Convergence of Gradient Descent for Low-Rank Matrix Approximation

Renaud-Alexandre Pitaval, Wei Dai and Olav Tirkkonen

***Abstract*—This paper provides a proof of global convergence of gradient search for low-rank matrix approximation. Such approximations have recently been of interest for large scale problems, as well as for dictionary learning for sparse signal representations and matrix completion. The proof is based on the interpretation of the problem as an optimization on the Grassmann manifold and Fubiny-Study distance on this space.**

I. Introduction

Consider the problem of approximating a matrix by another lower-rank matrix. The solution to this problem is well-known to be given by the truncated singular value decomposition (SVD) up to the desired rank [1], [2]. In this paper, we investigate if one is always able to find the optimum approximation through a classical first-order optimization algorithm such as a gradient search. The answer is shown to be positive almost surely (i.e. with probability one).

Low-rank matrix approximation is a ubiquitous problem in data processing. Gradient descent has been employed for truncated SVD in large scale problems [3]–[6] and in related matrix completion settings [7]–[9]. The considered low-rank matrix approximation has also application in dictionary learning for sparse signal representations. For some applications, it is desirable to use a learning-based approach adapting the dictionaries based on training data sets. Dictionary update can be formulated as an optimization problem on manifolds [10] generalizing the MOD [11] and K-SVD [12] algorithms. Furthermore, several approximation methods are based on the power method which is a gradient-descent type algorithm [13], including non-negative approximation [14], [15] and sparse approximation [16], [17].

Singular value decomposition is also directly related to Rayleigh quotient (RQ) maximization. RQ extremization is a long-standing problem and a number of RQ algorithms and their convergence properties have been discussed [13], [18]. The seminal work by Edelman, Arias and Smith [13] provides a taxonomy of such algorithms based on a unified geometric understanding, see also [18]. The algorithms are classified as variations of Newton, conjugate gradient, and gradient descent (a.k.a steepest descent) methods on Grassmann manifolds. These three methods differ in their speed of convergence. Conjugate gradient or Newton algorithms provide in general faster convergence, but also require higher complexity. Early approaches to RQ maximization are based on extrinsic Euclidean algorithms in embedding space with practical step sizes. These are locally equivalent to idealized intrinsic Riemannian methods. Among such algorithms, the Rayleigh quotient iteration (RQI) is a popular algorithm corresponding to a Newton method. In general, these algorithms are well-known to converge locally, however their global convergence properties seem to be less understood. Following the taxonomy of [13], some results on convergence of RQ algorithms can be classified as follows.

Gradient Descent: For a rank-one problem, a steepest gradient algorithm has been shown to converge globally to *an* eigenvalue [19]. This eigenvalue is the optimum one if the starting point is not orthogonal to the optimum eigenvector. The higher rank problem is considered in [20] but a proof of convergence is only given for rank one. Some intermediate results are given for higher rank where at least one dimension is shown to converge to the rank-one optimum and the other dimensions are converging to some *other* eigenvalues.

Congugate Gradient: The global convergence result of [19] is extended to a generalized RQ conjugate-gradient algorithm for rank one in [21]. Local convergence properties of conjugate-gradient algorithms are discussed in [22], [23], showing faster convergence of conjugate-gradient than steepest descent.

Newton Methods: According to [24], the global convergence properties of rank-one RQI are well understood. Rank-one RQI was shown to either converge to an eigenvector or converge to the bisector of a pair of eigenvectors in [25]. Later, the set of points for which the RQI does not converge to *an* eigenvector was shown to be a set of measure zero [26]. A Grassmann-RQI is presented in [24]. It generalizes the classical RQI to higher dimensional subspaces, along with its cubic local convergence. For higher-rank, good global convergence properties were also observed in [25], however the possibility of establishing a global convergence analysis along the lines of previous proofs was challenged. Local quadratic convergence of an intrinsic Newton method is also discussed in [27].

In this paper, we discuss the global convergence of an idealized gradient descent procedure. A true gradient search moves in the gradient direction with infinitesimal step sizes,

This work was done while the first author was with Aalto University, and has been supported in part by the Academy of Finland.

R.-A. Pitaval is with Huawei Technologies Sweden, Box 54, SE-164.94, Kista, Sweden (email: renaud.alexandre.pitaval@huawei.com).

W. Dai is with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom (email: wei.dai1@imperial.ac.uk).

O. Tirkkonen is with the Department of Communications and Networking, Aalto University, P.O. Box 13000, FI-00076, Finland (e-mail: olav.tirkkonen@aalto.fi).

DOI of published version: 10.1109/TIT.2015.2448695

such assumption has been previously used for convergence study e.g. in [28]. According to our knowledge, this is the first proof showing that an ideal gradient search on a Grassmann manifold almost surely solves the multiple-rank matrix approximation problem. Previously, it was known that there are multiple stationary points for the rank-one matrix approximation problem [29]. More recently, for the rank-one case, it was shown that a gradient descent method will not converge to any other stationary points than the global minimizer with probability one [10]. Our results show that this is generally true for any higher rank.

The proof consists in showing that the Fubiny-Study distance to the optimum is monotonically decreasing along the gradient path for almost all starting points on the Grassmann manifold. Since under the Fubiny-Study distance, all other stationary points are antipodal to the global optimum, by moving along the gradient path the algorithm is getting closer to the optimum while ultimately never approaching the other stationary points, i.e. the Grassmann manifold is almost everywhere the basin of attraction of the global optimum. The multiple-rank case is more delicate than rank-one as we have to handle multiple principal angles, which results in a non-unique notion of distance on the Grassmann manifold. As a result, the choice of distance in the proof is important. We make this argument explicit by showing that a similar argumentation with the chordal distance would not allow us to claim convergence of the algorithm to a global optimum but the decrement of the Fubini-Study distance does. Finally, our result provides theoretical support for application of optimization methods to low-rank matrix approximation problems.

II. Preliminaries on Manifolds

The gradient descent is performed on a curved surface consisting of constrained-norm matrices. The proof of convergence will rely on the notion of Stiefel and Grassmann manifolds.

