SIAM J. Sci. Comput.
Vol. 43, No. 5, pp. S448–S474

NONLINEAR POWER-LIKE AND SVD-LIKE ITERATIVE SCHEMES WITH APPLICATIONS TO ENTANGLED BIPARTITE RANK-1 APPROXIMATION[*]

MOODY T. CHU[†] AND MATTHEW M. LIN[‡]

Abstract. Gauging the distance between a mixed state and the convex set of separable states in a bipartite quantum mechanical system over the complex field is an important but challenging task. As a first step toward this difficult problem, this paper investigates the rank-1 approximation of a bipartite system over the real field where the entanglement is characterized in terms of the Kronecker product of density matrices. The approximation is recast in the form of a nonlinear eigenvalue problem and a nonlinear singular value problem for which two iterative methods are proposed, respectively. This study offers insight into and might serve as the building block for the more complicated multipartite systems and higher-rank approximation problems. The main focus is on the convergence analysis. Numerical experiments seem to suggest that these easily constructed solvers have higher efficiency when comparing with some state-of-the-art optimization techniques.

Key words. entanglement, separability, bipartite system, low-rank approximation, nonlinear eigenvalue problem, nonlinear singular value problem

AMS subject classifications. 65F10, 15A24, 65H10, 15A72, 58D19

DOI. 10.1137/20M1336059

1. Introduction.

Entanglement manifested in a system proves intricate but critical and is perhaps the most basic mode for characterizing a complicated phenomenon that involves components interacting with each other. Entanglement arises in nature and in almost all areas of disciplines whenever constituents, factors, parts, or subgroups interrelate with each other within the system.

Depending on how the parts in a system engage with each other, entanglement appears in different forms. It could be as simple as a few matrix multiplications if the rule of engagement is merely the causal nexus. In a Markov chain with memory where the evolution of states and memory must respect the Kolmogorov axioms, the entanglement involves more complicated tensor-tensor multiplications [38, 64]. Quantum entanglement, where particles in a composite quantum system generate, interact, or share properties in ways such that the variation of quantum properties of one particle will instantly change properties of another particle regardless of the distance, is another particularly interesting phenomenon with significant importance [22, 31]. Upon properly representing quantum states in terms of some suitable basis over the complex field, the rule of engagement in a bipartite quantum mechanics system can be cast as Kronecker products between density matrices of the subsystems. Our work in this paper concerns this kind of entanglement.

[*] Received by the editors May 8, 2020; accepted for publication (in revised form) April 12, 2021; published electronically June 24, 2021.

https://doi.org/10.1137/20M1336059

Funding: The work of the first author was partially supported by the National Science Foundation under grants DMS-1316779 and DMS-1912816. The work of the second author was partially supported by the National Center for Theoretical Sciences of Taiwan and by the Ministry of Science and Technology of Taiwan under grant MOST 110-2636-M-006-006.

[†] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 USA (chu@math.ncsu.edu).

[‡] Department of Mathematics, National Cheng Kung University, Tainan 701, Taiwan (matt.mlin@gmail.com).

Quantum entanglement plays an increasingly important role in modern quantum technologies. The extended lists of references in [22, 31] evince the vast research endeavors in the field. We mention quantum informatics [5, 21, 29, 44, 51, 55, 62] and quantum communication [3, 8, 25, 32] to exemplify applications that exploit the entanglement for faster and more secure delivery of information than classical algorithms. For better control of the underlying system, it is therefore of paramount importance to quantify the amount of entanglement. Such an undertaking, however, is known to be NP hard [24, 28]. Even so, there is one basic problem that can be used as a building block to carry out more advanced calculations. The purpose of this work is to investigate the mathematical formulation of this basic problem and its convergence analysis.

Without delving into details, it might be informative to outline some quantum mechanics background to motivate our problem. Even if only briefly, the following introduction might appear long but should be helpful in grasping the essential concepts. For a more formal and in-depth reading of the main ideas, we suggest [1, 30, 46] and the classic book [48]. Readers who are familiar with quantum mechanics and the notion of entanglement might skip to the next section immediately.

1.1. Density matrix. One of the basic postulates in quantum mechanics is that each quantum mechanical system is associated with a complex Hilbert space $\mathscr{H}$. Any unit vector in $\mathscr{H}$ is referred to as a pure state which typically is denoted by the Dirac's ket notation $|\mathbf{x}\rangle$. Two pure states $|\mathbf{x}\rangle$ and $|\mathbf{y}\rangle$ are considered equivalent if $|\mathbf{x}\rangle = c\,|\mathbf{y}\rangle$ for some $|c| = 1$. The inner product of two unit vectors $|\mathbf{x}\rangle, |\mathbf{y}\rangle \in \mathscr{H}$ is denoted as $\langle\mathbf{x}|\mathbf{y}\rangle$. The orthogonal projection $|\mathbf{x}\rangle\langle\mathbf{x}|\mathbf{z}\rangle$ of any $|\mathbf{z}\rangle \in \mathscr{H}$ onto a given pure state $|\mathbf{x}\rangle$ is an operator of significant importance. Such an operator $\mathcal{D} := |\mathbf{x}\rangle\langle\mathbf{x}|$ is called a density matrix. Phase equivalent pure states have the same density matrix. A mixed quantum state is a probabilistic ensemble of finitely many pure states. Since a mixed state can not always be described by a single ket vector, it is more convenient to describe a general state ρ as the probabilistic mixture

$$\rho := \sum_i \mu_i\,|\mathbf{x}_i\rangle\langle\mathbf{x}_i|\,; \quad \sum_i \mu_i = 1; \quad \mu_i \geq 0, \tag{1.1}$$

of the density matrices of some pure states $|\mathbf{x}_i\rangle \in \mathscr{H}$. Therefore, the density matrix ρ is a positive semidefinite operator with unit trace.

1.2. Bipartite system. Given two Hilbert spaces $\mathscr{H}_1$ and $\mathscr{H}_2$, the tensor product space is defined to be the set

$$\mathscr{H}_1 \otimes \mathscr{H}_2 := \left\{ \sum_{s,t} \mathbf{u}_s \otimes \mathbf{v}_t | \mathbf{u}_s \in \mathscr{H}_1, \mathbf{v}_t \in \mathscr{H}_2 \right\}, \tag{1.2}$$

where the summation is formal over any index subset with finite support and the symbol $\otimes$ represents a notional linkage between states from $\mathscr{H}_1$ and $\mathscr{H}_2$. The only property required of $\otimes$ is its bi-linearity. We stress the formal double summation because we have not limited the Hilbert spaces to finite dimension yet. We further stress the bilinearity because, in the physics world, the two quantum mechanical systems do not even have any shared relationship or common features. The bilinearity is to emphasize that each space contributes to the mixture linearly. This is not to be confused with the conjugate linearity required in the inner product of a complex Hilbert space. An inner product can be induced via the relationship

$$\langle\mathbf{x}\otimes\mathbf{y}|\mathbf{z}\otimes\mathbf{w}\rangle := \langle\mathbf{x}|\mathbf{z}\rangle\langle\mathbf{y}|\mathbf{w}\rangle\,. \tag{1.3}$$

Upon the completion (which is not needed over finite dimensional spaces), we may assume that $\mathscr{H}_1 \otimes \mathscr{H}_2$ is a Hilbert space. In this way, we obtain the state space of a bipartite system.

For simplicity, suppose $\mathscr{H}_1$ and $\mathscr{H}_2$ are finite dimensional quantum mechanical systems with orthonormal basis[1] states $\{\mathbf{e}_i\}_{i=1}^m$ and $\{\mathbf{f}_j\}_{j=1}^n$, respectively. Then, a natural orthonormal basis for $\mathscr{H}_1 \otimes \mathscr{H}_2$ is $\{\mathbf{e}_i \otimes \mathbf{f}_j\}$. We shall enumerate the basis in lexicographical order, i.e., $\mathbf{e}_1 \otimes \mathbf{f}_1, \mathbf{e}_1 \otimes \mathbf{f}_2, \ldots, \mathbf{e}_2 \otimes \mathbf{f}_1, \mathbf{e}_2 \otimes \mathbf{f}_2, \ldots, \mathbf{e}_m \otimes \mathbf{f}_n$. Elements in $\mathscr{H}_1$ and $\mathscr{H}_2$ can be interpreted as merely column vectors $\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{y} \in \mathbb{C}^n$ of their coordinates in terms of the bases, respectively, whereas an element in the bipartite system $\mathscr{H}_1 \otimes \mathscr{H}_2$ can be represented by a matrix[2] $C \in \mathbb{C}^{m\times n}$. In particular, if $\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{y} \in \mathbb{C}^n$, then the bilinear map among the coefficient vectors is equivalent to

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}^\top, \tag{1.4}$$

which is known as the outer product in the linear algebra literature. To distinguish (1.4) from the Kronecker product, usually the notation $\mathbf{x} \circ \mathbf{y}$ is preferred for the outer product.

A pure state $|C\rangle$ in $\mathscr{H}_1 \otimes \mathscr{H}_2$ is such that its matrix representation $C \in \mathbb{C}^{m\times n}$ has unit Frobenius norm. By the notion defined in (1.1), a mixed state ρ over $\mathscr{H}_1 \otimes \mathscr{H}_2$ should be a density matrix of the form

$$\rho = \sum_i \mu_i \,|C_i\rangle \langle C_i| \,; \quad \sum_i \mu_i = 1; \quad \mu_i \geq 0, \tag{1.5}$$

where each $|C_i\rangle$ represents a pure state in $\mathscr{H}_1 \otimes \mathscr{H}_2$. Since each $|C_i\rangle \langle C_i|$ is an operator acting on matrices in $\mathbb{C}^{m\times n}$, we can interpret ρ as an order-4 tensor represented by an $mn \times mn$ matrix.

Example 1. Consider the case $\mathscr{H}_i = \mathbb{C}^2$, $i = 1, 2$, where the standard basis is typically denoted by $|0\rangle = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and $|1\rangle = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. In quantum formalism, a tensor product $|\uparrow\rangle \otimes |\downarrow\rangle$ is often abbreviated as $|\uparrow\downarrow\rangle$. A natural basis for the tensor product space $\mathbb{C}^2 \otimes \mathbb{C}^2$ is

$$\{|00\rangle\,, |01\rangle\,, |10\rangle\,, |11\rangle\},$$

whose corresponding matrix representations by (1.4) are

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

respectively. In quantum information science, however, a more commonly used basis is the Bell states [4, 9, 48]

$$\begin{cases} \left|\Phi^+\right\rangle := \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle), \\ \left|\Phi^-\right\rangle := \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle), \\ \left|\Psi^+\right\rangle := \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle), \\ \left|\Psi^-\right\rangle := \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle), \end{cases}$$

[1] The notion remains true over infinite dimensional Hilbert spaces with countable orthonormal bases. In that case, the elements in $\mathscr{H}_1 \otimes \mathscr{H}_2$ can be represent by semi-infinite matrices. Recall that a Hilbert space is said to be separable if and only if it has a countable orthonormal basis. That topological "separability" is entirely different from the separability considered in the context of quantum mechanics.

[2] For beginners, it might be easier to regard C as a column vector $\mathbf{vec}(C) \in \mathbb{C}^{mn}$ and consider the Hilbert space $\mathscr{H} = \mathscr{H}_1 \otimes \mathscr{H}_2$ as a monopartite system, whence the notions in section 1.1 can be carried over.

representing the simplest example of (maximally) quantum entanglement whose notion will be explained in what follows. Note that the Bell states form an orthonormal basis. In particular, they are pure states in $\mathbb{C}^2 \otimes \mathbb{C}^2$. In terms of the natural basis, the matrix representations of $\left|\Phi^+\right\rangle$, $\left|\Phi^-\right\rangle$, $\left|\Psi^+\right\rangle$, and $\left|\Psi^-\right\rangle$ are respectively given by

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

whereas the corresponding density matrices $\rho_{|\Phi^+\rangle} = \left|\Phi^+\right\rangle\left\langle\Phi^+\right|$ and so on should be expressed respectively as

$$\tfrac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \tfrac{1}{2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix},$$

$$\tfrac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \tfrac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

If a pure state $|\boldsymbol{\psi}\rangle \in \mathscr{H}_1 \otimes \mathscr{H}_2$ can be expressed as

$$|\boldsymbol{\psi}\rangle = |\boldsymbol{\psi}_1\rangle \otimes |\boldsymbol{\psi}_2\rangle\,, \tag{1.6}$$

where $|\boldsymbol{\psi}_i\rangle \in \mathscr{H}_i$, $i = 1, 2$, are pure states, respectively, then we say that the pure state $|\boldsymbol{\psi}\rangle$ is separable; otherwise, it is said to be entangled. A pure state in the composite system can be entangled. For example, simple arithmetic shows that any of the Bell states cannot be expressed as the tensor product of two pure states in $\mathbb{C}^2$. However, a pure state can always be decomposed as a linear combination of separable states in the following way, known as the Schmidt decomposition in quantum mechanics [20], which is also readily recognizable as the singular value decomposition (SVD) from the linear algebra viewpoint.

