Non-smooth Optimization over Stiefel Manifolds with Applications to Dimensionality Reduction and Graph Clustering

Fariba Zohrizadeh, **Mohsen Kheirandishfard**, **Farhad Kamangar** and **Ramtin Madani**
The University of Texas at Arlington
{fariba.zohrizadeh,mohsen.kheirandishfard,farhad.kamangar,ramtin.madani}@uta.com

Abstract

This paper is concerned with the class of non-convex optimization problems with orthogonality constraints. We develop computationally efficient relaxations that transform non-convex orthogonality constrained problems into polynomial-time solvable surrogates. A novel penalization technique is used to enforce feasibility and derive certain conditions under which the constraints of the original non-convex problem are guaranteed to be satisfied. Moreover, we extend our approach to a feasibility-preserving sequential scheme that solves penalized relaxation to obtain near-globally optimal points. Experimental results on synthetic and real datasets demonstrate the effectiveness of the proposed approach on two practical applications in machine learning.

1 Introduction

Consider the following optimization problem

$$\underset{\boldsymbol{P}\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad \bar{f}_0(\boldsymbol{P}) + g_0(\boldsymbol{P}) \tag{1a}$$

$$\text{subject to} \quad \bar{f}_k(\boldsymbol{P}) \leq 0, \qquad k \in \{1,\dots,p\}, \tag{1b}$$

$$\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_m, \tag{1c}$$

where $g_0 : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a convex piecewise linear function and $\bar{f}_k : \mathbb{R}^{n\times m} \to \mathbb{R}$ is an arbitrary quadratic function of the form $\bar{f}_k(\boldsymbol{P}) \triangleq \langle \boldsymbol{M}_k, \boldsymbol{P}\boldsymbol{P}^\top\rangle + \langle \boldsymbol{N}_k, \boldsymbol{P}\rangle + q_k$, for every $k \in \{0,1,\dots,p\}$, and $\{\boldsymbol{M}_k \in \mathbb{S}_n\}_{k=0}^p$, $\{\boldsymbol{N}_k \in \mathbb{R}^{n\times m}\}_{k=0}^p$ and $\{q_k \in \mathbb{R}\}_{k=0}^p$ are given. With no loss of generality, we assume that $q_0 = 0$ and write g_0 in the form of $g_0(\boldsymbol{P}) = \|\alpha(\boldsymbol{P}) + \boldsymbol{b}\|_1$, where $\boldsymbol{b} \in \mathbb{R}^w$ is a given vector, $\alpha : \mathbb{R}^{n\times m} \to \mathbb{R}^w$ is a linear matrix function defined as $\alpha(\boldsymbol{Y}) \triangleq \sum_{i=1}^w \langle \boldsymbol{A}_i, \boldsymbol{Y}\rangle \boldsymbol{e}_i$, the matrices $\{\boldsymbol{A}_i \in \mathbb{R}^{n\times m}\}_{i=1}^w$ are given, and $\{\boldsymbol{e}_i \in \mathbb{R}^w\}_{i=1}^w$ represent the standard basis for $\mathbb{R}^w$. The formulation (1a)–(1c) encompasses a broad class of computationally-hard optimization problems with a variety of practical applications in discriminative dimensionality reduction [Bian and Tao, 2011], graph matching [Jiang *et al.*, 2017], feature selection [Tang and Liu, 2012; Yang *et al.*, 2011], compressed modes [Ozoliņš *et al.*, 2013; Chen *et al.*, 2016], among other areas of machine learning.

The majority of methods in the literature are focused on a special case of (1a)–(1c) that involves the minimization of a convex and smooth objective function over non-convex sets of the form $\mathcal{S}_{n,m} \triangleq \{\boldsymbol{P} \in \mathbb{R}^{n\times m} \mid \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_m\}$, known as the Stiefel manifolds. There are various iterative local search algorithms which preserve the structure of Stiefel manifolds via geodesics steps [Edelman *et al.*, 1998; Abrudan *et al.*, 2008] or retractions [Absil *et al.*, 2009; Wen and Yin, 2013]. Although these algorithms exhibit satisfactory performance in dealing with orthogonality constraints, they mostly restrict the objective function to the class of smooth functions and are not compatible with additional constraints [Gao *et al.*, 2018]. To overcome these limitations, general algorithms are proposed that work with either smooth or non-smooth objective functions [Bian and Tao, 2011; Chen *et al.*, 2016]. The paper [Bian and Tao, 2011] uses a family of semidefinite programming (SDP) problems to generate a converging sequence of points Stiefel manifolds. The paper [Chen *et al.*, 2016] introduces an inner-outer iteration scheme for solving ℓ_1-regularized optimization problems with orthogonality constraints based on the augmented Lagrangian method from [Fortin and Glowinski, 2000] and the proximal alternating minimization technique from [Attouch *et al.*, 2013]. Moreover, a series of splitting techniques are proposed in [Ozoliņš *et al.*, 2013; Lai and Osher, 2014; Kovnatsky *et al.*, 2016] that can efficiently handle non-smooth objective functions. They partition the problem into multiple sub-problems with analytical solutions and employ Bregman iterations [Yin *et al.*, 2008] or its variants [Boyd *et al.*, 2011] to obtain optimal solutions for orthogonality-constrained problems. In the more recent paper [Zhu *et al.*, 2017], an extended proximal alternating linearized minimization method is introduced to minimize convex functions subject to linear constraints and generalized orthogonality constraints.

The success of related sequential frameworks and penalized relaxations for non-convex optimization is demonstrated in [Ibaraki and Tomizuka, 2001; Kheirandishfard *et al.*, 2018b; Kheirandishfard *et al.*, 2018a]. In [Ibaraki and Tomizuka, 2001], a sequential framework is introduced for solving bilinear matrix inequalities without theoretical guarantees. In [Kheirandishfard *et al.*, 2018b; Kheirandishfard *et al.*, 2018a], this approach is further investigated and theoretical results are offered through the notion of generalized Mangasarian-Fromovitz regularity condition. Another sequential SDP-based algorithm for pattern recognition is introduced in [Bian and Tao, 2011] that is not feasibility preserving.

1.1 Contributions

Differentiated from the existing literature, we propose a computational approach with theoretical analysis for solving problems of the form (1a) – (1c), that guarantees the recovery of feasible points. The proposed approach generalizes the existing literature by including additional quadratic inequality constraints. The core of our approach is based on a novel and computationally efficient convex relaxation which transforms the non-convex problem (1a) – (1c) into a convex quadratically-constrained quadratic program (QCQP). To ensure that the solution of the relaxed problem is feasible for (1a) – (1c), we incorporate a penalty term into the objective function and derive certain conditions that guarantee the recovery of feasible points. Moreover, under certain conditions, we prove that by starting from any arbitrary initial point on a Stiefel manifold (not necessarily feasible), a sequence of penalized relaxations can be solved to find a feasible and near-optimal point. Unlike the existing algorithms, if mild assumptions are satisfied, the proposed sequential scheme is feasibility-preserving and improves the objective monotonically at every step. To corroborate the effectiveness of our method, we perform experiments on two practical applications with both synthetic and real datasets. The experimental results demonstrate that the proposed approach exhibits comparable results for both applications.

