}_{-\infty}$, $\widetilde{X}_\infty$, and $\widetilde{X}$ refer to the Theorem 3.4 can also be obtained, that is, $\widetilde{X}_{-\infty} = X_3 = G^6(Q^{\frac{1}{s}})$, $\widetilde{X}_{-\infty} = Y_3 = G^7(Q^{\frac{1}{s}})$, and $\widetilde{X} = \widetilde{X}_{-\infty} = \widetilde{X}_\infty = X_3$. Which implies that the equation $X - A^H X^{-2} A = I$ has an unique solution, i.e., $\widetilde{X}$. This result is true, which can be demonstrated according to the interpretation [5].

6. Conclusions

In this paper, we discuss a general nonlinear matrix equation, which include the existence and properties of the solutions, uniqueness of a solution, iterative method for solve the equations, perturbation analysis about the solutions. These results are more general than [5, 16–21], the iterative procedure there may be better for the special case under consideration there $F(X) = X^{-1}, F(X) = \pm X^{-2}$ or $F(X) = X^{-t}, (t = 1, 2, 3, \cdots)$, but not readily applied to the very general case we have under consideration here. In recent years, many authors have dedicated into finding a good method to solve the nonlinear matrix equation where $s = 1, F(X) = X^{-1}$, which is a special case of the discrete algebraic Riccati equation studied in [22, 23]. For example, they have proposed the structure-preserving doubling algorithm [21, 22, 24–26], cyclic reduction algorithm [4], Latouche-Ramaswami algorithm [27], these methods may have a better rate of convergence in some cases (where the critical case do not include in). Wether these methods can be applied for our general case or not is an open problem and remains to be further research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are very much indebted to the anonymous referees for their constructive and valuable comments and suggestions which greatly improved the original manuscript of this paper. This work was supported by Fujian Natural Science Foundation (Grant No.2022J01378, 2023J01131186) and Key Reform in Education (Grant No. FBJG20200310).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. S. M. El-Sayed, A. C. M. Ran, On an iteration method for solving a class of nonlinear matrix equations, *SIAM J. Matrix Anal.*, **23** (2002), 632–645. https://doi.org/10.1137/S0895479899345571
2. Q. Li, P. P. Liu, Positive definite solutions of a kind of nonlinear matrix equations, *J. Inform. Comput. Sci.*, **7** (2010), 1527–1533.
3. A. C. M. Ran, M. C. B. Reurings, On the nonlinear matrix equation $X + A^H F(X)A = Q$ solution and perturbation theory, *Linear Algebra Appl.*, **346** (2002), 15–26. https://doi.org/10.1016/S0024-3795(01)00508-0
4. B. Meini, Efficient computation of the extreme solutions of $X+A^H X^{-1}A = Q$ and $X-A^H X^{-1}A = Q$, *Math. Comp.*, **71** (2002), 1189–1204.
5. I. G. Ivanov, V. I. Hassanov, B. V. Minchev, On matrix equations $X \pm A^H X^{-2}A = I$, *Linear Algebra Appl.*, **326** (2001), 27–44.
6. I. G. Ivanov, S. M. El-Sayed, Properties of positive definite solutions of the equation $X+A^H X^{-2}A = I$, *Linear Algebra Appl.*, **297** (1998), 303–316. https://doi.org/10.1016/S0024-3795(98)00023-8
7. Y. T. Yang, The iterative method for solving nonlinear matrix equation $X^s + A^H X^{-t}A = Q$, *Appl. Math. Comput.*, **188** (2007), 46–53. https://doi.org/10.1016/j.amc.2006.09.085
8. X. G. Liu, H. Gao, On the positive definite solutions of the matrix equation$X^s \pm A^\top X^{-t}A = I_n$, *Linear Algebra Appl.*, **368** (2003), 83–97. https://doi.org/10.1016/S0024-3795(02)00661-4
9. X. F. Duan, A. Liao, B. Tang, On the nonlinear matrix equation $X - \sum_{i=1}^{m} A_i^H X^{\delta_i} A_i = Q$, *Linear Algebra Appl.*, **429** (2008), 110–121. https://doi.org/10.1016/j.laa.2008.02.014
10. X. F. Duan, A. Liao, On the existence of Hermitian positive definite solutions of the matrix equation $X^s + A^H X^{-t}A = Q$, *Linear Algebra Appl.*, **429** (2008), 673–687. https://doi.org/10.1016/j.laa.2008.03.019
11. W. Kulpa, The Schauder fixed point theorem, *Acta Univ. Carol., Math. Phys.*, **38** (1997), 39–44.
12. M. Parodi, *La Localisation Des Valeurs Caracterisiques Des Matrices Etses Applications*, Paris: Gauthier-Villars, 1959.
13. R. Bhatia, *Matrix Analysis*, New York: Springer-Verlag, 1997. https://doi.org/10.1007/978-1-4612-0653-8

