# A Dynamic Smoothing Technique for a Class of Nonsmooth Optimization Problems on Manifolds

Amir Beck* Israel Rosset †

February 16, 2023

**Abstract**

We consider the problem of minimizing the sum of a smooth nonconvex function and a nonsmooth convex function over a compact embedded submanifold. We describe an algorithm, which we refer to as "dynamic smoothing gradient descent on manifolds" (DSGM), that is based on applying Riemmanian gradient steps on a series of smooth approximations of the objective function that are determined by a diminishing sequence of smoothing parameters. The DSGM algorithm is simple and can be easily employed to a broad class of problems without any complex adjustments. We show that all accumulation points of the sequence generated by the method are stationary. We devise a convergence rate of $O(\frac{1}{k^{1/3}})$ in terms of an optimality measure that can be easily computed. Numerical experiments illustrate the potential of the DSGM method.

## 1 Introduction

This paper is concerned with minimization problems of the form

$$\min_{\mathbf{x}\in\mathcal{M}}\{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})\}, \tag{1.1}$$

where $\mathcal{M} \subseteq \mathbb{R}^n$ is a compact smooth embedded submanifold of $\mathbb{R}^n$, $f$ is a smooth function, $g$ is a real-valued convex nonsmooth function and $\mathbf{A}$ is a given matrix. We assume that $g$ is prox-tractable, i.e., $\text{prox}_{\mu g}(\mathbf{x})$ can be easily computed for any $\mu > 0$. Problems of this form arise often in machine learning, signal processing and scientific computing, see for example the many examples in [10]. We also refer the reader to the numerical section

*School of Mathematics Sciences, Tel Aviv University; email: becka@tauex.tau.ac.il. The research of Amir Beck is partially supported by ISF grant no. 926/21.

†School of Mathematics Sciences, Tel Aviv University; email: israelrosset5@gmail.com

(Section 5) that considers two models: sparse principal component analysis and robust subspace recovery.

Optimization of smooth functions over manifolds is a well studied topic. In their book [2], Absil et al. present and analyze several first-order methods aimed at solving optimization problems over manifolds. These methods heavily rely on the concept of retractions on manifolds. A retraction is a smooth mapping from the tangent bundle of the manifold to the manifold, $R : T\mathcal{M} \to \mathcal{M}$, where

$$R(0_{\mathbf{x}}) = \mathbf{x},\ DR_{\mathbf{x}}(0_{\mathbf{x}}) = Id_{T_{\mathbf{x}}\mathcal{M}}$$

for any $\mathbf{x} \in \mathcal{M}$. Here $0_{\mathbf{x}}$ denotes the zero element in the tangent linear space $T_{\mathbf{x}}\mathcal{M}$ and $R_{\mathbf{x}}$ is the retraction operator $R$ on $T_{\mathbf{x}}\mathcal{M}$. For a more comprehensive presentation, we refer to [2].

Given a retraction, one can replace the linear steps in Euclidean gradient descent with retraction-based steps over the manifold. Usually the definition of a retraction provides a mechanism to proving convergence to a stationary point as well as establishing rates of convergence. In order to receive better computational performances, the retraction can be chosen more carefully, see for example the work [24] in which it is suggested to use the Cayley transform on the Stiefel manifold to get a high performance retraction-based method. Many other Euclidean methods were generalized and illustrated in the Riemmanian case. Examples are the Riemmanian trust-region method [1], Newton methods [20] and much more, see also the comprehensive book [8].

Optimization of nonsmooth functions on manifolds attracts a growing attention in recent years. Subgradient-based algorithms on manifolds were proposed in [7, 12, 26] for geodesic-convex functions on Hadamard manifolds. Gradient sampling methods can be found for example in [13, 14]. In these methods, the chosen direction at each iteration is obtained as a solution of a minimization problem over the convex hull of sampled gradients. The convergence result (without a convergence rate) is obtained under mild assumptions; we note that each iteration comprises the evaluation of $dim(\mathcal{M})$ gradients and involves the solution of a quadratic programming problem.

Very recently, Li et al. [25] investigated the rate of convergence of the Riemmanian subgradient method for weakly convex functions over the Stiefel manifold. They used the gradient of an operator analogous to the Moreau envelope on the manifold to get a stationarity measure. They showed that their optimality measure converges to zero with a rate of $O(n^{-1/4})$ using a specific stepsize strategy. They also showed this result for two variants of an alternating subgradient method for a sum of weakly convex functions.

Early attempts to employ proximal point methods for nonsmooth optimization over manifolds include [6, 7, 11]. The convergence analysis assumes that the underlying manifold is Hadamard and the algorithms require the solution of rather complicated optimiza-

tion problems at each iteration.

Recently, the authors of [10] considered a proximal gradient approach to the model (1.1), where $\mathbf{A}$ is the identity matrix. Their method is an iterative method that generates a sequence of points $\mathbf{x}_k \in \mathcal{M}$. The update in each step is done using a direction that solves the following minimization problem:

$$\mathbf{v}_k := \underset{\mathbf{v}\in T_{\mathbf{x}_k}\mathcal{M}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{x}_k), \mathbf{v}\rangle + \frac{1}{2t}\|\mathbf{v}\|_F^2 + g(\mathbf{x}_k + \mathbf{v}) \right\}. \tag{1.2}$$

In order to solve (1.2), Chen et. al. [10] assumed that $\mathcal{M}$ is the Stiefel manifold and used a semi-smooth Newton method. In the scenario that (1.2) is solved exactly and $\mathcal{M}$ is the Stiefel manifold, convergence is established. Huang and Wei [15] use an update direction that is a stationary point for the following problem:

$$\mathbf{v}_k := \underset{\mathbf{v}\in T_{\mathbf{x}_k}\mathcal{M}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{x}_k), \mathbf{v}\rangle + \frac{1}{2t}\|\mathbf{v}\|_F^2 + g(R_{\mathbf{x}_k}(\mathbf{v})) \right\}, \tag{1.3}$$

for a retraction $R$. The convergence of the above algorithm in a setting where the objective function satisfies the Riemannian Kurdyka–Łojasiewicz property was shown.

The approach suggested in this paper uses a dynamic smoothing technique to overcome the nonsmooth part $g$. Optimization algorithms with proven convergence guarantees that use smoothing mechanisms for convex problems were first analyzed by Nesterov in [18] and later on in [5]. The idea of using a dynamic smoothing technique in which the smoothing parameter gradually decreases and tends to 0 was well studied in convex setting, see for example the works [9, 22] for a variety of complexity results.

