Linear Algebra and its Applications 664 (2023) 349–368

Contents lists available at ScienceDirect

Linear Algebra and its Applications

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

Numerical solution of a class of quasi-linear matrix equations [☆]

Margherita Porcelli [a,b], Valeria Simoncini [a,c,*]

[a] Dipartimento di Matematica, AM^2, Alma Mater Studiorum - Università di Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy
[b] ISTI–CNR, Via Moruzzi 1, Pisa, Italy
[c] IMATI-CNR, Via Ferrata 5/A, Pavia, Italy

ARTICLE INFO

Article history:
Received 11 August 2022
Accepted 29 January 2023
Available online 1 February 2023
Submitted by H. Fassbender

MSC:
65H10
65F10
65F45
15A06

Keywords:
Matrix equations
Sylvester equation
Matrix functions
Fixed point iteration

ABSTRACT

Given the matrix equation $\boldsymbol{AX} + \boldsymbol{XB} + f(\boldsymbol{X})\boldsymbol{C} = \boldsymbol{D}$ in the unknown $n \times m$ matrix $\boldsymbol{X}$, we analyze existence and uniqueness conditions, together with computational solution strategies for $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ being a linear or nonlinear function. We characterize different properties of the matrix equation and of its solution, depending on the considered classes of functions f. Our analysis mainly concerns small dimensional problems, though several considerations also apply to large scale matrix equations.

[☆] The authors are members of the INdAM Research Group GNCS that partially supported this work under Progetto di Ricerca GNCS-INdAM CUP_E55F22000270001.

[*] Corresponding author.
E-mail addresses: margherita.porcelli@unibo.it (M. Porcelli), valeria.simoncini@unibo.it (V. Simoncini).

https://doi.org/10.1016/j.laa.2023.01.024

1. The problem

We consider the following nonlinear equation

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{B} + f(\boldsymbol{X})\boldsymbol{C} = \boldsymbol{D}, \tag{1.1}$$

in the unknown matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, where $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a linear or nonlinear function, while $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{m\times m}$, and $\boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{n\times m}$ are given matrices. Throughout the paper we assume that $\boldsymbol{A}$ and $-\boldsymbol{B}$ have no common eigenvalues, so that the operator $\mathcal{L} : \boldsymbol{X} \mapsto \boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{B}$ is invertible. Since the nonlinear function in $\boldsymbol{X}$ yields a scalar contribution to the matrix equation, we will refer to this problem as a *quasi-linear* matrix equation. We also notice that depending on the type of function f, the condition $m = n$ may also hold, and this will be assumed throughout without explicit mention.

Equation (1.1) is among the simplest possible generalizations of the Sylvester equation to more than two terms in the unknown matrix $\boldsymbol{X}$. Yet, it provides different intriguing challenges for its numerical solution, that we aim to address. A natural further generalization is the inclusion of more quasi-linear terms. Our interest in this problem stems from certain applications with linear f, see section 2.1 and [8], however we believe that the general case of f nonlinear may find applications in different contexts where the given mathematical problem can be formulated in terms of a matrix equation. To the best of our knowledge, no numerical methods have been presented in the literature for the class of problems considered in (1.1).

To begin our analysis, we observe that by letting $\boldsymbol{N} = -\mathcal{L}^{-1}(\boldsymbol{C})$, $\boldsymbol{M} = \mathcal{L}^{-1}(\boldsymbol{D})$, problem (1.1) is mathematically equivalent to

$$\boldsymbol{X} = \boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N}. \tag{1.2}$$

This equation provides the ideal computational setting in case n, m are small, as attention can be put into the function f, assuming that $\boldsymbol{M}, \boldsymbol{N}$ can be computed accurately. In case of matrices with large dimensions, the equality in (1.2) will in general be replaced by an approximation.

We start by considering the case of linear f, which was motivated by an application in solid mechanics and civil engineering developed in [8]. A linear function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ can be defined as $f(\boldsymbol{X}) = \text{trace}(\boldsymbol{H}\boldsymbol{X})$ for some matrix $\boldsymbol{H}$ of appropriate dimensions. For instance, for $\boldsymbol{H}$ equal to the identity matrix and $\boldsymbol{X}$ square, $f(\boldsymbol{X}) = \text{trace}(\boldsymbol{X})$, while for $\boldsymbol{H} = \boldsymbol{u}\boldsymbol{v}^T$ with $\boldsymbol{u} \in \mathbb{R}^m, \boldsymbol{v} \in \mathbb{R}^n$, $f(\boldsymbol{X}) = \boldsymbol{v}^T\boldsymbol{X}\boldsymbol{u}$, where the properties of the trace have been used. We will derive a closed form for $\boldsymbol{X}$, and also observe that under certain hypotheses $f(\boldsymbol{X})$ may be obtained without explicitly computing $\boldsymbol{X}$.

We then analyze a more general setting where f is the composition of a linear and a nonlinear function. The order in which these two functions are combined significantly influences the analysis and results: as an example, different existence and uniqueness properties may hold. So, for instance, working with $f(\boldsymbol{X}) = \text{trace}(\exp(-\boldsymbol{X}))$ (linear combined with nonlinear) differs significantly from dealing with $f(\boldsymbol{X}) = \exp(-\mathrm{trace}(\boldsymbol{X}))$ (nonlinear combined with linear). Distinct computational procedures also need to be devised.

We will explore iterative techniques that appropriately handle both f and the matrices forming the linear part of the equation. The linear-nonlinear problem is more computationally involved as the iteration requires matrix function evaluations and matrix updates. For this problem we will derive convergence results for a natural fixed-point iteration. In the nonlinear-linear case, the nonlinear iteration is performed at the scalar level and classical results for nonlinear equations can be employed, while taking into account the properties of the given data.

The following notation is adopted: matrices (resp. vectors) are denoted by bold case capital (resp. small) roman letters, while small roman letters are used for real valued functions (and also for matrix indices and matrix dimensions), and Greek letters are used for scalars. The notation $\boldsymbol{A} \succ 0$ ($\boldsymbol{A} \succeq 0$) denotes a symmetric and positive definite (semidefinite) matrix $\boldsymbol{A}$; the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ is equivalent to $\boldsymbol{A} - \boldsymbol{B} \succeq 0$. For a given matrix $\boldsymbol{X}$, the operator $\mathrm{vec}(\boldsymbol{X})$ stacks all columns of $\boldsymbol{X}$ one below the other into a single long vector, while the Kronecker operator of two matrices $\boldsymbol{A} \in \mathbb{R}^{n_A \times m_A}$ and $\boldsymbol{B} \in \mathbb{R}^{n_B \times m_B}$, is given by

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1m_A}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2m_A}\boldsymbol{B} \\ \vdots & & & \vdots \\ a_{n_A1}\boldsymbol{B} & a_{n_A2}\boldsymbol{B} & \cdots & a_{n_Am_A}\boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{n_An_B \times m_Am_B}.$$

Finally, we say that $f(x) = o(g(x))$ as $x \to 0$ if $\lim_{x\to 0} \frac{f(x)}{g(x)} = 0$.

2. The case of linear f

The following proposition yields the solution of (1.1) in closed form when f is a linear function.

Proposition 2.1. *Let $\boldsymbol{M}, \boldsymbol{N}$ be the solutions to the Sylvester equations $\boldsymbol{AM} + \boldsymbol{MB} = \boldsymbol{D}$ and $\boldsymbol{AN} + \boldsymbol{NB} = -\boldsymbol{C}$, respectively. Assume that $1 - f(\boldsymbol{N}) \neq 0$. Then the solution to (1.1) is given by*

$$\boldsymbol{X} = \boldsymbol{M} + \sigma\boldsymbol{N}, \quad \sigma = \frac{f(\boldsymbol{M})}{1 - f(\boldsymbol{N})}.$$

Proof. The problem can be written as in (1.2). Applying the linear function f to both sides of (1.2) yields $f(\boldsymbol{X}) = f(\boldsymbol{M}) + f(\boldsymbol{X})f(\boldsymbol{N})$, that is $f(\boldsymbol{X}) = f(\boldsymbol{M})/(1 - f(\boldsymbol{N}))$. Substituting in (1.2) the expression for $\boldsymbol{X}$ follows. Finally, by linearity,

$$\boldsymbol{AX} + \boldsymbol{XB} + f(\boldsymbol{X})\boldsymbol{C} - \boldsymbol{D} = \boldsymbol{A}(\boldsymbol{M} + \sigma\boldsymbol{N}) + (\boldsymbol{M} + \sigma\boldsymbol{N})\boldsymbol{B} + f(\boldsymbol{X})\boldsymbol{C} - \boldsymbol{D}$$

$$= \sigma \boldsymbol{A}\boldsymbol{N} + \sigma \boldsymbol{N}\boldsymbol{B} + f(\boldsymbol{X})\boldsymbol{C} = -\sigma \boldsymbol{C} + f(\boldsymbol{X})\boldsymbol{C} = 0,$$

which verifies that $\boldsymbol{X}$ solves (1.1). □

In case it holds that $1 - f(\boldsymbol{N}) = 0$, the relation $f(\boldsymbol{X}) = f(\boldsymbol{M}) + f(\boldsymbol{X})f(\boldsymbol{N})$ shows that two possible scenarios arise: for $f(\boldsymbol{M}) = 0$ then $\boldsymbol{X} = \boldsymbol{M} + \sigma\boldsymbol{N}$ where σ can be any real number, yielding a nonunique solution; if $f(\boldsymbol{M}) \neq 0$ then no solutions exist.