14. V. I. Istratescu, *Fixed Point Theorem, Mathematics and Its Applications*, Dordrecht: Reidel, 1981.
15. A. Ambrosetti, G. Prodi, *A Primer of Nonlinear Analysis, Cambridge Studies in Advanced Mathematics*, Cambridge: Cambridge University Press, 1993.
16. J. C. Engwerda, A. C. M. Ran, A. L. Rijkeboer, Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation $X + A^H X^{-1} A = Q$, *Linear Algebra Appl.*, **186** (1993), 255–275. https://doi.org/10.1016/0024-3795(93)90295-Y
17. A. Ferrante, B. C. Levy, Hermitian solutions of the equation $X = Q + NX^{-1}N^H$, *Linear Algebra Appl.*, **247** (1996), 359–373. https://doi.org/10.1016/0024-3795(95)00121-2
18. C. H. Guo, P. Lancaster, Iterative solution of two matrix equations, *Math. Comput.*, **68** (1999), 1589–1603.
19. S. F. Xu, On the maximal solution of the matrix equation $X + A^T X^{-1} A = I$, *Acta Sci. Natur. Univ. Pekinensis*, **36** (2000), 29–38.
20. X. Zhan, J. Xie, On the matrix equation $X + A^T X^{-1} A = I$, *Linear Algebra Appl.*, **247** (1996), 337–345. https://doi.org/10.1016/0024-3795(95)00120-4
21. C. H. Guo, Y. C. Kuo, W. W. Lin, Numerical solution of nonlinear matrix equations arising from Green's function calculations in nano research, *J. Comput. Appl. Math.*, **236** (2012), 4166–4180. https://doi.org/10.1016/j.cam.2012.05.012
22. E. K. W. Chu, H. Y. Fan, W. W. Lin, C. S. Wang, A structure-preserving doubling algorithm for periodic discrete-time algebraic Riccati equations, *Int. J. Control*, **77** (2004), 767–788. https://doi.org/10.1080/00207170410001714988
23. C. H. Guo, Newton's method for discrete algebraic Riccati equations when the closed-loop matrix has eigenvalues on the unit circle, *SIAM J. Matrix Anal. Appl.*, **20** (1998), 279–294. https://doi.org/10.1137/S0895479897322999
24. P. C. Y. Weng, E. K. W. Chu, Y. C. Kuo, W. W. Lin Solving large-scale nonlinear matrix equations by doubling, *Linear Algebra Appl.*, **439** (2013), 914–932. https://doi.org/10.1016/j.laa.2012.08.008
25. C. Y. Chiang, E. K. Wan Chu, C. H. Guo, T. M. Huang, W. W. Lin, S. F. Xu, Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case, *SIAM J. Matrix Anal.*, 227–247. https://doi.org/10.1137/080717304
26. C. H. Guo, W. W. Lin, The matrix equation $X + A^\top X^{-1} A = Q$ and its application in nano research, *SIAM J. Sci. Comput.*, **32** (2010), 3020–3038. https://doi.org/10.1137/090758209
27. G. Latouche, V. Ramaswami, A logarithmic reduction algorithm for quasi-death-birth processes, *J. Appl. Probab.*, **30** (1993), 650–674. https://doi.org/10.2307/3214773