The complex Stiefel manifold $\mathcal{V}_{n,r}^{\mathbb{C}}$ is defined as the space of rectangular unitary matrices (with $r \leq n$):

$$\mathcal{V}_{n,r}^{\mathbb{C}} = \left\{ \boldsymbol{Y} \in \mathbb{C}^{n \times r} \quad | \quad \boldsymbol{Y}^H \boldsymbol{Y} = \boldsymbol{I}_r \right\}. \tag{1}$$

When $r = 1$, the Stiefel manifold can be identified as a unit hypersphere, and for $r = n$ as the unitary group $\mathcal{U}_r$. We denote by $\boldsymbol{I}_{n,r} \in \mathcal{V}_{n,r}^{\mathbb{C}}$ the truncation of the first r columns of the identity matrix $\boldsymbol{I}_n$.

The complex Grassmann manifold $\mathcal{G}_{n,r}^{\mathbb{C}}$ is the set of all r-dimensional subspaces of $\mathbb{C}^n$. This manifold can be expressed as the quotient space of the Stiefel manifold and the unitary group:

$$\mathcal{G}_{n,r}^{\mathbb{C}} \cong \mathcal{V}_{n,r}^{\mathbb{C}} / \mathcal{U}_r. \tag{2}$$

A point in the Grassmann manifold can thus be represented as the equivalence class of $n \times r$ Stiefel matrices whose columns span the same space:

$$[\boldsymbol{Y}] = \{ \boldsymbol{Y}\boldsymbol{U} \quad | \quad \boldsymbol{U} \in \mathcal{U}_r \}, \tag{3}$$

where $\boldsymbol{Y} \in \mathcal{V}_{n,r}^{\mathbb{C}}$.

Several different distances can be defined on the Grassmann manifold [13], [30] based on the notion of principal angles. We denote two subspaces of $\mathbb{C}^n$ as $[\boldsymbol{Y}]$, $[\boldsymbol{Z}] \in \mathcal{G}_{n,r}^{\mathbb{C}}$, with $\boldsymbol{Y}, \boldsymbol{Z} \in \mathcal{V}_{n,r}^{\mathbb{C}}$. The singular values of $\boldsymbol{Y}^H \boldsymbol{Z}$ are $\{\cos\theta_1, \cdots, \cos\theta_r\}$ where $\theta_1, \ldots, \theta_r \in [0, \frac{\pi}{2}]$ are the principal angles between these two subspaces [30]. We will make use of the two following distances

1) The **chordal distance** :

$$\begin{aligned} d_{\mathrm{CH}}([\boldsymbol{Y}], [\boldsymbol{Z}]) &= \frac{1}{\sqrt{2}} \|\boldsymbol{Y}\boldsymbol{Y}^H - \boldsymbol{Z}\boldsymbol{Z}^H\|_F & (4) \\ &= \left(r - \|\boldsymbol{Y}^H \boldsymbol{Z}\|_F^2\right)^{1/2} & (5) \\ &= \left(r - \sum_{i=1}^{r} \cos^2\theta_i\right)^{1/2} & (6) \\ &= \left(\sum_{i=1}^{r} \sin^2\theta_i\right)^{1/2}. & (7) \end{aligned}$$

2) The **Fubini-Study distance**

$$\begin{aligned} d_{\mathrm{FS}}([\boldsymbol{Y}], [\boldsymbol{Z}]) &= \arccos|\det[\boldsymbol{Y}^H \boldsymbol{Z}]| & (8) \\ &= \arccos\left(\prod_{i=1}^{r} \cos\theta_i\right). & (9) \end{aligned}$$

These two distances are non-equivalent in the sense they are derived from different embeddings thus corresponding to different Riemannian metrics. The chordal distance is obtained from the embedding of the Grassmann manifold to the set of n-by-n projection matrices of rank r [30]. The Fubini-Study distance is derived via the Plücker embedding [31].

III. Gradient Descent Procedure for Low-Rank Matrix Approximation

We describe a gradient descent procedure on the Grassmann manifold to solve a low-rank matrix approximation problem.

A. Low-rank Matrix Approximation

Consider a matrix $\boldsymbol{A} \in \mathbb{C}^{n \times p}$ with $p \leq n$, and an integer $r \leq p$. The rank-r matrix approximation problem is to find the approximation $\hat{\boldsymbol{A}}$ given by $\min_{\mathrm{rank}\hat{\boldsymbol{A}}=r} \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_F$. This problem can be reformulated equivalently as the following optimization on the singular subspaces of $\boldsymbol{A}$:

$$\min_{\boldsymbol{U} \in \mathcal{V}_{n,r}^{\mathbb{C}}} f(\boldsymbol{U}) \tag{10}$$

where

$$f(\boldsymbol{U}) = \min_{\boldsymbol{W} \in \mathbb{C}^{p \times r}} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^H\|_F^2, \quad \forall \boldsymbol{U} \in \mathcal{V}_{n,r}^{\mathbb{C}}. \tag{11}$$

The solution of (10) is known to be achieved by the left singular subspace associated with the r largest singular values [1], [2].

The optimization (10) is actually over the Grassmann manifold $\mathcal{G}_{n,r}^{\mathbb{C}}$. A simultaneous right unitary rotation of any pair $(\boldsymbol{U}, \boldsymbol{W})$ leads to the same value of the objective function $\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^H\|_F^2$. Then for any $\Omega \in \mathcal{U}_r$, one can verify that $f(\boldsymbol{U}\Omega) = f(\boldsymbol{U})$. For every fixed $\boldsymbol{U}$, the optimal solution in the least square problem (11) is $\boldsymbol{W} = \boldsymbol{A}^H\boldsymbol{U}$. We may then rewrite

$$\begin{aligned} f(\boldsymbol{U}) &= \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{U}^H\boldsymbol{A}\|_F^2 = \|(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^H)\boldsymbol{A}\|_F^2 \\ &= \|\boldsymbol{A}\|_F^2 - \|\boldsymbol{U}^H\boldsymbol{A}\|_F^2. \qquad (12) \end{aligned}$$

We note here that the power method is a gradient descent to minimize the objective function (12) [13], [29].