1.2 Notation

Throughout this paper, the scalars, vectors, and matrices are shown by italic, bold lower-case and bold upper-case letters, respectively. The symbols $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, $\mathbb{S}_n$, and $\mathbb{S}_n^+$ denote the set of real n-dimensional vectors, real $n \times m$ matrices, real symmetric $n \times n$ matrices, and real positive semidefinite matrices, respectively. The symbols $\mathrm{tr}\{.\}$ and $(.)^\top$ are indicative of the trace and transpose operators, respectively. Given a vector $\boldsymbol{a}$ and a matrix $\boldsymbol{A}$, the symbols a_i and A_{ij}, respectively, refer to the i^{th} element of $\boldsymbol{a}$ and the $(i,j)^{th}$ element of $\boldsymbol{A}$. The notation $\boldsymbol{A} \succeq 0$ states that $\boldsymbol{A}$ is symmetric positive semidefinite. Given matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle \triangleq \mathrm{tr}\{\boldsymbol{A}^\top \boldsymbol{B}\}$ and $\mathbf{A} \circ \mathbf{B}$, respectively, denote the Frobenius inner-product and the Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$. The operator $\mathrm{diag}(.)$ gets a vector and forms a diagonal matrix with its input on the diagonal elements. The notation $\|.\|_p$ refers to either matrix norm or vector norm depending on the context, $\|.\|_{\mathrm{F}}$ shows the Frobenius norm, and $|.|$ indicates the absolute value or the cardinality of a set depending on the context. The symbol $\boldsymbol{I}_m$ denotes the identity matrix of size m and the letter $\mathcal{K}$ is used as a shorthand for the set $\{1,\ldots,p\}$. The symbol $\mathcal{S}_{n,m}$ as the set of real $n \times m$ matrices with orthonormal columns, i.e., $\mathcal{S}_{n,m} \triangleq \{\boldsymbol{P} \in \mathbb{R}^{n\times m} \mid \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_m\}$. The projection operator $\mathrm{proj}_{\mathcal{S}_{n,m}} : \mathbb{R}^{n\times m} \to \mathcal{S}_{n,m}$ is defined as $\mathrm{proj}_{\mathcal{S}_{n,m}} \boldsymbol{H} = \arg\min\{\|\boldsymbol{P} - \boldsymbol{H}\|_{\mathrm{F}} \mid \boldsymbol{P} \in \mathcal{S}_{n,m}\}$.

2 Problem Formulation

Optimization problems of the form (1a) – (1c) can be computationally challenging due to the non-convexities of the objective function and constraints. In order to derive convex relaxations, we first lift the problem into a higher dimensional space by introducing an auxiliary variable $\boldsymbol{X} \in \mathbb{S}_n$, accounting for the quadratic term $\boldsymbol{P}\boldsymbol{P}^\top$. For every $k \in \{0\} \cup \mathcal{K}$, define $f_k : \mathbb{R}^{n\times m} \times \mathbb{S}_n \to \mathbb{R}$ as:

$$f_k(\boldsymbol{P}, \boldsymbol{X}) \triangleq \langle \boldsymbol{M}_k, \boldsymbol{X}\rangle + \langle \boldsymbol{N}_k, \boldsymbol{P}\rangle + q_k. \tag{2}$$

Using the auxiliary variable $\boldsymbol{X}$, the optimization problem (1a) – (1c) can be equivalently reformulated as

$$\begin{aligned}
&\underset{\substack{\boldsymbol{P}\in\mathbb{R}^{n\times m}\\ \boldsymbol{X}\in\mathbb{S}_n}}{\text{minimize}} && f_0(\boldsymbol{P}, \boldsymbol{X}) + g_0(\boldsymbol{P}) && && \text{(3a)}\\
&\text{subject to} && f_k(\boldsymbol{P}, \boldsymbol{X}) \leq 0 && k \in \mathcal{K}, && \text{(3b)}\\
& && \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_m, && && \text{(3c)}\\
& && \boldsymbol{P}\ \boldsymbol{P}^\top = \boldsymbol{X}\ , && && \text{(3d)}
\end{aligned}$$

with a convex objective function and convex linear inequality constraints (3b). The above formulation is still not convex due to the presence of the constraints (3c) and (3d) that capture all non-convexities of the problem.

2.1 Convex Relaxation

In order to convexify the lifted problem (3a) – (3d), we relax the constraints (3c) and (3d) to

$$\boldsymbol{I}_m - \boldsymbol{P}^\top \boldsymbol{P} \in \mathcal{C} \quad \wedge \quad \boldsymbol{X} - \boldsymbol{P}\boldsymbol{P}^\top \in \mathcal{D} \quad \wedge \quad \mathrm{tr}\{\boldsymbol{X}\} = m, \tag{4}$$

where $\mathcal{C} \subseteq \mathbb{S}_m$ and $\mathcal{D} \subseteq \mathbb{S}_n$ are convex cones to be defined. In this work, we consider the common-practice semidefinite programming (SDP) relaxation and introduce a novel convex relaxation that transforms the problem (3a) – (3d) into a convex quadratically-constrained quadratic program (QCQP).

Semidefinite Programming Relaxation

This relaxation provides a powerful method for tackling non-convex polynomial optimization problems [Boyd and Vandenberghe, 2004]. The SDP relaxation of the problem (3a) – (3d) can be derived by having $\mathcal{C} = \mathbb{S}_m^+$ and $\mathcal{D} = \mathbb{S}_n^+$. Despite the effectiveness of this relaxation in providing high-quality solutions, its applicability is limited to the problems of moderate size due to the computational cost of imposing high-dimensional conic constraints.

Convex Quadratic Relaxation

We propose a computationally efficient convex relaxation as an alternative to the SDP relaxation. In order to formulate the proposed relaxation for the problem (3a) – (3d), we need to have $\mathcal{C} = \mathcal{V}_m$ and $\mathcal{D} = \mathcal{V}_n$, where for every positive integer o, set $\mathcal{V}_o \subseteq \mathbb{S}_o$ is defined as follows

$$\mathcal{V}_o \triangleq \big\{\boldsymbol{H} \in \mathbb{S}_o \,\big|\, H_{ii} + H_{jj} \geq 2|H_{ij}|,\ \forall i,j \in \{1,\ldots,o\}\big\}.$$

Remark 1. *It can be easily observed that if* $(\mathcal{C},\mathcal{D}) = (\mathcal{V}_m,\mathcal{V}_n)$*, the constraints* (3c) *and* (3d) *boil down to convex quadratic inequalities. Hence, the proposed relaxation reduces (3a) – (3d) to a convex QCQP.*

Notice that either of the aforementioned relaxations may fail to produce a feasible point for (1a) – (1c), because in general, an optimal solution to a convex relaxation does not necessarily satisfy the constraints (3c) and (3d). In what follows, we propose a penalization technique that guarantees the recovery of feasible points for (1a) – (1c) under certain conditions.

