Low-Rank Riemannian Optimization on Positive Semidefinite Stochastic Matrices with Applications to Graph Clustering

Ahmed Douik [1] **Babak Hassibi** [1]

Abstract

This paper develops a Riemannian optimization framework for solving optimization problems on the set of symmetric positive semidefinite stochastic matrices. The paper first reformulates the problem by factorizing the optimization variable as $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$ and deriving conditions on p, i.e., the number of columns of $\mathbf{Y}$, under which the factorization yields a satisfactory solution. The reparameterization of the problem allows its formulation as an optimization over either an embedded or quotient Riemannian manifold whose geometries are investigated. In particular, the paper explicitly derives the tangent space, Riemannian gradients and retraction operator that allow the design of efficient optimization methods on the proposed manifolds. The numerical results reveal that, when the optimal solution has a known low-rank, the resulting algorithms present a clear complexity advantage when compared with state-of-the-art Euclidean and Riemannian approaches for graph clustering applications.

1. Introduction

MULTIPLE NP-hard and combinatorial optimization problems can be approximated using convex relaxation, e.g., non-negative matrix factorization (Ding et al., 2010; Yang & Oja, 2011), compressive sensing (Chandrasekaran et al., 2012), low-rank matrix completion (Boumal & Absil, 2011; Vandereycken, 2013; Cambier & Absil, 2015). A theoretical analysis of these problems allows one to obtain conditions under which the solution to the relaxed problem coincides with that of the original one. Nonetheless, despite their convexity, solving the relaxed problems can often be a computation bottleneck in large-scale applications.

In this paper, we consider solving optimization problems over the set of symmetric positive semidefinite stochastic matrices. Such optimization problems appear in multiple applications such as graph clustering and community detection (Zass & Shashua, 2006; Arora et al., 2011; Yang & Oja, 2012; Vinayak & Hassibi, 2016; Wang et al., 2016). Instead of solving the problem in the original $\frac{n(n+1)}{2}$-dimensional space, the factorization of the optimization variable $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$, with $\mathbf{Y}$ being an $n \times p$ matrix, allows us to reduce the dimension to np which is extremely attractive when $p \ll n$. However, while the factorization is convenient, it makes the problem non-convex with non-isolated solutions. Indeed, note that for any solution $\mathbf{Y}$ and an orthogonal $p \times p$ matrix $\mathbf{O}$, the matrix $\mathbf{YO}$ also represents a solution to the problem.

Factorizing a low-rank matrix and casting the convex problem into a non-convex one has been well studied in the literature, particularly for solving semidefinite programs (SDPs). For example, the factorization $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$ is suggested in (Homer & Peinado, 1997) to solve the maximum cut problem. The authors in (Helmberg & Rendl, 2000) exploit it to solve SDPs with fixed trace. More generally, (Burer & Monteiro, 2003) investigates low-rank factorization for solving SDPs in standard form.

Taking advantage of both the low-rank factorization and the optimization over Riemannian manifolds, the authors in (Grubišić & Pietersz, 2007) propose a first-order Riemannian algorithm for solving optimization problems on the *elliptope*, i.e., positive semidefinite matrices with ones on the diagonal. The quotient manifold is deeply investigated in (Bonnabel & Sepulchre, 2009) and an invariant metric that makes the manifold geodesically complete is derived. A simpler quotient structure is introduced in (Journée et al., 2010) to solve optimization problems with general trace constraints, including the *elliptope* and the *spectahedron*, by proposing a second-order algorithm with guaranteed quadratic convergence. This manuscript follows a similar approach to the aforementioned works in the sense that a new quotient structure is proposed to solve optimization problems in which the optimization variable is a low-rank positive semidefinite stochastic matrix.

Earlier work (Douik & Hassibi, 2018) studied the doubly stochastic and symmetric multinomial manifolds to represent doubly stochastic and symmetric stochastic matrices. This work extends the aforementioned results to

[1]Department of Electrical Engineering, California Institute of Technology, CA, USA. Correspondence to: Ahmed Douik <ahmed.douik@caltech.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

low-rank positive semidefinite stochastic matrices. To this end, Section 2 formulates the optimization problem of interest and derives conditions under which the non-convex reparametrization yields a satisfactory solution. Section 3 introduces the essential tools for optimization over Riemannian manifolds and their quotient. Geometries of the introduced manifold and its quotient are investigated in Section 4. Finally, before concluding in Section 6, Section 5 discusses various numerical results.

2. Low-Rank Optimization on the Set of Stochastic Matrices

2.1. Problem Formation

Let $\mathbf{X} \in \mathbb{R}^{n\times n}$ be a real symmetric $n \times n$ matrix and define the objective function $g : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$. For simplicity, this conference version assumes that the objective function is convex. We are interested in solving the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} g(\mathbf{X}) \tag{1a}$$

$$\text{s.t. } \mathbf{X}_{ij} \geq 0,\ 1 \leq i,j \leq n \tag{1b}$$

$$\sum_{j=1}^{m} \mathbf{X}_{ij} = 1,\ 1 \leq i \leq n, \tag{1c}$$

$$\mathbf{X} \succeq \mathbf{0}, \tag{1d}$$

wherein constraint (1b) underlines that the matrix is element-wise positive, constraint (1c) corresponds to the fact that the matrix is stochastic, and constraint (1d) insists that the matrix is positive semidefinite.

The optimization problem (1) requires searching for a solution in a space of dimension $\frac{n(n+1)}{2}$ which can be intractable for large-scale problems, e.g., community detection with millions of individuals. Nevertheless, the applications of interest in this paper share the intrinsic property that the optimal solution has a much smaller rank than the ambient dimension.

The rest of the paper assumes that the optimal solution $\mathbf{X}^*$ to the optimization problem (1) has a rank r. If such rank is known a priori through preliminary analysis, one may directly use the appropriate Riemannian geometry introduced in Section 3. Otherwise, one needs to increase the size of the model p until the optimality conditions derived in the rest of this section are satisfied.

2.2. Notation and Terminology

This paper uses the following matrix notations. Matrices and vectors are written with bold characters, e.g., $\mathbf{X}$. The identity matrix is represented by the symbol $\mathbf{I}$. The all one and all zeros vectors of length n are denoted by $\mathbf{1}_n$ and $\mathbf{0}_n$, respectively. The index n in $\mathbf{1}_n$ may be omitted if the dimension is clear from the context. The transpose and the (i,j)-th element of a matrix $\mathbf{X}$ are denoted by $\mathbf{X}^T$ and $\mathbf{X}_{ij}$, respectively. The inverse of a non-singular square matrix $\mathbf{X}$ is indicated by $\mathbf{X}^{-1}$. A positive semidefinite matrix $\mathbf{X}$ is denoted by $\mathbf{X} \succeq \mathbf{0}$. Similarly, the paper uses the notation $\mathbf{X} \geq \mathbf{0}$ to refer to an element-wise positive matrix.

Let the notation $\mathrm{Tr}(.)$ denote the trace operator. The notation $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}(\mathbf{Y}^T\mathbf{X})$ refers to the inner product of matrices $\mathbf{X}$ and $\mathbf{Y}$ on the space $\mathbb{R}^{n\times n}$, called herein the Frobenius inner product. The Frobenius norm of a matrix $\mathbf{X}$ is defined as $||\mathbf{X}||_F^2 = \mathrm{Tr}(\mathbf{X}^T\mathbf{X})$.