B. Global Minimizer on the Grassmann Manifold

The singular value decomposition of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{U}_A\Sigma\boldsymbol{V}_A^H$ with ordered singular values $\sigma_1 \geq \cdots \geq \sigma_p$, and left singular vectors $\boldsymbol{U}_A = (\boldsymbol{U}_{\text{opt}}\,\boldsymbol{U}_{\text{opt}\perp})$ where $\boldsymbol{U}_{\text{opt}} \in \mathcal{V}_{n,r}^{\mathbb{C}}$. The notion of Grassmann manifold is essential in the proof. We will assume the case $\sigma_r > \sigma_{r+1}$, so that all global minimizers of f are in the subspace spanned by $\boldsymbol{U}_{\text{opt}}$, i.e there is a unique global minimizer on the Grassmann manifold $[\boldsymbol{U}_{\text{opt}}] \in \mathcal{G}_{n,r}^{\mathbb{C}}$. For the degenerate case, there are several global minimizers on the Grassmann manifold, nevertheless the proof of convergence still holds.

C. Gradient Descent Procedure

The gradient of f at $\boldsymbol{U}$ is obtained by differentiating f and projecting onto the tangent space at $\boldsymbol{U}$. The definitions used for derivatives and gradients are given in Appendix A. The differential of f at $\boldsymbol{U}$ is $\mathrm{D}f(\boldsymbol{U}) = -\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{U}$, and the gradient restricted to the Grassmann manifold is

$$\nabla f(\boldsymbol{U}) = -(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^H)\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{U}. \qquad (13)$$

Note that here projecting to the tangent space at $\boldsymbol{U}$ of the Stiefel manifold, or to the tangent space at $[\boldsymbol{U}]$ of the Grassmann manifold leads to the same gradient, cf. Equations (2.53) and (2.70) in [13].

An ideal gradient search moves towards the optimum along an intrinsic path on the Grassmann manifold. Let $\boldsymbol{H}_f = -\nabla f(\boldsymbol{U})$ be the negative gradient matrix of the objective funtion f at $\boldsymbol{U}$. The geodesic in direction $\boldsymbol{H}_f$ emanating from $[\boldsymbol{U}]$ can be written from a matrix exponential [13],

$$\boldsymbol{U}(t) = (\boldsymbol{U}\;\boldsymbol{U}_\perp)\exp\left(t\begin{pmatrix} 0 & -\boldsymbol{B}_\nabla^H \\ \boldsymbol{B}_\nabla & 0 \end{pmatrix}\right)\boldsymbol{I}_{n,r} \qquad (14)$$

where $\boldsymbol{B}_\nabla = \boldsymbol{U}_\perp^H\boldsymbol{H}_f$ as $\boldsymbol{H}_f$ should satisfy [13, Eq. (2.63)] in order to be in the horizontal space at $\boldsymbol{U}$; and $\boldsymbol{U}_\perp$ is an orthogonal complement of $\boldsymbol{U}$ so that $\begin{pmatrix}\boldsymbol{U} & \boldsymbol{U}_\perp\end{pmatrix}$ is a unitary matrix. This follows from the embedding of the Grassmann manifold into the unitary group and taking the corresponding exponential map. There exists several practical methods to efficiently approximate the matrix exponential [32].

A gradient search with constant step size α proceeds as follows:

Gradient descent procedure:

- Given $\boldsymbol{U} \in \mathcal{V}_{n,r}^{\mathbb{C}}$, compute the negative gradient matrix $\boldsymbol{H}_f = -\nabla f(\boldsymbol{U})$.
- Move from $\boldsymbol{U}$ in the direction $\boldsymbol{H}_f$ to $\boldsymbol{U}(\alpha)$ according to (14).
- Repeat until convergence.

For an infinitesimal step size ϵ, the Riemann gradient update can be approximated by an Euclidean update in the tangent space, and one recovers the classical linear gradient procedure,

$$\begin{aligned} \boldsymbol{U}(\epsilon) &\approx [\boldsymbol{U}\,\boldsymbol{U}_\perp]\left(\boldsymbol{I} + \epsilon\begin{pmatrix} 0 & -\boldsymbol{B}_\nabla^H \\ \boldsymbol{B}_\nabla & 0 \end{pmatrix}\right)\boldsymbol{I}_{n,r} \qquad (15) \\ &= \boldsymbol{U} + \epsilon\boldsymbol{U}_\perp\boldsymbol{B}_\nabla = \boldsymbol{U} - \epsilon\boldsymbol{U}_\perp\boldsymbol{U}_\perp^H\nabla f \qquad (16) \\ &= \boldsymbol{U} - \epsilon\nabla f. \qquad (17) \end{aligned}$$

For completeness, we note that an alternative formulation of the geodesic formula was given in [13, Eq. (2.65)] in terms of the SVD of $\boldsymbol{H}_f = \boldsymbol{L}_{\text{H}}\Sigma_{\text{H}}\boldsymbol{R}_{\text{H}}^H$,

$$\boldsymbol{U}(t) = \boldsymbol{U}\boldsymbol{R}_{\text{H}}\cos(\Sigma_{\text{H}}t)\boldsymbol{R}_{\text{H}}^H + \boldsymbol{L}_{\text{H}}\sin(\Sigma_{\text{H}}t)\boldsymbol{R}_{\text{H}}^H \qquad (18)$$

Due to the quotient space structure, the right-multiplication by $\boldsymbol{R}_{\text{H}}^H$ can be omitted for simplification.

IV. Convergence Result

In this section, global convergence of Grassmannian gradient search for low-rank approximation is presented with the main lines of the proof. The crux of the proof is to polarize the Grassmann manifold with the Fubini-Study distance: with this choice of distance the global optimum is antipodal to all other stationary points. Then along the gradient descent path, it is shown that this distance is monotonically decreasing and thus guaranteeing not approaching any other stationary points. The result generalizes [10, Thm 1] to higher rank matrix approximation, but we note that the techniques used here, notably for the proof of Lemma 1, are rather different than in [10].

We consider an idealized gradient search with infinitesimal step size. By construction, the objective function is decreasing along the gradient path as

$$\nabla_{\boldsymbol{H}_f} f = -\|\nabla f\|^2 < 0. \qquad (19)$$

The existence of a convergent sequence is guaranteed by the smoothness of f. The assumption of an infinitesimal step insures a decrease in f at every step of the algorithm and hence convergence to a finite value. Since the overall step length is gradient-related, the convergence is guaranteed to be toward a stationary point, see Appendix B. Local convergence with more elaborated step rules for faster convergence are discussed in [29].