3 Penalization

In this section, we show that by including a penalty term in the objective, one can obtain feasible points for the non-convex problem (3a) – (3d). Given an arbitrary initial point $\check{\boldsymbol{P}} \in \mathcal{S}_{n,m}$, that is not necessarily feasible, we transform the problem (3a) – (3d) into the following convex relaxation with revised objective function:

$$\underset{\substack{\boldsymbol{P}\in\mathbb{R}^{n\times m}\\ \boldsymbol{X}\in\mathbb{S}_n}}{\text{minimize}} \quad f_0(\boldsymbol{P},\boldsymbol{X}) + g_0(\boldsymbol{P}) - \mu\langle \boldsymbol{P}, \check{\boldsymbol{P}}\rangle \tag{5a}$$

$$\text{subject to} \quad f_k(\boldsymbol{P},\boldsymbol{X}) \leq 0 \qquad k\in\mathcal{K}, \tag{5b}$$

$$\boldsymbol{I}_m - \boldsymbol{P}^\top\boldsymbol{P} \in \mathcal{C}, \tag{5c}$$

$$\boldsymbol{X} - \boldsymbol{P}\,\boldsymbol{P}^\top \in \mathcal{D}, \tag{5d}$$

$$\mathrm{tr}\{\boldsymbol{X}\} = m, \tag{5e}$$

where $(\mathcal{C},\mathcal{D}) \in \{(\mathbb{S}_m^+,\mathbb{S}_n^+),(\mathcal{V}_m,\mathcal{V}_n)\}$, and the fixed parameter $\mu > 0$ sets a trade-off between the original objective function and the linear penalty term $\langle \boldsymbol{P}, \check{\boldsymbol{P}}\rangle$.

Remark 2. *If an optimal solution* $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{X}})$ *of the problem (5a) – (5e) satisfies the constraints* (3c) *and* (3d)*, then* $\overset{*}{\boldsymbol{P}}$ *is feasible for (1a) – (1c).*

In the remainder of this section, certain conditions are introduced to guarantee that the penalized relaxation (5a) – (5e) produces feasible points for the non-convex problem (3a) – (3d).

Definition 1. *Define feasibility distance* $d_{\mathcal{F}}:\mathbb{R}^{n\times m}\to\mathbb{R}$ *as*

$$d_{\mathcal{F}}(\boldsymbol{P}) \triangleq \inf\{\|\boldsymbol{C}-\boldsymbol{P}\|_{\mathrm{F}} \mid \boldsymbol{C}\in\mathcal{F}\}, \tag{6}$$

where $\mathcal{F}$ *denotes the feasible set of the problem (1a) – (1c).*

Definition 2. *Define the singularity function* $s:\mathcal{S}_{n,m}\to\mathbb{R}$ *as:*

$$s(\boldsymbol{P}) \triangleq \sup_{\boldsymbol{D}\in\mathcal{Z}_{\boldsymbol{P}}}\Big\{\min_{k\in\mathcal{K}}\{-\langle 2\boldsymbol{M}_k\boldsymbol{P}+\boldsymbol{N}_k,\boldsymbol{D}\rangle\}\Big\}, \tag{7}$$

where $\mathcal{Z}_{\boldsymbol{P}} \triangleq \{\boldsymbol{D}\in\mathbb{R}^{n\times m} \mid \boldsymbol{P}^\top\boldsymbol{D} = \boldsymbol{0} \ \wedge\ \|\boldsymbol{D}\|_{\mathrm{F}} \leq 1\}$. *A point* $\boldsymbol{P}\in\mathcal{S}_{n,m}$ *is said to satisfy the Mangasarian-Fromovitz constraint qualification (MFCQ) condition if it is feasible for the problem (1a) – (1c) and* $s(\boldsymbol{P}) > 0$.

Theorem 1. *Define the constants*

$$\beta \triangleq \max_{\boldsymbol{P}\in\mathcal{S}_{m,n}}\{|g_0(\boldsymbol{P})+\langle\boldsymbol{M}_0,\boldsymbol{P}\boldsymbol{P}^\top\rangle+\langle\boldsymbol{N}_0,\boldsymbol{P}\rangle|\}, \tag{8a}$$

$$\psi \triangleq 2\|\boldsymbol{M}_0\|_{\mathrm{F}}+\|\boldsymbol{N}_0\|_{\mathrm{F}}+\sum_{i=1}^{w}\|\boldsymbol{A}_i\|_{\mathrm{F}}, \tag{8b}$$

$$\kappa \triangleq 4\max_{k\in\mathcal{K}}\{\|\boldsymbol{M}_k\|_{\mathrm{F}}\} + \max_{k\in\mathcal{K}}\{\|\boldsymbol{N}_k\|_{\mathrm{F}}\} \tag{8c}$$

and let $\check{\boldsymbol{P}}\in\mathcal{F}$ *be a feasible point for the problem (1a) – (1c) that satisfies the MFCQ condition. If*

$$\mu > \max\{\beta^{-1}\psi^2,\ \beta(26\kappa)^2 s(\check{\boldsymbol{P}})^{-2},\ 144\beta\}, \tag{9}$$

then the penalized relaxation (5a) – (5e) has a unique optimal solution $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{X}})$*, that satisfies* (3c) *and* (3d)*. Moreover,* $\overset{*}{\boldsymbol{P}}$ *is feasible for (1a) – (1c) and* $\bar{f}_0(\overset{*}{\boldsymbol{P}})+g_0(\overset{*}{\boldsymbol{P}}) \leq \bar{f}_0(\check{\boldsymbol{P}})+g_0(\check{\boldsymbol{P}})$.

Proof. See Section 5 for the proof. □

Remark 3. *For every point* $\boldsymbol{P}\in\mathcal{S}_{m,n}$*, it is straightforward to calculate* $s(\boldsymbol{P})$ *by solving the following convex problem:*

$$\underset{t\in\mathbb{R},\boldsymbol{D}\in\mathcal{Z}_{\mathcal{P}}}{\text{maximize}} \quad t$$

$$\text{subject to} \quad t \leq -\langle 2\boldsymbol{M}_k\boldsymbol{P}+\boldsymbol{N}_k,\boldsymbol{D}\rangle, \quad k\in\mathcal{K}.$$

Notice that β can be simply lower- and upper-bounded by any arbitrary member of the set $\mathcal{S}_{m,n}$ and the constant Ψ, respectively. This certifies the existence of a bounded μ that satisfies (9). In practice, there is no need to compute $s(\boldsymbol{P})$ for fine-tuning parameter μ, since (9) offers a conservative sufficient condition and there mostly exist smaller μ that satisfies Theorem 1. In Section 4, we assess the sensitivity of our approach with respect to different choices of μ.

Theorem 1 is concerned with the case where the initial point $\check{\boldsymbol{P}}$ is feasible for the original problem (1a) – (1c). However, finding a feasible starting point can be difficult due to the presence of the non-convex quadratic inequality constraints (1b). The next theorem states that even if $\check{\boldsymbol{P}}$ is not feasible, the proposed penalized relaxation can still result in a feasible point for the non-convex problem (1a) – (1c).

Theorem 2. *Consider an initial* $\check{\boldsymbol{P}}\in\mathcal{S}_{n,m}$ *that satisfies*

$$d_{\mathcal{F}}(\check{\boldsymbol{P}}) < 1, \tag{11a}$$

$$s(\check{\boldsymbol{P}}) > \kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})\left[1+(1-d_{\mathcal{F}}(\check{\boldsymbol{P}}))^{-1}\right], \tag{11b}$$

where κ *is defined in* (8c)*. If* μ *is sufficiently large, then the penalized convex relaxation (5a) – (5e) has a unique optimal solution* $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{X}})$ *that satisfies* (3c) *and* (3d)*. Moreover,* $\overset{*}{\boldsymbol{P}}$ *is feasible for (1a) – (1c).*

Proof. See Section 5 for the proof. □

3.1 Sequential Penalized Relaxation

Motivated by Theorems 1 and 2, this section presents a sequential approach that solves a sequence of penalized relaxations of the form (5a) – (5e) to infer high-quality feasible points for the non-convex problem (1a) – (1c). The proposed scheme starts from an initial point $\check{\boldsymbol{P}}$ on the Stiefel manifold. In each round, the solution of the penalized relaxation (5a) – (5e) is projected onto the Stiefel manifold and then the projected point is employed as an initialization for the next round. Once a feasible point for (1a) – (1c) is obtained, according to Theorem 1, the proposed scheme preserves feasibility and generates a sequence of points whose objective values monotonically improves. The details of the sequential scheme are delineated in Algorithm 1.