LEMMA 1.1. *Any pure state $|\boldsymbol{\psi}\rangle \in \mathscr{H}_1 \otimes \mathscr{H}_2$ can be written in the form*

$$|\boldsymbol{\psi}\rangle = \sum_j \sigma_j \left|\mathbf{u}_j\right\rangle \otimes \left|\mathbf{v}_j\right\rangle, \tag{1.7}$$

where $\left|\mathbf{u}_j\right\rangle \in \mathscr{H}_1$ and $\left|\mathbf{v}_j\right\rangle \in \mathscr{H}_2$ are orthonormal vectors, $\sigma_j \geq 0$, and $\sum_j \sigma_j^2 = 1$.

From the linear algebra point of view, especially when it is over finite dimensional spaces, the definition of a density matrix ρ in the form (1.5) and the decomposition of a pure state $|\boldsymbol{\psi}\rangle$ in the form of (1.7) are nothing but the spectral decomposition and SVD, respectively. There are well-developed numerical algorithms for handling these types of decompositions which, thus, do not impose computational difficulties. The real challenge is at the separability of the density matrices, which we describe below.

1.3. Entanglement. Given linear operators $\mathcal{A} : \mathscr{H}_1 \to \mathscr{H}_1$ and $\mathcal{B} : \mathscr{H}_2 \to \mathscr{H}_2$, there is a unique linear operation $\mathcal{T} : \mathscr{H}_1 \otimes \mathscr{H}_2 \to \mathscr{H}_1 \otimes \mathscr{H}_2$ such that [30]

$$\mathcal{T}(\mathbf{e}_i \otimes \mathbf{f}_j) := (\mathcal{A}\mathbf{e}_i) \otimes (\mathcal{B}\mathbf{f}_j). \tag{1.8}$$

If the basis $\{\mathbf{e}_i \otimes \mathbf{f}_j\}$ of $\mathscr{H}_1 \otimes \mathscr{H}_2$ is ordered lexicographically, then the matrix representation of $\mathcal{T}$ is precisely the Kronecker product of the matrix representations of

$\mathcal{A}$ and $\mathcal{B}$. For this reason, we denote $\mathcal{T} = \mathcal{A} \otimes \mathcal{B}$ to stress that $\mathcal{T}$ can be split as the tensor product of $\mathcal{A}$ and $\mathcal{B}$ in the sense of (1.8).

Different from the representation (1.5) where a mixed state density matrix ρ is always a statistical ensemble of density matrices of pure states in $\mathscr{H}_1 \otimes \mathscr{H}_2$, a more intriguing but difficult question is to determine whether a given density matrix ρ over $\mathscr{H}_1 \otimes \mathscr{H}_2$ can be decomposed as

$$\rho = \sum_k \eta_k \mathcal{D}_k^{(1)} \otimes \mathcal{D}_k^{(2)}, \quad \sum_k \eta_k = 1, \quad \eta_k \geq 0, \tag{1.9}$$

where $\{\mathcal{D}_k^{(1)}\}$ and $\{\mathcal{D}_k^{(2)}\}$ are density matrices of the subsystems $\mathscr{H}_1$ and $\mathscr{H}_2$, respectively. Note that the tensor product $\mathcal{D}_k^{(1)} \otimes \mathcal{D}_k^{(2)}$ is an operator in the sense of (1.8). Note also that, by definition, each $\mathcal{D}_k^{(1)}$ or $\mathcal{D}_k^{(2)}$ is itself a probabilistic ensemble of pure states in the form (1.1) but involves perhaps different numbers of terms and states. Upon further regrouping and relabeling, we may rewrite the expression (1.9) in the form

$$\rho = \sum_\ell \theta_\ell (|\mathbf{x}_\ell\rangle \langle \mathbf{x}_\ell|) \otimes (|\mathbf{y}_\ell\rangle \langle \mathbf{y}_\ell|), \tag{1.10}$$

where $\mathbf{x}_\ell \in \mathscr{H}_1$ and $\mathbf{y}_\ell \in \mathscr{H}_2$ are unit vectors, $\theta_\ell \geq 0$ and $\sum_\ell \theta_\ell = 1$. That is, a density matrix ρ over the bipartite space is separable if and only if it is the convex combination of tensor products of density matrices of pure states.

Entanglement detection and certification have been a subject attracting enormous research endeavors. The literature is very rich. We mention only [10, 12, 33, 35, 54, 55] that are appealing to our work. The three review articles [22, 27, 31] contain massive collection of references. One necessary condition for separability that can conveniently be checked by existent linear algebra techniques is the so-called realignment method [10] described below.

Lemma 1.2. *Given a density matrix $\rho \in \mathbb{C}^{mn \times mn}$, let $\mathscr{R}(\rho) \in \mathbb{C}^{m^2 \times n^2}$ denote the $\mathscr{R}$-folding*[3] *of ρ* [58, 59]. *If ρ is separable in the sense of* (1.10), *then necessarily the Ky Fan norm, i.e., the sum of all singular values of $\mathscr{R}(\rho)$, is less than* 1.

Example 2. We have already seen that the Bell state Φ^+ is not separable in the sense of (1.6). Its density matrix $\rho_{|\Phi^+\rangle}$, on one hand, is defined via (1.5) as a probabilistic mixture of density matrices of some pure states $|C_i\rangle$ (itself this time) in $\mathbb{C}^2 \otimes \mathbb{C}^2$ and is a 4×4 matrix. The $\mathscr{R}$-folding of the density matrix $\rho_{|\Phi^+\rangle}$ is $\frac{1}{2} I_4$ whose Ky Fan norm is 2 since all singular values are 1. By Lemma 1.2, $\rho_{|\Phi^+\rangle}$ is not separable in the sense of (1.10). Similar arguments can be applied to show that none of $\rho_{|\Phi^-\rangle}$, $\rho_{|\Psi^+\rangle}$, and $\rho_{|\Psi^-\rangle}$ is separable.

1.4. Approximation. If ρ is not separable, then seeking its nearest separable approximation is a problem of practical importance [16, 31, 42, 56]. Because the entanglement qualification depends on different operational paradigms and mathematical techniques, various metrics for gauging the entanglement have been proposed [11]. For example, the trace metric

$$D_T(\rho, \sigma) := \frac{1}{2} \mathrm{Tr} \sqrt{(\rho - \sigma)^2}$$

[3] Also defined in (2.2) in this paper.

is often employed to measure the maximum probability of distinguishability between two quantum states ρ and σ. It is the quantum version of the well-known Kolmogorov–Smirnov test for comparing random samples. The Bures distance

$$D_B(\rho,\sigma) := \sqrt{2 - 2\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}}$$

is used for parameter estimation of mixed quantum states based on repeated measurements just as the Fisher information is used in classical statistics. It allows the calculation of the minimum number of measurements to distinguish two different states. In our work, we measure the Frobenius norm

$$D_F(\rho,\sigma) = \frac{1}{2}\|\rho-\sigma\|_F = \frac{1}{2}\sqrt{\mathrm{Tr}(\rho-\sigma)^2}\,.$$

Not all distance formulas are easy to use for numerical computation. Taking the positive square root of a positive definite matrix, for example, is expensive, especially when it needs to be done repeatedly. There are theoretical discussions on computing the Bures formula without any diagonalization procedures [19]. Still, we find it difficult to implement numerically, e.g., it is hard to calculate the gradients of D_T and D_B. In contrast, using the Frobenius norm is perhaps the most convenient way since the square root is for scalars and the gradient of D_F is readily available by calculus. Different choices of metrics might lead to a different approximation result and the associated interpretation. A numerical comparison of various measures is worthy of further investigation but is beyond the scope of this paper. As a starter, our convergence analysis is based on the Frobenius norm.

Using the Frobenius norm, the proximity of a mixed state ρ to the convex set of separable states is estimated by [11, 43, 47, 50]

$$\min_{\substack{\mathbf{x}_\ell\in\mathbb{C}^m,\|\mathbf{x}_\ell\|_2=1\\ \mathbf{y}_\ell\in\mathbb{C}^n,\|\mathbf{y}_\ell\|_2=1\\ \theta_\ell\geq 0,\sum_{\ell=1}^R\theta_\ell=1}} \left\|\rho - \sum_{\ell=1}^{R}\theta_\ell(\mathbf{x}_\ell{\mathbf{x}_\ell}^*)\otimes(\mathbf{y}_\ell{\mathbf{y}_\ell}^*)\right\|_F^2, \tag{1.11}$$

where R is a predetermined positive integer and $\otimes$ denotes the Kronecker product. By the Carathéodory theorem, we need no more than $(mn)^2+1$ terms for the approximation, but often R is a much smaller number.

The quantum entanglement necessarily involves complex numbers [1, 36]. The approximation problem described above involves the optimization of a real-valued function over the complex field. This is a complicated task because, while we can identify $\mathbb{C}\equiv\mathbb{R}^2$, the multiplications of complex numbers entail a twist of real and imaginary parts, i.e., if $\mathbf{x}=\mathbf{u}+\imath\mathbf{v}\in\mathbb{C}^m$ and $\mathbf{y}=\mathbf{p}+\imath\mathbf{q}\in\mathbb{C}^n$ with $\mathbf{u},\mathbf{v}\in\mathbb{R}^m$ and $\mathbf{p},\mathbf{q}\in\mathbb{R}^n$, then

$$\mathbf{x}\otimes\mathbf{y} = (\mathbf{u}\otimes\mathbf{p}-\mathbf{v}\otimes\mathbf{q})+\imath(\mathbf{v}\otimes\mathbf{p}+\mathbf{u}\otimes\mathbf{q}).$$

Thus, not only do we have to deal with four real-valued vectors per $\mathbf{x}\in\mathbb{C}^m$ and $\mathbf{y}\in\mathbb{C}^n$, but also we have to consider their intertwinement, which complicates the cost function in (1.11).

1.5. Basic problem. It might be reasonable to consider the real analogue of the original complex problem as the first stepping stone. That is, we consider a simplified problem in the form

$$\min_{\mathbf{a}_\ell \in S^{m-1}, \mathbf{b}_\ell \in S^{n-1}, \theta_\ell \in \mathbb{R}_+} \left\| \rho - \sum_{\ell=1}^{R} \theta_\ell \big(\mathbf{a}_\ell \mathbf{a}_\ell^\top\big) \otimes \big(\mathbf{b}_\ell \mathbf{b}_\ell^\top\big) \right\|_F^2, \tag{1.12}$$

where $\rho \in \mathbb{R}^{mn \times mn}$ is a given symmetric and positive definite matrix, S^{m-1} stands for the unit sphere in $\mathbb{R}^m$, $\mathbb{R}_+$ denotes the half-array of nonnegative real numbers, and the sum-to-one condition of θ_ℓ is relaxed. One possible way to tackle (1.12) is through the greedy rank-1 update scheme that systematically adjusts one pair $(\mathbf{a}_j, \mathbf{b}_j)$, $j = 1, \ldots, R$, at a time [14, 15, 45, 57, 65]. Specifically, while advancing in p, we consider a sequence of subproblems by successively finding the triplet

$$\begin{aligned} &\left(\mathbf{a}_j^{[p+1]}, \mathbf{b}_j^{[p+1]}, \theta_j^{[p+1]}\right) \\ &\quad = \operatorname*{arg\,min}_{\mathbf{a}_j \in S^{m-1}, \mathbf{b}_j \in S^{n-1}, \theta_j \in \mathbb{R}_+} \left\| \rho_j^{[p+1]} - \theta_j \left(\mathbf{a}_j \mathbf{a}_j^\top\right) \otimes \left(\mathbf{b}_j \mathbf{b}_j^\top\right) \right\|_F^2, \quad j = 1, \ldots, R, \end{aligned} \tag{1.13}$$

with

$$\begin{aligned} \rho_j^{[p+1]} := \rho &- \sum_{\ell=1}^{j-1} \theta_\ell^{[p+1]} \left(\mathbf{a}_\ell^{[p+1]} \mathbf{a}_\ell^{[p+1]^\top}\right) \otimes \left(\mathbf{b}_\ell^{[p+1]} \mathbf{b}_\ell^{[p+1]^\top}\right) \\ &- \sum_{\ell=j+1}^{R} \theta_\ell^{[p]} \left(\mathbf{a}_\ell^{[p]} \mathbf{a}_\ell^{[p]^\top}\right) \otimes \left(\mathbf{b}_\ell^{[p]} \mathbf{b}_\ell^{[p]^\top}\right), \end{aligned}$$

where $(\mathbf{a}_\ell^{[p]}, \mathbf{b}_\ell^{[p]}, \theta_\ell^{[p]})$, $\ell = j+1, \ldots, R$, are previously known and $(\mathbf{a}_\ell^{[p+1]}, \mathbf{b}_\ell^{[p+1]}, \theta_\ell^{[p+1]})$, $\ell = 1, \ldots, j-1$, are newly updated.[4] Thus, for each j and p with $A := \rho_j^{[p+1]} \in \mathbb{R}^{mn \times mn}$, at the core is the basic rank-1 separability approximation problem

$$\min_{\mathbf{x} \in S^{m-1}, \mathbf{y} \in S^{n-1}, \lambda \in \mathbb{R}_+} \| A - \lambda \big(\mathbf{x}\mathbf{x}^\top\big) \otimes \big(\mathbf{y}\mathbf{y}^\top\big) \|_F^2. \tag{1.14}$$

Note that in this context, the target matrix A remains symmetric but may not be positive definite. The focus of this paper is on solving this relatively simpler rank-1 problem (1.14) for a given symmetric matrix A.