In this work, we suggest to employ a dynamic smoothing approach for the nonconvex model (1.1). At iteration $k$ we employ a Riemmanian gradient step on the problem

$$\min_{\mathbf{x}\in\mathcal{M}} \left\{ f(\mathbf{x}) + M_g^{\mu_k}(\mathbf{A}\mathbf{x}) \right\}, \tag{1.4}$$

where $M_g^{\mu}$ is the so-called Moreau envelope of $g$ with a smoothing parameter $\mu$, see more details in Section 2.1. The sequence $\{\mu_k\}_{k\geq 0}$ is positive and diminishing towards 0. The method, which we refer to as *Dynamic Smoothing Gradient descent on Manifolds (DSGM)* is extremely simple and requires at each iteration the calculation of four components: (1) gradient of $f$ (2) proximal operator of $g$ (3) orthogonal projection onto the tangent space at the given iterate vector (4) computation of a smooth path on the manifold passing in the direction of the given iterate vector.

In Section 3 we describe the model and the exact underlying assumptions; we also discuss the issue of stationarity and its connection to optimality in relation to (1.1). In Section 4 we study the convergence of the DSGM method—we show that all accumulation points of the sequence generated by the DSGM algorithm are stationary points of (1.1) and that a corresponding easily computed optimality measure converges in a rate of $O(\frac{1}{k^{1/3}})$.

# 2 Preliminaries and Notations

The underlying space in this paper is $\mathbb{R}^n$, where "$\mathbb{R}^n$" in this context can represent any finite-dimensional inner product linear space with an Euclidean norm (meaning $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$). Unless otherwise stated, the inner product and norm in $\mathbb{R}^n$ are the dot product and $\ell_2$-norm, while matrix spaces are endowed with the dot inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathrm{Tr}(\mathbf{x}^T \mathbf{y})$ and the induced Frobenius norm. The Frobenius and spectral norms are denoted by $\|\cdot\|_F$ and $\|\cdot\|_2$ respectively.

## 2.1 Optimization

We use standard terminology from convex and nonconvex optimization. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called $L$-smooth if it is differentiable and its gradient, denoted by $\nabla f$, is $L$-Lipschitz, meaning that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. For a proper closed and convex function $g : \mathbb{R}^n \to (-\infty, \infty]$, the celebrated *Moreau envelope* is given by

$$M_g^\mu(\mathbf{x}) = \min_{u \in \mathbb{R}^n} \left\{ g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{u}\|^2 \right\}. \tag{2.1}$$

It is well known that if $g$ is proper closed and convex, the Moreau envelope is a real valued, convex and $\frac{1}{\mu}$-smooth function [17].
The proximal mapping of a function $g : \mathbb{R}^n \to (-\infty, \infty]$ is defined as follows:

$$\mathrm{prox}_g(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^n} \left\{ g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}. \tag{2.2}$$

For a function $F : \mathbb{R}^n \to \mathbb{R}$, if the directional derivative at a point $\mathbf{x} \in \mathbb{R}^n$ along a direction $\mathbf{v} \in \mathbb{R}^n$ exists, it will be denoted $F'(\mathbf{x}; \mathbf{v})$. Another important notion is the **orthogonal projection mapping**. The projection onto a given nonempty closed set $C \subseteq \mathbb{R}^n$ is defined as the multi-valued mapping:

$$\mathcal{P}_C(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{z} \in C} \|\mathbf{z} - \mathbf{x}\|.$$

A well known result in the case that $C$ is nonempty closed and convex, is that this mapping is single-valued, and moreover, nonexpansive [4, Theorem 5.4], meaning that

$$\|\mathcal{P}_C(\mathbf{x}) - \mathcal{P}_C(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.3}$$

## 2.2 Manifolds

The standard and most general definition of a manifold is a topological space $\mathcal{M}$ endowed with a set of homeomorphisms,

$$\phi_i : U_i \to V,\ V \subseteq \mathbb{R}^d,\ i \in I,$$

where $I$ is a set of indices, $U_i \subseteq \mathcal{M}$ are open, $\bigcup_{i\in I} U_i = \mathcal{M}$ and $V \subseteq \mathbb{R}^d$ is open. These maps are called charts and $d$ is the so-called dimension of the manifold. Each such chart gives a local differentiable structure on $\mathcal{M}$, which yields to general concepts of a differentiable mapping on a manifold, the differential of a mapping and so on. For our purpose, we will focus on embedded submanifolds of $\mathbb{R}^n$. A manifold $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$ if it is a manifold, it is a topological subspace of $\mathbb{R}^n$, the inclusion $\iota : \mathcal{M} \to \mathbb{R}^n$ is a differentiable mapping and its differential $D\iota(\mathbf{x})$ is of full rank for every $\mathbf{x} \in \mathcal{M}$. For a more concise introduction to the theory of manifolds, we refer the reader to [2].

An important class of embedded submanifolds is the class of manifolds of the form

$$\mathcal{M} = \{\mathbf{x} \in \mathbb{R}^n \mid H(\mathbf{x}) = 0\},$$

where $H : \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable mapping such that $DH(\mathbf{x})$ is of rank $m$ for every $\mathbf{x} \in \mathcal{M}$. The dimension of $\mathcal{M}$ in this case is $n - m$.

Let $\mathcal{M}$ be an embedded submanifold and $\mathbf{x} \in \mathcal{M}$. A function $\gamma : [0, \infty) \to \mathcal{M}$ satisfying that it is smooth and that $\gamma(0) = \mathbf{x}$ is called a *path* from $\mathbf{x}$ in the direction $\gamma'(0)$, where the derivative of $\gamma$, when viewed as a function to $\mathbb{R}^n$ is denoted $\gamma'(t)$ for every $t \in [0, \infty)$. We note that $\gamma'(0)$ is actually the right derivative $\gamma'_+(0)$, but we will adopt this slight abuse of notation throughout the paper. The *tangent space* of an embedded submanifold $\mathcal{M} \subseteq \mathbb{R}^n$ at $\mathbf{x}$ can now be defined as follows:

$$T_{\mathbf{x}}\mathcal{M} = \{\gamma'(0) \mid \gamma : [0, \infty) \to \mathcal{M} \text{ smooth and } \gamma(0) = \mathbf{x}\} \subseteq \mathbb{R}^n.$$

This set turns up to be a linear subspace of $\mathbb{R}^n$. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we define the *Riemmanian gradient* of $f$ on $\mathcal{M}$ as follows:

$$\forall \mathbf{x} \in \mathcal{M} : \ \nabla_{\mathcal{M}} f(\mathbf{x}) = \mathcal{P}_{T_{\mathbf{x}}\mathcal{M}} \left(\nabla f(\mathbf{x})\right), \tag{2.4}$$

where $\mathcal{P}_V$ is the orthogonal projection on the set $V$. For more general and concise definitions of the tangent space and the differential of a mapping, which do not assume that $\mathcal{M}$ is an embedded submanifold, we again refer to [2].