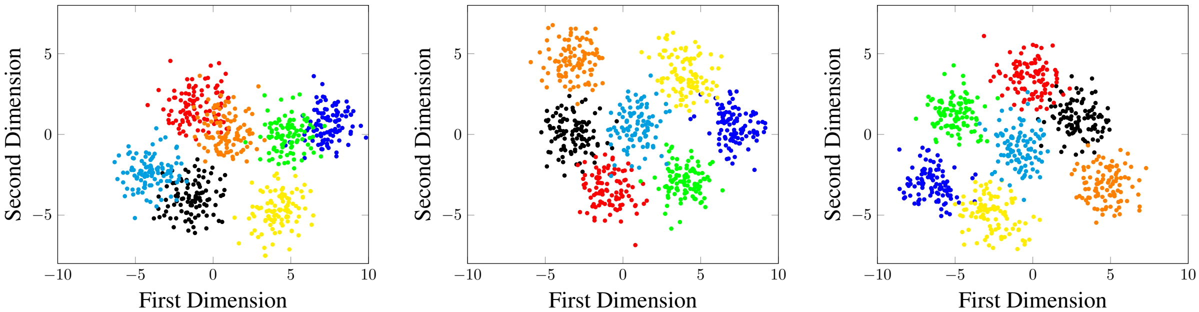

Figure 1: Two dimensional data representation on a training set from the synthetic data set. Left: MMDA [Bian and Tao, 2011], middle: SPR-S, right: SPR-Q. The results show that the SPR-S and SPR-Q algorithms have provided more discriminative 2D representations compared to the MMDA method.

Algorithm 1 Sequential Penalized Relaxation

Input: $\check{\boldsymbol{P}} \in \mathcal{S}_{n,m}$, a fixed parameter $\mu > 0$, and $k = 0$,
repeat
  $k \leftarrow k + 1$
  $\boldsymbol{P}^k \leftarrow$ solve (5a) – (5e) with the penalty $\mu\langle \boldsymbol{P}, \check{\boldsymbol{P}}\rangle$
  $\check{\boldsymbol{P}} \leftarrow \mathrm{proj}_{\mathcal{S}_{n,m}} \boldsymbol{P}^k$
until stopping criteria is met
Output: $\boldsymbol{P}^k$

4 Experimental Results

In this section, we conduct numerical experiments on real and synthetic datasets to verify the effectiveness of the proposed sequential approach, termed SPR, in solving non-convex optimization problems with orthogonality constraints. In Sections 4.1 and 4.2, we apply SPR on two practical problems involving orthogonality constraints. We use SPR-S and SPR-Q to refer to the combination of Algorithm 1 with the SDP relaxation and the proposed convex quadratic relaxation, respectively. To solve the penalized relaxations in each round of the algorithm, we use MOSEK version 7.0 [Mosek, 2015]. Through the experiments, we leverage the inherent sparsity patterns of the problems to reduce the computational cost of solving large-scale semidefinite programs. This enables us to break down large-scale conic constraints into a set of smaller ones [Nakata *et al.*, 2003]. Since finding a feasible point for (1a) – (1c) can be computationally demanding, we initialize Algorithm 1 with an arbitrary starting point on the Stiefel manifold and aim to improve the quality of the point. If the algorithm can recover a feasible point for (1a) – (1c), according to Theorem 1, it can generate a sequence of feasible points whose objective values monotonically improve. To measure the level of infeasibility, define $\mathrm{tr}\{\bar{\boldsymbol{X}} - \bar{\boldsymbol{P}}\bar{\boldsymbol{P}}^\top\}$ as the feasibility violation of an arbitrary feasible point $(\bar{\boldsymbol{P}}, \bar{\boldsymbol{X}})$ of the problem (5a) – (5e). We terminate the sequential algorithm once the feasibility violation and objective value improvement are less than 10^{-5} or if the round number exceeds 100. Notice that the Nesterov acceleration technique can be employed to improve the convergence behaviour of the SPR algorithm. However, in this case, the algorithm may fail to preserve the monotonically decreasing order of the objective values even if the initial point is feasible.

We apply the sequential algorithm on two fundamental machine learning problems of discriminative dimensionality reduction and graph clustering. Notice that each of these problems are well-studied in the literature and several approaches have been developed to efficiently target these applications. Therefore, it is not the intent of this work to compete with these state-of-the-art problem-specific approaches, but rather to demonstrate the potential of Algorithm 1 in solving the problems of form (1a) – (1c) that widely arise in different areas of machine learning.

4.1 Experiment I: Discriminative Dimensionality Reduction

Given a collection of high-dimensional data points from c different classes, the problem of discriminative dimensionality reduction aims to learn a low-dimensional subspace on which the projection of different classes are well-separated. To this end, [Bian and Tao, 2011] proposed a max-min distance analysis (MMDA) that maximizes the minimum distance between all class pairs. This problem can be cast as a non-convex and non-smooth optimization problem of form

$$\underset{\boldsymbol{P}\in\mathbb{R}^{n\times m}}{\text{maximize}} \quad \min_{1\leq i<j\leq c} \langle \boldsymbol{A}^{ij}, \boldsymbol{P}\boldsymbol{P}^\top \rangle \tag{12a}$$

$$\text{subject to} \quad \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_m, \tag{12b}$$

where each $\boldsymbol{A}^{ij} \in \mathbb{S}_n$ is a given weighted distance matrix between the i^{th} and j^{th} classes. In this experiment, we evaluate the performance of the SPR algorithm for solving the problems of form (12a) – (12b). Closely related to our work, [Bian and Tao, 2011] uses a sequence of local SDP relaxations to find the solution of problem (12a) – (12b). We benchmark the SPR method against the MMDA on both real and synthetic datasets. To ensure the comparison is fair, both methods use the same initial point and the same distance matrices $\boldsymbol{A}^{ij}$ which are computed based on [Bian and Tao, 2011]. Other parameter settings of the MMDA are set to their default values. Following [Bian and Tao, 2011], we conduct 100 independent experiments on 10-dimensional synthetic data from seven classes. For each class i, a mean vector $\boldsymbol{\eta}_i \in \mathbb{R}^{10}$ is sampled from 10-dimensional zero mean Gaussian distribution with co-variance matrix $2\boldsymbol{I}_{10}$ and then a pair of training and testing sets, each with 100 members, is generated based on the Gaussian distribution $\mathcal{N}(\boldsymbol{\eta}_i, \boldsymbol{I}_{10})$.