For fixed unit vectors $\mathbf{x}$ and $\mathbf{y}$, the optimal λ for the objective function in (1.14) is the component

$$\lambda(\mathbf{x}, \mathbf{y}) := \left\langle A, (\mathbf{x} \otimes \mathbf{y})(\mathbf{x} \otimes \mathbf{y})^\top \right\rangle \tag{1.15}$$

of A in the direction of $(\mathbf{x} \otimes \mathbf{y})(\mathbf{x} \otimes \mathbf{y})^\top$. Thus, the minimization in (1.14) is equivalent to the task of maximizing $|\lambda(\mathbf{x}, \mathbf{y})|$ over $S^{m-1} \times S^{n-1}$. However, unless A is positive semidefinite,[5] it is possible that this component λ is negative. In that case, the

[4] In practice, since such an iteration is to be repeated cyclically until convergence, it is not always needed to obtain the minimizer $(\mathbf{a}_j^{[p+1]}, \mathbf{b}_j^{[p+1]}, \theta_j^{[p+1]})$ to high precision per the subproblem.

[5] Definiteness over a tensor product space is a much more complicated notion than that over a Euclidean space. If A is partitioned into $m \times m$ blocks $A = [A_{ij}]$ with each block $A_{ij} \in \mathbb{R}^{n \times n}$, then we can write $\lambda(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top [\mathbf{y}^\top A_{ij} \mathbf{y}] \mathbf{x}$. If A in positive semidefinite over $\mathbb{R}^m \otimes \mathbb{R}^n$, then the matrix $[\mathbf{y}^\top A_{ij} \mathbf{y}] \in \mathbb{R}^{n \times n}$ should be positive semidefinite for all $\mathbf{y} \in S^{n-1}$. Such a matrix A must be very specific, including among others that all diagonal blocks A_{ii} must be positive semidefinite.

optimal value to the problem (1.14) is necessarily equal to $\|A\|_F^2$ with $\lambda = 0$. For our applications, because of the fact that

$$\lambda(\mathbf{x},\mathbf{y}) := \left\langle A + cI, (\mathbf{x}\otimes\mathbf{y})(\mathbf{x}\otimes\mathbf{y})^\top \right\rangle - c, \tag{1.16}$$

we may shift A by a sufficiently large scalar matrix to ensure that $A + cI$ is positive definite without tampering with the optimizer $(\mathbf{x},\mathbf{y})$ of the original objective function $\lambda(\mathbf{x},\mathbf{y})$. For the purpose of computation, therefore, it suffices to assume that A is symmetric and positive definite. Such an assumption also facilitates the convergence analysis for the algorithms proposed in this paper.

1.6. Related rank-1 tensor approximation.

The subject of low-rank tensor approximation has been intensively studied in recent years with many accomplished works [6, 15, 18, 34, 39, 40, 41, 57, 61, 65]. Most of the work, however, considers only the case when the factors are made of single states. For models where the factors themselves are high-order tensors, it is often necessary, and also for storage efficiency, to first break down each tensor factor as the outer product of single states. In our case, the basic problem (1.14) can be recast as a special type of rank-1 approximation with "shared" factors

$$\min_{\mathbf{x}\in S^{m-1},\mathbf{y}\in S^{n-1},\lambda\in\mathbb{R}_+} \|\mathfrak{A} - \lambda\,\mathbf{x}\circ\mathbf{x}\circ\mathbf{y}\circ\mathbf{y}\|_F^2, \tag{1.17}$$

where $\circ$ denotes the outer product and $\mathfrak{A}\in\mathbb{R}^{m\times m\times n\times n}$ is a special refolding of the original $A\in\mathbb{R}^{mn\times mn}$ into an order-4 tensor. This specially structured problem can still be handled by some conventional techniques, say, the Tensorlab toolbox [60]. In contrast, in this paper, we propose two new rank-1 approximation methods specifically for the bipartite systems. This is only a first step, but it plays an important role as a building block for the more general problems such as (1.12). The convergence analysis for the simplest bipartite systems is already quite involved and is of mathematical interest in its own right. Equally important is that the numerical comparison of our methods with various existing optimization packages as well as the different solvers available from the Tensorlab toolbox shows the advantages of our approach.

Finally, we outline the organization of this paper as follows. In section 2, we reformulate the rank-1 approximation problem as a nonlinear eigenvalue problem, propose a power-like iteration scheme, and prove its convergence. The scheme modifies one factor at a time. In section 3, we reformulate the approximation as a nonlinear singular value problem in which two factors are modified concurrently. By employing the conventional SVD as a black-box generating function, an abstract fixed-point iteration is proposed. The limiting behavior of this more sophisticated SVD-based scheme is also analyzed. We present experimental results in section 4 to demonstrate the working of the algorithms.

2. Nonlinear eigenvalue formulation.

The data stored in an order-4 tensor $\mathcal{T}\in\mathbb{R}^{m_2\times n_2\times m_1\times n_1}$ can be "visualized" in different ways. One way is to "flatten" an order-4 tensor $\mathcal{T}$ as an $m_1\times n_1$ block matrix T with blocks $T_{ij}\in\mathbb{R}^{m_2\times n_2}$,

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n_1} \\ T_{21} & T_{22} & \cdots & T_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ T_{m_11} & T_{m_12} & \cdots & T_{m_1n_1} \end{bmatrix} \in \mathbb{R}^{m_1m_2\times n_1n_2}. \tag{2.1}$$

Another way is to define the so-called $\mathscr{R}$-folding of T via the rearrangement

$$\mathscr{R}(T) := \begin{bmatrix} \mathbf{vec}(T_{11})^\top \\ \mathbf{vec}(T_{21})^\top \\ \vdots \\ \mathbf{vec}(T_{m_1 n_1})^\top \end{bmatrix} \in \mathbb{R}^{m_1 n_1 \times m_2 n_2}, \tag{2.2}$$

where **vec** denotes the conventional vectorization of a matrix by its columns. Such a rearrangement of T is particularly useful due to the following relationship [58, 59].

LEMMA 2.1. *The $\mathscr{R}$-folding of $T \in \mathbb{R}^{m_1 m_2 \times n_1 n_2}$ satisfies the relationship that*

$$\left\langle T, \sum_{k=1}^{r} X_k \otimes Y_k \right\rangle = \left\langle \mathscr{R}(T), \sum_{k=1}^{r} \mathbf{vec}(X_k)\mathbf{vec}(Y_k)^\top \right\rangle \tag{2.3}$$

for any $X_k \in \mathbb{R}^{m_1 \times n_1}$ and $Y_k \in \mathbb{R}^{m_2 \times n_2}$.

Our problem has the special structure that X_k and Y_k are symmetric rank-1 matrices. We now exploit this structure by recasting the maximization of (1.15) as a nonlinear eigenvalue problem.

2.1. Power-like iteration.

Partition $A \in \mathbb{R}^{mn \times mn}$ as $m \times m$ blocks with block size $n \times n$. Using (2.3), rewrite (1.15) as

$$\lambda(\mathbf{x}, \mathbf{y}) := \left\langle A, \left(\mathbf{x}\mathbf{x}^\top\right) \otimes \left(\mathbf{y}\mathbf{y}^\top\right) \right\rangle = \langle \mathscr{R}(A)(\mathbf{y} \otimes \mathbf{y}), \mathbf{x} \otimes \mathbf{x} \rangle, \tag{2.4}$$

where $\mathscr{R}(A) \in \mathbb{R}^{m^2 \times n^2}$. Define the bilinear operators

$$\mathscr{A}(\mathbf{y}, \widetilde{\mathbf{y}}) := \mathsf{reshape}(\mathscr{R}(A)(\mathbf{y} \otimes \widetilde{\mathbf{y}}), [m, m]), \tag{2.5}$$

$$\mathscr{B}(\mathbf{x}, \widetilde{\mathbf{x}}) := \mathsf{reshape}(\mathscr{R}(A)^\top(\mathbf{x} \otimes \widetilde{\mathbf{x}}), [n, n]), \tag{2.6}$$

over the respective unit spheres. It can be checked that

$$\begin{cases} \mathscr{A}(\mathbf{y}, \widetilde{\mathbf{y}}) = \mathscr{A}(\widetilde{\mathbf{y}}, \mathbf{y})^\top, \\ \mathscr{B}(\mathbf{x}, \widetilde{\mathbf{x}}) = \mathscr{B}(\widetilde{\mathbf{x}}, \mathbf{x})^\top, \end{cases} \tag{2.7}$$

and that both $\mathscr{A}(\mathbf{y}, \mathbf{y})$ and $\mathscr{B}(\mathbf{x}, \mathbf{x})$ are symmetric. Since we have assumed that A is positive definite, it can further be checked that $\mathscr{A}(\mathbf{y}, \mathbf{y})$ and $\mathscr{B}(\mathbf{x}, \mathbf{x})$ are also positive definite if $\mathbf{y} \neq 0$ and $\mathbf{x} \neq 0$, respectively. We can use (2.4) to calculate the projected gradients of $\lambda(\mathbf{x}, \mathbf{y})$ onto S^{m-1} and S^{n-1} easily, from which we obtain the optimality condition of $\lambda(\mathbf{x}, \mathbf{y})$ as follows, resulting in a nonlinear eigenvalue problem.

LEMMA 2.2. *The first-order necessary condition maximizing $\lambda(\mathbf{x}, \mathbf{y})$ is that*

$$\begin{cases} \mathscr{A}(\mathbf{y}, \mathbf{y})\mathbf{x} = \lambda(\mathbf{x}, \mathbf{y})\mathbf{x}, \\ \mathscr{B}(\mathbf{x}, \mathbf{x})\mathbf{y} = \lambda(\mathbf{x}, \mathbf{y})\mathbf{y}. \end{cases} \tag{2.8}$$

We are thus motivated to propose a power-like iterative scheme to obtain the (local) maximizer of $\lambda(\mathbf{x}, \mathbf{y})$. Starting from an initial value $(\mathbf{x}^{[0]}, \mathbf{y}^{[0]})$, we repeat the following process:

$$\begin{cases} \mathbf{x}^{[p+1]} := \frac{\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})\mathbf{x}^{[p]}}{\|\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})\mathbf{x}^{[p]}\|_2}, \\ \mathbf{y}^{[p+1]} := \frac{\mathscr{B}(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]})\mathbf{y}^{[p]}}{\|\mathscr{B}(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]})\mathbf{y}^{[p]}\|_2}, \end{cases} \quad p = 0, 1, 2, \ldots. \tag{2.9}$$

At first glance, the scheme resembles a conventional power method. We stress, however, that the matrices $\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})$ and $\mathscr{B}(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]})$ are not stationary. They are updated in a manner similar to the notion of the conventional Gauss–Seidel method. If the iteration ever converges, the limit point is a fixed-point and satisfies precisely the first-order optimality condition (2.8). When this happens, we have

$$\lambda(\mathbf{x}, \mathbf{y}) = \|\mathscr{A}(\mathbf{y}, \mathbf{y})\mathbf{x}\|_2 = \|\mathscr{B}(\mathbf{x}, \mathbf{x})\mathbf{y}\|_2 > 0.$$

It remains to study the dynamical behavior of this iterative scheme.

2.2. Convergence analysis. To facilitate the characterization of the dynamics of (2.9), define the functional $g : S^{m-1} \times S^{m-1} \times S^{n-1} \times S^{n-1} \to \mathbb{R}$ by

$$g(\mathbf{x}, \widetilde{\mathbf{x}}; \mathbf{y}, \widetilde{\mathbf{y}}) := \langle \mathscr{A}(\mathbf{y}, \widetilde{\mathbf{y}})\mathbf{x}, \widetilde{\mathbf{x}} \rangle. \tag{2.10}$$

Since

$$\langle \mathscr{A}(\mathbf{y}, \widetilde{\mathbf{y}})\mathbf{x}, \widetilde{\mathbf{x}} \rangle = \big\langle \mathscr{R}(A), (\mathbf{x} \otimes \widetilde{\mathbf{x}})(\mathbf{y} \otimes \widetilde{\mathbf{y}})^\top \big\rangle = \big\langle A, (\widetilde{\mathbf{x}}\mathbf{x}^\top) \otimes (\widetilde{\mathbf{y}}\mathbf{y}^\top) \big\rangle,$$

we can also write

$$g(\mathbf{x}, \widetilde{\mathbf{x}}; \mathbf{y}, \widetilde{\mathbf{y}}) = \langle \mathscr{B}(\mathbf{x}, \widetilde{\mathbf{x}})\mathbf{y}, \widetilde{\mathbf{y}} \rangle. \tag{2.11}$$

Clearly,

$$\lambda(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}, \mathbf{x}; \mathbf{y}, \mathbf{y}).$$

By (2.7), we see a symmetry in the sense that

$$g(\mathbf{x}, \widetilde{\mathbf{x}}; \mathbf{y}, \widetilde{\mathbf{y}}) = g(\widetilde{\mathbf{x}}, \mathbf{x}; \widetilde{\mathbf{y}}, \mathbf{y}).$$

We first establish a useful chain of variational relationships.