It is interesting to observe that when $f(\boldsymbol{X}) = \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{C} = \boldsymbol{v}\boldsymbol{v}^T$, it holds that $f(\boldsymbol{X})\boldsymbol{C} = (\boldsymbol{v}\boldsymbol{u}^T)\boldsymbol{X}\boldsymbol{u}\boldsymbol{v}^T$. The problem thus corresponds to the linear matrix equation

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{K}^T\boldsymbol{X}\boldsymbol{K} = \boldsymbol{D},$$

with $\boldsymbol{K} = \boldsymbol{u}\boldsymbol{v}^T$ of rank one; more generally, depending on f and $\boldsymbol{C}$, the term $\boldsymbol{K}^T\boldsymbol{X}\boldsymbol{K}$ may take the form $\boldsymbol{K}_1\boldsymbol{X}\boldsymbol{K}_2$, with $\boldsymbol{K}_2$ not necessarily the transpose of $\boldsymbol{K}_1$. With the previous choice of $\boldsymbol{K}$, the closed form solution in Proposition 2.1 is equivalent to the Sherman-Morrison-Woodbury (SMW) formula obtained by the vector form of the matrix equation. Indeed, the matrix equation above can be written as $(\boldsymbol{G} + \mathcal{V}\mathcal{U}^T)\boldsymbol{x} = \boldsymbol{d}$, where $\boldsymbol{G} = \boldsymbol{I} \otimes \boldsymbol{A} + \boldsymbol{B}^T \otimes \boldsymbol{I}$, $\mathcal{U} = \boldsymbol{u} \otimes \boldsymbol{u}$, $\mathcal{V} = \boldsymbol{v} \otimes \boldsymbol{v}$ and $\boldsymbol{d} = \mathrm{vec}(\boldsymbol{D})$; see, e.g., the specialized implementation presented in [3], together with the references in there, which also provide a historical recollection of the role of the SMW formula in the matrix equation context. Then the SMW formula reads

$$\begin{aligned}\boldsymbol{x} &= \boldsymbol{G}^{-1}\boldsymbol{d} - \boldsymbol{G}^{-1}\mathcal{V}(1 + \mathcal{U}^T\boldsymbol{G}^{-1}\mathcal{V})^{-1}\mathcal{U}^T\boldsymbol{G}^{-1}\boldsymbol{d} \\ &= \mathrm{vec}(\mathcal{L}^{-1}(\boldsymbol{D})) - \sigma\,\mathrm{vec}(\mathcal{L}^{-1}(\boldsymbol{C})),\end{aligned}$$

where $\sigma = (1+\mathcal{U}^T\boldsymbol{G}^{-1}\mathcal{V})^{-1}\mathcal{U}^T\boldsymbol{G}^{-1}\boldsymbol{d}$, which precisely corresponds to σ in Proposition 2.1, as $\mathcal{U}^T\boldsymbol{G}^{-1}\mathcal{V} = \boldsymbol{u}^T\mathcal{L}^{-1}(\boldsymbol{v}\boldsymbol{v}^T)\boldsymbol{u} = -f(\boldsymbol{N})$ and similarly for the other quantities. The cases where $\mathcal{U} = \boldsymbol{u}_1 \otimes \boldsymbol{u}_2$, $\mathcal{V} = \boldsymbol{v}_1 \otimes \boldsymbol{v}_2$ can be treated analogously.

It is also noticeable that for $f(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X})$ and $\boldsymbol{B} = \boldsymbol{A}$ with $\boldsymbol{A}$ nonsingular, the quantity $\mathrm{trace}(\boldsymbol{X})$ can be obtained without solving two Sylvester equations, but by only solving linear systems with $\boldsymbol{A}$. Indeed, from $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A} + \mathrm{trace}(\boldsymbol{X})\boldsymbol{C} = \boldsymbol{D}$ we write

$$\boldsymbol{X} + \boldsymbol{A}^{-1}\boldsymbol{X}\boldsymbol{A} + \mathrm{trace}(\boldsymbol{X})\boldsymbol{A}^{-1}\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{D}.$$

Applying the trace to all matrix terms we obtain

$$\mathrm{trace}(\boldsymbol{X}) = \frac{\mathrm{trace}(\boldsymbol{A}^{-1}\boldsymbol{D})}{2 + \mathrm{trace}(\boldsymbol{A}^{-1}\boldsymbol{C})},$$

where we have used the linearity and cyclic property of the trace. After this computation, the final $\boldsymbol{X}$ is obtained by solving the Sylvester equation $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A} = \boldsymbol{D} - \mathrm{trace}(\boldsymbol{X})\boldsymbol{C}$. The actual number of systems with $\boldsymbol{A}$ depends on the structure of $\boldsymbol{D}$ and $\boldsymbol{C}$. For instance, if $\boldsymbol{C} = \boldsymbol{C}_1\boldsymbol{C}_2^T$ has low rank equal to k and $\boldsymbol{C}_1 \in \mathbb{R}^{n\times k}$, then only k systems with $\boldsymbol{A}$

need to be solved to compute $\mathrm{trace}(\boldsymbol{A}^{-1}\boldsymbol{C}) = \mathrm{trace}(\boldsymbol{C}_2^T\boldsymbol{A}^{-1}\boldsymbol{C}_1)$. Other properties of the involved matrices can be exploited to lower the computational efforts.

Remark 2.2. The trace of the Sylvester solution matrix is of interest in its own right; see, e.g., [16], [14], [11] and their references. In particular, for $\boldsymbol{B} = \boldsymbol{A}$ nonsingular and $\boldsymbol{C} = 0$, the procedure discussed above can be used to compute the trace of the solution to $\boldsymbol{AX} + \boldsymbol{XA} = \boldsymbol{D}$, without explicitly computing or approximating the solution matrix.

This fact can be used for instance if in problem (1.1) one is interested in only computing the trace of $\boldsymbol{X}$, and not $\boldsymbol{X}$ itself. In this case, $\mathrm{trace}(\boldsymbol{M})$, $\mathrm{trace}(\boldsymbol{N})$ can be obtained without explicitly computing the two matrices $\boldsymbol{M}, \boldsymbol{N}$.

In a way similar to Proposition 2.1 one can treat the related problem

$$\boldsymbol{AX} + \boldsymbol{XB} + f_1(\boldsymbol{X})\boldsymbol{C}_1 + \ldots + f_\ell(\boldsymbol{X})\boldsymbol{C}_\ell = \boldsymbol{D}, \tag{2.1}$$

where f_i, $i = 1, \ldots, \ell$ are linear functions of their argument. Indeed, writing once again

$$\boldsymbol{X} = \boldsymbol{M} + \sum_{i=1}^{\ell} f_i(\boldsymbol{X})\boldsymbol{N}_i, \quad \boldsymbol{M} = \mathcal{L}^{-1}(\boldsymbol{D}),\, \boldsymbol{N}_i = -\mathcal{L}^{-1}(\boldsymbol{C}_i), \tag{2.2}$$

we can compute

$$f_j(\boldsymbol{X}) = f_j(\boldsymbol{M}) + \sum_{i=1}^{\ell} f_i(\boldsymbol{X}) f_j(\boldsymbol{N}_i), \qquad j = 1, \ldots, \ell.$$

Let $\sigma_j = f_j(\boldsymbol{X})$. Collecting all quantities, we obtain the $\ell \times \ell$ linear system

$$\begin{bmatrix} 1 - f_1(\boldsymbol{N}_1) & -f_1(\boldsymbol{N}_2) & \cdots & -f_1(\boldsymbol{N}_\ell) \\ -f_2(\boldsymbol{N}_1) & 1 - f_2(\boldsymbol{N}_2) & \cdots & -f_2(\boldsymbol{N}_\ell) \\ \vdots & \vdots & \ddots & \vdots \\ -f_\ell(\boldsymbol{N}_1) & \cdots & \cdots & 1 - f_\ell(\boldsymbol{N}_\ell) \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_\ell \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{M}) \\ \vdots \\ f_\ell(\boldsymbol{M}) \end{bmatrix} \Leftrightarrow \; (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{\sigma} = \boldsymbol{f}, \tag{2.3}$$

where $\boldsymbol{I}$ is the identity matrix of matching dimensions; see, e.g., [10, Th. 4.1] for a similar system. Solving this small linear system yields the coefficients in

$$\boldsymbol{X} = \boldsymbol{M} + \sum_{i=1}^{\ell} \sigma_i \boldsymbol{N}_i,$$

which generalizes the formula in Proposition 2.1. In general, the cost of solving this system remains moderate compared with all other computational costs as long as ℓ is significantly lower than n. Clearly, the solution uniqueness is related to the nonsingularity of $\boldsymbol{I} - \boldsymbol{F}$. A well known sufficient condition for the nonsingularity is that $\|\boldsymbol{F}\| < 1$ where $\|\cdot\|$ is any induced matrix norm.