# 3 The Manifold Composite Model

In this paper we study the following optimization problem.

**Manifold Composite Model**

$$\text{(P)} \quad \min_{\mathbf{x}\in\mathcal{M}} \{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})\}.$$

The following underlying assumption on the problem's data $(f, g, A, \mathcal{M})$ will be assumed throughout the paper.

**Assumption 1.**

- $\mathcal{M} \subseteq \mathbb{R}^n$ is a compact smooth embedded submanifold.
- $f : \mathbb{R}^n \to \mathbb{R}$ is an $L_f$-smooth function ($L_f > 0$).
- $g : \mathbb{R}^m \to \mathbb{R}$ is a convex function and $L_g$-Lipschitz continuous ($L_g > 0$).
- $\mathbf{A} \in \mathbb{R}^{m\times n}$.

We also define the function $h(\mathbf{x}) := g(\mathbf{A}\mathbf{x})$, and in this notation the objective function is $F = f + h$. We will frequently use the subdifferential calculus rule $\partial h(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x})$.

## 3.1 Stationarity

We start by defining the notion of stationarity.

**Definition 3.1** (stationarity)**.** A point $\mathbf{x} \in \mathcal{M}$ is stationary for (P) if there exists $\boldsymbol{\xi} \in \partial g(\mathbf{A}\mathbf{x})$ such that

$$\nabla f(\mathbf{x}) + \mathbf{A}^T \boldsymbol{\xi} \perp T_{\mathbf{x}}\mathcal{M}. \tag{3.1}$$

The above condition appeared in previous works such as [25]. To justify this condition, we prove Lemma 3.1 below stating that (3.1) is equivalent to saying that there are no descent directions belonging to the tangent space. We note that this result, as well as Theorem 3.1 that follows can be deduced from the work [27], which deals with more general settings. We provide here the simple proofs of the two results in our setting for the sake of completeness.

**Lemma 3.1.** A point $\mathbf{x} \in \mathcal{M}$ is a stationary point of (P) if and only if

$$F'(\mathbf{x}; \mathbf{v}) \geq 0 \text{ for all } \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}. \tag{3.2}$$

*Proof.* Let $\mathbf{x} \in \mathcal{M}$, and assume that (3.2) holds. Let $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. Then

$$F'(\mathbf{x}; \mathbf{v}) = \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle + h'(\mathbf{x}; \mathbf{v}) \geq 0.$$

Thus, by the convexity of $h$, we get

$$h(\mathbf{x} + \mathbf{v}) - h(\mathbf{x}) \geq h'(\mathbf{x}; \mathbf{v}) \geq \langle -\nabla f(\mathbf{x}), \mathbf{v} \rangle \text{ for any } \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}. \tag{3.3}$$

Denote $V = \mathbf{x} + T_{\mathbf{x}}\mathcal{M}$. Then (3.3) translates to $-\nabla f(\mathbf{x}) \in \partial(h + \delta_V)(\mathbf{x})$. By calculus of subdifferentials[1],

$$-\nabla f(\mathbf{x}) \in \partial h(\mathbf{x}) + N_V(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x}) + (T_{\mathbf{x}}\mathcal{M})^{\perp},$$

which is the same as condition (3.1).
In the opposite direction, assume that there exists $\boldsymbol{\xi} \in \partial h(\mathbf{x})$ such that $\nabla f(\mathbf{x}) + \boldsymbol{\xi} \perp T_{\mathbf{x}}\mathcal{M}$. Let $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. Then since $\boldsymbol{\xi} \in \partial h(\mathbf{x})$, it follows by [19, Theorem 23.2] that $h'(\mathbf{x}; \mathbf{v}) \geq \langle \boldsymbol{\xi}, \mathbf{v} \rangle$, and hence $F'(\mathbf{x}; \mathbf{v}) = \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle + h'(\mathbf{x}; \mathbf{v}) \geq \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle + \langle \boldsymbol{\xi}, \mathbf{v} \rangle = 0$. □

The following result shows that stationarity is a necessary condition for local optimality.

**Theorem 3.1.** Any local minimum of problem (P) is a stationary point of (P).

*Proof.* Let $\mathbf{x} \in \mathcal{M}$ be a local minimum point of (P) and let $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. Let $\gamma : [0, \infty) \to \mathcal{M}$ be a path such that $\gamma(0) = \mathbf{x}, \gamma'(0) = \mathbf{v}$. Since $\mathbf{x}$ is a local minimum of (P), there exists $\varepsilon > 0$ such that for every $t \in (0, \varepsilon)$,

$$F(\gamma(t)) \geq F(\mathbf{x}).$$

Recalling that $F = f + h$, and after some rearrangement of terms, we get that the above inequality is the same as

$$\frac{f(\gamma(t)) - f(\mathbf{x})}{t} \geq -\frac{h(\gamma(t)) - h(\mathbf{x})}{t}. \tag{3.4}$$

By the definition of the derivative,

$$\frac{f(\gamma(t)) - f(\mathbf{x})}{t} \xrightarrow{t \to 0^+} (f \circ \gamma)'(0) = \langle \nabla f(\gamma(0)), \gamma'(0) \rangle = \langle \nabla f(\mathbf{x}), \mathbf{v} \rangle = f'(\mathbf{x}; \mathbf{v}). \tag{3.5}$$

We wish to show that the right-hand side of (3.4) tends to $-h'(\mathbf{x}; \mathbf{v})$ as $t \to 0^+$. Indeed,

$$\frac{h(\gamma(t)) - h(\mathbf{x})}{t} = \frac{h(\gamma(t)) - h(\mathbf{x} + t\mathbf{v})}{t} + \frac{h(\mathbf{x} + t\mathbf{v}) - h(\mathbf{x})}{t}. \tag{3.6}$$

Since $g$ is $L_g$-Lipschitz continuous, it follows that $h$ is $\|\mathbf{A}\|L_g$-Lipschitz continuous. Thus,

$$\frac{|h(\gamma(t)) - h(\mathbf{x} + t\mathbf{v})|}{t} \leq \frac{\|\mathbf{A}\|L_g \|\gamma(t) - (\mathbf{x} + t\mathbf{v})\|}{t} \leq \|\mathbf{A}\|L_g \left\| \frac{\gamma(t) - \mathbf{x}}{t} - \mathbf{v} \right\| \xrightarrow[(\mathbf{v} = \gamma'(0))]{t \to 0^+} 0,$$

which combined with (3.6) and the definition of the directional derivative implies that $\frac{h(\gamma(t)) - h(\mathbf{x})}{t} \xrightarrow{t \to 0_+} h'(\mathbf{x}; \mathbf{v})$. Combining this with (3.4) and (3.5) implies that $F'(\mathbf{x}; \mathbf{v}) \geq 0$ for any $\mathbf{x} \in T_{\mathbf{x}}\mathcal{M}$, and by Lemma 3.1, we conclude that $\mathbf{x}$ is a stationary point of (P). □

[1] Given a set $C \subseteq \mathbb{R}^n$, $N_C(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \ \forall \mathbf{z} \in C\}$ is the normal cone of $C$ at $\mathbf{x}$

## 3.2 Smooth Paths Sets

We begin by defining the notion of "paths set".