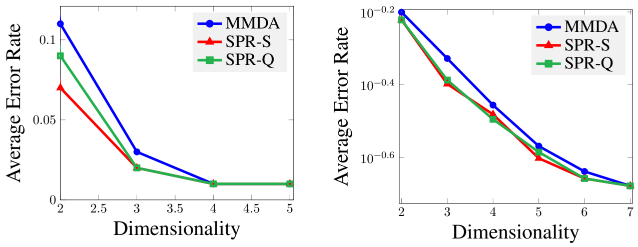

Figure 2: Performance of SPR comparing to MMDA [Bian and Tao, 2011] on synthetic dataset (left), YALE dataset (right) [Georghiades *et al.*, 2001]. *Best viewed in color.*

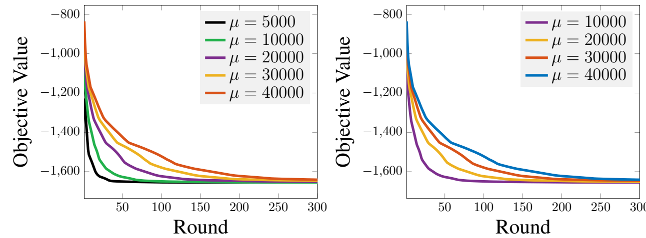

Figure 3: Sensitivity analysis of SPR-S (left) and SPR-Q (right) with respect to different choices of parameter μ for the discriminative dimensionality reduction problem, where $m = 2$. This experiment is performed on the YALE dataset. *Best viewed in color.*

To compare the classification error rate, we project each test set into subspaces with varying dimensions, learned on its corresponding training set. The projected instances are then classified using the nearest mean classifier. Figure 2 (left) shows the average classification error rate with respect to the reduced dimensionality on the synthetic datasets. To run the experiment on the synthetic datasets, we set μ to 100 and 200 for SPR-S and SPR-Q, respectively. Moreover, we conduct this experiment on the YALE dataset consisting of 165 frontal face images of 15 individuals under different illumination and lightening conditions [Belhumeur *et al.*, 1997]. Each image is of size 32×32 pixels. The results of this experiment are illustrated in Figure 2 (right). According to Figure 2, SPR-S and SPR-Q perform on par or better than the MMDA algorithm on both real and synthetic datasets in the problem of discriminative dimensionality reduction. In the experiment on the YALE dataset, we set μ to 5000 and 10000 for SPR-S and SPR-Q, respectively. To qualitatively compare the methods, Figure 1 visualizes the results of projecting a randomly chosen training set from the synthetic dataset on the 2D space. Observe that comparing to the MMDA method, the SPR-based algorithms learn more discriminative 2D representations that are suitable for classification tasks.

To assess the sensitivity of the SPR algorithm with respect to the parameter μ, we perform the discriminative dimensionality reduction experiment with $m = 2$ on YALE dataset and report the results in Figure 3 for various choices of μ. Observe that the final solution obtained by the proposed algorithm is not very sensitive to the choice of μ. According to the figure, the SPR-S requires smaller values of μ to recover feasible points, e.g. $\mu = 5000$, while SPR-Q fails to find feasible points for such choice of μ. Moreover, it can be seen that if μ exceeds a certain threshold, both SPR-S and SPR-Q provide the same sequence of feasible points.

4.2 Experiment II: Graph Clustering

Given a weighted graph $\mathcal{G}$ with n vertices, the graph clustering problem aims to partition $\mathcal{G}$ into a set of sub-graphs such that the vertices within each one are more densely connected to each other than those belonging to different sub-graphs. Inspired by the well-known spectral clustering technique [Ng *et al.*, 2002], this experiment incorporates a set of non-negative constraints to formulate the graph clustering problem as the following optimization [Han *et al.*, 2017]:

$$\underset{\boldsymbol{P}\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad \langle \boldsymbol{L}, \boldsymbol{P}\boldsymbol{P}^\top \rangle \tag{13a}$$

$$\text{subject to} \quad \boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_m, \tag{13b}$$

$$\boldsymbol{P} \geq 0, \tag{13c}$$

where $\boldsymbol{L}$ denotes the Laplacian matrix of the weighted graph $\mathcal{G}$ and $\geq$ is the element-wise inequality operator. Comparing to the spectral clustering, formulation (13a) – (13c) offers a more interpretable clustering framework which requires no further post-processing steps to identify the cluster members. Given $\overset{*}{\boldsymbol{P}}$, the optimal solution of the above problem, each vertex i is assigned to a cluster with label $\operatorname{argmax}_j \overset{*}{P}_{ij}$. [Han *et al.*, 2017] proposed a fast and scalable heuristic, denoted by ONGR, to solve large-scale instances of the form (13a) – (13c). Due to the fact that this problem is a special case of (1a) – (1c), we apply the SPR algorithm to find the solution of (13a) – (13c) and use the same procedure as [Han *et al.*, 2017] to create the Laplacian matrix $\boldsymbol{L}$. To make a fair comparison between the ONGR and SPR, we use the same initialization for both methods. Table 1 reports the clustering performance of the SPR against [Han *et al.*, 2017] on well-known datasets from the UCI machine learning repository [Dua and Graff, 2017] and shape sets [Gionis *et al.*, 2007; Jain and Law, 2005]. For each dataset, n, Dim, and m refer to the number of sample points, dimension of each point, and the number of classes, respectively. The scores for each method is computed by averaging over 30 independent runs for each dataset. As the results indicate, SPR-S and SPR-Q exhibit better performance compared to [Han *et al.*, 2017] on most of the datasets. Through this experiment, we set $\mu = 1000$ in the SPR algorithm and use the default parameter settings for the ONGR algorithm.

5 Proofs

This section presents the proof of Theorems 1 and 2. Before proceeding with the proofs, we provide some prerequisite lemmas. We defer the proofs of the lemmas to the full version of the paper which would be posted on the authors webpage.

Using the well-known epigraph technique [Boyd and Vandenberghe, 2004], the non-smooth term $g_0(\boldsymbol{P})$ in (3a) can be removed by adding a pair of linear constraints and incorporating an additional term into the objective function. This

Dataset	n	Dim.	m	ONGR	SPR-S	SPR-Q
Iris	150	4	3	79.84	**86.71**	81.23
Spiral	312	2	3	87.44	**95.76**	94.15
Jain	373	2	2	88.42	**92.33**	90.26
Compound	399	2	6	74.57	74.25	**76.48**
R15	600	2	15	86.07	85.36	**86.94**
Aggregation	788	2	7	**87.84**	86.39	84.66

Table 1: Clustering performance (%) on the UCI datasets [Dua and Graff, 2017] and shape sets [Gionis *et al.*, 2007; Chang and Yeung, 2008; Jain and Law, 2005].

reformulation of (3a) – (3d) leads to the following penalized non-convex problem:

$$\underset{\substack{\boldsymbol{P}\in\mathbb{R}^{n\times m}\\ \boldsymbol{t}\in\mathbb{R}^{w}}}{\text{minimize}} \quad \mathbf{1}^{\top}\boldsymbol{t}+\langle \boldsymbol{M}_0,\boldsymbol{P}\boldsymbol{P}^{\top}\rangle+\langle \boldsymbol{N}_0,\boldsymbol{P}\rangle-\mu\langle \check{\boldsymbol{P}},\boldsymbol{P}\rangle \tag{14a}$$

$$\text{subject to} \quad \bar{\boldsymbol{\gamma}}:+\alpha(\boldsymbol{P})+\boldsymbol{b}\leq \boldsymbol{t}, \tag{14b}$$

$$\underline{\boldsymbol{\gamma}}:-\alpha(\boldsymbol{P})-\boldsymbol{b}\leq \boldsymbol{t}, \tag{14c}$$

$$\boldsymbol{\lambda}:\langle \boldsymbol{M}_k,\boldsymbol{P}\boldsymbol{P}^{\top}\rangle+\langle \boldsymbol{N}_k,\boldsymbol{P}\rangle+q_k\leq 0,\ \ k\in\mathcal{K}, \tag{14d}$$

$$\boldsymbol{\Omega}:\boldsymbol{P}^{\top}\boldsymbol{P}=\boldsymbol{I}_m, \tag{14e}$$

with $\bar{\boldsymbol{\gamma}}\in\mathbb{R}^w$, $\underline{\boldsymbol{\gamma}}\in\mathbb{R}^w$, $\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{K}|}$, and $\boldsymbol{\Omega}\in\mathbb{S}_m$ as the dual variables associated with the constraints (14b), (14c), (14d), and (14e), respectively. Observe that the problems (14a) – (14e) and (1a) – (1c) are equivalent, if $\mu=0$. In what follows, we show that under certain conditions, the optimal solution of (14a) – (14e) can be obtained in polynomial time via convex relaxation.