2.1. An application to solid mechanics

Let $\mathcal{C}$ be an assigned positive definite symmetric linear map, associating $n \times n$ symmetric matrices with $n \times n$ symmetric matrices. The modeling of masonry-like materials calls for the computation of the projection of a symmetric matrix onto the cone of negative semidefinite symmetric matrices with respect to the inner product defined by $\mathcal{C}$. The map $\mathcal{C}$ contains the mechanical properties of the masonry material and can take different forms depending on the anisotropy of the material. When $\mathcal{C}$ models the elasticity tensor of an isotropic elastic material, it takes the form

$$\mathcal{C}(\boldsymbol{X}) = \frac{E}{1+\nu}\left(\boldsymbol{X} + \frac{\nu}{1-2\nu}\mathrm{trace}(\boldsymbol{X})\boldsymbol{I}\right), \tag{2.4}$$

where E is Young's modulus, $E > 0$, and ν is the Poisson ratio, satisfying $\nu \in (-1, 1/2)$. When, on the other hand, $\mathcal{C}$ represents a transversely isotropic elasticity tensor with respect to the direction $\mathbf{e}_3$, then it can be written as $\mathcal{C}(\boldsymbol{X}) = \sum_{i=1}^{\ell} \mathrm{trace}(\boldsymbol{H}_i\boldsymbol{X})\boldsymbol{K}_i$, for $\ell = n(n+1)/2$ and suitable symmetric matrices $\boldsymbol{H}_i, \boldsymbol{K}_i \in \mathbb{R}^{n\times n}$ for $i = 1, \ldots, \ell$ which depend on the scalars E and ν and on the spectral representation of $\mathcal{C}$ [7].

For a given symmetric matrix $\bar{\boldsymbol{Y}}$, in [8] the projection problem was reformulated as the following quadratic semidefinite programming problem

$$\min_{\boldsymbol{Y}\succeq 0} \quad \mathrm{trace}(\boldsymbol{Y}\mathcal{C}(\boldsymbol{Y}+\bar{\boldsymbol{Y}})) \tag{2.5}$$

and a primal-dual path-following interior point method was proposed. At each iteration of the interior-point method, one Newton step is computed for the following perturbed first-order optimality conditions for problem (2.5)

$$F_\mu(\boldsymbol{Y}, \boldsymbol{S}) = \begin{pmatrix} \boldsymbol{S} - \mathcal{C}(\boldsymbol{Y}+\bar{\boldsymbol{Y}}) \\ \boldsymbol{Y}\boldsymbol{S} - \mu\boldsymbol{I} \end{pmatrix} = \boldsymbol{0}, \qquad \boldsymbol{Y} \succ 0,\ \boldsymbol{S} \succ 0, \tag{2.6}$$

where the positive scalar μ is driven to zero as the method progresses. To ensure that the Newton steps produce symmetric matrices, different symmetrization schemes can be applied to the nonlinear equation $\boldsymbol{Y}\boldsymbol{S} - \mu\boldsymbol{I} = 0$ in (2.6): the popular Alizadeh-Haeberly-Overton (AHO) and Nesterov-Todd (NT) schemes have been explored in [8]. For a fixed $\mu > 0$ and given the current approximation $(\boldsymbol{Y}, \boldsymbol{S})$ of the solution of (2.6), let $\boldsymbol{X}$ denote the Newton step for the variable $\boldsymbol{Y}$. Consider first the AHO scheme: $\boldsymbol{X}$ solves the equation

$$\boldsymbol{S}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{S} + \mathcal{C}(\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{Y}\mathcal{C}(\boldsymbol{X}) = \boldsymbol{D}, \tag{2.7}$$

where the right-hand side $\boldsymbol{D} = 2\mu\boldsymbol{I} - (\boldsymbol{Y}\boldsymbol{S}+\boldsymbol{S}\boldsymbol{Y}) - (\boldsymbol{Y}(\mathcal{C}(\boldsymbol{Y}+\bar{\boldsymbol{Y}}) - \boldsymbol{S}) + (\mathcal{C}(\boldsymbol{Y}+\bar{\boldsymbol{Y}}) - \boldsymbol{S})\boldsymbol{Y})$ takes into account the value of the current $F_\mu(\boldsymbol{Y}, \boldsymbol{S})$ and the AHO symmetrization. When $\mathcal{C}$ is isotropic, inserting the form (2.4) into (2.7) yields

$$\left(\boldsymbol{S}+\frac{E}{1+\nu}\boldsymbol{Y}\right)\boldsymbol{X}+\boldsymbol{X}\left(\boldsymbol{S}+\frac{E}{1+\nu}\boldsymbol{Y}\right)+\text{trace}(\boldsymbol{X})\frac{\nu E}{(1+\nu)(1-2\nu)}\boldsymbol{Y}=\boldsymbol{D},$$

that corresponds to (1.1) with $\boldsymbol{A}=\boldsymbol{B}=\left(\boldsymbol{S}+\frac{E}{1+\nu}\boldsymbol{Y}\right)$ and $\boldsymbol{C}=\frac{\nu E}{(1+\nu)(1-2\nu)}\boldsymbol{Y}$. If $\mathcal{C}$ is transversely isotropic, the terms involving $\mathcal{C}$ in (2.7) are given by

$$\mathcal{C}(\boldsymbol{X})\boldsymbol{Y}+\boldsymbol{Y}\mathcal{C}(\boldsymbol{X})=\sum_{i=1}^{\ell}\text{trace}(\boldsymbol{H}_i\boldsymbol{X})(\boldsymbol{K}_i\boldsymbol{Y}+\boldsymbol{Y}\boldsymbol{K}_i)\equiv\sum_{i=1}^{\ell}f_i(\boldsymbol{X})\boldsymbol{C}_i,$$

yielding

$$\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}+\sum_{i=1}^{\ell}f_i(\boldsymbol{X})\boldsymbol{C}_i=\boldsymbol{D},$$

with $\boldsymbol{A}=\boldsymbol{S}$, which thus corresponds to (2.1).

In the case of the NT scheme, the Newton step solves the general equation $\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}+\mathcal{C}(\boldsymbol{X})=\boldsymbol{D}$ with $\boldsymbol{W}\succ 0$ being the geometric mean of $\boldsymbol{Y}^{-1}$ and $\boldsymbol{S}$, and $\boldsymbol{D}$ is suitably defined taking into account the residual $F_\mu(\boldsymbol{Y},\boldsymbol{S})$ and the NT scheme, see e.g. [13]. If $\mathcal{C}$ is isotropic, the equation above reads

$$\boldsymbol{W}\boldsymbol{X}\boldsymbol{W}+\frac{E}{1+\nu}\boldsymbol{X}+\frac{\nu E}{(1+\nu)(1-2\nu)}\text{trace}(\boldsymbol{X})\boldsymbol{I}=\boldsymbol{D}.$$

Multiplying by $\boldsymbol{W}^{-1}$,

$$\boldsymbol{X}\boldsymbol{W}+\frac{E}{1+\nu}\boldsymbol{W}^{-1}\boldsymbol{X}+\text{trace}(\boldsymbol{X})\frac{\nu E}{(1+\nu)(1-2\nu)}\boldsymbol{W}^{-1}=\boldsymbol{W}^{-1}\boldsymbol{D},$$

that is in the form (1.1) with $\boldsymbol{A}=\frac{E}{1+\nu}\boldsymbol{W}^{-1}$, $\boldsymbol{B}=\boldsymbol{W}$ and $\boldsymbol{C}=\frac{\nu E}{(1+\nu)(1-2\nu)}\boldsymbol{W}^{-1}$. Finally, for the transversely isotropic case one obtains the equation

$$\boldsymbol{X}=\boldsymbol{M}+\sum_{i=1}^{\ell}f_i(\boldsymbol{X})\boldsymbol{N}_i,\quad \boldsymbol{M}=\boldsymbol{W}^{-1}\boldsymbol{D}\boldsymbol{W}^{-1},\ \boldsymbol{N}_i=-\boldsymbol{W}^{-1}\boldsymbol{K}_i\boldsymbol{W}^{-1},$$

which has the form (2.2), with $f_i(\boldsymbol{X})=\text{trace}(\boldsymbol{H}_i\boldsymbol{X}), i=1,\dots,\ell$.

We remark that the explicit form of Newton step above within the NT scheme is a generalization of the formula given in [13, Lemma 5.1] for the case $\mathcal{C}(\boldsymbol{X})=\boldsymbol{K}\boldsymbol{X}\boldsymbol{K}$ and $\boldsymbol{K}\succeq 0$.

3. The trace of a matrix power

A first generalization to the nonlinear setting is given by the family of functions $f(\boldsymbol{X})=\text{trace}(\boldsymbol{X}^p)$, with $p\in\mathbb{N}$, $p>1$. For moderate p such as $p=2$, it is possible to give explicit solutions to the problem. We focus on the effect of f on the matrix equation, where we work with the form in (1.2).