**Definition 3.2** (paths set)**.** A paths set $\gamma$ on a manifold $\mathcal{M}$ is a set of paths on the manifold

$$\{\gamma_{\mathbf{x},\mathbf{v}} : [0,\infty) \to \mathcal{M} \mid \mathbf{x} \in \mathcal{M}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}, \|\mathbf{v}\| = 1\}$$

such that for every $\mathbf{x} \in \mathcal{M}$ and $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}, \|\mathbf{v}\| = 1$ the following holds:

- $\gamma_{\mathbf{x},\mathbf{v}}(0) = \mathbf{x}$;
- $\gamma'_{\mathbf{x},\mathbf{v}}(0) = \mathbf{v}$.

We consider only direction vectors of unit length because it suffices for the definition of our algorithm.
One way to construct paths set is using the notion of retractions (see for example [2, Definition 4.1.1]). Specifically, every retraction $R$ induces a paths set via the formula $\gamma_{\mathbf{x},\mathbf{v}}(t) = R_{\mathbf{x}}(t\mathbf{v})$. Note that for $R$ to be a retraction, one requires that the induced map on the tangent bundle is smooth. We will be satisfied using the notion of paths sets, as it is enough for our needs.

The analysis of our algorithm will be based on a uniform smoothness property for paths sets that is now defined.

**Definition 3.3** (smooth paths set)**.** A paths set $\gamma$ on a manifold $\mathcal{M}$ is $(L, M)$-*smooth* if for every $\mathbf{x} \in \mathcal{M}$ and $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ such that $\|\mathbf{v}\| = 1$, the following holds:

A. $\|\gamma_{\mathbf{x},\mathbf{v}}(t) - \gamma_{\mathbf{x},\mathbf{v}}(s)\| \leq M|t - s|$ for all $t, s \geq 0$;

B. $\|\gamma'_{\mathbf{x},\mathbf{v}}(t) - \mathbf{v}\| \leq Lt$ for all $t \geq 0$.

**Example 3.1.** Recall that the Stiefel manifold is defined as

$$\mathcal{M} = \{\mathbf{x} \in \mathbb{R}^{d\times p} \mid \mathbf{x}^T\mathbf{x} = \mathbf{I}_{p\times p}\}.$$

Define the following paths set on $\mathcal{M}$:

$$\gamma_{\mathbf{x},\mathbf{v}}(t) = \mathcal{P}_{\mathcal{M}}(\mathbf{x} + t\mathbf{v}),$$

where $P_{\mathcal{M}}$ is the orthogonal projection operator on $\mathcal{M}$. In Section 4.3 we show that $\{\gamma_{\mathbf{x},\mathbf{v}}\}$ constitutes a $(\sqrt{p}+1, 1)$-smooth paths set.

**Example 3.2.** Let $\mathcal{M} = \mathcal{M}_1 \times \cdots \times \mathcal{M}_r$ be a product of $r$ manifolds. This setting is relevant in many scenarios. For example, the Oblique manifold $\mathcal{OB}(d, r) := \overbrace{\mathbb{S}^{d-1} \times \cdots \times \mathbb{S}^{d-1}}^{r \text{ times}}$

appears in many applications, for example, in the context of optimization problems over unimodal constraints [21, 23].

Assume we have a paths set $\gamma^i$ on each manifold $\mathcal{M}_i$. Assume that for every $i = 1, ..., r$ the $\gamma^i$ paths set is $(L_i, M_i)$-smooth. We now use those paths sets to define a paths set $\gamma$ on $\mathcal{M}$. Let $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_r) \in \mathcal{M}$ be a point on $\mathcal{M}$, such that $\mathbf{x}_i \in \mathcal{M}_i$, and let $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. Using the natural isomorphism between $T_{\mathbf{x}}\mathcal{M}$ and $T_{\mathbf{x}_1}\mathcal{M}_1 \times \cdots \times T_{\mathbf{x}_r}\mathcal{M}_r$, denote $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_r)$ such that $\mathbf{v}_i \in T_{\mathbf{x}_i}\mathcal{M}_i$. Define $\gamma_{\mathbf{x},\mathbf{v}}$ as follows:

$$\gamma_{\mathbf{x},\mathbf{v}}(t) = (\gamma^1_{\mathbf{x}_1,\mathbf{v}_1}(t), ..., \gamma^r_{\mathbf{x}_r,\mathbf{v}_r}(t)).$$

The reader can easily verify that $\gamma$ is a $(\sqrt{\sum L_i^2}, \sqrt{\sum M_i^2})$-smooth paths set.

Combining this discussion with the previous example, we can conclude that $\gamma_{\mathbf{x},\mathbf{v}}(t) = (\mathcal{P}_{\mathbb{S}^{d-1}}(\mathbf{x}_1 + t\mathbf{v}_1), \ldots, \mathcal{P}_{\mathbb{S}^{d-1}}(\mathbf{x}_r + t\mathbf{v}_r))$ is a $(2\sqrt{r}, \sqrt{r})$-smooth paths set on the oblique manifold $\mathcal{OB}(d, r)$.

# 4 Dynamic Smoothing Gradient on Manifolds

## 4.1 The Method

To describe our algorithm, we will use the following smoothing operation on $F$. For every $\mu > 0$ and for every $x \in \mathcal{M}$, denote

$$F^{(\mu)}(\mathbf{x}) \equiv f(\mathbf{x}) + M_g^{\mu}(\mathbf{A}\mathbf{x}), \tag{4.1}$$

where $M_g^{\mu}$ is the Moreau envelope given in (2.1). At iteration $j$ of the algorithm ($j = 0, 1, 2, \ldots$), one gradient step of Riemmanian gradient descent is taken with respect to $M_g^{\mu_j}$ with a positive and decreasing sequence of smoothing parameters $\{\mu_j\}_{j\geq 0}$.