LEMMA 2.3. *Assume that A is symmetric and positive definite. Then the sequence $\{(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\} \subset S^{m-1} \times S^{n-1}$ generated by the scheme (2.9) satisfies the inequalities*

$$\begin{aligned} g(\mathbf{x}^{[p]}, \mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}) &\leq g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) &&\leq g\big(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) \\ &\leq g\big(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p+1]}\big) &&\leq g\big(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p+1]}, \mathbf{y}^{[p+1]}\big). \end{aligned} \tag{2.12}$$

The sequence $\{\lambda(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\}$ converges.

Proof. The first inequality follows from the definitions of $\mathbf{x}^{[p+1]}$. That is,

$$\begin{aligned} g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) &= \big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]} \big\rangle \\ &= \|\mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}\|_2 \geq g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) \end{aligned}$$

by the Cauchy–Schwarz inequality. Similarly, by the definitions of $\mathbf{y}^{[p+1]}$, the third inequality also holds. Only the second and fourth inequalities need proof. We shall argue for the second inequality only, as the argument for the fourth inequality is similar.

Write

$$\Delta\mathbf{x}^{[p]} := \mathbf{x}^{[p+1]} - \mathbf{x}^{[p]}. \tag{2.13}$$

Then

$$g\big(\Delta\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) = g\big(\Delta\mathbf{x}^{[p]}, \Delta\mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) + g\big(\Delta\mathbf{x}^{[p]}, \mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big).$$

The first term on the right side is nonnegative because $\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})$ is symmetric and positive semidefinite. We rewrite the second term as

$$\begin{aligned} g\big(\Delta\mathbf{x}^{[p]}, \mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) &= \mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\big(\mathbf{x}^{[p+1]} - \mathbf{x}^{[p]}\big) \\ &= \mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) \left(\frac{\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})\mathbf{x}^{[p]}}{\|\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})\mathbf{x}^{[p]}\|_2} \right) - \mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]} \\ &= \big(\mathbf{x}^{[p]^\top} \mathscr{A}^2\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}\big)^{\frac{1}{2}} - \mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}. \end{aligned}$$

Observe that

$$\begin{aligned} &\mathbf{x}^{[p]^\top} \mathscr{A}^2\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]} - \big(\mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}\big)^2 \\ &\quad = \mathbf{x}^{[p]^\top} \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\big(I - \mathbf{x}^{[p]}\mathbf{x}^{[p]^\top}\big)\mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]} \geq 0. \end{aligned}$$

We have thus proved that $g(\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}) \leq g(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]})$. Since the monotone sequence $\{\lambda(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\}$ is bounded, it must converge. ∎

We next argue that the gaps between successive iterates of $\{(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\}$ are diminishing to zero.

LEMMA 2.4. *Suppose that A is symmetric and positive definite. Then $\Delta\mathbf{x}^{[p]}$ and $\Delta\mathbf{y}^{[p]}$ converge to zero.*

Proof. By using the first and the second inequalities in (2.12), we see that the differences

$$\begin{aligned} g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) - g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) &= \big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p]}, \Delta\mathbf{x}^{[p]} \big\rangle, \\ g\big(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) - g\big(\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]}; \mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big) &= \big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\Delta\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]} \big\rangle \end{aligned}$$

converge to zero. By the symmetry of $\mathscr{A}(\mathbf{y}^{[p]}, \mathbf{y}^{[p]})$, the right side of the second equation above can be replaced by

$$\big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\Delta\mathbf{x}^{[p]}, \mathbf{x}^{[p+1]} \big\rangle = \big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\mathbf{x}^{[p+1]}, \Delta\mathbf{x}^{[p]} \big\rangle.$$

Taking the difference, it follows that

$$\big\langle \mathscr{A}\big(\mathbf{y}^{[p]}, \mathbf{y}^{[p]}\big)\Delta\mathbf{x}^{[p]}, \Delta\mathbf{x}^{[p]} \big\rangle \to 0. \tag{2.14}$$

Similarly,

$$\big\langle \mathscr{B}\big(\mathbf{x}^{[p+1]}, \mathbf{x}^{[p+1]}\big)\Delta\mathbf{y}^{[p]}, \Delta\mathbf{y}^{[p]} \big\rangle \to 0. \tag{2.15}$$

By the assumption of positive definiteness, the increments $\Delta\mathbf{x}^{[p]}$ and $\Delta\mathbf{y}^{[p]}$ converge to zero. ∎

It is well known in algebraic geometry that almost every square system of polynomial equations over the complex field has finitely many solutions [23]. Furthermore, if $F(\mathbf{z}; \mathbf{q})$ is a system of polynomials in both the variables $\mathbf{z}$ and the parameters $\mathbf{q}$ and is square in $\mathbf{z}$, then for almost all parameters $\mathbf{q}$ the number of isolated solutions[6] to

[6] Some clarification on the terminology "isolated solution" is due. Based on [37], a solution $\mathbf{x} = \mathbf{x}_0$ of $F(\mathbf{x}) = 0$ is said to be "isolated" if the Fréchet derivative $F'(\mathbf{x}_0)$ is nonsingular. The "isolation" therefore is implicitly implied in [52, Theorem 7.1.1]. The solution $\mathbf{x}_0$ is said to be "geometrically isolated" if no other solution is in the neighborhood $\{\mathbf{x} | \|\mathbf{x} - \mathbf{x}_0 < \epsilon\}$ for some $\epsilon > 0$. Isolated solutions are always geometrically isolated, but the converse is not true. A nonisolated solution, e.g., a double root, may also be geometrically isolated.

this polynomial system is finite [52, Theorem 7.1.1]. The phrase "almost all" means that those values of parameters that fail to produce finitely many and geometrically isolated solutions constitute a nowhere dense and measure zero subset in the ambient space. These cases of exceptions are referred to as "nongeneric."

In our problem, observe that the optimality condition (2.8) constitutes a system of polynomials in the variables $(\mathbf{x}, \mathbf{y})$ with highest degree 5. Any limit point of the iteration (2.9), if it exists, is necessarily a solution to this polynomial system. The system is not homogeneous but can be regarded as being parameterized by the matrix A. Although the problem is structured, recall that the intersection of a nowhere dense and measure zero subset with any other nonempty open subset remains nowhere dense and measure zero. Therefore, we may claim that for almost all matrices A in our problem[7] the corresponding set of limit points for the iterative scheme (2.9) contains finitely many and geometrically isolated points. The following definition is for easy reference later.

DEFINITION 2.5. *We say that the matrix A satisfies Condition* P *if the corresponding polynomial system* (2.8) *has finitely many and geometrically isolated real-valued solutions.*

THEOREM 2.6. *Assume that A is symmetric, positive definite and satisfies Condition* P. *Then the sequence of the iterates $\{(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\} \subset S^{m-1} \times S^{n-1}$ generated by the scheme* (2.9) *converges to a single limit point which satisfies the system* (2.8).

Proof. It has been established that if a bounded sequence $\{a_s\}$ of real numbers has the properties that $|a_{s+1} - a_s| \to 0$ as $s \to \infty$ and that its accumulation points are isolated, then the sequence $\{a_s\}$ converges to a unique limit point. See [13, Lemma 4.3] and [26, Lemma 2.6]. Under the assumption that A is generic, we already know that the set of limit points is finite and isolated. We also see in Lemma 2.4 that the difference between two consecutive iterates diminishes to zero. These two criteria, i.e., finitely many isolated limit points and diminishing successive increments, are enough to guarantee the convergence of $\{(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\}$ to a single point. □

3. Nonlinear singular value formulation.

In the preceding section, the system (2.8) has the appearance of an eigenvalue system, albeit the operators $\mathscr{A}(\mathbf{y}, \mathbf{y})$ and $\mathscr{B}(\mathbf{x}, \mathbf{x})$ depend nonlinearly on the unknowns $\mathbf{y}$ and $\mathbf{x}$, respectively. We now consider an equivalent singular value formulation.

Define

$$\mathscr{C}(\mathbf{x}, \mathbf{y}) := \mathsf{reshape}(A(\mathbf{x} \otimes \mathbf{y}), [n, m]) \in \mathbb{R}^{n \times m}. \tag{3.1}$$

Then

$$\lambda(\mathbf{x}, \mathbf{y}) = \langle \mathscr{C}(\mathbf{x}, \mathbf{y})\mathbf{x}, \mathbf{y} \rangle, \tag{3.2}$$

whose extreme values resemble the variational formulation for the singular values of the matrix $\mathscr{C}(\mathbf{x}, \mathbf{y})$. Not surprisingly, we should have the following characterization of critical points.

[7] By the fundamental theorem of algebra, the closure of any infinite subset is always dense in $\mathbb{C}$ under the Zariski topology. In particular, $\mathbb{R}$ is dense in $\mathbb{C}$. The concept remains true in multidimensional spaces. A real-valued matrix A that fails to produce finitely many and geometrically isolated solutions is a member in the nongeneric set in the complex field. The collection of such real-valued matrices is nowhere dense and of measure zero over the complex space. Regarding the real space as a cross section of the complex space, such a subset is nongeneric over the real space with respect to the induced measure. This can also be seen from the transversality theorem. In other words, almost all real-valued matrices A are generic over the real space for our problem.

LEMMA 3.1. *The first-order necessary condition for maximizing* (3.2) *under the unit length constraints is that* $\mathbf{x}$ *and* $\mathbf{y}$ *are the right and left singular vectors of the matrix* $\mathscr{C}(\mathbf{x},\mathbf{y})$*, respectively. The local minima to* (1.14) *are attained at singular vector pairs of* $\mathscr{C}(\mathbf{x},\mathbf{y})$*.*

Proof. Using (3.2), the partial gradients of $\lambda(\mathbf{x},\mathbf{y})$ projected onto the unit spheres S^{m-1} and S^{n-1} are given by

$$\begin{cases} \operatorname{Proj}_{S^{m-1}} \frac{\partial \lambda}{\partial \mathbf{x}} = 2(\mathscr{C}(\mathbf{x},\mathbf{y})^\top \mathbf{y} - (\mathbf{y}^\top \mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x})\mathbf{x}), \\ \operatorname{Proj}_{S^{n-1}} \frac{\partial \lambda}{\partial \mathbf{y}} = 2(\mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x} - (\mathbf{y}^\top \mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x})\mathbf{y}), \end{cases}$$

restrictively. Thus, a critical point $(\mathbf{x},\mathbf{y})$ necessarily satisfies the relationship

$$\begin{cases} \mathscr{C}(\mathbf{x},\mathbf{y})^\top \mathbf{y} = (\mathbf{y}^\top \mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x})\mathbf{x}, \\ \mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x} = (\mathbf{y}^\top \mathscr{C}(\mathbf{x},\mathbf{y})\mathbf{x})\mathbf{y}, \end{cases} \tag{3.3}$$

which translates to the role of singular vectors for the matrix $\mathscr{C}(\mathbf{x},\mathbf{y})$. At such a critical point, we also see that $\lambda(\mathbf{x},\mathbf{y})$ plays the role of the dominant singular value since (3.2) is being maximized. □

Define also the functional $h : S^{m-1} \times S^{n-1} \times S^{m-1} \times S^{n-1} \to \mathbb{R}$ by

$$h(\mathbf{x},\mathbf{y};\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) := \langle A(\mathbf{x}\otimes\mathbf{y}), \widetilde{\mathbf{x}}\otimes\widetilde{\mathbf{y}}\rangle = \langle \mathscr{C}(\mathbf{x},\mathbf{y})\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}\rangle, \tag{3.4}$$

where the pair $(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})$ will be referred to as parameters in the constrained maximization of $h(\mathbf{x},\mathbf{y};\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})$. By the symmetry of A, we see that

$$h(\mathbf{x},\mathbf{y};\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) = h(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}};\mathbf{x},\mathbf{y}). \tag{3.5}$$

LEMMA 3.2. *For any* $\mathbf{x},\widetilde{\mathbf{x}} \in S^{m-1}$ *and* $\mathbf{y},\widetilde{\mathbf{y}} \in S^{n-1}$*, the following identities hold:*

$$\begin{cases} h(\mathbf{x},\mathbf{y};\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) = g(\mathbf{x},\widetilde{\mathbf{x}};\mathbf{y},\widetilde{\mathbf{y}}), \\ \mathscr{C}(\mathbf{x},\mathbf{y})\widetilde{\mathbf{x}} = \mathscr{B}(\mathbf{x},\widetilde{\mathbf{x}})\mathbf{y}, \\ \mathscr{C}(\mathbf{x},\mathbf{y})^\top \widetilde{\mathbf{y}} = \mathscr{A}(\mathbf{y},\widetilde{\mathbf{y}})\mathbf{x}. \end{cases} \tag{3.6}$$

Proof. The first equation is obvious from the definitions of g in (2.10) and h in (3.4). We prove the third identity only. With respect to a given $\mathbf{z} \in \mathbb{R}^m$, observe that

$$\begin{aligned} \langle \mathscr{C}(\mathbf{x},\mathbf{y})^\top \widetilde{\mathbf{y}}, \mathbf{z}\rangle &= \langle A(\mathbf{x}\otimes\mathbf{y}), \mathbf{z}\otimes\widetilde{\mathbf{y}}\rangle = \langle A, (\mathbf{z}\mathbf{x}^\top)\otimes(\widetilde{\mathbf{y}}\mathbf{y}^\top)\rangle \\ &= \langle \mathscr{R}(A), (\mathbf{x}\otimes\mathbf{z})(\mathbf{y}\otimes\widetilde{\mathbf{y}})^\top\rangle = \langle \mathscr{A}(\mathbf{y},\widetilde{\mathbf{y}})\mathbf{x}, \mathbf{z}\rangle. \end{aligned}$$

Since $\mathbf{z}$ is arbitrary, the identity must hold. □

It is now clear that, because of (3.6), the condition (3.3) is identical to (2.8). The difference is that (3.3) is more like a singular value setting, whereas (2.8) is more like an eigenvalue value setting. It is interesting to note that, though we already are familiar with the relationship between eigenvalues and singular values of a matrix, in the context of entanglement via the Kronecker product the equivalence between the two settings is obtained by simply rearranging the order of multiplication.