Let $p = 2$. We have

$$\begin{aligned} f(\boldsymbol{X}) &= \mathrm{trace}(\boldsymbol{X}^2) = \mathrm{trace}((\boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N})(\boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N})) \\ &= \mathrm{trace}(\boldsymbol{M}^2) + 2\,\mathrm{trace}(\boldsymbol{M}\boldsymbol{N})f(\boldsymbol{X}) + f(\boldsymbol{X})^2\mathrm{trace}(\boldsymbol{N}^2) \\ &= f(\boldsymbol{M}) + 2\,\mathrm{trace}(\boldsymbol{M}\boldsymbol{N})f(\boldsymbol{X}) + f(\boldsymbol{X})^2 f(\boldsymbol{N}). \end{aligned}$$

Let $\beta = 2\,\mathrm{trace}(\boldsymbol{M}\boldsymbol{N}) - 1$. The equation above corresponds to the following (scalar) quadratic algebraic equation in the variable $r = f(\boldsymbol{X})$,

$$r^2 f(\boldsymbol{N}) + \beta r + f(\boldsymbol{M}) = 0.$$

If $f(\boldsymbol{N}) = 0$ and $\beta \neq 0$ then the solution is $r = -f(\boldsymbol{M})/\beta$, giving $\boldsymbol{X} = \boldsymbol{M} + r\boldsymbol{N}$. If $f(\boldsymbol{N}) \neq 0$ then the following two solutions are derived,

$$r_{1,2} = \frac{1}{2f(\boldsymbol{N})}\left(-\beta \pm \sqrt{\beta^2 - 4f(\boldsymbol{N})f(\boldsymbol{M})}\right).$$

The two final solution matrices $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}$ are obtained as

$$\boldsymbol{X}_{(1)} = \boldsymbol{M} + r_1\boldsymbol{N}, \qquad \boldsymbol{X}_{(2)} = \boldsymbol{M} + r_2\boldsymbol{N}.$$

For higher powers of $\boldsymbol{X}$, correspondingly larger degree scalar polynomial equations are obtained, from which the corresponding *numerical* solution matrices can be derived, in case the roots can only be computed numerically. The procedure may also yield complex (conjugate) values for r even for real data, from which complex (conjugate) solutions will follow.

Powers of affine functions can also be considered, such as $f(\boldsymbol{X}) = \mathrm{trace}((\boldsymbol{X} + \boldsymbol{H})^p)$, for a fixed matrix $\boldsymbol{H}$. A similar solution procedure can be devised for other, related functions such as the Frobenius norm, that is

$$f(\boldsymbol{X}) = \|\boldsymbol{X}\|_F^2 = \mathrm{trace}(\boldsymbol{X}^T\boldsymbol{X}).$$

A second generalization for which explicit solutions can be obtained under certain hypotheses is the function $f(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X}^{-1})$. We remark that these results apply to problem (1.2) and not necessarily to problem (1.1), as the hypothesis that either $\boldsymbol{M}$ or $\boldsymbol{N}$ is rank-one is unlikely to hold for $\boldsymbol{M}$ or $\boldsymbol{N}$ being the solution to Sylvester equations.

Proposition 3.1. *Let $\boldsymbol{M} = \boldsymbol{m}_1\boldsymbol{m}_2^T$ be a rank-one matrix and $\boldsymbol{N}$ be invertible. Let the nonlinear function be $f(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X}^{-1})$. If the matrix equation $\boldsymbol{X} = \boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N}$ admits nonsingular solutions, then these solutions are given as $X_{(i)} = \boldsymbol{M} + r_i\boldsymbol{N}$, $i = 1, 2, 3$ where r_i are the roots of the polynomial equation*

$$r^3 + \eta_2 r^2 + \eta_1 r + \eta_0 = 0,$$

with $\eta_2 = \boldsymbol{m}_2^T \boldsymbol{N}^{-1}\boldsymbol{m}_1$, $\eta_1 = -f(\boldsymbol{N})$ *and* $\eta_0 = \eta_1\eta_2 + \boldsymbol{m}_2^T \boldsymbol{N}^{-2}\boldsymbol{m}_1$.

Proof. We first note that if $\boldsymbol{X}$ is a nonsingular solution to the given equation, then $f(\boldsymbol{X}) \neq 0$ must hold, otherwise $\boldsymbol{X} = \boldsymbol{M}$ would not be invertible. Using the SMW formula we obtain

$$\begin{aligned}\boldsymbol{X}^{-1} &= (\boldsymbol{m}_1\boldsymbol{m}_2^T + f(\boldsymbol{X})\boldsymbol{N})^{-1}\\ &= \frac{1}{f(\boldsymbol{X})}\left(\boldsymbol{N}^{-1} - \boldsymbol{N}^{-1}\boldsymbol{m}_1(f(\boldsymbol{X}) + \boldsymbol{m}_2^T\boldsymbol{N}^{-1}\boldsymbol{m}_1)^{-1}\boldsymbol{m}_2^T\boldsymbol{N}^{-1}\right).\end{aligned}$$

Using $f(\boldsymbol{X}) = \text{trace}((\boldsymbol{m}_1\boldsymbol{m}_2^T + f(\boldsymbol{X})\boldsymbol{N})^{-1})$, we obtain

$$f(\boldsymbol{X}) = \frac{1}{f(\boldsymbol{X})}\left(f(\boldsymbol{N}) - (f(\boldsymbol{X}) + \boldsymbol{m}_2^T\boldsymbol{N}^{-1}\boldsymbol{m}_1)^{-1}\boldsymbol{m}_2^T\boldsymbol{N}^{-1}\boldsymbol{N}^{-1}\boldsymbol{m}_1\right).$$

Reordering terms, the third degree polynomial in $r = f(\boldsymbol{X})$ is obtained. □

A similar result can be obtained by exchanging the role of $\boldsymbol{M}$ and $\boldsymbol{N}$, that is, requiring that $\boldsymbol{N}$ is rank-one and $\boldsymbol{M}$ nonsingular, giving rise to at most two distinct solutions. More precisely, given the problem $\boldsymbol{X} = \boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{n}_1\boldsymbol{n}_2^T$, similar algebraic steps show that the solutions are given as $X_{(i)} = \boldsymbol{M} + r_i\boldsymbol{n}_1\boldsymbol{n}_2^T$, where r_i are the roots of the polynomial

$$\eta_2 r^2 + \eta_1 r + \eta_0 = 0,$$

with $\eta_0 = -f(\boldsymbol{M})$, $\eta_2 = \boldsymbol{n}_2^T\boldsymbol{M}^{-1}\boldsymbol{n}_1$ and $\eta_1 = 1 + \eta_0\eta_2 + \boldsymbol{n}_2^T\boldsymbol{M}^{-2}\boldsymbol{n}_1$.

If $\boldsymbol{M}$ or $\boldsymbol{N}$ have larger rank, then the procedure described above cannot be directly generalized.

Example 3.2. In Fig. 3.1 we give a general test code in Matlab [5] for the case of $f(\boldsymbol{X}) = \text{trace}(\boldsymbol{X}^{-1})$ and the solution formulas in Proposition 3.1 and the subsequent discussion. The left hand side implements the case of $\boldsymbol{M}$ rank-one, while the right-hand side refers to the case of $\boldsymbol{N}$ rank-one. The obtained computational results are

```
[1.5543e-15 5.0626e-14 2.4425e-15]     [8.3313e-16 8.3313e-16]
[9.0555e-15 2.9400e-13 1.4315e-14]     [7.4289e-15 7.4289e-15]
```

According to the code in Fig. 3.1, the first row of numbers shows the absolute error between the true value of $f(\boldsymbol{X}_i)$ and the computed ones, for both settings, as described in Fig. 3.1. The second set of numbers shows the corresponding matrix equation residual norms. At least for this setting, the procedure appears accurate. In general, the accuracy will be largely influenced by that of the root finder and by the computations with the inverses of the given matrices.

```
f=@(X)(trace(inv(X)));
n=10; rng(2)
%X=m1*m2’+f(X) N;
m1=randn(n,1); m2=randn(n,1);
N=randn(n,n);

t2=m2’/N*m1;
t1=-trace(inv(N));
t0=t1*t2+m2’/N^2*m1;

r=roots([1 t2 t1 t0]);

X1=m1*m2’+r(1)*N;
X2=m1*m2’+r(2)*N;
X3=m1*m2’+r(3)*N;

[norm(f(X1)-r(1)) norm(f(X2)-r(2)) norm(f(X3)-r(3))]

res=@(X)(norm(X-(m1*m2’+f(X)*N)));
[res(X1) res(X2) res(X3)]
```

```
rng(1)
%X=M+f(X) n1*n2’;
n1=randn(n,1); n2=randn(n,1);
M=randn(n,n);

f2=n2’/M*n1;
f1=1-f(M)*f2+ n2’/M^2*n1;
f0=-f(M);

r=roots([f2, f1, f0]);

X1=M+r(1)*n1*n2’;
X2=M+r(2)*n1*n2’;

[norm(f(X1)-r(1)), norm(f(X2)-r(2))]

res=@(X)(norm(X-(M+f(X)*n1*n2’)));
[res(X1) res(X2) ]
```

Fig. 3.1. Matlab code for Example 3.2.

4. The nonlinear case. Linear-nonlinear composition

As already shown in section 3 for matrix integer powers, the problem changes significantly from the linear setting in case the function f has the general form

$$f(\boldsymbol{X}) = \phi(\psi(\boldsymbol{X})), \quad \phi : \mathbb{R}^{n\times n} \to \mathbb{R}, \quad \psi : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n},$$

where ϕ is linear, and ψ is a (nonlinear) matrix function [4]. This is the case for instance for $f(\boldsymbol{X}) = \mathrm{trace}(\exp(-\boldsymbol{X}))$. We focus on the small size case, and use the form in (1.2) derived from (1.1) after the application of the inverse Sylvester operator.