Given two matrices of the same dimensions, the element-wise product, known as the Hadamard product, is denoted by the symbol $\odot$. Let $\mathcal{S}^n = \{\mathbf{X} \in \mathbb{R}^{n\times n} | \mathbf{X} = \mathbf{X}^T\}$ denote the set of symmetric $n \times n$ matrices. Similarly, let $\mathcal{S}^p_{\text{skew}}$ be the set of $p \times p$ skew-symmetric matrices, i.e., $\mathbf{X} = -\mathbf{X}^T$ for all $\mathbf{X} \in \mathcal{S}^p_{\text{skew}}$. A full rank $n \times p$ matrix $\mathbf{Y}$ is an element of the set $\mathbb{R}^{n\times p}_*$. The set $\mathcal{O}^p = \{\mathbf{O} \in \mathbb{R}^{p\times p}_* | \mathbf{O}\mathbf{O}^T = \mathbf{I}\}$ groups the orthogonal $p \times p$ matrices.

Consider a smooth function $f : \mathcal{E} \longrightarrow \mathbb{R}$ from some Euclidean space $\mathcal{E}$. The Euclidean gradient of $f(\mathbf{X})$ at $\mathbf{X}$, i.e., matrix whose (i,j) entry is $\frac{\delta f(\mathbf{X})}{\delta \mathbf{X}_{ij}}$, is denoted by $\mathrm{Grad}_{\mathbf{X}}(f(\mathbf{X}))$ which can be abbreviated, unless confusion prevents, as $\mathrm{Grad}(f(\mathbf{X}))$. The directional derivative of $f(\mathbf{X})$ in the direction $\mathbf{Z} \in \mathcal{E}$, denoted by $\mathrm{D}(f(\mathbf{X}))[\mathbf{Z}]$, is defined as follows:

$$\mathrm{D}(f(\mathbf{X}))[\mathbf{Z}] = \lim_{t\to 0} \frac{f(\mathbf{X}+t\mathbf{Z}) - f(\mathbf{X})}{t}.$$

2.3. The Non-Convex Reparametrization

As stated previously, solving the optimization problem (1) requires searching for a solution in an $\frac{n(n+1)}{2}$-dimensional space. To alleviate such computation burden, this section proposes using the low-rank decomposition $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$ wherein $\mathbf{Y} \in \mathbb{R}^{n\times p}$. Therefore, the reformulated optimization problem is the following:

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times p}} g(\mathbf{Y}\mathbf{Y}^T) \tag{2a}$$

$$\text{s.t. } \mathbf{Y}\mathbf{Y}^T \geq \mathbf{0} \tag{2b}$$

$$\mathbf{Y}\mathbf{Y}^T\mathbf{1} = \mathbf{1}. \tag{2c}$$

The reparametrized problem requires searching for a solution in an $np \ll \frac{n(n+1)}{2}$ dimensional space. However, even under the assumption that the optimization problem (1) is convex, the reformulated problem (1d) is non-convex. The next subsection derives conditions under which the reformulated problem and the original one are equivalent.

2.4. Optimality Conditions

This section derives conditions under which an extreme point $\mathbf{Y}$ of (2) corresponds to an extreme point $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$

of the optimization problem (1). To this end, the first order optimality conditions are derived. These results extend the findings in (Burer & Monteiro, 2003; Journée et al., 2010) to problems with inequality constraints.

The following lemma characterizes the first-order optimality condition of the optimization problem (1).

Lemma 1. *A solution $\mathbf{X} \in \mathbb{R}^{n\times n}$ is an extreme point of the optimization problem (1) if there exists the unique dual variables $\sigma \in \mathbb{R}^n$, $\mathbf{S_X}, \mathbf{\Psi} \in \mathcal{S}^n$ such that the following equations hold:*

$$Grad_{\mathbf{X}}(g(\mathbf{X})) + \sigma\mathbf{1}^T + \mathbf{1}\sigma^T = \mathbf{S_X} + \mathbf{\Psi} \quad (3a)$$
$$\mathbf{X} \succeq \mathbf{0} \quad (3b)$$
$$\mathbf{X} \geq \mathbf{0} \quad (3c)$$
$$\mathbf{X1} = \mathbf{1} \quad (3d)$$
$$\mathbf{S_X} \succeq \mathbf{0} \quad (3e)$$
$$\mathbf{\Psi} \geq \mathbf{0} \quad (3f)$$
$$\mathbf{X} \odot \mathbf{\Psi} = \mathbf{0} \quad (3g)$$
$$\mathbf{S_X X} = \mathbf{0}, \quad (3h)$$

with constraint (3a) translating the fact $(\mathbf{X}, \sigma, \mathbf{S_X}, \mathbf{\Psi})$ is a saddle point for the Lagrangian, (3b)-(3d) stating that the solution is a feasible point, (3e) and (3f) representing the positiveness of the dual variables and (3g) and (3h) expressing the complementary slackness.

Similarly, the optimality conditions for the non-convex reformulation and (2) are given in the following lemma:

Lemma 2. *A solution $\mathbf{Y} \in \mathbb{R}^{n\times p}$ is an extreme point of the optimization problem (2) if there exists the unique dual variables $\lambda \in \mathbb{R}^n$ and $\mathbf{\Phi} \in \mathcal{S}^n$ satisfying the following:*

$$(Grad_{\mathbf{X}}(g(\mathbf{YY}^T)) + \lambda\mathbf{1}^T + \mathbf{1}\lambda^T - \mathbf{\Phi})\mathbf{Y} = \mathbf{0} \quad (4a)$$
$$\mathbf{YY}^T \geq \mathbf{0} \quad (4b)$$
$$\mathbf{YY}^T\mathbf{1} = \mathbf{1} \quad (4c)$$
$$\mathbf{\Phi} \geq \mathbf{0} \quad (4d)$$
$$\mathbf{YY}^T \odot \mathbf{\Phi} = \mathbf{0}, \quad (4e)$$

with constraint (4a) translating the fact $(\mathbf{Y}, \lambda, \mathbf{\Phi})$ is a saddle point for the Lagrangian, (4b) and (4c) stating that the solution is a feasible point, (4d) representing the positiveness of the dual variable and (4e) expressing the complementary slackness.

Given the first order optimality conditions of both problems, the following theorem derives the conditions under which an extreme value of (2) coincides with an extreme one of the original optimization problem (1):

Theorem 1. *An extreme point $\mathbf{Y}$ of the optimization problem (2) produces an extreme point $\mathbf{X} = \mathbf{YY}^T$ of the problem (1) if and only if the following holds:*

$$Grad_{\mathbf{X}}(g(\mathbf{YY}^T)) + \lambda\mathbf{1}^T + \mathbf{1}\lambda^T - \mathbf{\Phi} = \mathbf{S_Y} \succeq \mathbf{0}, \quad (5)$$

wherein λ and $\mathbf{\Phi}$ are the dual variable associated with $\mathbf{Y}$.

Algorithm 1 Gradient Descent on Riemannian Manifolds

Require: Manifold $\mathcal{M}$, function f, and retraction R.
1: Initialize $\mathbf{Y} \in \mathcal{M}$.
2: **while** $||\text{Grad } f(\mathbf{Y})|| > 0$ **do**
3: Set descent direction

$$\xi_{\mathbf{Y}} = -\text{grad } f(\mathbf{Y})/||\text{grad } f(\mathbf{Y})||_{\mathbf{Y}}.$$

4: Compute step size α using the backtracking method.
5: Update $\mathbf{Y}$ by retraction $\mathbf{Y} = \text{R}_{\mathbf{Y}}(\alpha\xi_{\mathbf{Y}})$.
6: **end while**
7: Output $\mathbf{Y}$.