LEMMA 3.3. *Assume that A is symmetric, positive definite and satisfies Condition* P. *Then the limit point $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ of the sequence $\{(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})\} \subset S^{m-1} \times S^{n-1}$ generated by the power-like scheme* (2.9) *is a dominant singular vector pair of the matrix $\mathscr{C}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$.*

To determine the exact number of critical points, i.e., real-valued solutions to the polynomial system (3.3), is an interesting but challenging question in the realm of real algebraic geometry. On the other hand, note that the dominant singular vectors $(\mathbf{x}, \mathbf{y})$ of $\mathscr{C}(\mathbf{x}, \mathbf{y})$ serve only as a local solution to (1.14) (equivalently, (1.15)). To increase the likelihood of an absolute best approximation, we might need to resort to some global optimization techniques, which will not be explored in this work. Just like the nonlinear system (2.8), a solution to the first-order optimality condition (3.3) is not readily available because the matrix $\mathscr{C}(\mathbf{x}, \mathbf{y})$ itself depends on $\mathbf{x}$ and $\mathbf{y}$. This is a nonlinear singular value problem. We now propose an interesting numerical method to handle it.

3.1. SVD-type iteration.

The power-like iterative scheme (2.9) is developed initially on the basis of (2.8). Through the interchangeable relationship (3.6), we can rewrite the power-like iterative scheme as

$$\text{(3.7)} \qquad \begin{cases} \mathbf{x}^{[p+1]} = \frac{\mathscr{C}(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})^\top \mathbf{y}^{[p]}}{\|\mathscr{C}(\mathbf{x}^{[p]}, \mathbf{y}^{[p]})^\top \mathbf{y}^{[p]}\|_2}, \\ \mathbf{y}^{[p+1]} = \frac{\mathscr{C}(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}) \mathbf{x}^{[p+1]}}{\|\mathscr{C}(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}) \mathbf{x}^{[p+1]}\|_2}, \end{cases} \qquad p = 0, 1, 2, \ldots,$$

on the basis of (3.3). Since the two iterative schemes are equivalent, we have exactly the same dynamics. For instance, in terms of the function h, the first four inequalities in (2.12) are readily translated to

$$\text{(3.8)} \quad \begin{aligned} h\big(\mathbf{x}^{[p]}, \mathbf{y}^{[p]}; \mathbf{x}^{[p]}, \mathbf{y}^{[p]}\big) &\le h\big(\mathbf{x}^{[p]}, \mathbf{y}^{[p]}; \mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}\big) && \le h\big(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}; \mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}\big) \\ &\le h\big(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p]}; \mathbf{x}^{[p+1]}, \mathbf{y}^{[p+1]}\big) && \le h\big(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p+1]}; \mathbf{x}^{[p+1]}, \mathbf{y}^{[p+1]}\big). \end{aligned}$$

There is nothing new up to this point.

However, since the ultimate goal of the iteration (3.7) is to solve the SVD of $\mathscr{C}(\mathbf{x}, \mathbf{y})$, it is appealing to solve (3.3) by directly obtaining the dominant singular value triplet at every step. That is, employing any dominant singular value triplet finder, say, the MATLAB routine svds, as a black-box generating function, we propose a fixed-point iteration

$$\text{(3.9)} \qquad \begin{aligned} &\big(\operatorname{sgn}\big(\overline{x}_1^{[p+1]}\big)\overline{\mathbf{y}}^{[p+1]}, \lambda^{[p+1]}, \operatorname{sgn}\big(\overline{x}_1^{[p+1]}\big)\overline{\mathbf{x}}^{[p+1]}\big) \\ &\quad = \mathsf{svds}\big(\mathscr{C}\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big), 1\big), \quad p = 0, 1, 2, \ldots, \end{aligned}$$

where sgn is meant to ensure the continuity by keeping the first entry $\overline{x}_1^{[p+1]}$ of $\overline{\mathbf{x}}^{[p+1]}$ positive. To distinguish the iterates from those power-like iterates described earlier, we have denoted the SVD-like iterates by $(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$. It is interesting to note that we can rewrite (3.9) as an implicit power-like iterative scheme:

$$\text{(3.10)} \quad \begin{cases} \overline{\mathbf{x}}^{[p+1]} := \frac{\mathscr{A}(\overline{\mathbf{y}}^{[p]}, \overline{\mathbf{y}}^{[p+1]}) \overline{\mathbf{x}}^{[p]}}{\|\mathscr{A}(\overline{\mathbf{y}}^{[p]}, \overline{\mathbf{y}}^{[p+1]}) \overline{\mathbf{x}}^{[p]}\|_2} = \frac{\mathscr{C}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})^\top \overline{\mathbf{y}}^{[p+1]}}{\|\mathscr{C}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})^\top \overline{\mathbf{y}}^{[p+1]}\|_2}, \\ \overline{\mathbf{y}}^{[p+1]} := \frac{\mathscr{B}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{x}}^{[p+1]}) \overline{\mathbf{y}}^{[p]}}{\|\mathscr{B}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{x}}^{[p+1]}) \overline{\mathbf{y}}^{[p]}\|_2} = \frac{\mathscr{C}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}) \overline{\mathbf{x}}^{[p+1]}}{\|\mathscr{C}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}) \overline{\mathbf{x}}^{[p+1]}\|_2}, \end{cases} \qquad p = 0, 1, 2, \ldots,$$

in comparison with (2.9) and (3.7), respectively.

The variational property indwelt in the conventional SVD certainly gives rise to an optimality property for the scheme (3.9). Using the Berge's maximum theorem, we gain some additional insight which is characterized as follows. Regarding $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ as parameters in (3.4), the maximum value function

$$\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) := \max_{\mathbf{x} \in S^{m-1}, \mathbf{y} \in S^{n-1}} h(\mathbf{x}, \mathbf{y}; \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \tag{3.11}$$

is well defined and continuous, and the so-called optimal policy correspondence

$$\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) := \{(\mathbf{x}, \mathbf{y}) \in S^{m-1} \times S^{n-1} | h(\mathbf{x}, \mathbf{y}; \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})\} \tag{3.12}$$

is nonempty, compact valued, and upper hemicontinuous.[8] Based on the variational property of the SVD, the scheme (3.9) can be interpreted as a fixed-point iteration of the two maps μ and Ξ:

$$\begin{cases} \lambda^{[p+1]} = \mu\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big), \\ \big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) \in \Xi\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big). \end{cases} \tag{3.13}$$

In fact, by the continuity property inherited in the SVD, we can choose the representative $(\mathbf{x}, \mathbf{y})$ in $\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ so that the pair of singular vectors $(\mathbf{x}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}), \mathbf{y}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))$ varies continuously in $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$. In this way, the set-valued correspondence Ξ is interpreted as a "map" of $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$. Furthermore, for parameters $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ where the dominant singular value $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ of $\mathscr{C}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is simple, it can be proved that the singular value function $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is analytic [7, 63].

We have already explained that the subset of matrices satisfying Condition P is open and dense [23, 52]. It is also an established fact that symmetric matrices with multiple eigenvalues form an algebraic variety of codimension 2 [17]. Together, we conclude that matrices whose largest singular value is simple form an open and dense subset. Thus, similar to Condition P, the matrices satisfying Condition S defined below are generic.

DEFINITION 3.4. *We say that the matrix A satisfies Condition* S *if the corresponding polynomial system* (3.3) *has finitely many, isolated, real solutions and that the associated $\mathscr{C}(\mathbf{x}, \mathbf{y})$ has simple dominant singular value.*[9]

We now concentrate on the action of Ξ on the parameters $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ and the selection of the parameters to maximize the maximum value function $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$.

LEMMA 3.5. *Under Condition* S, *the parameters that maximize $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ must satisfy the system of equations:*

$$\begin{cases} \mathscr{C}(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))^\top \widetilde{\mathbf{y}} = (\widetilde{\mathbf{y}}^\top \mathscr{C}(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))\widetilde{\mathbf{x}})\widetilde{\mathbf{x}}, \\ \mathscr{C}(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))\widetilde{\mathbf{x}} = (\widetilde{\mathbf{y}}^\top \mathscr{C}(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))\widetilde{\mathbf{x}})\widetilde{\mathbf{y}}. \end{cases} \tag{3.14}$$

Proof. Under Condition S, we may assume that the optimal policy correspondence $\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is continuous. By the envelope theorem [2, 53], the change in the maximum

[8] That is, Ξ is a set-valued correspondence that maps any sequence that converges to $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ to a convergent sequence with its limit point in $\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$, but those sequences are not guaranteed to produce all possible $\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ as their limits.

[9] Corresponding to the same dominant singular value, the singular vectors can be the same or antipode to each other. The sign check entailed in the algorithm ensures that antipode is ruled out.

value function is given by the partial derivative of the Lagrangian with respect to the parameters. Since the constraint $S^{m-1} \times S^{n-1}$ is independent of the parameters, we conclude that the gradient of $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ is given by

$$\left\{ \begin{array}{l} \frac{\partial \mu(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})}{\partial \widetilde{\mathbf{x}}} = \frac{\partial h(\Xi(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}});\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})}{\partial \widetilde{\mathbf{x}}} = \mathscr{C}(\Xi(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}))^{\top}\widetilde{\mathbf{y}}, \\ \frac{\partial (\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})}{\partial \widetilde{\mathbf{y}}} = \frac{\partial h(\Xi(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}});\widetilde{\mathbf{x}},\widetilde{\mathbf{y}})}{\partial \widetilde{\mathbf{y}}} = \mathscr{C}(\Xi(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}))\widetilde{\mathbf{x}}. \end{array} \right. \tag{3.15}$$

Taking into account the constraints that $\widetilde{\mathbf{x}} \in S^{m-1}$ and $\widetilde{\mathbf{y}} \in S^{n-1}$, $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ reaches its maximum when its projected gradient vanishes, which is exactly the system of equations given in (3.14). ∎

Corollary 3.6. *Under Condition* S*, suppose that* $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ *is a maximizer of the maximum value function* μ*. Then*

$$\Xi(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})) = (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}). \tag{3.16}$$

Proof. The optimality condition (3.14) for $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ asserts that the pair $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ itself is the dominant singular vectors of $\mathscr{C}(\Xi(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}))$. ∎

Recall that if A is symmetric, positive definite, and satisfies Condition P, then the power-like scheme (2.9) is guaranteed to converge by Theorem 2.6. Its limit point $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$, according to the reformulation (3.7), forms the pair of the dominant singular vectors of $\mathscr{C}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$. It follows that

$$\Xi(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = (\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \tag{3.17}$$

and that the criteria in Lemma 3.5 are met. In this case,

$$\mu(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) = h(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}; \widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \tag{3.18}$$

is a maximal value.[10] However, at the moment, we have proved only that the maximizer $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ of the maximum value function μ satisfied the composition relationship (3.16). It is not clear whether such a pair is a fixed-point for the map Ξ alone because the limiting behavior of the new SVD-based iteration (3.9) (or any of the equivalent schemes (3.10) and (3.13)) has not been proven to converge yet.

3.2. Convergence analysis. In contrast to the power-like iterative scheme (2.9) which is explicit, the sequence $\{(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})\}$ satisfies an implicit relationship (3.10) and is obtained through a black-box fixed-point iteration (3.9). Thus, it is even more imperative to understand its limiting behavior. Similar to Theorem 2.6, we shall argue ultimately that the positive definiteness of A plays a critical role.

Even without the positive definiteness of A, the convergence of the dominant singular values is quite straightforward by using the properties of h.