Let us consider the case when $\phi(\boldsymbol{Y}) = \mathrm{trace}(\boldsymbol{Y})$, and assume that $\boldsymbol{N}$ is diagonalizable, so that $\boldsymbol{N} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$. Then (1.2) is equivalent to

$$\boldsymbol{Q}^{-1}\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q} + f(\boldsymbol{X})\boldsymbol{\Lambda}.$$

We then note that

$$f(\boldsymbol{X}) = \mathrm{trace}(\psi(\boldsymbol{X})) = \mathrm{trace}(\psi(\boldsymbol{Q}^{-1}\boldsymbol{X}\boldsymbol{Q})) = f(\boldsymbol{Q}^{-1}\boldsymbol{X}\boldsymbol{Q}),$$

as the trace is invariant under similarity transformations.

Let $\boldsymbol{X}_1 \equiv \boldsymbol{Q}^{-1}\boldsymbol{X}\boldsymbol{Q}$ and $\boldsymbol{M}_1 {\equiv} \boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}$, so that

$$\boldsymbol{X}_1 = \boldsymbol{M}_1 + f(\boldsymbol{X}_1)\boldsymbol{\Lambda}. \tag{4.1}$$

This form shows that the scalar value $f(\boldsymbol{X}_1)$ only appears in the diagonal elements of $\boldsymbol{X}_1$, while the off-diagonal part of $\boldsymbol{X}_1$ coincides with $\boldsymbol{M}_1$. In spite of this simple relation, it is hard to determine expressions for the solution in closed form for general matrix functions ψ, since computing trace$(\psi(\boldsymbol{X}_1))$ still involves the whole matrix $\boldsymbol{X}_1$, and the nonlinearity of ψ does not allow for algebraic simplifications. We are thus led to consider classical iterative schemes for solving (4.1).

Using the formulation in (4.1), starting with some $\boldsymbol{X}_1^{(0)}$, a fixed point iteration can be written as

$$\boldsymbol{X}_1^{(k+1)} = \boldsymbol{M}_1 + f(\boldsymbol{X}_1^{(k)})\boldsymbol{\Lambda}, \tag{4.2}$$

for $k \geq 0$, where, it is apparent that only the diagonal elements of $\boldsymbol{X}_1^{(k+1)}$ are updated at each iteration k, while the off-diagonal elements of $\boldsymbol{X}_1$ still coincide with those of $\boldsymbol{M}_1$ and they never change through the iteration. Combining two consecutive iterations, we obtain for $k \geq 1$,

$$\boldsymbol{X}_1^{(k+1)} = \boldsymbol{X}_1^{(k)} + (f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{(k-1)}))\boldsymbol{\Lambda},$$

which only updates the diagonal elements of the matrices. Hence, setting $\boldsymbol{\lambda} = \mathrm{diag}(\boldsymbol{\Lambda})$ where the function diag extracts the diagonal elements of a matrix, we can write

$$\mathrm{diag}(\boldsymbol{X}_1^{(k+1)}) = \mathrm{diag}(\boldsymbol{X}_1^{(k)}) + (f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{(k-1)}))\boldsymbol{\lambda}. \tag{4.3}$$

The final approximate solution is obtained as $\boldsymbol{X}^{(k)} = \boldsymbol{Q}\boldsymbol{X}_1^{(k)}\boldsymbol{Q}^{-1}$. We observe that for $\boldsymbol{N}$ non-symmetric, the conditioning of $\boldsymbol{Q}$ influences the error norm in the final approximate solution. More precisely, let $\boldsymbol{X}_1^\star$ be the exact solution to (4.1) and $\boldsymbol{X}^\star = \boldsymbol{Q}\boldsymbol{X}_1^\star\boldsymbol{Q}^{-1}$. Then $\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^\star\| \leq \|\boldsymbol{Q}\|\,\|\boldsymbol{Q}^{-1}\|\|\boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^\star\|$, so that the final $\boldsymbol{X}^{(k)}$ may be less accurate than the iteration in $\boldsymbol{X}_1^{(k)}$ would grant.

We first report on an algebraic characterization of the iteration, and then focus on error norm bounds for a selection of well known matrix functions. To this end, we will focus on the form (4.1), in which only the diagonal elements are modified by the iteration, when taking $\boldsymbol{X}_1^{(0)} = \boldsymbol{M}_1$. The same occurs for the error matrix.

Proposition 4.1. *Let* $\{\boldsymbol{X}_1^{(k)}\}_{k\geq 1}$ *be the sequence of iterates from (4.2), with* $\boldsymbol{M}_1 \succ 0$ *and* $\boldsymbol{\Lambda} \succeq 0$*, and let* $\boldsymbol{X}_1^{(0)} = \boldsymbol{M}_1$*.*

i) If f *is a nonnegative function satisfying* $f(\boldsymbol{X}) \leq f(\boldsymbol{Y})$ *for* $\boldsymbol{Y} \succeq \boldsymbol{X}$*, then* $\boldsymbol{X}_1^{(k+1)} \succeq \boldsymbol{X}_1^{(k)}$ *for all ks.*

ii) If f *is a nonnegative function satisfying* $f(\boldsymbol{X}) \geq f(\boldsymbol{Y})$ *for* $\boldsymbol{Y} \succeq \boldsymbol{X}$*, then the iterates* $\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{(k)}$ *alternate definiteness at each k.*

Proof. For $k = 0$, $\boldsymbol{X}_1^{(1)} = \boldsymbol{M}_1 + f(\boldsymbol{M}_1)\boldsymbol{\Lambda} \succeq \boldsymbol{M}_1 = \boldsymbol{X}_1^{(0)}$. For the subsequent iterates we have $\boldsymbol{X}_1^{(k+1)} = \boldsymbol{M}_1 + f(\boldsymbol{X}_1^{(k)})\boldsymbol{\Lambda}$ and $\boldsymbol{X}_1^{(k)} = \boldsymbol{M}_1 + f(\boldsymbol{X}_1^{(k-1)})\boldsymbol{\Lambda}$. Subtracting we obtain

$$\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{(k)} = (f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{(k-1)}))\boldsymbol{\Lambda}.$$

Hence, for any $k > 1$, if $f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{(k-1)}) > 0$ then $\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{(k)} \succeq 0$ because $\boldsymbol{\Lambda} \succeq 0$; this shows (i). For (ii), if $\boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^{(k-1)} \succeq 0$ then $(f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{(k-1)})) < 0$ so that $\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{(k)} \preceq 0$, and vice versa. □

We observe that the hypothesis that $\boldsymbol{X}_1^{(0)} = \boldsymbol{M}_1$ is used in the proof to ensure that $\boldsymbol{X}_1^{(1)} \succeq \boldsymbol{X}_1^{(0)}$, since $\boldsymbol{X}_1^{(1)} \succeq \boldsymbol{M}_1 = \boldsymbol{X}_1^{(0)}$. Nonetheless, if $\boldsymbol{X}_1^{(0)}$ is any symmetric matrix such that $\boldsymbol{M}_1 + f(\boldsymbol{X}_1^{(0)})\boldsymbol{\Lambda} \succeq \boldsymbol{X}_1^{(0)}$, the result will still hold.

Recalling relation (4.3), the definiteness explored in Proposition 4.1 refers to the way the diagonal elements of the iteration matrix change. In the first case, these elements grow monotonically as the iterations proceed; in case of convergence, the diagonal elements reach the final value from below. On the other hand, if f grows monotonically, the diagonal entries may showcase an alternating leapfrog behavior, which in case of convergence will terminate with the exact solution.

For instance, the function trace$(\boldsymbol{X}^{1/2})$ satisfies (i) while the function trace$(\exp(-\boldsymbol{X}))$ satisfies (ii). The different behavior is reported in Fig. 4.1 for the iteration in (4.3), with the data created below in Matlab [5]. The values correspond to the $(n/2, n/2)$ diagonal element, however all diagonal elements behave similarly, as they change by the same factor.

```
n=10; rng(1);randn(n,n);   %first rand matrix to avoid first terms in the sequence
N=randn(n,n); N=sqrtm(N'*N);                             N=rand(n,n); N=0.2*sqrtm(N'*N);
f=@(X)(trace(expm(-X)));                                 f=@(X)(trace(sqrtm(X)))
Xstar=2*n*randn(n,n); Xstar=sqrtm(Xstar'*Xstar);
M=Xstar-f(Xstar)*N;
```

In the case of the exponential, the true solution `Xstar` was constructed so as to ensure that trace$(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^{\star})) = \sigma < 1$. For this particular selection, $\sigma = 0.52215 < 1$. This property will be used in Theorem 4.3 as a sufficient condition on the exponential function to prove convergence of the iteration. We also remark that, although the starting guess was chosen to comply with the hypotheses of Proposition 4.1, a negative definite starting matrix would not do much harm in the case of the exponential. Indeed, assuming $0 \succeq \boldsymbol{X}_1^{(0)}$, it holds that $f(\boldsymbol{X}_1^{(0)}) > 0$ and $\boldsymbol{X}_1^{(1)} \succeq 0$, so that from then on the iteration generates only non-negative definite matrices. In our experiments, the iteration then converges to the same solution. In the case of the matrix square root, positive semi-definiteness is required to ensure real iterates.