Theorem 1. *Starting from a uniformly randomly chosen point on the Stiefel manifold, the gradient descent procedure on low-rank matrix approximation* (10) *with infinitesimal steps converges to a global minimizer with probability one.*

Sketch of Proof: First, with infinitesimal steps, the gradient search converges locally to a stationary point, see Appendix B.

Then, let define the Fubini-Study distance between the subspace spanned by $\boldsymbol{U}$ and the subspace spanned by the optimum $\boldsymbol{U}_{\text{opt}}$, denoted by

$$\text{dist}_{\text{FS}}(\boldsymbol{U}) = d_{\text{FS}}([\boldsymbol{U}], [\boldsymbol{U}_{\text{opt}}]). \qquad (20)$$

Accordingly, define the set

$$\mathcal{B} = \{\boldsymbol{U} \in \mathcal{V}_{n,r}^{\mathbb{C}} \quad | \quad \text{dist}_{\text{FS}}(\boldsymbol{U}) = \frac{\pi}{2}\}. \qquad (21)$$

corresponding to matrices generating Grassmannian planes with a maximal principal angle w.r.t. $[\boldsymbol{U}_{\text{opt}}]$ equal to $\frac{\pi}{2}$, so that the FS-distance attains its maximal value $\frac{\pi}{2}$. This is a set of measure zero [33] of 'bad' starting points. If $\boldsymbol{U} \in \mathcal{B}$, the subspace $[\boldsymbol{U}]$ has a dimension orthogonal with $[\boldsymbol{U}_{\text{opt}}]$, i.e. $\text{rank}(\boldsymbol{U}_{\text{opt}}^H \boldsymbol{U}) < r$ or $\det[\boldsymbol{U}_{\text{opt}}^H \boldsymbol{U}] = 0$. Conversely, if $\boldsymbol{U} \notin \mathcal{B}$, by definition one satisfies $\text{dist}_{\text{FS}}(\boldsymbol{U}) < \frac{\pi}{2}$.

Recall that $\boldsymbol{H}_f = -\nabla f(\boldsymbol{U})$ is the negative gradient matrix of the objective funtion f at $\boldsymbol{U}$. The following result is proved in Appendix C.

Lemma 1. *Starting from $\boldsymbol{U} \notin \mathcal{B}$, the Fubini-Study distance to the optimum solution is strictly monotonically decreasing along the gradient descent path, i.e. $\nabla_{\boldsymbol{H}_f} \text{dist}_{\text{FS}} < 0$, $\forall \boldsymbol{U}$ satisfying $\text{dist}_{\text{FS}}(\boldsymbol{U}) \in (0,\, \frac{\pi}{2})$.*

Lemma 1 implies that first starting from $\boldsymbol{U} \notin \mathcal{B}$, the gradient procedure will step away from $\mathcal{B}$ and thus never enter $\mathcal{B}$, since $\text{dist}_{\text{FS}}(\boldsymbol{U}) < \frac{\pi}{2}$ will hold along the gradient path. Secondly, it verifies that $\nabla f \neq 0$ for all $\boldsymbol{U}$ satisfying $\text{dist}_{\text{FS}}(\boldsymbol{U}) \in (0,\, \frac{\pi}{2})$[1].

Hence, the only stationary points in $\mathcal{V}_{n,r}^{\mathbb{C}} \setminus \mathcal{B}$ belong to the equivalence class $[\boldsymbol{U}_{\text{opt}}]$. So if the gradient procedure converges to a stationary point, it can only be the global optimum on the Grassmann manifold. Finally, since $\mathcal{B}$ is a zero-measure set, Theorem 1 is proved.

V. Discussions

A. Chordal Distance

To emphasize the importance of the choice of distance in the proof of Theorem 1, we provide a similar result than Lemma 1 for the chordal distance. The proof is in Appendix D.

Lemma 2. *The chordal distance to the optimum solution is monotonically decreasing along the gradient descent path, i.e. $\nabla_{\boldsymbol{H}_f} \text{dist}_{\text{CH}} \leq 0$. If $\boldsymbol{U}$ is not a stationary point, there is a strict decrease in the chordal distance w.r.t. the optimum.*

Replacing Lemma 1 by Lemma 2 would not be sufficient to prove the convergence to a global optimum. A strict decrease in the chordal distance w.r.t. the optimum does not guarantee that none of the principal angles is converging to $\frac{\pi}{2}$, meaning that the algorithm would enter $\mathcal{B}$ and converge to another stationary point than the global minimum. On the other hand, when starting from $\mathcal{B}$, the chordal distance to the optimum will decrease but be strictly lower-bounded by $\sqrt{K}$, where $K > 0$ is the number of dimensions in the starting plane $[\boldsymbol{U}]$ orthogonal to the optimum. The gradient search will converge to the closest point to the optimum in $\mathcal{B}$ as only the principal angles not equal to $\pi/2$ will converge to zero.

B. Comments for Non-Infinitesimal Steps

Theorem 1 implies that there exists a small step size α that guarantee global convergence. In practice, for fast convergence rate, one desires to use the largest possible steps, rather than infinitesimal steps. In the rank-one case, starting from any point in $\mathcal{B}$ a gradient descent path would stay in $\mathcal{B}$ [10] with any step size. However, for ranks higher than one and with non-infinitesimal steps, the set $\mathcal{B}$ is not anymore an absolute bottleneck. If the starting point is not a stationary point, the search is able to escape $\mathcal{B}$ with a large-enough step size. On the other hand, starting from $\boldsymbol{U} \notin \mathcal{B}$ and following the geodesic direction of the gradient at $\boldsymbol{U}$, one enters $\mathcal{B}$ only for some discrete, periodic values of step size α. The periodicity comes from the fact that the path goes around the Grassmann manifold which is a closed curved surface.

VI. Conclusion

A proof of the global convergence of an ideal gradient search for low-rank matrix approximations has been presented. This generalizes a recently shown result for rank-one approximation to higher rank.

Appendix

A. Definitions of Derivatives and Gradients

Since the result is presented for complex matrices, the generalized definition of complex derivative as in [34]–[36] is used.