3. Optimization on Riemannian Embedded and Quotient Manifolds

The fundamental idea of optimization algorithms on manifolds is to locally approximate the manifold by its *tangent space* at $\mathbf{Y}$. Afterwards, unconstrained optimization is performed on the tangent space. In particular, a descent direction is computed by deriving the *Riemannian gradient*, denoted by grad $f(\mathbf{Y})$. Finally, the point on the tangent space is retracted to the manifold using the *retraction* $\text{R}_{\mathbf{Y}}$. The steps of the gradient descent algorithm on both the embedded and quotient Riemannian manifolds can be found in Algorithm 1. Performing these steps requires the computation of the tangent space, Riemannian gradient, and retraction operator which we derive in Section 4 for both the embedded and the quotient manifolds.

This section defines the above relevant concepts from differential and Riemannian geometry. In particular, Subsection 3.1 introduces the essential tools for optimization over Riemannian embedded manifolds. Subsection 3.2 provides the tools for optimization over quotient Riemannian manifolds. The definitions and notations used herein can be found in the book (Absil et al., 2008). In the rest of the manuscript, variables relative to the quotient manifold, e.g., equivalence classes, are denoted by overline characters.

3.1. Manifold Optimization: Definitions and Notation

Let $\mathcal{M}$ be a matrix manifold embedded in the set of matrices $\mathbb{R}^{n\times p}$, known as the embedding space. The manifold is said to have a dimension d, also known as the number of degrees of freedom, if there exists a mapping from the manifold to an open subset of $\mathbb{R}^d$. Given a point $\mathbf{Y}$ on the manifold, the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ is a d-dimensional Euclidean space that approximates the manifold $\mathcal{M}$ at $\mathbf{Y}$. Such tangent space is generated by computing the speed at the origin of all curves in $\mathcal{M}$ going through and rooted at $\mathbf{Y}$.

Tangent spaces play a primary role in optimization methods over Riemannian manifolds as they allow to locally transform the curved manifold into a smooth vector space to

which common unconstrained optimization techniques can be applied. However, one needs the notion of distance and length on these tangent spaces to apply optimization algorithms. Such a notion is provided by a bilinear, symmetric, positive, and smoothly varying form, known as the Riemannian metric. The restriction of the Riemannian metric to the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ forms an inner-product denoted by $\langle ., . \rangle_{\mathbf{Y}}$. Although this paper uses the induced Frobenius inner product as a Riemannian metric, the subscript $\mathbf{Y}$ in $\langle ., . \rangle_{\mathbf{Y}}$ is kept to further clarify which tangent space is being considered. In the rest of the manuscript, tangent vectors are denoted by Greek letters wherein the point on the manifold in which the tangent is computed is given as a subscript. In particular, the norm of a tangent vector $\xi_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}$ is denoted by the following:

$$||\xi_{\mathbf{Y}}||^2_{\mathbf{Y}} = \langle \xi_{\mathbf{Y}}, \xi_{\mathbf{Y}} \rangle_{\mathbf{Y}}, \ \forall \, \xi_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}$$

The first order derivative, i.e., the Riemannian gradient, is obtained by taking the component of the Euclidean gradient in the tangent space. This can be mathematically formalized by finding the unique tangent vector grad $f(\mathbf{Y})$ in the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ such that:

$$\langle \text{grad } f(\mathbf{Y}), \xi_{\mathbf{Y}} \rangle_{\mathbf{Y}} = \mathrm{D}(f(\mathbf{Y}))[\xi_{\mathbf{Y}}], \ \forall \, \xi_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}.$$

As stated earlier, this manuscript considers the induced Frobenius inner product as a Riemannian metric which allows the simplification of the above equation. Indeed, let $\mathcal{P}_{\mathbf{Y}} : \mathbb{R}^{n \times p} \to \mathcal{T}_{\mathbf{Y}}\mathcal{M}$ denote the orthogonal projection from the ambient space to the tangent one, then the Riemannian gradient grad f can be written as a function of the Euclidean Grad f one as follows:

$$\text{grad } f(\mathbf{Y}) = \mathcal{P}_{\mathbf{Y}}(\text{Grad } f(\mathbf{Y})).$$

After choosing the descent direction, the step size is chosen according to Wolfs conditions, i.e., the Armijo and curvature conditions (Absil et al., 2008). Finally, the tangent vector is retracted to the manifold using the retraction operator R whose restriction $\mathrm{R}_{\mathbf{Y}}$ to the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ satisfies the centering, i.e., $\mathrm{R}_{\mathbf{Y}}(\mathbf{0}) = \mathbf{Y}$, and local rigidity, i.e., $\frac{d\mathrm{R}_{\mathbf{Y}}(\tau\xi_{\mathbf{Y}})}{d\tau}\Big|_{\tau=0} = \xi_{\mathbf{Y}}$, properties (Absil et al., 2008).

3.2. Optimization on Quotient Riemannian Manifolds

Let $\sim$ be an equivalence relationship and define the set $\overline{\mathcal{M}} = \mathcal{M}/\sim$ as the quotient of the manifold $\mathcal{M}$ by $\sim$. The set $\overline{\mathcal{M}}$ admits a manifold structure. In other words, the quotient manifold $\overline{\mathcal{M}}$ groups all elements of $\mathcal{M}$ in the same equivalence class as a single point. Let π be the natural projection that associates to each $\mathbf{Y} \in \mathcal{M}$ its equivalence class $\pi(\mathbf{Y}) = [\mathbf{Y}] = \overline{\mathbf{Y}} \in \mathcal{M}$. These three notations for equivalence classes are used interchangeably in this paper depending on the context.

Let $\langle ., . \rangle_{\mathbf{Y}}$ be the Riemannian metric on the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ of the embedding space $\mathcal{M}$. The quotient $\overline{\mathcal{M}} = \mathcal{M}/\sim$ admits a Riemmanian structure for the induced Riemannian metric if and only if the metric is compatible with the equivalence relationship $\sim$, i.e., it does not depend on the chosen representative of the equivalence class. To express the compatibility of the metric, we first introduce the horizontal lift.

For a point $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}$, let $\xi_{\overline{\mathbf{Y}}} \in \mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}$ be a tangent vector. In a similar manner that $\overline{\mathbf{Y}}$ can be represented by multiple $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$, the tangent vector $\xi_{\overline{\mathbf{Y}}}$ can be represented by multiple predecessors for each $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$. Indeed, fix $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$, then any tangent vector $\xi_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}$ satisfying $\mathrm{D}(\pi(\mathbf{Y}))[\xi_{\mathbf{Y}}] = \xi_{\overline{\mathbf{Y}}}$ can be considered as a valid representation of the tangent vector $\xi_{\overline{\mathbf{Y}}}$. To circumvent the aforementioned problem and obtain a unique representation of $\xi_{\overline{\mathbf{Y}}}$ for each predecessor $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$, we use the fact that $\pi^{-1}(\overline{\mathbf{Y}})$ represents a manifold. Therefore, one can obtain a unique representation by orthogonally (in the Riemannian metric sense) decomposing the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ into a vertical space $\mathcal{V}_{\mathbf{Y}}\mathcal{M}$ and a horizontal space $\mathcal{H}_{\mathbf{Y}}\mathcal{M}$ such that:

$$\mathcal{V}_{\mathbf{Y}}\mathcal{M} = \mathcal{T}_{\mathbf{Y}}\pi^{-1}(\overline{\mathbf{Y}})$$
$$\mathcal{T}_{\mathbf{Y}}\mathcal{M} = \mathcal{V}_{\mathbf{Y}}\mathcal{M} \oplus \mathcal{H}_{\mathbf{Y}}\mathcal{M}$$