**Dynamic Smoothing Gradient descent on Manifolds (DSGM) Method**

**Model Input:** $(\mathcal{M}, f, g, A)$ satisfying Assumption 1.
**Algorithm Input:** $\gamma$ – $(L_\gamma, M_\gamma)$-smooth paths set on $\mathcal{M}$; $\{\mu_j\}_{j=0}^{\infty}$ – decreasing sequence of positive smoothing parameters.
**Initialization:** Pick $\mathbf{x}_0 \in \mathcal{M}$.
**General step**: for any $j = 0, 1, 2, \ldots$ execute the following steps:

- Set $\mathbf{v}_j = -\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)/\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|$ and $\gamma = \gamma_{\mathbf{x}_j, \mathbf{v}_j}$;
- Set $\mathbf{x}_{j+1} = \gamma\big(t_j \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|\big)$ for some stepsize $t_j > 0$;
- Set $\mathbf{X}_j = \mathbf{x}_{N_j}$ where $N_j \in \underset{k=\lfloor \frac{j}{2} \rfloor, \ldots, j}{\operatorname{argmin}} \left\|\nabla_{\mathcal{M}} F^{(\mu_k)}(\mathbf{x}_k)\right\|$.

**Intermediate sequence** – $\{\mathbf{x}_j\}_{j\geq 0}$.
**Associated smoothing parameters sequence** – $\{\tilde{\mu}_j\}_{j\geq 0}$ where $\tilde{\mu}_j = \mu_{N_j}$.
**Output sequence** – $\{\mathbf{X}_j\}_{j\geq 0}$.

The output sequence $\{\mathbf{X}_j\}_{j\geq 0}$ is the actual output of the method. The intermediate sequence $\{\mathbf{x}_j\}_{j\geq 0}$ will be important in the convergence analysis to follow, but is not considered as the "output" of the method. The associated smoothing parameters sequence $\{\tilde{\mu}_j\}_{j\geq 0}$, which will be used in the convergence analysis, comprises the parameters corresponding to the output sequence.

Two missing important ingredients in the above description of the method are (1) the choice of smoothing parameters sequence $\{\mu_j\}_{j\geq 0}$ and (2) strategy for choosing the stepsize sequence $\{t_j\}_{j\geq 0}$. The smoothing parameters will be determined during the convergence analysis in Section 4.2 (specifically, Theorem 4.1). Given the choice of the smoothing parameters, we now propose two possible stepsize strategies.

- **Predefined diminishing stepsize.** $t_j = \frac{1}{\kappa(\mu_j)}$, where $\kappa : \mathbb{R}_{++} \to \mathbb{R}_{++}$ is a decreasing function.
- **Backtracking stepsize**. This procedure requires three parameters $s > 0, \alpha \in (0,1), \beta \in (0,1)$; $t_j$ is chosen as the largest element in the set $\{s\beta^k\}_{k=0}^{\infty}$ such that

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}\big(\gamma\big(t_j \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|\big)\big) \geq \alpha t_j \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2. \tag{4.2}$$

Since the sequence of smoothing parameters $\{\mu_j\}_{j\geq 0}$ is decreasing, and $\kappa$ is a decreasing function, it follows that the sequence of stepsize sequence $t_j = \frac{1}{\kappa(\mu_j)}$ in the first strategy ("predefined diminishing stepsize") is also decreasing. The function $\kappa$ will be determined in the sequel, see equation (4.4).

## 4.2 Convergence Analysis

First, we recall some properties of the Moreau envelope (see (2.1)) .

**Lemma 4.1.** Let $g : \mathbb{R}^m \to \mathbb{R}$ be an $L_g$-continuous convex function, $\mathbf{z} \in \mathbb{R}^m$ and let $\mathbf{y} = \text{prox}_{\mu g}(\mathbf{z})$ for some $\mu > 0$. Then,

(a) $M_g^\mu$ is $\frac{1}{\mu}$-smooth;

(b) $\nabla M_g^\mu(\mathbf{z}) = \frac{1}{\mu}(\mathbf{z} - \mathbf{y}) \in \partial g(\mathbf{y})$;

(c) $\|\mathbf{z} - \mathbf{y}\| \leq L_g \mu$ and $\|\nabla M_g^\mu(\mathbf{z})\| \leq L_g$.

*Proof.* (a) Follows [4, Theorem 6.60]. (b) The formula for the gradient of the Moreau envelope is well known, see for example [4, Theorem 6.60]; the membership in the sub-differential set is the result [4, Theorem 6.39(ii)]. (c) Since $g$ is $L_g$-Lipschitz continuous, it follows by [4, Theorem 3.61] that $\partial g(\mathbf{y}) \subseteq B[0, L_g]$ for any $\mathbf{y}$. Thus, since by part (b), $\nabla M_g^\mu(\mathbf{z}) \in \partial g(\mathbf{y})$, it follows $\|\nabla M_g^\mu(\mathbf{z})\| \leq L_g$ and since $\nabla M_g^\mu(\mathbf{z}) = \frac{1}{\mu}(\mathbf{z} - \mathbf{y})$, we also conclude that $\|\mathbf{z} - \mathbf{y}\| \leq L_g \mu$. □

Using part (b) of Lemma 4.1, we can rephrase the descent direction that is chosen in the DSGM algorithm at iteration $j$ as

$$-\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j) = -\nabla_{\mathcal{M}} f(\mathbf{x}_j) - \frac{1}{\mu_j} \mathcal{P}_{T_{\mathbf{x}_j}\mathcal{M}}(\mathbf{A}^T \mathbf{A}\mathbf{x}_j - \mathbf{A}^T \,\text{prox}_{\mu_j g}(\mathbf{A}\mathbf{x}_j)).$$

To prove the convergence of the algorithm, we quantify the smoothness of the composition $F^{(\mu)} \circ \gamma_{\mathbf{x},\mathbf{v}}$, where $\gamma$ is a given smooth paths set. The result is expressed in terms of an upper bound on the norm of $\nabla f$ over the manifold, whose existence is warranted by the compactness of the manifold (Assumption 1). The bound is denoted by $U_f$, meaning that

$$\max_{\mathbf{z} \in \mathcal{M}} \|\nabla f(\mathbf{z})\| \leq U_f. \tag{4.3}$$