Given a real function f of complex matrix input $\boldsymbol{X}$, we define the complex derivative as

$$\mathrm{D}f(\boldsymbol{X}) = \frac{df}{d\boldsymbol{X}^*} \tag{22}$$

where the derivative for matrix input is defined componentwise, i.e. such that $[df/d\boldsymbol{X}^*]_{k,l} = df/d[\boldsymbol{X}^*]_{k,l}$; and the complex derivative of a real-valued scalar function f with complex input x is defined as

$$\frac{df}{dx^*} = \frac{1}{2}\left(\frac{\partial f}{\partial \Re[x]} + i\frac{\partial f}{\partial \Im[x]}\right) \tag{23}$$

The variables x and x^* can be treated as independent variables, leading e.g. to

$$\mathrm{D}\text{Tr}[\boldsymbol{X}^H \boldsymbol{M} \boldsymbol{X}] = \boldsymbol{M}\boldsymbol{X}. \tag{24}$$

Note that the derivative $\mathrm{D}f$ is the conventional gradient in the ambient space of our manifold problem. The functions considered are rather functions acting on the Grassmann manifold rather than its linear representation. For computational purpose, it is appropriated, with a small abuse of notation, to express derivatives on the Grassmann manifold by derivatives on the Stiefel manifold, i.e. with respect to the matrix $\boldsymbol{X}$ rather than its column space $[\boldsymbol{X}]$ [37]. Define the tangent space at $\boldsymbol{X}$ by $\mathcal{T}_{\boldsymbol{X}}$, we then introduce the notion of directional derivative: the derivative of f along the direction $\boldsymbol{V}$ at $\boldsymbol{X}$ is defined by

$$\nabla_{\boldsymbol{V}} f(\boldsymbol{X}) = \lim_{t \to 0} \frac{f(\boldsymbol{X} + t\boldsymbol{V}) - f(\boldsymbol{X})}{t}. \tag{25}$$

The gradient of f at $\boldsymbol{X}$ is the unique tangent vector ∇f satisfying

$$\langle \nabla f(\boldsymbol{X}), \boldsymbol{V} \rangle = \nabla_{\boldsymbol{V}} f(\boldsymbol{X}) \tag{26}$$

for all $\boldsymbol{V} \in \mathcal{T}_{\boldsymbol{X}}$. This can be computed by projecting the complex derivative $\mathrm{D}f$ on $\mathcal{T}_{\boldsymbol{X}}$ as given by

$$\nabla f(\boldsymbol{X}) = (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^H)\mathrm{D}f(\boldsymbol{X}). \tag{27}$$

[1] Otherwise one would have $\nabla_{\boldsymbol{H}_f} \text{dist}_{\text{FS}} = -\langle \nabla \text{dist}_{\text{FS}}, \nabla f \rangle = 0$.

B. Local Convergence

Local convergence of gradient-related search are discussed e.g. in [38]. Define a sequence $\{\boldsymbol{U}_k\}$ emanating from the gradient procedure. Note that the function f is infinitely differentiable. Using the Taylor expansion, we have the zero-order and first order term:

$$f(\boldsymbol{U}_{k+1}) = f(\boldsymbol{U}_k)+\langle \mathrm{D}f(\boldsymbol{U}_k), (\boldsymbol{U}_{k+1}-\boldsymbol{U}_k)\rangle+o(\|\boldsymbol{U}_{k+1}-\boldsymbol{U}_k\|)$$

combined with an expansion of the matrix exponential for $\boldsymbol{U}_{k+1} = \boldsymbol{U}_k(\epsilon)$ according to (14)

$$\begin{aligned}&\boldsymbol{U}_{k+1}-\boldsymbol{U}_k = \\ &\quad \epsilon\boldsymbol{H}_f+\sum_{n=1}^{\infty}(-1)^n\left(\frac{\epsilon^{2n}}{2n!}\boldsymbol{U}_k+\frac{\epsilon^{2n+1}}{(2n+1)!}\boldsymbol{H}_f\right)(\boldsymbol{H}_f^H\boldsymbol{H}_f)^n\end{aligned}$$

leads to

$$\begin{aligned}f(\boldsymbol{U}_{k+1}) &= f(\boldsymbol{U}_k)-\epsilon\langle \mathrm{D}f(\boldsymbol{U}_k), \nabla f(\boldsymbol{U}_k)\rangle+o(\epsilon)\\ &= f(\boldsymbol{U}_k)-\epsilon\|\nabla f(\boldsymbol{U}_k)\|^2+o(\epsilon) \qquad (28)\end{aligned}$$

where the last equality comes from the property of a projector, $(\boldsymbol{I}-\boldsymbol{U}_k\boldsymbol{U}_k^H)^2 = (\boldsymbol{I}-\boldsymbol{U}_k\boldsymbol{U}_k^H)$, and thus

$$\begin{aligned}\mathrm{Tr}[\mathrm{D}f(\boldsymbol{U}_k)^H\nabla f] &= \mathrm{Tr}[((\boldsymbol{I}-\boldsymbol{U}_k\boldsymbol{U}_k^H)\mathrm{D}f(\boldsymbol{U}_k))^H\\ &\qquad\times(\boldsymbol{I}-\boldsymbol{U}_k\boldsymbol{U}_k^H)\mathrm{D}f(\boldsymbol{U}_k)]\\ &= \mathrm{Tr}[\nabla f^H\nabla f] = \|\nabla f(\boldsymbol{U}_k)\|^2. \qquad (29)\end{aligned}$$

Therefore with $\epsilon > 0$ sufficiently small, one has $f(\boldsymbol{U}_k) \geq f(\boldsymbol{U}_{k+1}) \geq 0$ and

$$f(\boldsymbol{U}_k)-f(\boldsymbol{U}_{k+1}) \geq \epsilon\|\nabla f(\boldsymbol{U}_k)\|^2. \qquad (30)$$

Since the sequence $\{f(\boldsymbol{U}_k)\}$ is decreasing and lower bounded by zero, it converges to a finite value $\{\overline{f(\boldsymbol{U}_k)}\}$. By continuity of f, the sequence $\{\boldsymbol{U}_k\}$ is converging to a finite value $\overline{\boldsymbol{U}_k}$ and $f\left(\overline{\boldsymbol{U}_k}\right) = \overline{f(\boldsymbol{U}_k)}$.