Lemma 3.7. *Given any symmetric matrix* $A \in \mathbb{R}^{mn \times mn}$*, the sequence* $\{\lambda^{[p]}\}$ *generated by the scheme* (3.9) *is monotone nondecreasing and, hence, converges.*

Proof. Observe that for $p \geq 0$,

$$\begin{aligned} \lambda^{[p+1]} = \mu\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) &= h\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) \\ &= h\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}; \overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) \leq \mu\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) = \lambda^{[p+2]}. \end{aligned} \tag{3.19}$$

Since $\{\lambda^{[p]}\}$ is bounded above by $\|A\|_F$, it must converge. ∎

[10] We remark that the maximum of the functional h occurs at the "diagonal" of its domain, i.e., the first set of variables is identical to the second set of variables. This is a special case of the so-called symmetric criticality [49].

If the equality $\lambda^{[p+1]} = \lambda^{[p+2]}$ ever happens for the first time at a finite value of p in (3.19), then the singular vector pair $(\overline{\mathbf{x}}^{[p+2]}, \overline{\mathbf{y}}^{[p+2]})$ associated with the dominant singular value $\lambda^{[p+2]}$ of the matrix $\mathscr{C}(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]})$ in (3.9) can be taken to be the singular vector pair $(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$ already in existence. Since the pair $(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$ is used in $\mathscr{C}(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$ to generate the next iterate $(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]})$, the iteration may become cyclic between $(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$ and $(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]})$ or stays invariant for any further processes.

Still under the assumption that A is symmetric only, together with Condition S, we observe the optimal values of the accumulation point as follows.

Lemma 3.8. *Suppose that A is symmetric and generic in the sense of satisfying Condition* S. *If $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$ is an accumulation point of the iterates $\{(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})\}$ generated by* (3.9)*, then both $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$ and $\Xi(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$ are maximizers of the maximum value function $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ with the same optimal value.*

Proof. Let $\{(\overline{\mathbf{x}}^{[p_j]}, \overline{\mathbf{y}}^{[p_j]})\}$ be a convergent subsequence such that

$$\lim_{j\to\infty} (\overline{\mathbf{x}}^{[p_j]}, \overline{\mathbf{y}}^{[p_j]}) = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0).$$

Then, by continuity,

$$\mu(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) = \lim_{j\to\infty} \mu(\overline{\mathbf{x}}^{[p_j]}, \overline{\mathbf{y}}^{[p_j]}) = \lim_{j\to\infty} \lambda^{[p_j+1]} = \widehat{\lambda},$$

where $\widehat{\lambda}$ is the limit point guaranteed by Lemma 3.7. By the definition of μ, the pair $(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1)$ at which the maximum

$$\mu(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) = h(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1; \widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) \tag{3.20}$$

is attained must be the singular vectors associated with the dominant singular value $\widehat{\lambda}$ of the matrix $\mathscr{C}(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$.

By the algorithm (3.9), the subsequence $\{(\overline{\mathbf{x}}^{[p_j+1]}, \overline{\mathbf{y}}^{[p_j+1]})\}$ consists of the singular vectors of the convergent subsequence $\{\mathscr{C}(\overline{\mathbf{x}}^{[p_j]}, \overline{\mathbf{y}}^{[p_j]})\}$, so it must also converge. Under Condition S, the dominant singular value triplet is unique. So it must be the case that

$$\lim_{j\to\infty} (\overline{\mathbf{x}}^{[p_j+1]}, \overline{\mathbf{y}}^{[p_j+1]}) = (\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1).$$

We may repeat this process and conclude that

$$\lim_{j\to\infty} (\overline{\mathbf{x}}^{[p_j+2]}, \overline{\mathbf{y}}^{[p_j+2]}) = (\widehat{\mathbf{x}}_2, \widehat{\mathbf{y}}_2), \tag{3.21}$$

where the pair $(\widehat{\mathbf{x}}_2, \widehat{\mathbf{y}}_2)$ is the dominant singular vectors of the matrix $\mathscr{C}(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1)$ with the same dominant singular value $\widehat{\lambda}$. On the other hand, observe that the relationship (3.20) can be expressed as

$$\begin{aligned} \widehat{\mathbf{y}}_1^\top \mathscr{C}(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)\widehat{\mathbf{x}}_1 &= \widehat{\lambda}, \\ \widehat{\mathbf{y}}_0^\top \mathscr{C}(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1)\widehat{\mathbf{x}}_0 &= \widehat{\lambda}. \end{aligned}$$

Since $\widehat{\lambda}$ is the dominant singular value, we see that the pair $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$ is also the dominant singular vector pair of $\mathscr{C}(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1)$. By uniqueness, it must be that $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) = (\widehat{\mathbf{x}}_2, \widehat{\mathbf{y}}_2)$. We thus conclude that

$$\begin{cases} \Xi(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) = (\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1), \\ \Xi(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1) = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0). \end{cases} \tag{3.22}$$

In particular, we have established the relationship

$$\begin{cases} \mathscr{C}(\Xi(\widehat{\mathbf{x}}_\ell, \widehat{\mathbf{y}}_\ell))^\top \widehat{\mathbf{y}}_\ell = \widehat{\lambda}\widehat{\mathbf{x}}_\ell, \\ \mathscr{C}(\Xi(\widehat{\mathbf{x}}_\ell, \widehat{\mathbf{y}}_\ell))\widehat{\mathbf{x}}_\ell = \widehat{\lambda}\widehat{\mathbf{y}}_\ell, \end{cases} \quad \ell = 0, 1, \tag{3.23}$$

which implies, by Lemma 3.5, that $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0)$ and $(\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1)$ are maximizers for the maximum value function $\mu(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ with optimal value $\widehat{\lambda}$. □

It is worthwhile to point out that the argument used in Lemma 3.8 for the subsequence $\{(\overline{\mathbf{x}}^{[p_j+2]}, \overline{\mathbf{y}}^{[p_j+2]})\}$ can actually be extended to $\{(\overline{\mathbf{x}}^{[p_j+\ell]}, \overline{\mathbf{y}}^{[p_j+\ell]})\}$ for $\ell = 3, 4, \dots$ and, hence,

$$\lim_{j\to\infty} (\overline{\mathbf{x}}^{[p_j+\ell]}, \overline{\mathbf{y}}^{[p_j+\ell]}) = \begin{cases} (\widehat{\mathbf{x}}_0, \widehat{\mathbf{y}}_0) & \text{if } \ell \text{ is even}, \\ (\widehat{\mathbf{x}}_1, \widehat{\mathbf{y}}_1) & \text{if } \ell \text{ is odd}. \end{cases} \tag{3.24}$$

Since the collection $\{p_j + \ell | j, \ell = 0, 1, 2, \dots\}$ contains all integers greater than or equal to p_0, we become curious about whether there are other accumulation points. In our numerical experiments, we have observed persistently that the iterates have at most two accumulation points if A is only symmetric. However, we do not have a mathematical proof at present.

We do claim that if, in addition, A is positive definite, then there is only one accumulation point. Toward that end, we first extend Lemma 3.7 to show the interlacing of $\lambda(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$ and $\mu(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$.

LEMMA 3.9. *Suppose that A is symmetric and positive definite. Then*

$$\lambda\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) \le \lambda^{[p+1]} \le \lambda\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) \le \lambda^{[p+2]}, \tag{3.25}$$

or equivalently

$$\begin{aligned} h\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) &\le h\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) \\ &\le h\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}; \overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) \\ &\le h\big(\overline{\mathbf{x}}^{[p+2]}, \overline{\mathbf{y}}^{[p+2]}; \overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big). \end{aligned} \tag{3.26}$$

Proof. The first inequality follows from the fact that

$$\mu\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) = h\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) = \lambda^{[p+1]}$$

is the global maximum per given parameters $(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})$. The third inequality can be argued similarly. It only remains to prove the second inequality.

Introduce the abbreviations $\mathbf{a} := \overline{\mathbf{x}}^{[p+1]} \otimes \overline{\mathbf{y}}^{[p+1]}$ and $\mathbf{b} := \overline{\mathbf{x}}^{[p]} \otimes \overline{\mathbf{y}}^{[p]}$ and rewrite the difference

$$\lambda\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) - \mu\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) = \langle A\mathbf{a}, \mathbf{a} - \mathbf{b}\rangle.$$

Observe that if A is symmetric and positive definite, then we have

$$0 \le \langle A(\mathbf{a} - \mathbf{b}), \mathbf{a} - \mathbf{b}\rangle = \langle A\mathbf{a}, \mathbf{a}\rangle + \langle A\mathbf{b}, \mathbf{b}\rangle - 2\langle A\mathbf{a}, \mathbf{b}\rangle.$$

Therefore,

$$\langle A\mathbf{a}, \mathbf{a} - \mathbf{b}\rangle \ge \frac{\langle A\mathbf{a}, \mathbf{a}\rangle - \langle A\mathbf{b}, \mathbf{b}\rangle}{2}.$$

To prove the second inequality, it suffices to prove

$$\lambda\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}\big) - \lambda\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) \geq 0. \tag{3.27}$$

Let $f : \mathbb{R}^{mn} \to \mathbb{R}$ denote the quadratic form

$$f(\mathbf{w}) := \langle A\mathbf{w}, \mathbf{w}\rangle.$$

Because A is symmetric and positive definite, the level sets

$$\{\mathbf{w} \in \mathbb{R}^{mn} | f(\mathbf{w}) \equiv c\}$$

form a family of concentric hyperellipsoids in $\mathbb{R}^{mn}$. The higher the level c is, the larger the ellipsoid becomes. Regarding the vectors $\mathbf{a}$ and $\mathbf{b}$ as two special points in $\mathbb{R}^{mn}$, we are interested in comparing which levels of hyperellipsoids they reside on. To show (3.27) is to show that $\mathbf{a}$ is at a higher level than $\mathbf{b}$. Observe that

$$\langle \mathbf{a}-\mathbf{b}, \nabla f(\mathbf{b})\rangle = \langle \mathbf{a}-\mathbf{b}, A\mathbf{b}\rangle = h\big(\overline{\mathbf{x}}^{[p+1]}, \overline{\mathbf{y}}^{[p+1]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) - h\big(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}; \overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]}\big) \geq 0,$$

implying that the vector $\mathbf{a}-\mathbf{b}$ is forming an acute angle with the steep ascent direction $\nabla f(\mathbf{b})$ at the point $\mathbf{b}$. That is, the point $\mathbf{a}$ is indeed pointing outward and, hence, at a higher level than $\mathbf{b}$. The second inequality in (3.26) is therefore proved. □

THEOREM 3.10. *Suppose that A is positive definite and satisfies Condition* S. *Suppose also that the sequence $\{\lambda^{[p]}\}$ is strictly increasing. Then the iterates $\{(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})\}$ generated by the scheme* (3.9) *converge.*

Proof. Rewrite

$$\langle \mathbf{a}-\mathbf{b}, A(\mathbf{a}-\mathbf{b})\rangle = (\langle \mathbf{a}, A\mathbf{a}\rangle - \langle \mathbf{a}, A\mathbf{b}\rangle) - (\langle \mathbf{a}, A\mathbf{b}\rangle - \langle \mathbf{b}, A\mathbf{b}\rangle).$$

Therefore, by the interlacing property and Lemma 3.7, we see that

$$\overline{\mathbf{x}}^{[p+1]} \otimes \overline{\mathbf{y}}^{[p+1]} - \overline{\mathbf{x}}^{[p]} \otimes \overline{\mathbf{y}}^{[p]} \to 0.$$

Defining

$$\Delta\overline{\mathbf{x}}^{[p]} := \overline{\mathbf{x}}^{[p+1]} - \overline{\mathbf{x}}^{[p]},$$
$$\Delta\overline{\mathbf{y}}^{[p]} := \overline{\mathbf{y}}^{[p+1]} - \overline{\mathbf{y}}^{[p]},$$

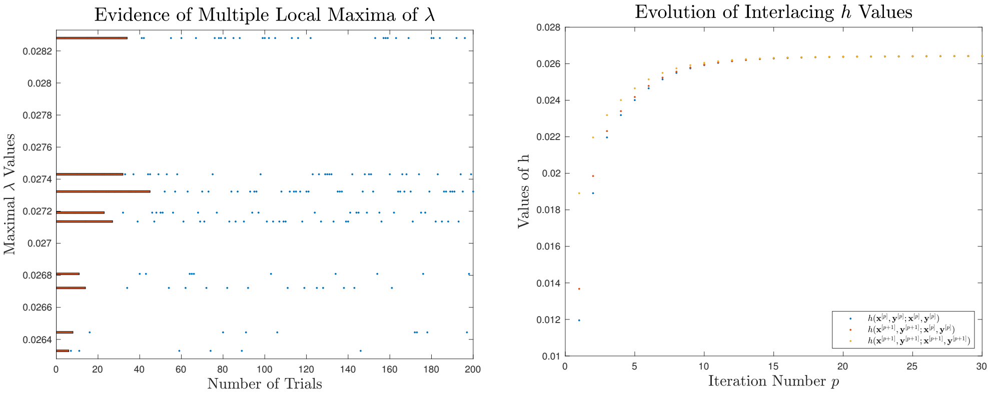
it can be verified that

$$\|\overline{\mathbf{x}}^{[p+1]} \otimes \overline{\mathbf{y}}^{[p+1]} - \overline{\mathbf{x}}^{[p]} \otimes \overline{\mathbf{y}}^{[p]}\|_2^2 = \|\Delta\overline{\mathbf{x}}^{[p]}\|_2^2 + \|\Delta\overline{\mathbf{y}}^{[p]}\|_2^2 - \frac{1}{2}\|\Delta\overline{\mathbf{x}}^{[p]}\|_2^2\|\Delta\overline{\mathbf{y}}^{[p]}\|_2^2.$$

Therefore, we find that $\Delta\overline{\mathbf{x}}^{[p]}$ and $\Delta\overline{\mathbf{y}}^{[p]}$ converge to zero. Using an argument similar to that in Theorem 2.6, these are sufficient to guarantee the convergence of $\{(\overline{\mathbf{x}}^{[p]}, \overline{\mathbf{y}}^{[p]})\}$. □

4. Numerical experiments.

Now that we have completed the theory for the bipartite systems, it is illuminating to consider a few numerical experiments in this section to further demonstrate the working of the two methods we have proposed in this paper. The following experiments are performed on a Windows 10 Pro desktop with Intel Core i7-8700 @ 3.20GHz processor and 8GB RAM by using MATLAB, version 2019a, as the computing platform.