As a first consideration for the convergence analysis of the iteration $\{\boldsymbol{X}^{(k)}\}_{k \geq 0}$, we first notice that

$$\boldsymbol{E}^{(k+1)} \equiv \boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{\star} = (f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^{\star}))\boldsymbol{\Lambda} \equiv \eta_k \boldsymbol{\Lambda}, \tag{4.4}$$

showing that the error is a scalar multiple of a constant, diagonal matrix, and only the scalar η_k changes with k. Moreover, for any matrix norm $\|\cdot\|$,

$$\|\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^{\star}\| = |f(\boldsymbol{X}_1^{\star}) - f(\boldsymbol{X}_1^{(k)})| \, \|\boldsymbol{\Lambda}\|$$

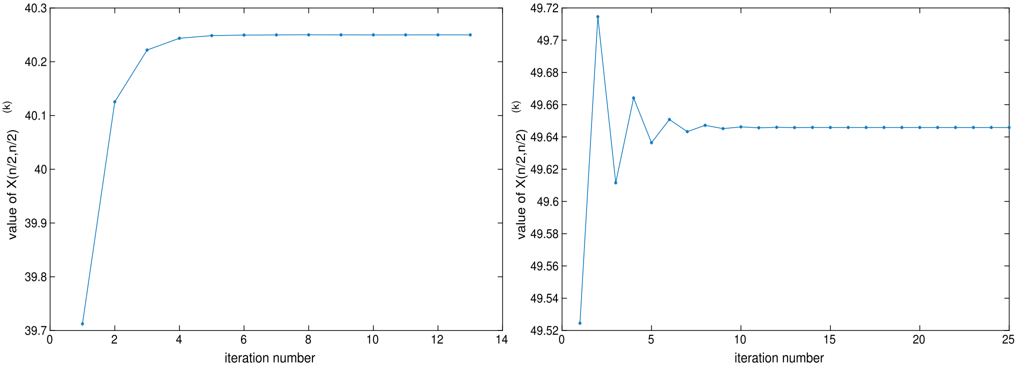

Fig. 4.1. Convergence of the $(n/2, n/2)$ diagonal element of $\boldsymbol{X}_1^{(k)}$. Left: $f(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X}^{1/2})$. Right: $f(\boldsymbol{X}) = \mathrm{trace}(\exp(-\boldsymbol{X}))$.

$$= \left(\frac{|f(\boldsymbol{X}_1^\star) - f(\boldsymbol{X}_1^{(k)})|}{\|\boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^\star\|} \|\boldsymbol{\Lambda}\| \right) \|\boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^\star\|.$$

Due to the linearity of the trace, the quotient in parentheses is closely related to the differential of the considered matrix function ψ. To make more precise statements about convergence we thus need to focus on specific examples of f. We will make use of the Frechet derivative of a matrix function ψ, defined as the linear function $L(\boldsymbol{X}, \boldsymbol{E})$ such that we have $\psi(\boldsymbol{X}+\boldsymbol{E}) - \psi(\boldsymbol{X}) = L(\boldsymbol{X}, \boldsymbol{E}) + o(\|\boldsymbol{E}\|)$ for $\|\boldsymbol{E}\|$ sufficiently small [4, section 3.1]. If the Frechet derivative of the given function ψ in $\boldsymbol{X}$ exists, then ψ is said to be Frechet differentiable at $\boldsymbol{X}$.

Lemma 4.2. *Let $f(\boldsymbol{X}) = \mathrm{trace}(\exp(-\boldsymbol{X}))$, and let $\{\boldsymbol{X}_1^{(k)}\}_{k\geq 0}$ be the sequence of iterates from (4.2), with $\boldsymbol{M}_1 \succeq 0$ and $\boldsymbol{\Lambda} \succeq 0$. If $\boldsymbol{X}_1^\star$ is a solution to (4.1) and we let $\boldsymbol{E}^{(k)} = \boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^\star$, then for $\boldsymbol{X}_1^{(k)}$ sufficiently close to $\boldsymbol{X}_1^\star$ we have*

$$\boldsymbol{E}^{(k+1)} = \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{E}^{(k)} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda}.$$

Proof. The matrix exponential is Frechet differentiable at any $\boldsymbol{X}$, and its Frechet derivative is given by $L(\boldsymbol{X}, \boldsymbol{E}) = \int_0^1 \exp(-\boldsymbol{X}(1-s))\boldsymbol{E}\exp(-\boldsymbol{X}s)ds$ [4, formula (10.15)].

The differential of f thus corresponds to $\mathrm{trace}(L(\boldsymbol{X}, \boldsymbol{E}))$, which can be written as

$$\begin{aligned} \mathrm{trace}(L(\boldsymbol{X}, \boldsymbol{E})) &= \int_0^1 \mathrm{trace}(\exp(-\boldsymbol{X}(1-s))\boldsymbol{E}\exp(-\boldsymbol{X}s))ds \\ &= \int_0^1 \mathrm{trace}(\boldsymbol{E}\exp(-\boldsymbol{X}s)\exp(-\boldsymbol{X}(1-s)))ds \\ &= \int_0^1 \mathrm{trace}(\boldsymbol{E}\exp(-\boldsymbol{X}))ds = \mathrm{trace}(\boldsymbol{E}\exp(-\boldsymbol{X})). \end{aligned}$$

Hence, for $\boldsymbol{X}_1^{(k)}$ sufficiently close to $\boldsymbol{X}_1^\star$, using (4.4) we have

$$\begin{aligned}\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^\star &= (f(\boldsymbol{X}_1^{(k)}) - f(\boldsymbol{X}_1^\star))\boldsymbol{\Lambda} \\ &= \mathrm{trace}(L(\boldsymbol{X}_1^\star, \boldsymbol{E}^{(k)}))\boldsymbol{\Lambda} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda} \\ &= \mathrm{trace}(\boldsymbol{E}^{(k)} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{\Lambda} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda} \\ &= \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\eta_{k-1}\boldsymbol{\Lambda} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda} \\ &= \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{E}^{(k)} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda},\end{aligned}$$

and the proof is complete. □

The above expression for the error allows us to give a sufficient condition for convergence. The proof follows the usual steps of the Ostrowski Theorem; see, e.g., [6, Th. 10.1.3].

Theorem 4.3. *Assume that the notation and hypotheses of Lemma 4.2 hold.*

If $\mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star)) = \sigma < 1$ *then there exist an* $\boldsymbol{X}_1^{(0)}$ *and a* $\sigma_1 \in [0, 1)$ *such that*

$$\|\boldsymbol{E}^{(k+1)}\| \leq \sigma_1 \|\boldsymbol{E}^{(k)}\|,$$

for $k \geq 0$, *for any matrix norm* $\|\cdot\|$.

Proof. We proceed by induction. Let $k = 0$. The differentiability of the exponential function and Lemma 4.2 ensure that for an arbitrary $\epsilon > 0$ there exists a $\boldsymbol{X}_1^{(0)}$ sufficiently close to $\boldsymbol{X}_1^\star$ such that

$$||\boldsymbol{E}^{(1)} - \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{E}^{(0)}\| \leq \epsilon \|\boldsymbol{E}^{(0)}\|,$$

so that

$$\begin{aligned}||\boldsymbol{E}^{(1)}\| &\leq ||\boldsymbol{E}^{(1)} - \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{E}^{(0)}\| + |\mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))| \, \|\boldsymbol{E}^{(0)}\| \\ &\leq (\epsilon + \sigma)\|\boldsymbol{E}^{(0)}\|.\end{aligned}$$

By taking ϵ so that $\sigma_1 = \epsilon + \sigma < 1$ the result follows for $k = 0$. Assuming now that the result holds for $\|\boldsymbol{E}^{(k)}\|$ with $k > 0$, we can write again

$$||\boldsymbol{E}^{(k+1)} - \mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))\boldsymbol{E}^{(k)}\| \leq \epsilon \|\boldsymbol{E}^{(k)}\|,$$

and proceed as for $k = 0$, to obtain the final bound. □

Example 4.4. We analyze the convergence of the fixed point iteration with respect to the condition of Theorem 4.3 on the derivative $\mathrm{trace}(\boldsymbol{\Lambda} \exp(-\boldsymbol{X}_1^\star))$. To this end, we consider $f(\boldsymbol{X}) = \mathrm{trace}(\exp(-\boldsymbol{X}))$ and the matrix $\boldsymbol{X}^\star = \sqrt{\alpha}\boldsymbol{G}$ with $\boldsymbol{G} = (\boldsymbol{G}_0^T \boldsymbol{G}_0)^{\frac{1}{2}}$,

Table 4.1
Example 4.4. $f(\boldsymbol{X}) = \mathrm{trace}(\exp(-\boldsymbol{X}))$. Behavior of the iteration (4.3) as α varies.