Assuming the ambient space is a vector space, it can be composed into a tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}$ and its orthogonal complement $\mathcal{T}^{\perp}_{\mathbf{Y}}\mathcal{M}$. In particular, for each $\mathbf{Y} \in \mathcal{M}$, the embedding space $\mathbb{R}^{n \times p}$ can be uniquely decomposed into a direct sum of the above defined linear space, i.e.,:

$$\mathbb{R}^{n \times p} = \mathcal{H}_{\mathbf{Y}}\mathcal{M} \oplus \mathcal{V}_{\mathbf{Y}}\mathcal{M} \oplus \mathcal{T}^{\perp}_{\mathbf{Y}}\mathcal{M}.$$

The representation of $\xi_{\overline{\mathbf{Y}}} \in \mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}$ at $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$, denoted by $\bar{\xi}_{\mathbf{Y}}$ and referred to as the horizontal lift of a tangent vector $\xi_{\overline{\mathbf{Y}}}$ at $\mathbf{Y}$, is the unique element in the horizontal space $\mathcal{H}_{\mathbf{Y}}\mathcal{M}$ satisfying $\mathrm{D}(\pi(\mathbf{Y}))[\bar{\xi}_{\mathbf{Y}}] = \xi_{\overline{\mathbf{Y}}}$. Such representation as horizontal lift allows to get a unique parameterization of tangent vectors in a quotient manifold.

The manifold $\overline{\mathcal{M}}$ represents a Riemannian manifold for the Riemmanian metric $\langle ., . \rangle_{\overline{\mathbf{Y}}}$ on $\mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}$ if and only if for all tangent vectors $\xi_{\overline{\mathbf{Y}}}, \eta_{\overline{\mathbf{Y}}} \in \mathcal{T}_{\overline{\mathbf{Y}}}\mathcal{M}$ the following holds:

$$\langle \bar{\xi}_{\mathbf{Y}_1}, \bar{\eta}_{\mathbf{Y}_1} \rangle_{\mathbf{Y}_1} = \langle \bar{\xi}_{\mathbf{Y}_2}, \bar{\eta}_{\mathbf{Y}_2} \rangle_{\mathbf{Y}_2}, \ \forall \, \mathbf{Y}_1, \mathbf{Y}_2 \in \pi^{-1}(\overline{\mathbf{Y}}).$$

Under the above assumption, the operator $\langle ., . \rangle_{\overline{\mathbf{Y}}}$ on $\mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}$ defined by $\langle \xi_{\overline{\mathbf{Y}}}, \xi_{\overline{\mathbf{Y}}} \rangle_{\overline{\mathbf{Y}}} = \langle \bar{\xi}_{\mathbf{Y}}, \bar{\eta}_{\mathbf{Y}} \rangle_{\mathbf{Y}}$ for any $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$ represents a well-defined Riemannian metric for the quotient manifold $\overline{\mathcal{M}}$.

Let $\mathcal{P}^{\mathcal{H}}_{\mathbf{Y}}$ be the orthogonal projection, in the Riemannian inner product sense, from the ambient space $\mathbb{R}^{n \times p}$ to the horizontal space $\mathcal{H}_{\mathbf{Y}}\mathcal{M}$. Let $f : \mathcal{M} \to \mathbb{R}$ be a function that is constant on each equivalence class $[\mathbf{Y}]$ for all $\mathbf{Y} \in \mathcal{M}$. The above function, said to be compatible with the equivalence relationship, induces a function $\overline{f} : \overline{\mathcal{M}} \to \mathbb{R}$ such that $\overline{f}(\overline{\mathbf{Y}}) = f(\mathbf{Y})$ for any predecessor $\mathbf{Y}$ of the equivalence class $\overline{\mathbf{Y}}$. Under the above assumptions, the

Riemannian gradient is obtained by projecting the Euclidean one onto the horizontal space of any predecessor. In other words, the Riemannian gradient is given by:

$$\text{grad } \overline{f}(\overline{\mathbf{Y}}) = \mathcal{P}^{\mathcal{H}}_{\mathbf{Y}}(\text{Grad } f(\mathbf{Y})),\ \mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$$

Let $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}$ and let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be any two representatives in $\pi^{-1}(\overline{\mathbf{Y}})$. Assume that the retractions $\text{R}_{\mathbf{Y}_1}$ and $\text{R}_{\mathbf{Y}_2}$ on the tangent spaces $\mathcal{T}_{\mathbf{Y}_1}\mathcal{M}$ and $\mathcal{T}_{\mathbf{Y}_2}\mathcal{M}$ of the manifold $\mathcal{M}$ satisfy the property:

$$\pi(\text{R}_{\mathbf{Y}_1}(\overline{\xi}_{\mathbf{Y}_1})) = \pi(\text{R}_{\mathbf{Y}_2}(\overline{\xi}_{\mathbf{Y}_2})),\ \forall\ \xi_{\overline{\mathbf{Y}}} \in \mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}.$$

A retraction that satisfy the above equation for all representatives in $\pi^{-1}(\overline{\mathbf{Y}})$ is said to be compatible with the equivalence relationship and generate a retraction on the quotient manifold as follow:

$$\text{R}_{\overline{\mathbf{Y}}}(\xi_{\overline{\mathbf{Y}}}) = \pi(\text{R}_{\mathbf{Y}}(\overline{\xi}_{\mathbf{Y}})), \mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}}). \tag{6}$$

4. Geometry of the Embedded and Quotient Low-Rank Positive Multinomial Manifolds

This section studies the geometry of the embedded and quotient low-rank positive multinomial manifolds. Subsection 4.1 derives the expression of the tangent space, Riemannian gradient and retraction for the embedded low-rank positive multinomial manifold. Similarly, Subsection 4.2 derives these ingredients for the quotient low-rank positive multinomial manifold.

4.1. The Low-Rank Positive Multinomial Manifold

In the rest of the paper, we use the notation $\mathcal{M}_p^n$ to refer to the embedded low-rank positive multinomial manifold defined as:

$$\mathcal{M}_p^n = \{\mathbf{Y} \in \mathbb{R}_*^{n\times p} \mid \mathbf{Y}\mathbf{Y}^T > \mathbf{0} \text{ and } \mathbf{Y}\mathbf{Y}^T\mathbf{1} = \mathbf{1}\}.$$

It is easy to see that the above set represents a well-defined manifold as it can be mapped to an open set in $\mathbb{R}^{n(p-1)}$ by vectorizing the first $(p-1)$ columns of $\mathbf{Y}$. The above manifold is seen as an embedded manifold in the set of non-singular matrices in $\mathbb{R}^{n\times p}$. In other words, the manifold is regarded as an embedded structure in the non-compact Stiefel manifold $\mathbb{R}_*^{n\times p}$. Define the function $f : \mathcal{M}_p^n \to \mathbb{R}$ by $f(\mathbf{Y}) = g(\mathbf{Y}\mathbf{Y}^T)$ wherein the function $g : \mathbb{R}^{n\times n} \to \mathbb{R}$ is defined in (1) and (2). Using the aforementioned definitions, this section proposes a first order Riemannian optimization algorithm to solve the following problem:

$$\min_{\mathbf{Y}\in\mathcal{M}_p^n} f(\mathbf{Y}). \tag{7}$$

The following proposition provides the expression of the tangent space of embedded low-rank positive multinomial manifold:

Proposition 1. *The tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$ for a point $\mathbf{Y} \in \mathcal{M}_p^n$ is given by the following $n(p-1)$-dimensional Euclidean space:*

$$\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n = \{\xi_{\mathbf{Y}} \in \mathbb{R}^{n\times p} \mid (\xi_{\mathbf{Y}}\mathbf{Y}^T + \mathbf{Y}\xi_{\mathbf{Y}}^T)\mathbf{1} = \mathbf{0}\}$$

Let the embedding space $\mathbb{R}_*^{n\times p}$ be equipped with the Frobenius inner product, defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \text{Tr}(\mathbf{A}^T\mathbf{B})$ for all matrices $\mathbf{A}$ and $\mathbf{B}$ in $\mathbb{R}_*^{n\times p}$. This paper considers that the embedded manifold inherits the inner product of the embedding space. In other words, the induced norm $\langle ., .\rangle_{\mathbf{Y}}$ on the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$ for $\mathbf{Y} \in \mathcal{M}_p^n$ is given by:

$$\langle \xi_{\mathbf{Y}}, \eta_{\mathbf{Y}}\rangle_{\mathbf{Y}} = \langle \xi_{\mathbf{Y}}, \eta_{\mathbf{Y}}\rangle,\ \forall\ \xi_{\mathbf{Y}}, \eta_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$$

Given the above Riemannian metric and the tangent space definition in Proposition 1, the expression of the Riemannian gradient is given in the following theorem:

Theorem 2. *Let Grad $f(\mathbf{Y})$ be the Euclidean gradient of f at $\mathbf{Y}$. The Riemannian gradient grad $f(\mathbf{Y})$ is given by:*

$$\textit{grad } f(\mathbf{Y}) = \textit{Grad } f(\mathbf{Y}) - (\alpha\mathbf{1}^T + \mathbf{1}\alpha^T)\mathbf{Y}, \tag{8}$$

with α being the n-dimensional vector obtained by:

$$\alpha = \frac{1}{n}(\mathbf{I} - \frac{1}{2n}\mathbf{1}\mathbf{1}^T)(\mathbf{I} + \mathbf{Y}\mathbf{Y}^T)^{-1}\Big(\textit{Grad } f(\mathbf{Y})\mathbf{Y}^T + \mathbf{Y}\textit{Grad } f(\mathbf{Y})^T\Big)\mathbf{1}.$$

Let $\text{R}_{\mathbf{Y}}$ denote a retraction from the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$ to the manifold $\mathcal{M}_p^n$. In order to derive an expression of such operator, first recall the DAD theorem (Csima & Datta, 1972) which extends the Sinkhorn's theorem for symmetric matrices (Sinkhorn, 1964).

Theorem 3. *Let $\mathbf{A} \in \mathcal{S}^n$ be an entry-wise positive matrix, there exists a unique diagonal matrix $\mathbf{D}$ with strictly positive entries such that $\mathbf{S} = \mathbf{DAD}$ is a doubly stochastic matrix. Such matrix is obtained by the DAD algorithm (Csima & Datta, 1972).*

Let $\mathbb{R}_{+/2}^{n\times p} = \{\mathbf{Z} \in \mathbb{R}^{n\times p} \mid \mathbf{Z}\mathbf{Z}^T > 0\}$ and introduce the projection $\Pi : \mathbb{R}_{+/2}^{n\times p} \to \mathcal{M}_p^n$ defined by $\Pi(\mathbf{Z}) = \mathbf{DZ}$ wherein the diagonal matrix $\mathbf{D}$ is obtained from applying the DAD algorithm to the matrix $\mathbf{Z}\mathbf{Z}^T$. This paper suggests the following retraction to project tangent vectors to the manifold:

Theorem 4. *Let $R_{\mathbf{Y}} : \mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n \to \mathcal{M}_p^n$ be defined by:*

$$R_{\mathbf{Y}}(\xi_{\mathbf{Y}}) = \Pi\Big(\mathbf{Y} + \mathbf{1}_n\mathbf{1}_p^T - \exp(-\xi_{\mathbf{Y}})\Big), \tag{9}$$

with $\exp(\xi_{\mathbf{Y}})$ begin the entry-wise exponential of the entries of the matrix $\xi_{\mathbf{Y}}$. The operator $R_{\mathbf{Y}}$ is a well-defined retraction from the neighborhood $\mathcal{N}_{\mathbf{0}}$ of $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$ to $\mathcal{M}_p^n$.

Given the expression of the tangent space, the Riemannian gradient, and the retraction, the per-iteration complexity of the gradient descent in Algorithm 1 on the embedded positive multinomial manifold scales as $\mathcal{O}(n^2p)$ which considerably reduces the $\mathcal{O}(n^3)$ complexity of solving the original constrained problem.

4.2. The Quotient Low-Rank Positive Multinomial Manifold

As stated earlier, the considered problem exhibits non-isolated solutions. Indeed, given a solution $\mathbf{Y} \in \mathcal{M}_p^n$ and an orthogonal matrix $\mathbf{O} \in \mathcal{O}^p$, the point $\mathbf{YO}$ represent another solution. Therefore, define the relationship $\sim$ on $\mathcal{M}_p^n$ such that:

$$\mathbf{Y}_1 \sim \mathbf{Y}_2 \Leftrightarrow \exists\, \mathbf{O} \in \mathcal{O}^p \text{ s.t. } \mathbf{Y}_1\mathbf{O} = \mathbf{Y}_2$$

Clearly, the relationship $\sim$ defines an equivalence relationship. Let the set $\overline{\mathcal{M}}_p^n = \mathcal{M}_p^n/\sim$, or equivalently $\overline{\mathcal{M}}_p^n = \mathcal{M}_p^n/\mathcal{O}^p$, be the quotient manifold of $\mathcal{M}_p^n$ by the above equivalence relationship. Points on $\overline{\mathcal{M}}_p^n$ are seen as equivalence classes denoted by $[\mathbf{Y}] = \overline{\mathbf{Y}}$ for $\mathbf{Y} \in \mathcal{M}_p^n$. Let $\pi : \mathcal{M}_p^n \to \overline{\mathcal{M}}_p^n$ be the canonical, or natural, projection of points to their equivalence class, i.e., $\pi(\mathbf{Y}) = \overline{\mathbf{Y}}$.

Note that $f(\mathbf{Y}) = g(\mathbf{YY}^T)$ is invariant under $\sim$ as $f(\mathbf{Y}_1) = f(\mathbf{Y}_2)$ for all $\mathbf{Y}_1 \sim \mathbf{Y}_2$. Therefore, there exists a unique function $\overline{f} : \overline{\mathcal{M}}_p^n \to \mathbb{R}$, known as the projection of f, such that $f(\mathbf{Y}) = \overline{f} \circ \pi(\mathbf{Y})$ for all $\mathbf{Y} \in \mathcal{M}_p^n$. The rest of this section is interested in studying the geometry of the quotient low-rank positive multinomial manifold in order to solve the following optimization problem:

$$\min_{\overline{\mathbf{Y}} \in \overline{\mathcal{M}}_p^n} \overline{f}(\overline{\mathbf{Y}}).$$

Let $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}_p^n$, the equivalence class $\pi^{-1}(\overline{\mathbf{Y}})$ can be represented by the following set $\pi^{-1}(\overline{\mathbf{Y}}) = \{\mathbf{YO} \mid \mathbf{O} \in \mathcal{O}^p\}$. Recall that the vertical space $\mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n$ of $\overline{\mathbf{Y}}$ at $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$ is given by $\mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n = \mathcal{T}_{\mathbf{Y}}\pi^{-1}(\overline{\mathbf{Y}})$. The expression of the vertical space is given in the following lemma:

Lemma 3. *The vertical space $\mathcal{V}_{\overline{\mathbf{Y}}}\mathcal{M}_p^n$ is given by:*

$$\mathcal{V}_{\overline{\mathbf{Y}}}\mathcal{M}_p^n = \{\mathbf{YM} \mid \mathbf{M} \in \mathcal{S}_{skew}^p\}. \tag{10}$$

Proof. Let the function $F : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times n}$ defined by $F(\mathbf{Z}) = \mathbf{YY}^T - \mathbf{ZZ}^T$. Note that $\pi^{-1}(\mathbf{Y})$ is given by the level set of F at $\mathbf{0}_{n\times n}$. Indeed, each $\mathbf{Z}$ satisfying $F(\mathbf{Z}) = \mathbf{0}$ implies that $\mathbf{YY}^T = \mathbf{ZZ}^T$, i.e., $\mathbf{Z} = \mathbf{YO}$ for some orthogonal matrix $\mathbf{O}$. Furthermore, it is straightforward to conclude that F is a constant-rank function from either the fact that $\mathbf{0}_{n\times n}$ is a regular value or by noting that F is a submersion onto the set of positive matrices. Therefore, the tangent space at $\mathbf{Y}$ is given by the kernel of the indefinite directional derivative, i.e.,

$$\mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n = \mathcal{T}_{\mathbf{Y}}\pi^{-1}(\overline{\mathbf{Y}}) = \{\xi_{\mathbf{Y}} \mid \xi_{\mathbf{Y}}\mathbf{Y}^T + \mathbf{Y}\xi_{\mathbf{Y}}^T = \mathbf{0}\}.$$

Recall that $\mathbf{Y} \in \mathbb{R}_*^{n\times p}$ is a full rank matrix and define $\mathbf{Y}^\perp$ as any $n \times n - p$ matrix orthogonal complement of $\mathbf{Y}$ satisfying $\mathbf{Y}^T\mathbf{Y}^\perp = \mathbf{0}$, then any matrix $\xi_{\mathbf{Y}}$ can be written as $\mathbf{YM} + \mathbf{Y}^\perp\mathbf{K}$ for some $p \times p$ matrix $\mathbf{M}$ and some $n - p \times p$ matrix $\mathbf{K}$. Using the decomposition above, the characterization $\xi_{\mathbf{Y}}\mathbf{Y}^T + \mathbf{Y}\xi_{\mathbf{Y}}^T = \mathbf{0}$ of $\mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n$ can be rewritten as follows:

$$\mathbf{YMY}^T + \mathbf{Y}^\perp\mathbf{KY}^T + \mathbf{YM}^T\mathbf{Y}^T + \mathbf{YK}^T(\mathbf{Y}^\perp)^T = \mathbf{0}$$

Post and pre-multiplying the above equation by $\mathbf{Y}^T$ and $\mathbf{Y}$, respectively gives $(\mathbf{Y}^T\mathbf{Y})\mathbf{M}(\mathbf{Y}^T\mathbf{Y}) = -(\mathbf{Y}^T\mathbf{Y})\mathbf{M}^T(\mathbf{Y}^T\mathbf{Y})$ which, after noting that $(\mathbf{Y}^T\mathbf{Y})$ is invertible, gives the the alternate representation in (10) of the vertical space of $\mathcal{M}_p^n$ at $\mathbf{Y}$. ■

The following theorem endows the manifold $\overline{\mathcal{M}}_p^n$ with a compatible Riemannian metric in order to product a Riemannian quotient manifold.

Theorem 5. *Consider the $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}_p^n$. The bi-linear form defined on $\mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}_p^n \times \mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}_p^n$ by*

$$\langle \xi_{\overline{\mathbf{Y}}}, \eta_{\overline{\mathbf{Y}}} \rangle_{\overline{\mathbf{Y}}} = Tr\left(\bar{\xi}_{\mathbf{Y}}^T \bar{\eta}_{\mathbf{Y}}\right), \mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}}) \tag{11}$$

is a well-defined Riemannian metric that is compatible with $\sim$ which turns $\overline{\mathcal{M}}_p^n$ into a Riemannian quotient manifold. The horizontal distribution of $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}_p^n$ at $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$ is given by:

$$\mathcal{H}_{\mathbf{Y}}\mathcal{M}_p^n = \{\eta_{\mathbf{Y}} \in \mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n \mid \eta_{\mathbf{Y}}^T\mathbf{Y} = \mathbf{Y}^T\eta_{\mathbf{Y}}\}.$$

Let $\mathcal{P}_{\mathbf{Y}}^{\mathcal{V}}$ and $\mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}$ be the orthogonal projections, in the Riemmanian metric sense, from the ambient space to the vertical space $\mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n$ and horizontal space $\mathcal{H}_{\mathbf{Y}}\mathcal{M}_p^n$, respectively. Furthermore, let $\mathcal{P}_{\mathbf{Y}}$ be the orthogonal projection from the ambient space to the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$. Recall that the ambient space can be decomposed as $\mathbb{R}^{n\times p} = \mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n \oplus \mathcal{T}_{\mathbf{Y}}^\perp\mathcal{M}_p^n$ wherein the tangent space can be expressed as $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n = \mathcal{V}_{\mathbf{Y}}\mathcal{M}_p^n \oplus \mathcal{H}_{\mathbf{Y}}\mathcal{M}_p^n$. Therefore, $\mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}(\mathbf{Z})$ can be written as $\mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}(\mathcal{P}_{\mathbf{Y}}(\mathbf{Z}))$ which reduces the study of $\mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}$ to the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$. The expression of the orthogonal projection onto the horizontal space is given in the below proposition:

Proposition 2. *The orthogonal projection of $\mathbf{Z}$ from the ambient space $\mathbb{R}^{n\times p}$ to the horizontal $\mathcal{H}_{\mathbf{Y}}\mathcal{M}_p^n$ is given by the following:*

$$\mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}(\mathbf{Z}) = \mathcal{P}_{\mathbf{Y}}(\mathbf{Z}) - \mathbf{YM},$$

with $\mathcal{P}_{\mathbf{Y}}(\mathbf{Z})$ being the orthogonal projection from the ambient space to the tangent one, and $\mathbf{M}$ being the solution to the following Sylvester equation:

$$(\mathbf{Y}^T\mathbf{Y})\mathbf{M} + \mathbf{M}(\mathbf{Y}^T\mathbf{Y}) = \mathbf{Y}^T\mathcal{P}_{\mathbf{Y}}(\mathbf{Z}) - \mathcal{P}_{\mathbf{Y}}(\mathbf{Z})^T\mathbf{Y}.$$

Following the definition given in Section 3 of the Riemannian gradient on the quotient space, the Riemannian gradient can be written as a function of the Euclidean gradient and its Riemannian counterpart on the embedded manifold as follows:

$$\text{grad}\, \overline{f}(\overline{\mathbf{Y}}) = \mathcal{P}_{\mathbf{Y}}^{\mathcal{H}}(\text{Grad}\, f(\mathbf{Y})) = \text{grad}\, f(\mathbf{Y}) - \mathbf{YM},$$

with $\mathbf{M}$ being the solution to the Sylvester equation

$$\mathbf{Y}^T\mathbf{YM} + \mathbf{MY}^T\mathbf{Y} = \mathbf{Y}^T\text{grad}\, f(\mathbf{Y}) - \text{grad}\, f(\mathbf{Y})^T\mathbf{Y}.$$