(a) Multiple optimal λ values via 200 random trials

(b) Monotone increasing and interlacing behavior of h values.

FIG. 4.1. *Distribution of optimal λ values and dynamics of h values.*

Example 1. Consider the scenario where $A \in \mathbb{R}^{80\times 80}$ is a randomly generated symmetric and positive definite matrix with unit trace and search for unit vectors $\mathbf{x} \in \mathbb{R}^{16}$ and $\mathbf{y} \in \mathbb{R}^{5}$ to maximize (1.15). We carry out the iteration in the SVD scheme (3.9) until the difference between two consecutive iterates is less than 10^{-10}. With this fixed A, we repeat this experiment 200 times with randomly generated starting values.

We find in one of the test data that the iterates produce nine optimal values, each of which can be reached with significant probability, as shown in Figure 4.1(a), where the frequency of occurrence is marked along the left margin in the horizontal histogram. Such a phenomenon should not be a surprise because we are dealing with a nonconvex optimization problem. Multiple local maxima are normally expected for nonlinear optimization problems. What is not clear is how the number of optimal values, which is nine in this case but may vary in other cases, depends on the problem data A. This question concerns the real-valued solutions to the polynomial system (3.3), which is an important subject in the realm of real algebraic geometry.

The interlacing property (3.26), which is essential to our proof of convergence, is manifested by the right graph in Figure 4.1(b). The result is from only one run of the iteration but is typical in all other runs. For clarity, we display the evolution of the h values defined in (3.4) for the first 30 iterates only. Stacked vertically on top of each other for each p and gradually increased to a common limit point are the first three h values in (3.26), where the variables are updated one pair at a time.

Example 2. While developing the convergence analysis for both iterative schemes is of theoretical interest in its own right, ultimately it is of practical significance to compare the performance of the power-like iteration (2.9) and the SVD-like iteration (3.9) against some of the existing constrained optimization packages. Since we employ MATLAB as the computing platform, we adopt the various solvers available in the MATLAB Optimization Toolbox for comparison.

Some precautions should be taken when conducting such an experiment. First, one iteration in each method may mean significant disparities in the complexities. One iteration of the power-like scheme (2.9) clearly is straightforward, but one iteration of the SVD-like scheme (3.9) involves many iterations within the Lanczos algorithm used by svds, whereas one iteration reported by the MATLAB Optimization Toolbox may involve many bookkeeping and internal chores of which a precise account of cost is difficult. Second, because each method, including the different solvers within MATLAB, has its own special characteristics, there is no obvious way to impose absolutely fair and unified stopping criteria. To maintain a reasonable degree of fairness, we make our comparison by imposing the following specifics across all methods:

1. For test data, we randomly generate a list of density matrices of sizes $n^2 \times n^2$ with $n = 5, 10, 20, 40$, respectively, and look for the unit vectors $\mathbf{x}$ and $\mathbf{y} \in S^{n-1}$ that maximize (1.15).
2. For each size of the test data, We repeat our experiments with 20 different starting values, while applying the same initial values to all methods.
3. For each method, we terminate the iteration whenever the first-order optimality condition approximated by the calculation

$$\left\| \begin{bmatrix} \mathscr{C}(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p+1]})^\top \mathbf{y}^{[p+1]} - \lambda^{[p+1]} \mathbf{x}^{[p+1]} \\ \mathscr{C}(\mathbf{x}^{[p+1]}, \mathbf{y}^{[p+1]}) \mathbf{x}^{[p+1]} - \lambda^{[p+1]} \mathbf{y}^{[p+1]} \end{bmatrix} \right\|_F < 10^{-8}$$

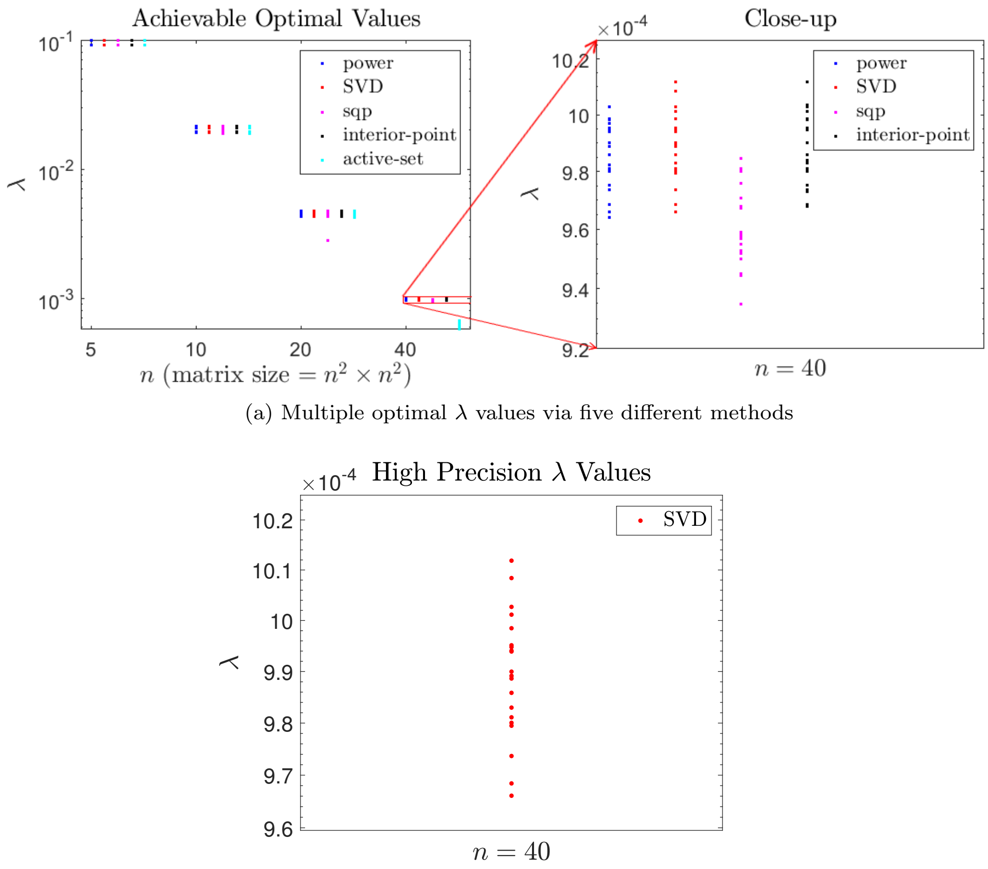
   is satisfied.
4. We compare our two iterative methods with the conventional MATLAB routine fmincon employing solvers sqp, interior-point, and active-set, respectively. For these solvers, we choose the parameter

$$\mathsf{OptimalityTolerance} = 10^{-8}$$

   as the secondary stopping criteria. Also, it might be redundant for other solvers, but the active-set method requires a box constraint that all variables are bounded in the interval $[-1, 1]$.

We first report the achievable optimal λ values. The left diagram of Figure 4.2(a) suggests that all five methods return approximately the same (multiple but discrete) optimal values with the exception that the active-set algorithm returns inferior (smaller) optimal values when the test matrix is of size 1600×1600. In that case, if we disregard the results obtained by the active-set algorithm and zoom in the remaining four results, then we see further details in the right diagram of Figure 4.2(a). First, it seems that there are multiple optimal values, but they are clustered around 10^{-3}. Are these actual different values or are they simply computational artifacts? To address this question, we demand a high precision by refining the stopping criteria to 10^{-14}. Depicted in Figure 4.2(b) is the result by the SVD-like method under the more stringent stopping criteria, which yields almost the same number of optimal solutions as that under more the more relaxed stopping criteria. Thus, these clustered λ values are indeed distinct but close-by optimal values. Second, there is a considerable overlap of achievable optimal values by all methods, confirming that these are valid optimal values. However, there are also values attained by one method but not by others within the 20 trials, even though all methods start with the same initial values. This might have something to do with the innate features unique to different algorithms.

Our codes are direct implementation of the schemes (2.9) and (3.9). They are not as fine tuned as those carefully coded routines in the MATLAB Optimization Toolbox. However, the above observations suggest that our simple schemes are as capable of finding a maximizer for (1.15) as those more sophisticated methods employed in the MATLAB Optimization Toolbox. The next question is which method is more efficient.



(a) Multiple optimal λ values via five different methods

(b) Optimal values obtained by the SVD-like method by refining the stopping criteria to 10^{-14}

FIG. 4.2. *Comparison of optimal λ values.*

The most objective means for gauging the computational complexity used to be directly counting the theoretical number of floating-point operations (flops). However, given advanced computing technologies and highly optimized computer software libraries, counting flops is no longer a good indicator to measure the efficiency of an algorithm. Instead, we compare the overhead by measuring the elapsed time taken by the underlying method to finish the task, i.e., from starting the iteration to meeting the stopping criteria. For fairness, we do not count the time needed to prepare the data. Neither do we furnish analytic Hessian information to some of the MATLAB solvers. For calls of routines such as svds or fmincon, which are already highly optimized, we simply measure the time around the calls without interfering with any internal maneuvers. Depending on the loading of the CPU, the time measurement might fluctuate. So, for each of the methods and each of the dimensions, we repeat the experiment 20 times with randomly generated starting values and measure the CPU time individually. We understand that the face value of time measurement is machine dependent, but the trend should be generally indicative.

The performance in terms of the CPU time is depicted in Figure 4.3. We see that for bipartite systems of sizes up to $n = 20$, the power-like method not only provides a

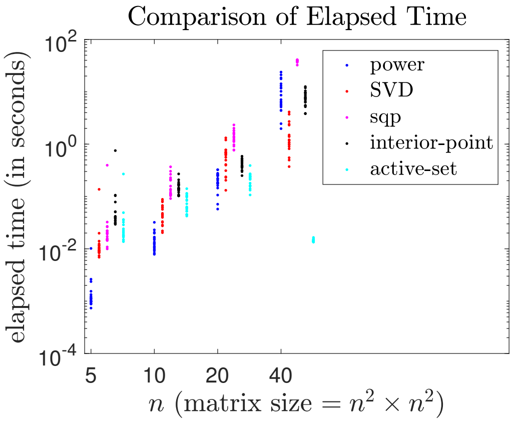


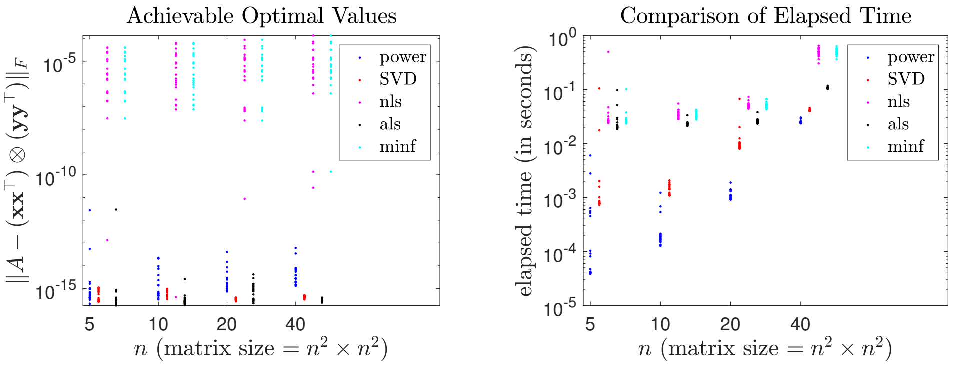
FIG. 4.3. *Comparison of elapsed time (*20 *runs per* n*).*

good approximation but requires relatively less time to reach the state of convergence. On the other hand, the time needed by the SVD-like method seems to grow slowly as the sizes of the matrices increase. For the case of $n = 40$, in almost all trials, it requires one order less time than the other four methods (the active-set method is no longer competitive because it returns a smaller optimal value as is seen in the left diagram of Figure 4.2(a)) to reach convergence.

Example 3. We point out in (1.17) that our problem can be cast as a fourth-order rank-1 approximation problem with shared factors. In this example, we compare our methods with the various solvers available in the Tensorlab toolbox which is designed to handle such a structure effectively. To our knowledge, the current release of the Tensorlab toolbox has not yet implemented the mechanism to compute the norm of the gradient iterate by iterate. Thus, the option OptimalityTolerance is not available. We modify our test as follows.