The next lemma guarantees the existence of Lagrange multipliers corresponding optimal solutions of the problem (14a) – (14e).

Lemma 1. *Consider an arbitrary $\check{\boldsymbol{P}}\in\mathcal{S}_{n,m}$ that satisfies*

$$s(\check{\boldsymbol{P}})-\kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})>0. \tag{15}$$

If the following inequality holds true,

$$\mu>4\beta[\kappa^{-1}s(\check{\boldsymbol{P}})-d_{\mathcal{F}}(\check{\boldsymbol{P}})]^{-2}, \tag{16}$$

then for every primal optimal pair $(\mathring{\boldsymbol{P}},\mathring{\boldsymbol{t}})$ of (14a) – (14e), there exists Lagrange multipliers $(\mathring{\bar{\boldsymbol{\gamma}}},\mathring{\underline{\boldsymbol{\gamma}}},\mathring{\boldsymbol{\lambda}},\mathring{\boldsymbol{\Omega}})\in\mathbb{R}^w\times\mathbb{R}^w\times\mathbb{R}^{|\mathcal{K}|}\times\mathbb{S}_m$ that satisfy the following Karush–Kuhn–Tucker (KKT) conditions

$$\nabla_{\boldsymbol{P}}\mathcal{L}(\mathring{\boldsymbol{P}},\mathring{\boldsymbol{t}},\mathring{\bar{\boldsymbol{\gamma}}},\mathring{\underline{\boldsymbol{\gamma}}},\mathring{\boldsymbol{\lambda}},\mathring{\boldsymbol{\Omega}})=\boldsymbol{0}, \tag{17a}$$

$$\mathbf{1}+\mathring{\bar{\boldsymbol{\gamma}}}+\mathring{\underline{\boldsymbol{\gamma}}}=\boldsymbol{0}, \tag{17b}$$

$$\mathring{\bar{\boldsymbol{\gamma}}}\circ(+\alpha(\mathring{\boldsymbol{P}})+\boldsymbol{b}-\mathring{\boldsymbol{t}})=\boldsymbol{0}, \tag{17c}$$

$$\mathring{\underline{\boldsymbol{\gamma}}}\circ(-\alpha(\mathring{\boldsymbol{P}})-\boldsymbol{b}-\mathring{\boldsymbol{t}})=\boldsymbol{0}, \tag{17d}$$

$$\mathring{\lambda}_k(\langle \boldsymbol{M}_k,\mathring{\boldsymbol{P}}\mathring{\boldsymbol{P}}^{\top}\rangle+\langle \boldsymbol{N}_k,\mathring{\boldsymbol{P}}\rangle+q_k)=0 \qquad k\in\mathcal{K}, \tag{17e}$$

$$\mathring{\bar{\boldsymbol{\gamma}}}\ \leq\boldsymbol{0}, \qquad \mathring{\underline{\boldsymbol{\gamma}}}\ \leq\boldsymbol{0}, \qquad \mathring{\boldsymbol{\lambda}}\leq 0, \tag{17f}$$

where $\mathcal{L}(\boldsymbol{P},\boldsymbol{t},\bar{\boldsymbol{\gamma}},\underline{\boldsymbol{\gamma}},\boldsymbol{\lambda},\boldsymbol{\Omega})$ represents the Lagrangian function of (14a) – (14e), defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{P},\boldsymbol{t},\bar{\boldsymbol{\gamma}},\underline{\boldsymbol{\gamma}},\boldsymbol{\lambda},\boldsymbol{\Omega})&\triangleq\mathbf{1}^{\top}\boldsymbol{t}+\langle \boldsymbol{M}_0,\boldsymbol{P}\boldsymbol{P}^{\top}\rangle+\langle \boldsymbol{N}_0,\boldsymbol{P}\rangle-\mu\langle \check{\boldsymbol{P}},\boldsymbol{P}\rangle\\ &-\bar{\boldsymbol{\gamma}}^{\top}(\alpha(\boldsymbol{P})+\boldsymbol{b}-\boldsymbol{t})+\underline{\boldsymbol{\gamma}}^{\top}(\alpha(\boldsymbol{P})+\boldsymbol{b}+\boldsymbol{t})\\ &-\sum_{k\in\mathcal{K}}\lambda_k(\langle \boldsymbol{M}_k,\boldsymbol{P}\boldsymbol{P}^{\top}\rangle+\langle \boldsymbol{N}_k,\boldsymbol{P}\rangle+q_k)-\langle \boldsymbol{\Omega},\boldsymbol{P}^{\top}\boldsymbol{P}-\boldsymbol{I}_m\rangle.\end{aligned}$$

and, β and κ are defined in (8a) *and* (8c).

The next lemma provides an upper bound on the Lagrange multipliers of the problem (14a) – (14e), that will be used to show that this problem can be relaxed to (5a) – (5e) with no effect on the solution.

Lemma 2. *Consider an arbitrary $\check{\boldsymbol{P}}\in\mathcal{S}_{n,m}$ that satisfies* (15) *and let μ satisfy* (16)*. For every solution $(\mathring{\boldsymbol{P}},\mathring{\boldsymbol{t}})$ of (14a) – (14e), there exist Lagrange multipliers $(\mathring{\bar{\boldsymbol{\gamma}}},\mathring{\underline{\boldsymbol{\gamma}}},\mathring{\boldsymbol{\lambda}},\mathring{\boldsymbol{\Omega}})\in\mathbb{R}^w\times\mathbb{R}^w\times\mathbb{R}^p\times\mathbb{S}_m$ that satisfy the KKT conditions (17a) – (17f) as well as the inequalities:*

$$-\frac{\mathbf{1}^{\top}\mathring{\boldsymbol{\lambda}}}{\mu}\leq\frac{d_{\mathcal{F}}(\check{\boldsymbol{P}})+\mu^{-1}\psi+2\sqrt{\beta\mu^{-1}}}{s(\check{\boldsymbol{P}})-\kappa(d_{\mathcal{F}}(\check{\boldsymbol{P}})+2\sqrt{\beta\mu^{-1}})} \tag{18a}$$

$$\Big\|\frac{2}{\mu}\mathring{\boldsymbol{\Omega}}+\boldsymbol{I}_m\Big\|_{\mathrm{F}}\leq\kappa_2\Big(-\frac{\mathbf{1}^{\top}\mathring{\boldsymbol{\lambda}}}{\mu}\Big)+d_{\mathcal{F}}(\check{\boldsymbol{P}})+\mu^{-1}\psi+2\sqrt{\beta\mu^{-1}} \tag{18b}$$

where constant κ_2 is given by

$$\kappa_2\triangleq 2\max_{k\in\mathcal{K}}\{\|\boldsymbol{M}_k\|_{\mathrm{F}}\}+\max_{k\in\mathcal{K}}\{\|\boldsymbol{N}_k\|_{\mathrm{F}}\}, \tag{19}$$

and β, ψ and κ are defined in (8a) – (8c).