By definition, one has $\|\nabla f(\boldsymbol{U})\| = 0$ only if $\boldsymbol{U}$ is a stationary point. Now, let assume that the accumulation point $\overline{\boldsymbol{U}_k}$ is not stationary. By convergence, one has $f(\boldsymbol{U}_k)-f(\boldsymbol{U}_{k+1}) \to 0$ which implies that $\epsilon\|\nabla f(\boldsymbol{U}_k)\|^2 \to 0$, which leads to a contradiction since the finite step $\epsilon > 0$ is strictly positive.

C. Proof of Lemma 1

Consider the directional derivative $\nabla_{\boldsymbol{H}_f}|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2$, which by definition is

$$\begin{aligned}&\nabla_{\boldsymbol{H}_f}|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2 = \\ &\quad\lim_{\epsilon\to 0}\frac{|\det[(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)^H\boldsymbol{U}_{\mathrm{opt}}]|^2-|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2}{\epsilon}. \qquad (31)\end{aligned}$$

This is equal to $\nabla_{\boldsymbol{H}_f}\det[\boldsymbol{U}^H\Pi_{\mathrm{opt}}\boldsymbol{U}^H]$ where for simplicity we have defined the projector $\Pi_{\mathrm{opt}} = \boldsymbol{U}_{\mathrm{opt}}\boldsymbol{U}_{\mathrm{opt}}^H$. We shall also use $\Pi_u = \boldsymbol{U}\boldsymbol{U}^H$.

A direct expansion of the first term in the limit in (31) can be written as

$$\begin{aligned}&|\det[(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)^H\boldsymbol{U}_{\mathrm{opt}}]|^2\\ &\qquad= \det[(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)^H\Pi_{\mathrm{opt}}(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)]\\ &\qquad= \det[\boldsymbol{M}_1]\det[\boldsymbol{I}+\epsilon\boldsymbol{M}_1^{-1}\boldsymbol{M}_2+\epsilon^2\boldsymbol{M}_1^{-1}\boldsymbol{M}_3] \qquad (32)\end{aligned}$$

where $\boldsymbol{M}_1 = \boldsymbol{U}^H\Pi_{\mathrm{opt}}\boldsymbol{U}$ is an invertible matrix since $\boldsymbol{U} \notin \mathcal{B}$,

$$\boldsymbol{M}_2 = \boldsymbol{U}^H\Pi_{\mathrm{opt}}\boldsymbol{H}_f+\boldsymbol{H}_f^H\Pi_{\mathrm{opt}}\boldsymbol{U}, \qquad (33)$$

and

$$\boldsymbol{M}_3 = \boldsymbol{H}_f^H\Pi_{\mathrm{opt}}\boldsymbol{H}_f. \qquad (34)$$

Given a matrix $\boldsymbol{X}$, the Taylor series expansion of the function $\det[\boldsymbol{I}+t\boldsymbol{X}]$ with real parameter t, at $t \to 0$, yields

$$\det[\boldsymbol{I}+t\boldsymbol{X}] = \det[\boldsymbol{I}]+\left.\frac{d\det[\boldsymbol{I}+t\boldsymbol{X}]}{dt}\right|_{t=0}t+o(t). \qquad (35)$$

The coefficient of t in this polynomial can be computed from Jacobi's formula for the derivative of the determinant:

$$\frac{d\det[\boldsymbol{I}+t\boldsymbol{X}]}{dt} = \det[\boldsymbol{I}+t\boldsymbol{X}]\,\mathrm{Tr}\left[(\boldsymbol{I}+t\boldsymbol{X})^{-1}\boldsymbol{X}\right]. \qquad (36)$$

So for an infinitesimal ϵ, we have the following well-known approximation of the determinant close to identity

$$\det[\boldsymbol{I}+\epsilon\boldsymbol{X}] = 1+\epsilon\mathrm{Tr}[\boldsymbol{X}]+o(\epsilon). \qquad (37)$$

From this, we can reformulate (32) as

$$\begin{aligned}&|\det[(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)^H\boldsymbol{U}_{\mathrm{opt}}]|^2\\ &\qquad= (1+\epsilon\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{M}_2]+o(\epsilon))\det[\boldsymbol{M}_1]. \qquad (38)\end{aligned}$$

For further simplification, let us write from the SVD of $\boldsymbol{A}$,

$$\boldsymbol{A}\boldsymbol{A}^H = \boldsymbol{U}_{\mathrm{opt}}\Sigma_{\mathrm{max}}^2\boldsymbol{U}_{\mathrm{opt}}^H+\boldsymbol{U}_{\mathrm{opt}\perp}\Sigma_{\mathrm{min}}^2\boldsymbol{U}_{\mathrm{opt}\perp}^H \qquad (39)$$

and Σ_{max} is a diagonal matrix containing the r-largest singular values of $\boldsymbol{A}$. Similarly Σ_{min} contains the remaining singular values of $\boldsymbol{A}$ in decreasing order. Then one has $\boldsymbol{U}_{\mathrm{opt}}^H\boldsymbol{A}\boldsymbol{A}^H = \Sigma_{\mathrm{max}}^2\boldsymbol{U}_{\mathrm{opt}}^H$ and the first-order term in the right-hand side of (38) simplies to

$$\begin{aligned}&\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{M}_2]\\ &= 2\Re\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{U}^H\Pi_{\mathrm{opt}}\boldsymbol{H}_f]\\ &= 2\Re\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{U}^H\Pi_{\mathrm{opt}}\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{U}]\\ &\qquad\qquad-2\Re\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{U}^H\Pi_{\mathrm{opt}}\Pi_u\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{U}]\\ &= 2\Re\mathrm{Tr}[\Sigma_{\mathrm{max}}^2\boldsymbol{U}_{\mathrm{opt}}^H\boldsymbol{U}\boldsymbol{M}_1^{-1}\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]\\ &\qquad\qquad-2\Re\mathrm{Tr}[\boldsymbol{M}_1^{-1}\boldsymbol{M}_1\boldsymbol{U}^H\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{U}]\\ &= 2\mathrm{Tr}[\Sigma_{\mathrm{max}}^2]-2\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u] \qquad (40)\end{aligned}$$

where in the last equality we have used the fact $\boldsymbol{M}_1^{-1} = (\boldsymbol{U}_{\mathrm{opt}}^H\boldsymbol{U})^{-1}(\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}})^{-1}$ and that the traces are real since matrices inside are Hermitian. In the intermediate steps, the real parts have been used for simplicity but it could be verified also that all traces in (40) are actually real.