1. For $n = 5, 10, 20, 40$, we randomly generate an exact, separable, unit trace, rank-1 matrix $A \in \mathbb{R}^{n^2 \times n^2}$ which is then converted via $\mathfrak{A} := \mathsf{reshape}(\mathbf{vec}(\mathscr{R}(A)), [n, n, n, n])$ to an order-4 tensor to be used as the target matrix. Consequently, an ideal calculation should have produced a nearly zero residual in (1.17) with optimal value $\lambda \approx 1$.
2. We compare our two methods with the built-in solvers nls, als, and minf in the Tensorlab toolbox, which invoke the nonlinear least squares, the alternating least squares, and the unconstrained nonlinear optimization techniques, respectively.
3. We use the option model.factorizations.symm.cpd = {'A', 'A', 'B', 'B'} in the Tensorlab toolbox to specify the symmetric structure embedded in (1.17).
4. To stay compatible with the Tensorlab toolbox, we adopt the function tolerance, TolFun $= 10^{-8}$, as the stopping criteria. However, note that the Tensorlab toolbox gauges only the relative changes of function values with respect to the starting residual, whereas we measure the absolute residuals.
5. We repeat our experiments with 20 different starting values, while keeping the same target matrix $\mathfrak{A}$ and starting values for all methods.

In Figure 4.4(a), we plot the final residual values obtained by the various methods. It should be obvious that our SVD-like scheme consistently produces nearly perfect approximations as the solver als does in all tests and the power-like method comes to



(a) Multiple optimal residual values via five different methods

(b) Comparison of elapsed time (20 runs per n)

FIG. 4.4. *Performance comparison.*

the second. However, when we compare CPU time need to meet the stopping criteria, the results plotted in Figure 4.4(b) clearly indicate that our two methods have better efficiency than any of those solvers currently available in Tensorlab.

5. Conclusion.

Quantum entanglement is an indispensable resource for many salient applications as it is capable of delivering fast, concurrent, and secure communication in quantum computing. One critical task in the construction of a quantum system is to measure the "absolute gap" between a mixed state and its nearest separable state. The nonlinearity due to the entanglement among subsystems over the complex field makes the task extremely challenging. This work investigates the entangled problem over the real field as an important first step toward this endeavor.

The rank-1 approximation to bipartite systems is recast in the form of a nonlinear eigenvalue problem and a nonlinear singular value problem. A power-like iteration and an SVD-like iteration are proposed as numerical means to tackle these problems, respectively. These methods bear only a resemblance to the conventional power method and the SVD method, but the iterative schemes are nonstationary and nonlinear in nature. This paper focuses on analyzing the limiting behavior of these two methods. Convergence is guaranteed under generic conditions. Implementation of the proposed iterative methods is straightforward, but experimental results suggest that they are favorably comparable in both precision and efficiency with the more sophisticated optimization routines available in the MATLAB Optimization Toolbox and Tensorlab.

This study serves as the building block for the low-rank approximation to the general entangled systems. Future work includes a generalization of the theory to low-rank approximation to bipartite systems, to the general multipartite systems, and ultimately to systems over the complex field.

REFERENCES

[1] S. Aaronson, *Quantum Computing Since Democritus*, Cambridge University Press, Cambridge, UK, 2013.

[2] S. N. Afriat, *Theory of maxima and the method of Lagrange*, SIAM J. Appl. Math., 20 (1971), pp. 343–357.

[3] J. Barrett, A. Kent, and S. Pironio, *Maximally nonlocal and monogamous quantum correlations*, Phys. Rev. Lett., 97 (2006), 170409.

[4] J. S. Bell, *On the Einstein Podolsky Rosen paradox*, Physics Physique Fizika, 1 (1964), pp. 195–200.

[5] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, *Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels*, Phys. Rev. Lett., 70 (1993), pp. 1895–1899.

[6] J. Brachat, P. Comon, B. Mourrain, and E. Tsigaridas, *Symmetric tensor decomposition*, Linear Algebra Appl., 433 (2010), pp. 1851–1872.

[7] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.

[8] N. J. Cerf, M. Bourennane, A. Karlsson, and N. Gisin, *Security of quantum key distribution using d-level systems*, Phys. Rev. Lett., 88 (2002), 127902.

[9] N. Chandra, *Quantum Entanglement in Electron Optics: Generation, Characterization, and applications*, At. Opt. Plasma Phys., Springer, Berlin, 2013.

[10] K. Chen and L.-A. Wu, *A matrix realignment method for recognizing entanglement*, Quantum Inf. Comput., 3 (2003), pp. 193–202.

[11] L. Chen, M. Aulbach, and M. Hajdušek, *Comparison of different definitions of the geometric measure of entanglement*, Phys. Rev. A, 89 (2014), 042305.

[12] E. Chitambar, C. A. Miller, and Y. Shi, *Matrix pencils and entanglement classification*, J. Math. Phys., 51 (2010), 072205.

[13] M. T. Chu and J. L. Watterson, *On a multivariate eigenvalue problem.* I. *Algebraic theory and a power method*, SIAM J. Sci. Comput., 14 (1993), pp. 1089–1106.

[14] P. Comon, *Tensor Decompositions: State of the art and applications*, in Mathematics in Signal Processing, V (Coventry, 2000), Inst. Math. Appl. Conf. Ser. New Ser. 71, Oxford University Press, Oxford, UK, 2002, pp. 1–24.

[15] P. Comon, X. Luciani, and A. L. F. de Almeida, *Tensor decompositions, alternating least squares, and other tales*, J. Chemometrics, 23 (2009), pp. 393–405.

[16] G. Dahl, J. M. Leinaas, J. Myrheim, and E. Ovrum, *A tensor product matrix approximation problem in quantum physics*, Linear Algebra Appl., 420 (2007), pp. 711–725.

[17] M. Dana and K. D. Ikramov, *On the codimension of the variety of symmetric matrices with multiple eigenvalues*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 323 (2005), pp. 34–46, 224.

[18] L. De Lathauwer, B. De Moor, and J. Vandewalle, *On the best rank-*1 *and rank-(r_1, r_2,..., r_n) approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342.

[19] J. Dittmann, *Explicit formulae for the Bures metric*, J. Phys. A, 32 (1999), pp. 2663–2670.

[20] A. Ekert and P. L. Knight, *Entangled quantum systems and the Schmidt decomposition*, Amer. J. Phys., 63 (1995), pp. 415–423.

[21] A. K. Ekert, *Quantum cryptography based on Bell's theorem*, Phys. Rev. Lett., 67 (1991), pp. 661–663.

[22] N. Friis, G. Vitagliano, M. Malik, and M. Huber, *Entanglement certification from theory to experiment*, Nature Reviews Physics, 1 (2019), pp. 72–87.

[23] C. B. García and T.-Y. Li, *On the number of solutions to polynomial systems of equations*, SIAM J. Numer. Anal., 17 (1980), pp. 540–546.

[24] S. Gharibian, *Strong NP-hardness of the quantum separability problem*, Quantum Inf. Comput., 10 (2010), pp. 343–360.

[25] S. Gröblacher, T. Jennewein, A. Vaziri, G. Weihs, and A. Zeilinger, *Experimental quantum cryptography with qutrits*, New J. Phys., 8 (2006), pp. 75–75.

[26] Y. Guan, M. T. Chu, and D. Chu, *SVD-based algorithms for the best rank-*1 *approximation of a symmetric tensor*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1095–1115.

[27] O. Gühne and G. Tóth, *Entanglement detection*, Phys. Rep., 474 (2009), pp. 1–75.

[28] L. Gurvits, *Classical complexity and quantum entanglement*, J. Comput. System Sci., 69 (2004), pp. 448–484.

[29] M. Hayashi, *Quantum Information Theory*: *Mathematical Foundation*, 2nd ed., Grad. Texts Phys., Springer-Verlag, Berlin, 2017.

[30] F. Hiai and D. Petz, *Introduction to Matrix Analysis and Applications*, Universitext, Springer, Cham, 2014.

[31] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, *Quantum entanglement*, Rev. Mod. Phys., 81 (2009), pp. 865–942.

[32] M. Huber and M. Pawłowski, *Weak randomness in device-independent quantum key distribution and the advantage of using high-dimensional entanglement*, Phys. Rev. A, 88 (2013), 032309.

[33] L. M. Ioannou, B. C. Travaglione, D. Cheung, and A. K. Ekert, *Improved algorithm for quantum separability and entanglement detection*, Phys. Rev. A, 70 (2004), 060303.

[34] M. Ishteva, P.-A. Absil, and P. Van Dooren, *Jacobi algorithm for the best low multilinear rank approximation of symmetric tensors*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 651–672.

[35] N. Johnston, *Norm duality and the cross norm criteria for quantum entanglement*, Linear Multilinear Algebra, 62 (2014), pp. 648–658.

[36] R. Karam, *Why are complex numbers needed in quantum mechanics? Some answers for the introductory level*, Amer. J. Phys., 88 (2020), pp. 39–45.

[37] H. B. Keller, *Geometrically isolated nonisolated solutions and their approximation*, SIAM J. Numer. Anal., 18 (1981), pp. 822–838.

[38] B. N. Khoromskij, *Structured rank-$(R_1, \ldots, R_D)$ decomposition of function-related tensors in $\mathbb{R}^D$*, Comput. Methods Appl. Math., 6 (2006), pp. 194–220.

[39] E. Kofidis and P. A. Regalia, *On the best rank-*1 *approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2001/02), pp. 863–884.

[40] T. G. Kolda, *Numerical optimization for symmetric tensor decomposition*, Math. Program., 151 (2015), pp. 225–248.

[41] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[42] S.-H. Kye, *Necessary conditions for optimality of decomposable entanglement witnesses*, Rep. Math. Phys., 69 (2012), pp. 419–426.

[43] J. M. Leinaas, J. Myrheim, and E. Ovrum, *Geometrical aspects of entanglement*, Phys. Rev. A (3), 74 (2006), 012313.

[44] M. Melucci, *Introduction to Information Retrieval and Quantum Mechanics*, Information Retrieval Series 35, Springer, Heidelberg, 2015.

[45] M. J. Mohlenkamp, *Musings on multilinear fitting*, Linear Algebra Appl., 438 (2013), pp. 834–852.

[46] M. Nakahara and T. Ohmi, *Quantum Computing: From Linear Algebra to Physical Realizations*, CRC Press, Boca Raton, FL, 2008.

[47] G. Ni, L. Qi, and M. Bai, *Geometric measure of entanglement and U-eigenvalues of tensors*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 73–87.

[48] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information:* 10*th Anniversary Edition*, Cambridge University Press, Cambridge, UK, 2010.

[49] R. S. Palais, *The principle of symmetric criticality*, Comm. Math. Phys., 69 (1979), pp. 19–30.

[50] M. B. Plbnio and S. Virmani, *An introduction to entanglement measures*, Quantum Info. Comput., 7 (2007), pp. 1–51.

[51] R. Raussendorf and H. J. Briegel, *A one-way quantum computer*, Phys. Rev. Lett., 86 (2001), pp. 5188–5191.

[52] A. J. Sommese and C. W. Wampler, II, *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific, Hackensack, NJ, 2005.

[53] A. Takayama, *Mathematical Economics*, 2nd ed., Cambridge University Press, Cambridge, UK, 1985.

[54] S. Tamaryan, *Completely mixed state is a critical point for three-qubit entanglement*, Phys. Lett. A, 375 (2011), pp. 2224–2229.

[55] B. M. Terhal, *Detecting quantum entanglement*, Theoret. Comput. Sci., 287 (2002), pp. 313–335, natural computing.

[56] W. Thirring, R. A. Bertlmann, P. Köhler, and H. Narnhofer, *Entanglement or separability: The choice of how to factorize the algebra of a density matrix*, Eur. Phys. J. D, 64 (2011), pp. 181–196.

[57] A. Uschmajew, *Local convergence of the alternating least squares algorithm for canonical tensor approximation*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 639–652.

[58] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.

[59] C. F. Van Loan and N. Pitsianis, *Approximation with Kronecker Products*, in Linear Algebra for Large Scale and Real-Time Applications (Leuven, 1992), NATO Adv. Sci. Inst. Ser. E Appl. Sci. 232, Kluwer, Dordrecht, 1993, pp. 293–314.

[60] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, *Tensorlab* 3.0, 2016, https://www.tensorlab.net/.

[61] L. Wang and M. T. Chu, *On the global convergence of the alternating least squares method for rank-one approximation to generic tensors*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1058–1072.

[62] M. M. Wilde, *Quantum Information Theory*, 2nd ed., Cambridge University Press, Cambridge, UK, 2017.

[63] K. Wright, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math., 3 (1992), pp. 283–295.

[64] S.-J. Wu and M. T. Chu, *Markov chains with memory, tensor formulation, and the dynamics of power iteration*, Appl. Math. Comput., 303 (2017), pp. 226–239.

[65] T. Zhang and G. H. Golub, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.