Using Lemma 2, the next lemma offers conditions to guarantee that penalized relaxations give feasible points for (14a) – (14e).

Lemma 3. *Consider an initial point $\check{\boldsymbol{P}}\in\mathcal{S}_{n,m}$ and $\mu>0$. Let $(\mathring{\boldsymbol{P}},\mathring{\boldsymbol{t}})$ be a primal optimal solution of (14a) – (14e) with the corresponding Lagrange multipliers $(\mathring{\bar{\boldsymbol{\gamma}}},\mathring{\underline{\boldsymbol{\gamma}}},\mathring{\boldsymbol{\lambda}},\mathring{\boldsymbol{\Omega}})$ that satisfy the KKT conditions (17a) – (17e). Define*

$$\varepsilon\triangleq\frac{1}{4}\Big(1-d_{\mathcal{F}}(\check{\boldsymbol{P}})-\frac{\kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})}{s(\check{\boldsymbol{P}})-\kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})}\Big). \tag{20}$$

If the following inequalities hold true

$$2\mu^{-1}\|\boldsymbol{M}_0\|\leq\varepsilon, \tag{21a}$$

$$-\frac{\mathbf{1}^{\top}\mathring{\boldsymbol{\lambda}}}{\mu}\leq\frac{d_{\mathcal{F}}(\check{\boldsymbol{P}})}{s(\check{\boldsymbol{P}})-\kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})}+\frac{\varepsilon}{2\kappa_2}, \tag{21b}$$

$$\Big\|\frac{2}{\mu}\mathring{\boldsymbol{\Omega}}+\boldsymbol{I}_m\Big\|_{\mathrm{F}}\leq\kappa_2\Big(-\frac{\mathbf{1}^{\top}\mathring{\boldsymbol{\lambda}}}{\mu}\Big)+d_{\mathcal{F}}(\check{\boldsymbol{P}})+\varepsilon, \tag{21c}$$

then the pair $(\mathring{\boldsymbol{P}},\mathring{\boldsymbol{P}}\mathring{\boldsymbol{P}}^{\top})$ is the unique primal solution of the penalized convex relaxation (5a) – (5d), where κ and κ_2 are defined in (8c) *and* (19)*, respectively.*

Proof of Theorem 1. Due to the main assumption, it is straightforward to verify the following three inequalities:

$$\mu^{-1}\psi < \sqrt{\beta\mu^{-1}}, \tag{22a}$$

$$2\kappa\sqrt{\beta\mu^{-1}} < 13^{-1}s(\check{\boldsymbol{P}}), \tag{22b}$$

$$\sqrt{\beta\mu^{-1}} < 12^{-1}. \tag{22c}$$

Consider an arbitrary optimal solution $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{t}})$ of (14a) – (14e). The point $\overset{*}{\boldsymbol{P}}$ is consequently feasible for (1a) – (1c). Therefore $d_{\mathcal{F}}(\check{\boldsymbol{P}}) = 0$ and the inequalities (15) and (16) are satisfied. According to Lemma 2, there exist Lagrange multipliers $(\overset{*}{\bar{\boldsymbol{\gamma}}}, \overset{*}{\underline{\boldsymbol{\gamma}}}, \overset{*}{\boldsymbol{\lambda}}, \overset{*}{\boldsymbol{\Omega}}) \in \mathbb{R}^w \times \mathbb{R}^w \times \mathbb{R}^p \times \mathbb{S}_m$ corresponding to $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{t}})$ that satisfy the KKT conditions (17a) – (17f) as well as the inequalities (18a) and (18b). Based on Lemma 3 and since $d_{\mathcal{F}}(\check{\boldsymbol{P}}) = 0$, in order to prove the theorem, it suffices to show that:

$$2\mu^{-1}\|\boldsymbol{M}_0\| \leq 4^{-1} \tag{23a}$$

$$-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu} \leq \frac{4^{-1}}{2\kappa_2} \tag{23b}$$

$$\Big\|\frac{2}{\mu}\overset{*}{\boldsymbol{\Omega}} + \boldsymbol{I}_m\Big\|_{\mathrm{F}} \leq \kappa_2\Big(-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu}\Big) + 4^{-1}. \tag{23c}$$

- (23a) is the direct consequence of (22a):

$$2\mu^{-1}\|\boldsymbol{M}_0\| \leq \mu^{-1}\psi \leq \sqrt{\beta\mu^{-1}} \leq 12^{-1} < 4^{-1}. \tag{24}$$

- (23b) is the direct consequence of (18a), (22b), and (22c):

$$-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu} \leq \frac{\mu^{-1}\psi + 2\sqrt{\beta\mu^{-1}}}{s(\check{\boldsymbol{P}}) - 2\kappa\sqrt{\beta\mu^{-1}}} \leq \frac{\sqrt{\beta\mu^{-1}} + 2\sqrt{\beta\mu^{-1}}}{s(\check{\boldsymbol{P}}) - 2\kappa\sqrt{\beta\mu^{-1}}} \tag{25a}$$

$$\leq \frac{\sqrt{\beta\mu^{-1}} + 2\sqrt{\beta\mu^{-1}}}{s(\check{\boldsymbol{P}}) - 13^{-1}s(\check{\boldsymbol{P}})} = \frac{3\sqrt{\beta\mu^{-1}}}{(1 - 13^{-1})s(\check{\boldsymbol{P}})} \tag{25b}$$

$$\leq \frac{3 \times 13^{-1}(2\kappa)^{-1}s(\check{\boldsymbol{P}})}{(1 - 13^{-1})s(\check{\boldsymbol{P}})} = \frac{4^{-1}}{2\kappa} < \frac{4^{-1}}{2\kappa_2}. \tag{25c}$$

- (23c) can be concluded from (18b), (22a), and (22c):

$$\Big\|\frac{2}{\mu}\overset{*}{\boldsymbol{\Omega}} + \boldsymbol{I}_m\Big\|_{\mathrm{F}} \leq \kappa_2\Big(-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu}\Big) + \mu^{-1}\psi + 2\sqrt{\beta\mu^{-1}} \tag{26a}$$

$$\leq \kappa_2\Big(-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu}\Big) + 3\sqrt{\beta\mu^{-1}} \tag{26b}$$

$$\leq \kappa_2\Big(-\frac{\boldsymbol{1}^\top\overset{*}{\boldsymbol{\lambda}}}{\mu}\Big) + 4^{-1}. \tag{26c}$$