$\mathrm{trace}(\boldsymbol{\Lambda}\exp(-\boldsymbol{X}_1^\star))$	α	k	$\frac{\|\boldsymbol{X}^{(k+1)}-(\boldsymbol{M}+f(\boldsymbol{X}^{(k+1)})\boldsymbol{N})\|}{\|\boldsymbol{M}\|}$
0.079	12.589	3	8.3190e-08
0.176	10.000	6	3.4123e-08
0.335	7.9433	11	3.7944e-08
0.570	6.3096	23	6.9902e-08
0.889	5.0119	117	9.6324e-08
1.296	3.9811	500	3.5943e-01
1.789	3.1623	500	1.2832e+00

where $\boldsymbol{G}_0$ =`randn(n,n)` (Matlab seed `rng(1)`). By varying α a different magnitude of the Frechet derivative can be obtained. The matrix $\boldsymbol{N}$ is defined in the same way as $\boldsymbol{G}$, and $\boldsymbol{M} = \boldsymbol{X}^\star - f(\boldsymbol{X}^\star)\boldsymbol{N}$. In Table 4.1 we report the results of the fixed point iteration $\boldsymbol{X}_1^{(0)} = \boldsymbol{M}_1$, $\boldsymbol{X}_1^{(k+1)} = \boldsymbol{M}_1 + f(\boldsymbol{X}_1^{(k)})\boldsymbol{\Lambda}$, $k = 0, 1, \ldots$. The iteration stops either for $\|\boldsymbol{X}^{(k+1)} - (\boldsymbol{M} + f(\boldsymbol{X}^{(k+1)})\boldsymbol{N})\|/\|\boldsymbol{M}\| < 10^{-7}$ or for $k = 500$. The numbers in the table show lack of convergence as soon as the condition on the derivative fails, as is typical of Ostrowski type theorems.

We next derive similar results for a nonlinear function involving the matrix square root.

Theorem 4.5. *Let $f(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X}^{\frac{1}{2}})$, and let $\{\boldsymbol{X}_1^{(k)}\}_{k\geq 0}$ be the sequence of iterates from (4.1), with $\boldsymbol{M}_1 \succeq 0$ and $\boldsymbol{\Lambda} \succeq 0$, so that the exact solution $\boldsymbol{X}_1^\star$ to (4.1) is symmetric and positive definite. Let $\boldsymbol{E}^{(k)} = \boldsymbol{X}_1^{(k)} - \boldsymbol{X}_1^\star$. Then*

$$\boldsymbol{E}^{(k+1)} = \frac{1}{2}\mathrm{trace}(\boldsymbol{\Lambda}(\boldsymbol{X}_1^\star)^{-1})\boldsymbol{E}^{(k)} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda}.$$

If $\frac{1}{2}\mathrm{trace}(\boldsymbol{\Lambda}(\boldsymbol{X}_1^\star)^{-1}) = \sigma < 1$ then there exist an $\boldsymbol{X}_1^{(0)}$ and a $\sigma_1 \in [0, 1)$ such that

$$\|\boldsymbol{E}^{(k+1)}\| \leq \sigma_1\|\boldsymbol{E}^{(k)}\|,$$

for any matrix norm $\|\cdot\|$.

Proof. For the matrix square root we have $L(\boldsymbol{X}, \boldsymbol{E}) = \mathcal{L}_{\boldsymbol{X}}^{-1}(\boldsymbol{E})$, where $\mathcal{L}_{\boldsymbol{X}}$ is the linear operator $\mathcal{L}_{\boldsymbol{X}} : \boldsymbol{Z} \mapsto \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{X}$ [4, p.134]. Using Remark 2.2 and (4.4) we can write

$$\begin{aligned}\boldsymbol{X}_1^{(k+1)} - \boldsymbol{X}_1^\star &= \mathrm{trace}(\mathcal{L}_{\boldsymbol{X}_1^\star}^{-1}(\boldsymbol{E}^{(k)}))\boldsymbol{\Lambda} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda} \\ &= \frac{1}{2}\mathrm{trace}((\boldsymbol{X}_1^\star)^{-1}\boldsymbol{E}^{(k)})\boldsymbol{\Lambda} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda} \\ &= \frac{1}{2}\mathrm{trace}((\boldsymbol{X}_1^\star)^{-1}\boldsymbol{\Lambda})\boldsymbol{E}^{(k)} + o(\|\boldsymbol{E}^{(k)}\|)\boldsymbol{\Lambda},\end{aligned}$$

and the first result follows. The proof of the final bound follows the same lines as the corresponding bound in Theorem 4.3. □

5. The nonlinear case. Nonlinear-linear composition

The procedure described in Proposition 2.1 can be employed within a procedure for solving (1.1) when the nonlinear function has the form $f(\boldsymbol{X}) = g(h(\boldsymbol{X}))$ where $g : [\alpha, \beta] \to \mathbb{R}$ and h is a real valued *linear* function with image in $[\alpha, \beta]$. The problem becomes nonlinear in $\boldsymbol{X}$, hence uniqueness of the solution is in general not guaranteed.

To analyze the new setting, consider again equation (1.2), that is $\boldsymbol{X} = \boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N}$, and apply the linear function h to both sides,

$$h(\boldsymbol{X}) = h(\boldsymbol{M}) + f(\boldsymbol{X})h(\boldsymbol{N}). \tag{5.1}$$

For $\gamma_1 \equiv h(\boldsymbol{M})$, $\gamma_2 \equiv h(\boldsymbol{N})$ and setting $y \equiv h(\boldsymbol{X})$, the equation above corresponds to the nonlinear scalar equation

$$\gamma_1 + g(y)\gamma_2 - y = 0, \qquad y \in [\alpha, \beta]. \tag{5.2}$$

We next formalize the fact that if this equation has a solution y^* in the considered interval, then (5.2) yields a solution to (1.2).

Proposition 5.1. *With the previous notation, assume that y^* is a solution to (5.2) in $[\alpha, \beta]$. Then $\boldsymbol{X} \equiv \boldsymbol{M} + g(y^*)\boldsymbol{N}$ is a solution to (1.1) with $f = g \circ h$. If y^* is unique, then $\boldsymbol{X}$ is also the unique solution to (1.1).*

Proof. Let $\boldsymbol{X} \equiv \boldsymbol{M} + g(y^*)\boldsymbol{N}$. Applying the linear function h to both sides we obtain $h(\boldsymbol{X}) = h(\boldsymbol{M}) + g(y^*)h(\boldsymbol{N})$. We recall that $h(\boldsymbol{M}) + g(y^*)h(\boldsymbol{N}) = y^*$, therefore it must be that $h(\boldsymbol{X}) = y^*$, that is, $\boldsymbol{X} = \boldsymbol{M} + f(\boldsymbol{X})\boldsymbol{N}$, which is equivalent to (1.1). The uniqueness of y^* together with the fact that g is a proper (single-valued) function ensure the uniqueness of $\boldsymbol{X} = \boldsymbol{M} + g(y^*)\boldsymbol{N}$. □

The quantities γ_1, γ_2 play a crucial role in the existence of (at least) one solution to (5.2). In turn, these scalars depend on the eigenvalues of the two Sylvester solutions, and thus on $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$. We abstain from exploring all possible cases of the nonlinear scalar problem, as our focus is on the difficulties stemming from the matrix setting. Below we give a sample of theoretical and computational considerations that can be of help in solving the final problem, keeping in mind that several other strategies could be used.

To explore the influence of the data on the nonlinear scalar equation, we assume $\gamma_2 \neq 0$ and rewrite (5.2) as

$$g(y) = -\frac{\gamma_1}{\gamma_2} + \frac{1}{\gamma_2}y, \tag{5.3}$$

and set $\ell(y) \equiv -\frac{\gamma_1}{\gamma_2} + \frac{1}{\gamma_2}y$, where the function ℓ is linear and defined on the whole real line. Hence, y^* is a solution to (5.2) in $[\alpha, \beta]$ if and only if the two functions g and ℓ intersect (at y^*). For simplicity, let us assume that $[\alpha, \beta] \equiv \mathbb{R}$. If for instance g (resp. ℓ) is monotonically decreasing (resp. increasing) in $\mathbb{R}$, then y^* exists and is unique. This behavior depends on the choice of g, but also on the sign of γ_1 and γ_2, which in turn depends on the properties of the matrices $\boldsymbol{M}, \boldsymbol{N}$. Examining all possible combinations of these properties would be cumbersome. We provide here a typical setting. To make the treatment simpler, we assume that $h(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X})$. The general case $h(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{H}\boldsymbol{X})$ will also depend on the spectral and structural properties of the matrix $\boldsymbol{H}$.

Proposition 5.2. *Assume that $\boldsymbol{M}$ and $\boldsymbol{N}$ are symmetric and positive definite. Assume also that $h(\boldsymbol{X}) = \mathrm{trace}(\boldsymbol{X})$ and $g(y) \geq 0$, $g''(y) \geq 0$ for all $y \geq 0$, g at least C^2 and monotonically decreasing. Then the Newton iteration $\{y_k\}$ applied to $F(y) = 0$ with $F(y) = \gamma_1 + g(y)\gamma_2 - y$ will converge for all $y_0 \geq 0$.*

Proof. Note that the hypothesis on $\boldsymbol{M}, \boldsymbol{N}$ implies that γ_1, γ_2 are both positive real values. Moreover, the hypotheses on g also imply that F is at least C^2, $F'(y) < 0$ and $F''(y) > 0$ for all positive y, so that F is convex in $[0, +\infty)$. Moreover, $\lim_{y \to +\infty} F(y) = -\infty$. Since $F(0) > 0$, a zero y^* must exist. The tangent passing through $y = 0$ encounters the first coordinate axis at $y_1 = -F(0)/F'(0) > 0$. Convexity ensures that $y_1 < y^*$. The tangent passing through $y = b$ for some $b > y^*$ encounters the first coordinate axis at $y_1 = b - F(b)/F'(b) = (bg'(b)\gamma_2 - \gamma_1 - g(b)\gamma_2)/F'(b) > 0$ for all $b > 0$. Hence, on geometric grounds, it follows that the Newton iteration converges in any interval $[0, b]$ with $y^* \in [0, b]$; see, e.g., [2, Example p.276]. □

Example 5.3. Let $g(t) = \exp(-t)$, so that $f(\boldsymbol{X}) = \exp(-\mathrm{tr}(\boldsymbol{X}))$. Then (5.1) becomes $\gamma_1 - e^{-y}\gamma_2 - y = 0$, for $y \in \mathbb{R}$.