Let the following retraction $\text{R}_{\mathbf{Y}}$ be defined on the tangent space $\mathcal{T}_{\mathbf{Y}}\mathcal{M}_p^n$ of the embedded manifold $\mathcal{M}_p^n$ by $\text{R}_{\mathbf{Y}}(\xi_{\mathbf{Y}}) =$

$\Pi(\mathbf{Y}+\xi_{\mathbf{Y}})$. The proof that the above operator represents a retraction on $\mathcal{M}_p^n$ is omitted as it mirrors the steps used in the proof of Theorem 4. Indeed, note that the proposed retraction represents the first order approximation of the retraction in (9). Finally, consider the following retraction $\overline{\mathrm{R}}_{\overline{\mathbf{Y}}} : \mathcal{T}_{\overline{\mathbf{Y}}}\overline{\mathcal{M}}_p^n \to \overline{\mathcal{M}}_p^n$ be defined by:

$$\overline{\mathrm{R}}_{\overline{\mathbf{Y}}}(\xi_{\overline{\mathbf{Y}}}) = \pi\Big(\Pi(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})\Big), \tag{12}$$

for $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$. The aforementioned operator represents a well-defined function as it does not depend on the representative $\mathbf{Y} \in \pi^{-1}(\overline{\mathbf{Y}})$. Indeed, let $\overline{\mathbf{Y}} \in \overline{\mathcal{M}}_p^n$ and consider a couple of representatives $\mathbf{Y}$ and $\mathbf{YO}$ in $\pi^{-1}(\overline{\mathbf{Y}})$. Let $\mathbf{D}$ be the diagonal matrix such that $\mathbf{D}(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})^T\mathbf{D}$ is doubly stochastic and note the following:

$$\begin{aligned}&\mathbf{D}(\mathbf{YO}+\overline{\xi}_{\mathbf{YO}})(\mathbf{YO}+\overline{\xi}_{\mathbf{YO}})^T\mathbf{D}\\ &\quad= \mathbf{D}(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})^T\mathbf{D}\end{aligned}$$

We have $\Pi(\mathbf{YO}+\overline{\xi}_{\mathbf{YO}}) = \Pi(\mathbf{Y}+\overline{\xi}_{\mathbf{Y}})$ which shows that $\overline{\mathrm{R}}_{\overline{\mathbf{Y}}}$ satisfies (6) and thus concludes the proof.

5. Numerical Results

This section extensively investigates the performance of the proposed Riemannian manifolds. Subsection 5.1 exploits the proposed framework to provide recovery of a similarity clustering, also known as affinity in the clustering literature, via convex programming. This paper uses real-world data obtained through crowdsourcing on Amazon Mechanical Turk (Buhrmester et al., 2011). Subsection 5.2 performs a similar comparison for a large dimension using synthetic data generated from a stochastic block model which approximate the real-world data (Vinayak & Hassibi, 2016).

We compare the performance of the proposed conjugate gradient (CG) method on both the embedded and the quotient manifolds. Unlike the steepest descent, the conjugate gradient algorithm requires a vector transport $\mathcal{T}$. The expression of such operator can be obtained by exploiting the linear structure of the embedding space as $\mathcal{T}_{\eta_{\mathbf{Y}}}(\xi_{\mathbf{Y}}) = \mathcal{P}_{\mathrm{R}_{\mathbf{Y}}(\eta_{\mathbf{Y}})}(\xi_{\mathbf{Y}})$ (see Proposition 8.1.2 (Absil et al., 2008)). The performance of the proposed algorithms is tested against the generic convex solver CVX (Grant & Boyd, 2014), a specialized approximate solver (Lin et al., 2010), and the symmetric multinomial $\mathcal{M}^n = \{\mathbf{X} \in \mathcal{S}^n \mid \mathbf{X} \succ \mathbf{0}, \mathbf{X} > \mathbf{0}, \mathbf{X1} = \mathbf{1}\}$ (Douik & Hassibi, 2018). In Subsection 5.2, the problems are evaluated over a large number of iterations, and the mean value is presented. All simulations are carried out using the Matlab toolbox Manopt (Boumal et al., 2014) on an Intel Xeon Processor E5-1650 v4 (15M cache, 3.60 GHz) computer with 32Gb 2.4 GHz DDR4 RAM.

5.1. Similarity Clustering via Convex Programming

This part suggests retrieving the cluster structure of an adjacency matrix obtained from crowdsourcing on Amazon Mechanical Turk. Images of $n = 473$ dogs from the Stanford Dogs Dataset (Khosla et al., 2011) of $p = 3$ different breeds, i.e., Norfolk Terrier (172 images), Toy Poodle (151 images) and Bouvier des Flandres (150 images), are used in the experiment. At each trial, non-expert workers are required to determine if the pair of dogs presented on the screen has the same breed or not. Each worker is given a set of 30 pair images, and around 600 responses have been collected. Out of the total possible edges $\frac{n(n-1)}{2} = 111628$, only 17260 edges, i.e., around 15% of the total number of entries, are queried and used to construct the adjacency matrix $\mathbf{A}$. Out of these 17260 queried edges, 3941 responses are wrong which gives a 22% error rate. This part reveals the cluster structure by solving the following convex optimization problem whose theoretical guarantees are studied in (Vinayak & Hassibi, 2016):

$$\min_{\mathbf{X}\in\mathcal{M}} \frac{1}{2}||\hat{\mathbf{A}} - \mathbf{X}||_F^2 + \lambda \mathrm{Tr}(\mathbf{X}),$$

with $\hat{\mathbf{A}}$ being the similarity matrix obtained from the partially observed $\mathbf{A}$ by replacing the unknown entries by 0.5. While this is not the best thing to do, the purpose of the current paper is to show the numerical superiority of the proposed method, not the best way to extend $\mathbf{A}$ to $\hat{\mathbf{A}}$. The optimization problem is solved using the numerical optimization toolbox CVX (Grant & Boyd, 2014), a specialized approximate and fast algorithm (Lin et al., 2010), known as augmented Lagrange multipliers (ALM), and the symmetric multinomial. Afterwards, the same problem is solved by reformulating $\mathbf{X} = \mathbf{YY}^T$ and using our proposed methods on the embedded and the quotient manifold.

Table 1. Performance of the Proposed Methods for Clustering

Algorithm	Run. Time	Var. of Inf.	Error Rate
CVX	3183.060 s	0.5404	6.3%
ALM	2.848651 s	0.8688	12.68%
CG on $\mathcal{M}^n$	6.121646 s	0.5543	6.7%
CG on $\mathcal{M}_p^n$	4.777171 s	0.5403	6.3%
CG on $\overline{\mathcal{M}}_p^n$ s	3.813541	0.5501	6.5%

The quality of the recovery is attested through the computation of the variation of information (Meilă, 2003) between the reached cluster structure and the ground truth. Table 1 illustrates the running time and the performance of the above-mentioned optimization methods. From the table, one can see that our method provides 3 orders of magnitude improvement as compared to CVX. Furthermore, it improves the running time of the symmetric multinomial by a factor of 2. With the same running time as ALM, our framework provides twice as better accuracy than ALM. The simulation also shows that the quotient manifold provides better results than its embedded counterpart which is expected as the quotient manifold reduces the dimension of the ambient space by grouping all equivalent solutions.