This leads directly to

$$\begin{aligned}&|\det[(\boldsymbol{U}+\epsilon\boldsymbol{H}_f)^H\boldsymbol{U}_{\mathrm{opt}}]|^2 = \det[\boldsymbol{M}_1]\\ &\qquad\times(1+2\epsilon(\mathrm{Tr}[\Sigma_{\mathrm{max}}^2]-\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u])+o(\epsilon)) \qquad (41)\end{aligned}$$

and the directional derivative (31) is given by

$$\begin{aligned}&\nabla_{\boldsymbol{H}_f}|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2 = \\ &\qquad 2(\mathrm{Tr}[\Sigma_{\mathrm{max}}^2]-\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u])\,|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2. \qquad (42)\end{aligned}$$

As Π_u is a projector of rank r, it has only r non-zero singular values all equal to one. Using the Von Neuman trace inequality, we can upper bound the last term by

$$\begin{align} \mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u] &\leq \sum_{i=1}^{p} \mathrm{s}_i[\boldsymbol{A}\boldsymbol{A}^H]\mathrm{s}_i[\Pi_u] \tag{43}\\ &= \sum_{i=1}^{r} \sigma_i^2 \times 1 = \mathrm{Tr}[\Sigma^2_{\max}] \tag{44}\end{align}$$

where $\mathrm{s}_i[\boldsymbol{M}]$, $i = 1, \ldots, r$ are the singular values of the matrix $\boldsymbol{M}$ in decreasing order. The equality holds if and only if $\boldsymbol{A}\boldsymbol{A}^H$ can be diagonalized simultaneously with Π_u, which would happen only if it projects on some left-singular subspaces of $\boldsymbol{A}$ [39]. For $\boldsymbol{U} \notin \mathcal{B}$, this condition is fulfilled if and only if $\boldsymbol{U} \in [\boldsymbol{U}_{\mathrm{opt}}]$. For $\boldsymbol{U}$ satysfying $\mathrm{dist}_{\mathrm{FS}}(\boldsymbol{U}) \in (0, \frac{\pi}{2})$, one can conclude that

$$\nabla_{\boldsymbol{H}_f} |\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2 > 0.$$

Finally, using the chain rule of the directional derivative, with the fact that $\arccos$ is a strictly decreasing function, implies that $\nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{FS}} < 0$ for $\mathrm{dist}_{\mathrm{FS}}(\boldsymbol{U}) \in (0, \frac{\pi}{2})$. Namely, one has

$$\begin{align} &\nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{FS}}(\boldsymbol{U}) \nonumber\\ &\quad = \left.\frac{d \arccos\sqrt{z}}{dz}\right|_{|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2} \times \nabla_{\boldsymbol{H}_f} |\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2 \nonumber\\ &\quad = \frac{-\nabla_{\boldsymbol{H}_f} |\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2}{2\sqrt{|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2 - |\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^4}} \nonumber\\ &\quad = (\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u] - \mathrm{Tr}[\Sigma^2_{\max}])\frac{|\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|}{\sqrt{1 - |\det[\boldsymbol{U}^H\boldsymbol{U}_{\mathrm{opt}}]|^2}}. \tag{45}\end{align}$$

D. Proof of Lemma 2

For ease of notation, the squared chordal distance between the subspace spanned by $\boldsymbol{U}$ and the subspace spanned by the optimum $\boldsymbol{U}_{\mathrm{opt}}$ is denoted by

$$\mathrm{dist}_{\mathrm{CH}}(\boldsymbol{U}) = d^2_{\mathrm{CH}}([\boldsymbol{U}], [\boldsymbol{U}_{\mathrm{opt}}]). \tag{46}$$

Similarly as for ∇f, the gradient of $\mathrm{dist}_{\mathrm{CH}}$ at $\boldsymbol{U}$ is given by

$$\nabla\mathrm{dist}_{\mathrm{CH}} = -(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^H)\boldsymbol{U}_{\mathrm{opt}}\boldsymbol{U}^H_{\mathrm{opt}}\boldsymbol{U}. \tag{47}$$

The relationship between the directional derivative $\nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{CH}}$ and the gradient is given by the inner product between $\nabla\mathrm{dist}_{\mathrm{CH}}$ and $\boldsymbol{H}_f$. We then need to show that the following quantity is negative

$$\begin{align} \nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{CH}} &= \langle\nabla\mathrm{dist}_{\mathrm{CH}}, \boldsymbol{H}_f\rangle \tag{48}\\ &= -2\mathrm{Tr}[\Pi_{\mathrm{opt}}(\boldsymbol{I} - \Pi_u)\boldsymbol{A}\boldsymbol{A}^H\Pi_u]. \tag{49}\end{align}$$

First, note that since the term inside the trace is a product of Hermitian matrices, this inner product is real. After some simple algrabra, we have

$$\begin{align} \nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{CH}} &= 2\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_u\Pi_{\mathrm{opt}}\Pi_u] \nonumber\\ &\quad - 2\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\frac{1}{2}(\Pi_u\Pi_{\mathrm{opt}} + \Pi_{\mathrm{opt}}\Pi_u)]. \tag{50}\end{align}$$

The equality arises from the identity $2\Re\mathrm{Tr}[\boldsymbol{Z}] = \mathrm{Tr}[\boldsymbol{Z}] + \mathrm{Tr}[\boldsymbol{Z}^H]$ combined with the fact that the last term in the first line is real since the matrices inside the trace are Hermitian. Using an appropriate factorization, we simplify it further as

$$\begin{align} \nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{CH}} &= \frac{1}{2}\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H(\boldsymbol{I} - 2\Pi_u)\Pi_{\mathrm{opt}}(\boldsymbol{I} - 2\Pi_u)] \nonumber\\ &\quad - \mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_{\mathrm{opt}}]. \tag{51}\end{align}$$

Since $(\boldsymbol{I} - 2\Pi_u)$ is Hermitian and unitary, it follows that $(\boldsymbol{I} - 2\Pi_u)\Pi_{\mathrm{opt}}(\boldsymbol{I} - 2\Pi_u)$ is a projector of same rank as Π_{opt}. By construction, these projectors have r eigenvalues (and thus also singular values) equal to one, while the others are equal to zero.