**Lemma 4.2.** Let $\gamma$ be an $(L_\gamma, M_\gamma)$-smooth paths set on $\mathcal{M}$ and let $\kappa : \mathbb{R}_{++} \to \mathbb{R}_{++}$ be given by

$$\kappa(\mu) := L_f M_\gamma^2 + L_\gamma(\|\mathbf{A}\| L_g + U_f) + M_\gamma^2 \frac{\|\mathbf{A}\|^2}{\mu}, \tag{4.4}$$

where $U_f$ satisfies (4.3). Then for every $\mathbf{x} \in \mathcal{M}$ and $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$, it holds that

$$|(F^{(\mu)} \circ \gamma_{\mathbf{x},\mathbf{v}})'(t) - (F^{(\mu)} \circ \gamma_{\mathbf{x},\mathbf{v}})'(0)| \leq \kappa(\mu) t \quad \text{for all } t \geq 0.$$

*Proof.* First, we upper bound the norm of the gradient of $F^{(\mu)}$ as follows:

$$\max_{\mathbf{x} \in \mathcal{M}} \|\nabla F^{(\mu)}(\mathbf{x})\| \leq U_f + \max_{\mathbf{x} \in \mathcal{M}} \|\mathbf{A}^T \nabla M_g^\mu(\mathbf{A}\mathbf{x})\| \leq U_f + \|\mathbf{A}\| L_g, \tag{4.5}$$

where we used Lemma 4.1(c) in the second inequality. Since $F^{(\mu)}$ is sum of $f$, which is $L_f$-smooth, and $\mathbf{x} \mapsto M_g^\mu(\mathbf{A}\mathbf{x})$, which is $\frac{\|\mathbf{A}\|^2}{\mu}$-smooth (follows from Lemma 4.1(a)), we get that $F^{(\mu)}$ is $(L_f + \frac{\|\mathbf{A}\|^2}{\mu})$-smooth. For the sake of simplicity of notation, we omit the super/subscripts in $F^{(\mu)}$ and $\gamma_{\mathbf{x},\mathbf{v}}$ and write $F$ and $\gamma$ instead. Thus, for example, in this terminology we have shown that the $F$ is $L_F$-smooth with $L_F = L_f + \frac{\|\mathbf{A}\|^2}{\mu}$.

Let $\mathbf{x} \in \mathcal{M}$ and $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. By property A of smooth paths sets (Definition 3.3), we have for any $t \geq 0$,

$$\|\gamma'(t)\| = \lim_{\delta \to 0} \frac{\|\gamma(t+\delta) - \gamma(t)\|}{\delta} \leq M_\gamma. \tag{4.6}$$

Denote $\phi = F \circ \gamma$. For every $t \geq 0$, the following holds:

$$\begin{aligned}
|\phi'(t) - \phi'(0)| &= |\nabla F(\gamma(t))^T \gamma'(t) - \nabla F(\mathbf{x})^T \mathbf{v}| \\
&= |\nabla F(\gamma(t))^T \gamma'(t) - \nabla F(\mathbf{x})^T \gamma'(t) + \nabla F(\mathbf{x})^T \gamma'(t) - \nabla F(\mathbf{x})^T \mathbf{v}| \\
&\leq |\nabla F(\gamma(t))^T \gamma'(t) - \nabla F(\mathbf{x})^T \gamma'(t)| + |\nabla F(\mathbf{x})^T \gamma'(t) - \nabla F(\mathbf{x})^T \mathbf{v}| \\
&\leq \|\nabla F(\gamma(t)) - \nabla F(\mathbf{x})\| \cdot \|\gamma'(t)\| + \|\nabla F(\mathbf{x})\| \cdot \|\gamma'(t) - \mathbf{v}\| \\
&\leq L_F \|\gamma(t) - \mathbf{x}\| \cdot \|\gamma'(t)\| + \|\nabla F(\mathbf{x})\| \cdot \|\gamma'(t) - \mathbf{v}\| \\
&= L_F \|\gamma(t) - \gamma(0)\| \cdot \|\gamma'(t)\| + \|\nabla F(\mathbf{x})\| \cdot \|\gamma'(t) - \mathbf{v}\| \\
&\overset{(4.5),(4.6)}{\leq} \left(L_F M_\gamma^2 + (U_f + \|\mathbf{A}\| L_g) L_\gamma\right) t.
\end{aligned}$$

Plugging $L_F = L_f + \frac{\|\mathbf{A}\|^2}{\mu}$ in the above, the result follows. □

Lemma 4.4 below establishes a decrease property of two consecutive iterates $\mathbf{x}_j$ and $\mathbf{x}_{j+1}$ generated by the DSGM method with respect to the smoothed functions $F^{(\mu_j)}$ and $F^{(\mu_{j+1})}$ respectively. The result requires the following technical lemma.

**Lemma 4.3.** Suppose $\phi : [0, \infty) \to \mathbb{R}$ is differentiable and satisfies that $\phi'(0) < 0$ and that for any $t \geq 0$

$$|\phi'(t) - \phi'(0)| \leq Kt$$

for some $K > 0$. Then for any $t \geq 0$

$$\phi(0) - \phi(-t\phi'(0)) \geq t\left(1 - \frac{Kt}{2}\right)\phi'(0)^2.$$

*Proof.* By the premise of the lemma we get that for any $s \geq 0$, $\phi'(s) \leq \phi'(0) + Ks$. Thus,

$$\phi(0) - \phi(-t\phi'(0)) = -\int_0^{-t\phi'(0)} \phi'(s)ds \geq \int_0^{-t\phi'(0)} (-\phi'(0) - Ks)ds = t\left(1 - \frac{Kt}{2}\right)\phi'(0)^2.$$

□

**Lemma 4.4.** Let $\{\mathbf{x}_j\}_{j\geq 0}$ be the sequence generated by the DSGM method using a decreasing sequence of smoothing parameters $\{\mu_j\}_{j\geq 0}$. Assume that the stepsizes are

chosen using either the predefined procedure with the function $\kappa$ given in (4.4) or the backtracking procedure with parameters $s > 0, \alpha, \beta \in (0,1)$. Then for any $j \geq 0$,

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_{j+1})}(\mathbf{x}_{j+1}) \geq \frac{Q}{\kappa(\mu_j)} \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2 - \frac{L_g^2}{2}(\mu_j - \mu_{j+1}), \tag{4.7}$$

where $Q = \frac{1}{2}$ in the predefined stepsize setting and $Q = \alpha \min\{s\kappa(\mu_0), 2(1-\alpha)\beta\}$ when the backtracking scheme is employed.