Hence, according to Lemma 3, the point $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{P}}\overset{*}{\boldsymbol{P}}^\top)$ is the unique optimal solution for the penalized relaxation (5a) – (5e), for which the relaxed constraints (3c) and (3d) are satisfied. Finally, due to the feasibility of pair $(\check{\boldsymbol{P}}, \check{\boldsymbol{P}}\check{\boldsymbol{P}}^\top)$, we have:

$$\bar{f}_0(\check{\boldsymbol{P}}) + g_0(\check{\boldsymbol{P}}) - \mu m = f_0(\check{\boldsymbol{P}}, \check{\boldsymbol{P}}\check{\boldsymbol{P}}^\top) + g_0(\check{\boldsymbol{P}}) - \mu\langle\check{\boldsymbol{P}}, \check{\boldsymbol{P}}\rangle \tag{27a}$$

$$\geq f_0(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{P}}\overset{*}{\boldsymbol{P}}^\top) + g_0(\overset{*}{\boldsymbol{P}}) - \mu\langle\overset{*}{\boldsymbol{P}}, \check{\boldsymbol{P}}\rangle \tag{27b}$$

$$\geq \bar{f}_0(\check{\boldsymbol{P}}) + g_0(\check{\boldsymbol{P}}) - \mu m \tag{27c}$$

and the proof is completed. □

Proof of Theorem 2. Consider an arbitrary optimal solution $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{t}})$ of (14a) – (14e). Due to the main assumption, (15) is satisfied and if μ is large, then (16) is satisfied as well. Moreover, according to Lemma 2, there exist Lagrange multipliers $(\overset{*}{\bar{\boldsymbol{\gamma}}}, \overset{*}{\underline{\boldsymbol{\gamma}}}, \overset{*}{\boldsymbol{\lambda}}, \overset{*}{\boldsymbol{\Omega}}) \in \mathbb{R}^w \times \mathbb{R}^w \times \mathbb{R}^p \times \mathbb{S}_m$ corresponding to $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{t}})$ that satisfy the KKT conditions (17a) – (17f) as well as the inequalities (18a) and (18b). According to Lemma 3, the proof follows directly from the fact that

$$\varepsilon = \frac{1}{4}\Big(1 - d_{\mathcal{F}}(\check{\boldsymbol{P}}) - \frac{\kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})}{s(\check{\boldsymbol{P}}) - \kappa\, d_{\mathcal{F}}(\check{\boldsymbol{P}})}\Big) > 0, \tag{28}$$

and therefore, if μ is sufficiently large, the inequalities (18a) and (18b) conclude (21a) – (21c). As a result, if μ is large, $(\overset{*}{\boldsymbol{P}}, \overset{*}{\boldsymbol{P}}\overset{*}{\boldsymbol{P}}^\top)$ is the unique primal solution of the penalized convex relaxation (5a) – (5d). □

6 Conclusions

This work introduces convex relaxations for solving a broad class of non-convex and non-smooth optimization problems involving orthogonality constraints. The proposed approach relies on solving a sequence of penalized convex relaxations to find feasible and near-globally optimal points for a given non-convex orthogonality-constrained problem. Experimental results on two fundamental problems in machine learning demonstrate the potential and effectiveness of the proposed approach in solving practical problems.

Acknowledgements

This work is in part supproted by a University of Texas System STARs award.

References

[Abrudan *et al.*, 2008] Traian E Abrudan, Jan Eriksson, and Visa Koivunen. Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE T SIGNAL PROCES*, 56(3):1134–1147, 2008.

[Absil *et al.*, 2009] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[Attouch *et al.*, 2013] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *MATH PROGRAM*, 137(1-2):91–129, 2013.

[Belhumeur *et al.*, 1997] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, (7):711–720, 1997.

[Bian and Tao, 2011] Wei Bian and Dacheng Tao. Max-min distance analysis by using sequential SDP relaxation for dimension reduction. *IEEE T PATTERN ANAL*, 33(5):1037–1050, 2011.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Chang and Yeung, 2008] Hong Chang and Dit-Yan Yeung. Robust path-based spectral clustering. *PATTERN RECOGN*, 41(1):191–203, 2008.

[Chen *et al.*, 2016] Weiqiang Chen, Hui Ji, and Yanfei You. An augmented Lagrangian method for ℓ_1-regularized optimization problems with orthogonality constraints. *SIAM J SCI COMPUT*, 38(4):B570–B592, 2016.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Edelman *et al.*, 1998] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM J MATH ANAL*, 20(2):303–353, 1998.

[Fortin and Glowinski, 2000] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*, volume 15. Elsevier, 2000.

[Gao *et al.*, 2018] Bin Gao, Xin Liu, Xiaojun Chen, and Yaxiang Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM J OPTIMIZ*, 28(1):302–332, 2018.

[Georghiades *et al.*, 2001] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE T PATTERN ANAL*, 23(6):643–660, 2001.

[Gionis *et al.*, 2007] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *ACM T KNOWL DISCOV D*, 1(1):4, 2007.

[Han *et al.*, 2017] Junwei Han, Kai Xiong, and Feiping Nie. Orthogonal and nonnegative graph reconstruction for large scale clustering. In *IJCAI*, 2017.

[Ibaraki and Tomizuka, 2001] Soichi Ibaraki and Masayoshi Tomizuka. Rank minimization approach for solving BMI problems with random search. In *ACC*, volume 3, pages 1870–1875. IEEE, 2001.

[Jain and Law, 2005] Anil K Jain and Martin HC Law. Data clustering: A user's dilemma. In *ICPRAI*, 2005.

[Jiang *et al.*, 2017] Bo Jiang, Jin Tang, Chris HQ Ding, and Bin Luo. Nonnegative orthogonal graph matching. In *AAAI*, 2017.

[Kheirandishfard *et al.*, 2018a] Mohsen Kheirandishfard, Fariba Zohrizadeh, Muhammad Adil, and Ramtin Madani. Convex relaxation of bilinear matrix inequalities part II: Applications to optimal control synthesis. In *CDC*, 2018.

[Kheirandishfard *et al.*, 2018b] Mohsen Kheirandishfard, Fariba Zohrizadeh, and Ramtin Madani. Convex relaxation of bilinear matrix inequalities part I: Theoretical results. In *CDC*, 2018.

[Kovnatsky *et al.*, 2016] Artiom Kovnatsky, Klaus Glashoff, and Michael M Bronstein. MADMM: a generic algorithm for non-smooth optimization on manifolds. In *ECCV*, 2016.

[Lai and Osher, 2014] Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *SIAM J SCI COMPUT*, 58(2):431–449, 2014.

[Mosek, 2015] ApS Mosek. The MOSEK optimization toolbox for MATLAB manual. *Version 7.1*, 2015.

[Nakata *et al.*, 2003] Kazuhide Nakata, Katsuki Fujisawa, Mituhiro Fukuda, Masakazu Kojima, and Kazuo Murota. Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results. *Mathematical Programming*, 95(2):303–327, 2003.

[Ng *et al.*, 2002] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2002.

[Ozoliņš *et al.*, 2013] Vidvuds Ozoliņš, Rongjie Lai, Russel Caflisch, and Stanley Osher. Compressed modes for variational problems in mathematics and physics. *PNAS*, 110(46):18368–18373, 2013.

[Tang and Liu, 2012] Jiliang Tang and Huan Liu. Unsupervised feature selection for linked social media data. In *SIGKDD*, 2012.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *MATH PROGRAM*, 142(1-2):397–434, 2013.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, 2011.

[Yin *et al.*, 2008] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J IMAGING SCI*, 1(1):143–168, 2008.

[Zhu *et al.*, 2017] Hong Zhu, Xiaowei Zhang, Delin Chu, and Li-Zhi Liao. Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method. *SIAM J SCI COMPUT*, pages 1–42, 2017.