Using a similar argument than in the proof of Lemma 1, we can upper bound the first term with the Von Neuman trace inequality

$$\begin{align} &\mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H(\boldsymbol{I} - 2\Pi_u)\Pi_{\mathrm{opt}}(\boldsymbol{I} - 2\Pi_u)] \nonumber\\ &\quad \leq \sum_{i=1}^{r} \mathrm{s}_i[\boldsymbol{A}\boldsymbol{A}^H]\mathrm{s}_i[(\boldsymbol{I} - 2\Pi_u)\Pi_{\mathrm{opt}}(\boldsymbol{I} - 2\Pi_u)] \tag{52}\\ &\quad = \sum_{i=1}^{r} \sigma_i^2 = \mathrm{Tr}[\boldsymbol{A}\boldsymbol{A}^H\Pi_{\mathrm{opt}}]. \tag{53}\end{align}$$

This proves the claim $\nabla_{\boldsymbol{H}_f}\mathrm{dist}_{\mathrm{CH}} \leq 0$. Again, equality holds if and only if $\boldsymbol{A}\boldsymbol{A}^H$ can be diagonalized simultaneously with $(\boldsymbol{I} - 2\Pi_u)\Pi_{\mathrm{opt}}(\boldsymbol{I} - 2\Pi_u)$, which would happen only if Π_u projects on some singular space of $\boldsymbol{A}$ [39], i.e. a stationary point.

References

[1] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[2] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," *Quart. J. Math. Oxford*, vol. 11, pp. 50–59, 1966.

[3] S. Funk. (2006) The evolution of cybernetics blog. [Online]. Available: http://sifter.org/~simon/journal/20061211.html

[4] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proc. KDD Cup and Workshop*, 2007, pp. 5–8.

[5] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2008, pp. 426–434.

[6] G. Takacs, I. Pilszy, B. Nemeth, and D. Tikk, "Scalable collaborative filtering approaches for large recommender systems," *J. Mach. Learn. Res.*, vol. 10, pp. 623–656, Dec. 2009.

[7] R. H. Keshavan and S. Oh, "A gradient descent algorithm on the grassman manifold for matrix completion," *CoRR*, vol. abs/0910.5260, 2009. [Online]. Available: http://arxiv.org/abs/0910.5260

[8] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. on Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[9] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Jul. 2010.

[10] W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (SimCO) for dictionary update and learning," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6340–6353, Dec. 2012.

[11] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, 1999, pp. 2443–2446.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[13] A. Edelman, T. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562.

[15] G. Zhou, A. Cichocki, and S. Xie, "Fast nonnegative matrix/tensor factorization based on low-rank approximation," *IEEE Trans. Signal Process*, vol. 60, no. 6, pp. 2928–2940, Jun. 2012.

[16] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J Mach. Learn. Res.*, vol. 11, pp. 517–553, Feb. 2010.

[17] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *J Mach. Learn. Res.*, vol. 14, pp. 899–925, Apr. 2013.

[18] A. Edelman and S. T. Smith, "On conjugate gradient-like methods for eigen-like problems," *BIT Numer. Math.*, vol. 36, no. 3, pp. 494–508, 1996.

[19] D. K. Faddeev and V. N. Faddeeva, *Computational methods of linear algebra*. San Francisco London: Freeman, 1963.

[20] D. E. Longsine and S. F. McCormick, "Simultaneous Rayleigh-quotient minimization methods for ax=λbx," *Linear Algebra Appl.*, vol. 34, pp. 195–234, 1980.

[21] H. Yang, "Conjugate gradient methods for the Rayleigh quotient minimization of generalized eigenvalue problems," *Computing*, vol. 51, no. 1, pp. 79–94, 1993.

[22] E. E. Ovtchinnikov, "Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation I: Computing an extreme eigenvalue," *SIAM J. Numer. Anal.*, vol. 46, no. 5, pp. 2567–2592, 2008.

[23] ——, "Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation II: Computing several extreme eigenvalues," *SIAM J. Numer. Anal.*, vol. 46, no. 5, pp. 2593–2619, 2008.

[24] P.-A. Absil, R. Mahony, R. Sepulchre, and P. Van Dooren, "A Grassmann–Rayleigh quotient iteration for computing invariant subspaces," *SIAM Review*, vol. 44, no. 1, pp. 57–73, 2002.

[25] B. N. Parlett, *The symmetric eigenvalue problem*. SIAM, 1980, vol. 7.

[26] S. Batterson and J. Smillie, "The dynamics of Rayleigh quotient iteration," *SIAM J. Numer. Anal.*, vol. 26, no. 3, pp. 624–636, 1989.

[27] K. Hüper and J. Trumpf, "Newton-like methods for numerical optimization on manifolds," in *Proc. Asilomar Conf. on Sig., Systems and Comp.*, vol. 1, Nov. 2004, pp. 136–139.

[28] S. Singh, M. Kearns, and Y. Mansour, "Nash convergence of gradient dynamics in general-sum games," in *Proc. Conf. Uncertainty Artif. Intell.*, 2000, pp. 541–548.

[29] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton Univ. Press, 2008.

[30] J. Conway, R. Hardin, and N. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian space," *Exp. Math*, vol. 5, pp. 139–159, 1996.

[31] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*. Wiley, New York, 1969.

[32] T. Abrudan, J. Eriksson, and V. Koivunen, "Steepest descent algorithms for optimization under unitary matrix constraint," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1134–1147, Mar. 2008.

[33] P.-A. Absil, A. Edelman, and P. Koev, "On the largest principal angle between random subspaces," *Linear Algebra Appl.*, vol. 414, no. 1, pp. 288–294, 2006.

[34] D. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.-F*, vol. 130, no. 1, Feb. 1983.

[35] D. Palomar and S. Verdu, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.

[36] A. Hjørungnes, *Complex-Valued Matrix Derivatives*. Cambridge University Press, 2011.

[37] N. Boumal and P.-A. Absil, "Low-rank matrix completion via preconditioned optimization on the Grassmann manifold," *Linear Algebra and its Applications*, vol. 475, pp. 200–239, 2015.

[38] D. P. Bertsekas, *Nonlinear Programming: 2nd Edition*. Athena Scientific, 1999.

[39] D. Rhea, "The case of equality in the Von Neumann trace inequality," 2011, preprint, available at http://www.drhea.net/wp-content/uploads/2011/01/vonNeumann.pdf.