*Proof.* We first show the following:

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}(\mathbf{x}_{j+1}) \geq \frac{Q}{\kappa(\mu_j)} \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2, \tag{4.8}$$

Denote $\gamma_j = \gamma_{\mathbf{x}_j, \mathbf{v}_j}$, where $\mathbf{v}_j = -\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j) / \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|$ as defined in the DSGM method. By Lemma 4.2, we have that the function $\phi = F^{(\mu_j)} \circ \gamma_j$ satisfies $|\phi'(t) - \phi'(0)| \leq \kappa(\mu_j) t$ for any $t \geq 0$. Thus, invoking Lemma 4.3 with $K = \kappa(\mu_j)$ we have for any $t \geq 0$,

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}\left(\gamma_j\left(t\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|\right)\right) \geq t\left(1 - \frac{\kappa(\mu_j)t}{2}\right)\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2, \tag{4.9}$$

where we used the following computation in the above:

$$\phi'(0) = \nabla F^{(\mu_j)}(\mathbf{x}_j)^T \gamma_j'(0) = -\frac{\nabla F^{(\mu_j)}(\mathbf{x}_j)^T \nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)}{\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|} = -\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|.$$

Consider first the case where the stepsizes are chosen by the predefined diminishing rule $t_j = \frac{1}{\kappa(\mu_j)}$. Substituting $t = t_j = \frac{1}{\kappa(\mu_j)}$ in (4.9) and using the relation $\mathbf{x}_{j+1} = \gamma(t_j \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|)$, we obtain that

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}(\mathbf{x}_{j+1}) \geq \frac{1}{2\kappa(\mu_j)} \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2,$$

meaning that (4.8) holds with $Q = \frac{1}{2}$.

Now assume that the backtracking procedure is employed. Suppose that the stepsize at the $j$th iteration is $t_j = s\beta^r$ and denote $\gamma_j = \gamma_{\mathbf{x}_j, \mathbf{v}_j}$. Then by the construction of the backtracking procedure, either $t_j = s$, or the following holds:

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}\left(\gamma_j\left(\frac{t_j}{\beta}\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|\right)\right) < \alpha \frac{t_j}{\beta}\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2. \tag{4.10}$$

Substituting $t = t_j/\beta$ in (4.9) yields

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}\left(\gamma_j\left(\frac{t_j}{\beta}\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|\right)\right) \geq \frac{t_j}{\beta}\left(1 - \frac{\kappa(\mu_j)t_j}{2\beta}\right)\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2. \tag{4.11}$$

Combining inequalities (4.10) and (4.11), we get that

$$t_j \geq \min\left\{s, \frac{2(1-\alpha)\beta}{\kappa(\mu_j)}\right\},$$

which together with (4.2) yields

$$F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}(\mathbf{x}_{j+1}) \geq \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{\kappa(\mu_j)}\right\} \|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2. \tag{4.12}$$

Note that

$$\alpha \min\left\{s, \frac{2(1-\alpha)\beta}{\kappa(\mu_j)}\right\} \overset{\kappa(\mu_j)\geq\kappa(\mu_0)}{\geq} \alpha \min\left\{s\frac{\kappa(\mu_0)}{\kappa(\mu_j)}, \frac{2(1-\alpha)\beta}{\kappa(\mu_j)}\right\} = \frac{\alpha \min\{s\kappa(\mu_0), 2(1-\alpha)\beta\}}{\kappa(\mu_j)},$$

which together with (4.12) implies that (4.8) holds with $Q = \alpha \min\{s\kappa(\mu_0), 2(1-\alpha)\beta\}$. To prove (4.7), note that

$$\begin{aligned} F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_{j+1})}(\mathbf{x}_{j+1}) &= \left[F^{(\mu_j)}(\mathbf{x}_j) - F^{(\mu_j)}(\mathbf{x}_{j+1})\right] + \left[F^{(\mu_j)}(\mathbf{x}_{j+1}) - F^{(\mu_{j+1})}(\mathbf{x}_{j+1})\right] \\ &\overset{(4.8)}{\geq} \frac{Q}{\kappa(\mu_j)}\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2 + F^{(\mu_j)}(\mathbf{x}_{j+1}) - F^{(\mu_{j+1})}(\mathbf{x}_{j+1}). \end{aligned} \tag{4.13}$$

Finally, by [9, p. 130], we have that for any $\mathbf{x} \in \mathbb{R}^n$, it holds that $M_g^{\mu_j}(\mathbf{A}\mathbf{x}) - M_g^{\mu_{j+1}}(\mathbf{A}\mathbf{x}) \geq -\frac{L_g^2}{2}(\mu_j - \mu_{j+1})$, which combined with (4.13) leads to the desired result (4.7). □

Summing inequality (4.7) over $j = 0, 1, \ldots, N$ yields

$$\sum_{j=0}^{N} \frac{Q}{\kappa(\mu_j)}\|\nabla_{\mathcal{M}} F^{(\mu_i)}(\mathbf{x}_j)\|^2 \leq F^{(\mu_0)}(\mathbf{x}_0) - F^{(\mu_{N+1})}(\mathbf{x}_{N+1}) + \frac{L_g^2}{2}\mu_0. \tag{4.14}$$

Using [9, p. 130], we have that for any $\mathbf{x} \in \mathcal{M}$ and $\mu > 0$,

$$F^{(\mu)}(\mathbf{x}) \geq F(\mathbf{x}) - \frac{L_g^2}{2}\mu. \tag{4.15}$$

Plugging inequality (4.15) with $\mathbf{x} = \mathbf{x}_{N+1}$ and $\mu = \mu_{N+1}$ into (4.14) yields that

$$\begin{aligned} \sum_{j=0}^{N} \frac{1}{\kappa(\mu_j)}\|\nabla_{\mathcal{M}} F^{(\mu_j)}(\mathbf{x}_j)\|^2 &\leq \frac{1}{Q}\left(F^{(\mu_0)}(\mathbf{x}_0) - F(\mathbf{x}_{N+1}) + \frac{1}{2}L_g^2\mu_{N+1} + \frac{1}{2}L_g^2\mu_0\right) \\ &\overset{\mu_{N+1}\leq\mu_0}{\leq} \frac{1}{Q}\left(F^{(\mu_0)}(\mathbf{x}_0) - F_* + L_g^2\mu_0\right) =: C_F \end{aligned} \tag{4.16}$$

where $F_* := \min_{x\in\mathcal{M}} F(\mathbf{x})$. We are now ready to state and prove the main convergence theorems. Theorem 4.1 establishes a rate of convergence result for two expressions that form an optimality measure for the problem. Theorem 4.2 shows that any accumulation point of the sequence is a stationary point.