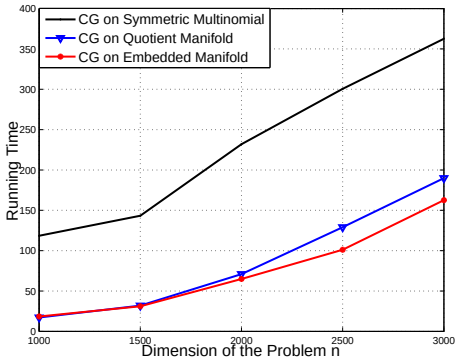

Figure 1. Performance of the proposed optimization scheme in clustering in terms of running time against the system dimension n for a number of clusters $p = 4n/1000$.

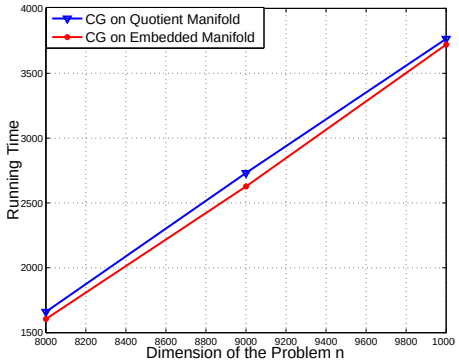

Figure 2. Running time of the proposed first and second order optimization methods in a system with large dimension for a number of clusters $p = n/1000$.

5.2. High Dimension Community Detection

This part proposes solving the clustering problem for a large number of entries n, e.g., a large number of dogs in Subsection 5.1, using synthetic data. In particular, the crowdsourcing part is simulated by sampling from a stochastic block model to obtain a similarity matrix. The number of clusters, e.g., the number of breeds of dogs in Subsection 5.1, is also variable so as to study multiple scenarios. The size of clusters is chosen randomly from a set of predefined sizes for each dimension such that the recovery is theoretically guaranteed. Furthermore, the parameters of the stochastic block model are selected so that the theoretical guarantees proposed in (Vinayak & Hassibi, 2016) are valid which is further confirmed by an almost null variation of information between the ground truth and the reached solution.

The first part of these simulations compares the time performance of the proposed methods on the embedded and quotient manifolds against the performance achieved by a first-order method on the symmetric multinomial $\mathcal{M}^n$. The second part shows the performance of the proposed solution against a system of huge dimension. For such large dimension, neither the generic CVX nor the specialized Riemannian symmetric multinomial are applicable. The final part plots the running time of the suggested methods against the number of clusters p for a fixed dimension n.

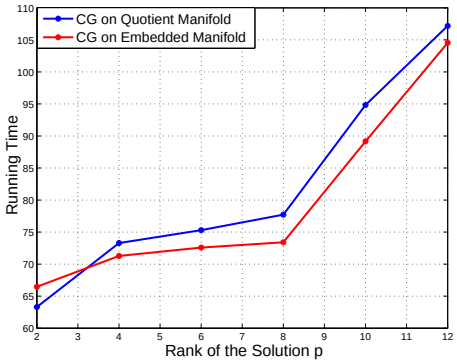

Figure 3. Performance of the proposed Riemannian optimization algorithm against the number of clusters p for a system of dimension $n = 2000$.

Figure 1 plots the running time of the proposed methods again the dimension of the problem n for clusters scaling as $p = 4n/1000$. As a base of comparison, this section plots the performance of the conjugate gradient algorithm on the symmetric multinomial. Figure 1 clearly displays that the proposed methods achieve the same performance with drastically lower running time. The behavior is further illustrated in Figure 2 wherein the system dimension is very large $8000 \leq n \leq 10000$ for a number of clusters $p = n/1000$. The configuration of Figure 1 is prohibitively complex to run either CVX or the symmetric multinomial. Nevertheless, our proposed methods achieve the optimal solution in reasonable running time.

Figure 3 plots the performance of the proposed algorithms in clustering large data sets, $n = 2000$, versus the number of clusters p. As shown in the analysis in the manuscript, the dimension of the suggested manifold increases with the rank p. Such fact is attested by Figure 3.

6. Conclusion

This manuscript designs efficient optimization algorithms for solving optimization problems on the set of symmetric positive semidefinite stochastic matrices. Assuming that the optimal solution has a much lower rank than the ambient dimension, the paper reformulates the problem by introducing the factorization of the optimization variable $\mathbf{X} = \mathbf{Y}\mathbf{Y}^T$. Theoretical guarantees under which the reparametrized problem produces satisfactory solution are derived. The paper introduced an embedded and a quotient Riemannian manifolds in order to solve the reparameterized problem. The efficiency of the proposed framework is attested using both real-world and synthetic data.

Acknowledgements

The authors would like to thank Ramya Korlakai Vinayak for collecting and providing the real-world data used in this manuscript. The authors also extend their thank to the reviewers for their insightful and valuable comments.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

Arora, Raman, Gupta, Maya, Kapila, Amol, and Fazel, Maryam. Clustering by left-stochastic matrix factorization. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, pp. 761–768, 2011.

Bonnabel, Silvere and Sepulchre, Rodolphe. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2009.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL http://www.manopt.org.

Boumal, Nicolas and Absil, Pierre-antoine. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pp. 406–414, 2011.

Buhrmester, Michael, Kwang, Tracy, and Gosling, Samuel D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

Burer, Samuel and Monteiro, Renato DC. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Cambier, Léopold and Absil, P-A. Robust low-rank matrix completion via riemannian optimization. *To appear in SIAM Journal on Scientific Computing*, 2015.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A, and Willsky, Alan S. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

Csima, J and Datta, B.N. The DAD theorem for symmetric non-negative matrices. *Journal of Combinatorial Theory, Series A*, 12(1):147 – 152, 1972. ISSN 0097-3165.

Ding, Chris HQ, Li, Tao, and Jordan, Michael I. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32 (1):45–55, 2010.

Douik, Ahmed and Hassibi, Babak. Manifold Optimization Over the Set of Doubly Stochastic Matrices: A Second-Order Geometry. *ArXiv e-prints arXiv:1802.02628*, 2018.

Grant, Michael and Boyd, Stephen. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Grubišić, Igor and Pietersz, Raoul. Efficient rank reduction of correlation matrices. *Linear algebra and its applications*, 422(2-3):629–653, 2007.

Helmberg, Christoph and Rendl, Franz. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

Homer, Steven and Peinado, Marcus. Design and performance of parallel and distributed approximation algorithms for maxcut. *Journal of Parallel and Distributed Computing*, 46(1):48–61, 1997.

Journée, Michel, Bach, Francis, Absil, P-A, and Sepulchre, Rodolphe. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20 (5):2327–2351, 2010.

Khosla, Aditya, Jayadevaprakash, Nityananda, Yao, Bangpeng, and Li, Fei-Fei. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, pp. 1, 2011.

Lin, Zhouchen, Chen, Minming, and Ma, Yi. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Meilă, Marina. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pp. 173–187. Springer, 2003.

Sinkhorn, Richard. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Vandereycken, Bart. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Vinayak, R. K. and Hassibi, B. Similarity clustering in the presence of outliers: Exact recovery via convex program. In *2016 IEEE International Symposium on Information Theory (ISIT' 2016)*, pp. 91–95, July 2016.

Wang, Xiaoqian, Nie, Feiping, and Huang, Heng. Structured doubly stochastic matrix for graph based clustering: Structured doubly stochastic matrix. In *Proc. of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1245–1254, 2016.

Yang, Z. and Oja, E. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22 (12):1878–1891, Dec 2011.

Yang, Zhirong and Oja, Erkki. Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 707–714, USA, 2012.

Zass, Ron and Shashua, Amnon. Doubly stochastic normalization for spectral clustering. In *NIPS*, pp. 1569–1576, 2006.