As an alternative to the Newton method, one can resort once again to a fixed point iteration. A natural choice, but not necessarily the best one, is given by

$$y^{(k+1)} = \gamma_1 + g(y^{(k)})\gamma_2 \equiv \Phi(y^{(k)}).$$

If a zero y^* exists such that $|\Phi'(y^*)| < 1$ then Ostrowski's theorem ensures that there exists an open interval centered in y^* such that this iteration will converge for any $y^{(0)}$ taken in this interval. Hence, the sufficient condition is that $|g'(y^*)\gamma_2| < 1$.

As an example, let use take $g(y) = \ln(y)$, for $y > 0$, so that $f(\boldsymbol{X}) = \ln(\mathrm{trace}(\boldsymbol{X}))$. Then $g'(y) = 1/y$ and $|\Phi'(y^*)| < 1$ as long as $y^* > \gamma_2$. The existence of y^* depends on whether the curves $g(y)$ and $\ell(y)$ intersect, and as discussed around (5.3), this depends on the mutual values of γ_1, γ_2.

6. Considerations on the large scale case

Problem (1.1) becomes computationally very challenging if the given matrices have large dimensions. Let $\widetilde{\boldsymbol{M}}, \widetilde{\boldsymbol{N}}$ be the approximations to the solutions $\boldsymbol{M}$ and $\boldsymbol{N}$ respectively, of the Sylvester equations. If f is linear, say $f(\boldsymbol{X}) = \text{trace}(\boldsymbol{X})$, then from Proposition 2.1 an approximate solution is obtained as

$$\widetilde{\boldsymbol{X}} \equiv \widetilde{\boldsymbol{M}} + \sigma \widetilde{\boldsymbol{N}}, \quad \sigma = \frac{f(\widetilde{\boldsymbol{M}})}{1 - f(\widetilde{\boldsymbol{N}})},$$

with a clear dependence of the error $\boldsymbol{X} - \widetilde{\boldsymbol{X}}$ on the error committed in approximating $\widetilde{\boldsymbol{M}}, \widetilde{\boldsymbol{N}}$.

For the approximation of $\boldsymbol{M}, \boldsymbol{N}$ different methods can be considered, especially in case the right-hand sides $\boldsymbol{D}$ and $\boldsymbol{C}$ have low rank [12]; see also [9] for the sparse setting. Structural or sparsity properties are in fact a crucial hypothesis to be able to store $\widetilde{\boldsymbol{M}}, \widetilde{\boldsymbol{N}}$ and thus $\widetilde{\boldsymbol{X}}$ in a memory saving, factored format. Evaluating the trace can also profit from a factored form. If iterative methods are used to determine $\widetilde{\boldsymbol{M}}, \widetilde{\boldsymbol{N}}$ [12], then the same type of strategy could be applied directly to (1.1), so that the residual can be monitored explicitly. For the approximation $\widetilde{\boldsymbol{X}}$ the associated residual is

$$\boldsymbol{R} = \boldsymbol{A}\widetilde{\boldsymbol{X}} + \widetilde{\boldsymbol{X}}\boldsymbol{B} + f(\widetilde{\boldsymbol{X}})\boldsymbol{C} - \boldsymbol{D},$$

which yields the following relation with the error matrix $\boldsymbol{E} \equiv \widetilde{\boldsymbol{X}} - \boldsymbol{X}^\star$,

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{B} + (f(\widetilde{\boldsymbol{X}}) - f(\boldsymbol{X}^\star))\boldsymbol{C}.$$

If f is linear, then $f(\widetilde{\boldsymbol{X}}) - f(\boldsymbol{X}^\star) = f(\boldsymbol{E})$, hence it follows

$$\boldsymbol{E} = \mathcal{L}^{-1}(\boldsymbol{R}) + f(\boldsymbol{E})\boldsymbol{N},$$

which is the natural (linear) generalization of the known expression for the error matrix in terms of the residual in linear algebraic equations.

Dealing with a nonlinear-linear f is similar to the linear case, since the matrix $h(\boldsymbol{X})$ is a scalar, after which the nonlinear function acts as in section 5. The linear-nonlinear case analyzed in section 4 with large matrices is far more complicated. Assuming that the problem to be solved can again be written as $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{M}} + f(\widetilde{\boldsymbol{X}})\widetilde{\boldsymbol{N}}$, a fixed-point iteration could be considered, possibly taking into account memory saving representations of $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{M}}$ and $\widetilde{\boldsymbol{N}}$, that is

$$\widetilde{\boldsymbol{X}}^{(k+1)} \equiv \widetilde{\boldsymbol{M}} + f(\widetilde{\boldsymbol{X}}^{(k)})\widetilde{\boldsymbol{N}}.$$

However, how to *approximate* $f(\widetilde{\boldsymbol{X}}^{(k)})$ remains complicated. Consider for instance $f(\widetilde{\boldsymbol{X}}) = \text{trace}(\psi(\widetilde{\boldsymbol{X}}))$. The approximation of this function is a problem in its own, and different, mostly iterative, approaches have been devised. This will give rise to an inner-outer procedure for the fixed point scheme above. Now popular choices for approximating trace$(\psi(\widetilde{\boldsymbol{X}}))$ include randomized, Monte-Carlo and probing methods, which replace the trace computation with the product $\mathbf{z}_k{}^T\psi(\boldsymbol{X})\mathbf{z}_k$ for a selection of vectors $\{\mathbf{z}_k\}$; see, e.g., [1], [15] and their references. Since in general we cannot expect high accuracy in this computation at each iteration, the quality of the outer iteration may be considerably affected. A detailed analysis and experimental study of these approaches is left for future research.

7. Conclusions

We have analyzed a new class of quasi-linear matrix equations, devising solutions in closed form for the linear case. In the quasi-linear framework, we have proposed numerical methods and theoretically studied their convergence under hypotheses that are satisfied for a wide class of problem data. The large scale problem remains particularly challenging, especially when involving the computation of matrix functions, for which further work is required.

Declaration of competing interest

No conflict of interest exists.

Data availability

No data was used for the research described in the article.

Acknowledgements

We thank two anonymous reviewers for their insightful remarks.

References

[1] A. Cortinovis, D. Kressner, On randomized trace estimates for indefinite matrices with an application to determinants, Found. Comput. Math. 22 (2022) 875–903.
[2] W. Gautschi, Numerical Analysis, second ed., Birkhäuser, Boston, 2012.
[3] Y. Hao, V. Simoncini, The Sherman-Morrison-Woodbury formula for generalized linear matrix equations and applications, Numer. Linear Algebra Appl. 28 (2021) e2384.
[4] N.J. Higham, Functions of Matrices - Theory and Computation, SIAM, Philadelphia, USA, 2008.
[5] Inc The MathWorks, MATLAB 7, r2020b ed., 2020.
[6] J. Ortega, W. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Classics in Applied Mathematics, SIAM, Philadelphia, USA, 2000.
[7] C. Padovani, Strong ellipticity of transversely isotropic elasticity tensors, Meccanica 37 (2002) 515–525.
[8] C. Padovani, M. Porcelli, A semidefinite programming approach for the projection onto the cone of negative semidefinite symmetric tensors with applications to solid mechanics, Calcolo 59 (2022) 1–31.

[9] D. Palitta, V. Simoncini, Numerical methods for large-scale Lyapunov equations with symmetric banded data, SIAM J. Sci. Comput. 40 (2018) A3581–A3608.

[10] E. Ringh, G. Mele, J. Karlsson, E. Jarlebring, Sylvester-based preconditioning for the waveguide eigenvalue problem, Linear Algebra Appl. 542 (2018) 441–463.

[11] S. Savov, I. Popchev, New generalized upper trace bound for the solution of the Lyapunov equation, Int. J. Pure Appl. Math. 49 (3) (2008) 381–389.

[12] V. Simoncini, Computational methods for linear matrix equations, SIAM Rev. 58 (2016) 377–441.

[13] M. Todd, R. Tütüncü, K. Toh, Inexact primal-dual path-following algorithms for a special class of convex quadratic SDP and related problems, Pac. J. Optim. 3 (2007) 135–164.

[14] N. Truhar, K. Veselić, Bounds on the trace of a solution to the Lyapunov equation with a general stable matrix, Syst. Control Lett. 56 (2007) 493–503.

[15] S. Ubaru, J. Chen, Y. Saad, Fast estimation of tr(f(A)) via stochastic Lanczos quadrature, SIAM J. Matrix Anal. Appl. 38 (2017) 1075–1099.

[16] S.-D. Wang, T.-S. Kuo, C.-F. Hsu, Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equation, IEEE Trans. Autom. Control AC-31 (1986).