**Theorem 4.1.** Let $\{\mathbf{X}_j\}_{j\geq 0}$ be the output sequence generated by the DSGM method with smoothing parameters $\mu_j = \frac{D}{\max\{j^v,1\}}$ for some $v \in (0,1)$ and $D > 0$. Assume that the stepsizes were chosen using either the predefined procedure with the function $\kappa$ given in (4.4) or the backtracking procedure with parameters $s > 0, \alpha, \beta \in (0,1)$. Then for any $j \geq 2$ there exists $\mathbf{y}_j \in \mathbb{R}^m$ and $\boldsymbol{\xi}_j \in \partial g(\mathbf{y}_j)$ such that

(a) $\|\mathcal{P}_{T_{\mathbf{X}_j}\mathcal{M}}(\nabla f(\mathbf{X}_j)+\mathbf{A}^T\boldsymbol{\xi}_j)\|^2 \le 4C_F\max\{c_1 j^{-1}, \frac{c_2}{D}j^{v-1}\}$;

(b) $\|\mathbf{y}_j-\mathbf{A}\mathbf{X}_j\| \le \frac{DL_g}{\lfloor \frac{j}{2}\rfloor^v}$,

where

$$c_1 = L_f M_\gamma^2 + L_\gamma(\|\mathbf{A}\|L_g+U_f), c_2 = M_\gamma^2\|\mathbf{A}\|^2, \tag{4.17}$$

with $U_f$ being a constant satisfying (4.3) and $C_F$ is given in (4.16). Lastly, $Q=\frac{1}{2}$ in the case of predefined stepsize and $\alpha\min\{s\kappa(\mu_0), 2(1-\alpha)\beta\}$ in the backtracking setting.

*Proof.* Let $\{\mathbf{x}_j\}_{j\ge 0}$ be the intermediate sequence generated by the method and $\{\tilde{\mu}_j\}_{j\ge 0}$ be the associated smoothing parameters sequence. We begin by noting that

$$\begin{aligned} C_F &\overset{(4.16)}{\ge} \sum_{i=\lfloor j/2\rfloor}^{j}\frac{1}{\kappa(\mu_i)}\|\nabla_{\mathcal{M}}F^{(\mu_i)}(\mathbf{x}_i)\|^2 \ge \frac{j}{2}\min_{i=\lfloor j/2\rfloor,\ldots,j}\frac{1}{\kappa(\mu_i)}\|\nabla_{\mathcal{M}}F^{(\mu_i)}(\mathbf{x}_i)\|^2 \\ &\ge \frac{j}{2}\cdot\frac{\min_{i=\lfloor j/2\rfloor,\ldots,j}\|\nabla_{\mathcal{M}}F^{(\mu_i)}(\mathbf{x}_i)\|^2}{\max_{i=\lfloor j/2\rfloor,\ldots,j}\kappa(\mu_i)} \ge \frac{j}{2\kappa(\mu_j)}\|\nabla_{\mathcal{M}}F^{(\tilde{\mu}_j)}(\mathbf{X}_j)\|^2, \end{aligned}$$

where the last inequality follows from the definition of $\mathbf{X}_j$ and the monotonicity of the sequence $\{\kappa(\mu_j)\}_{j\ge 0}$. By the expression of $\kappa$ given in (4.4), it follows that $\kappa(\mu_j)=c_1+\frac{c_2}{\mu_j}$ where $c_1$ and $c_2$ are given in (4.17). Thus, for any $j\ge 2$,

$$\|\nabla_{\mathcal{M}}F^{(\tilde{\mu}_j)}(\mathbf{X}_j)\|^2 \le \frac{2C_F(c_1+c_2\frac{1}{\mu_j})}{j} \le 4C_F\max\left\{\frac{c_1}{j},\frac{c_2}{j\mu_j}\right\} = 4C_F\max\left\{c_1 j^{-1},\frac{c_2}{D}j^{v-1}\right\}. \tag{4.18}$$

Define $\mathbf{y}_j=\operatorname{prox}_{\tilde{\mu}_j g}(\mathbf{A}\mathbf{X}_j)$ and $\boldsymbol{\xi}_j=\nabla M_g^{\tilde{\mu}_j}(\mathbf{X}_j)\in\partial g(\mathbf{y}_j)$. Since $F^{(\mu)}=f+M_g^{\mu}$, it follows by the formula for the Riemmanian gradient (2.4) that

$$\nabla_{\mathcal{M}}F^{(\tilde{\mu}_j)}(\mathbf{X}_j)=\mathcal{P}_{T_{\mathbf{X}_j}\mathcal{M}}\big(\nabla f(\mathbf{X}_j)+\mathbf{A}^T\boldsymbol{\xi}_j\big).$$

Plugging this equality into (4.18) proves part (a) of the theorem. To prove part (b), note that by Lemma 4.1(c) and the monotonicity of $\mu_j$,

$$\|\mathbf{y}_j-\mathbf{A}\mathbf{X}_j\| \le L_g\tilde{\mu}_j \le L_g\mu_{\lfloor\frac{j}{2}\rfloor} = \frac{DL_g}{\lfloor\frac{j}{2}\rfloor^v},$$

establishing part (b). □

**Remark 4.1.** Note that for the choice $v=\frac{1}{3}$, both expressions $\|\mathcal{P}_{T_{\mathbf{X}_j}\mathcal{M}}(\nabla f(\mathbf{X}_j)+\mathbf{A}^T\boldsymbol{\xi}_j)\|^2$ and $\|\mathbf{y}_j-\mathbf{A}\mathbf{X}_j\|$ are of an order of $O(\frac{1}{k^{1/3}})$, providing a rate of $O(\frac{1}{k^{1/3}})$ for the two expressions. Note that these two expressions form together a proxy for measuring the stationarity of the given point. Indeed, they are both nonnegative and they are equal to zero if and only if $\mathbf{X}_j$ is a stationary point of the main problem, see Definition 3.1.

**Theorem 4.2.** Assume the settings of Theorem 4.1. Then any accumulation point of $\{\mathbf{X}_j\}_{j\ge 0}$ is a stationary point of problem (1.1).

*Proof.* Let $\{\mathbf{x}_j\}_{j\geq 0}$ be the intermediate sequence generated by the method and $\{\tilde{\mu}_j\}_{j\geq 0}$ be the associated smoothing parameters sequence. Let $\mathbf{X}\in\mathcal{M}$ be an accumulation point of the sequence $\{\mathbf{X}_j\}_{j\geq 0}$. Then there exists a subsequence $\{\mathbf{X}_j\}_{j\in K}$ converging to $\mathbf{X}$. By Theorem 4.1, we have that for any $j\geq 2$ there exist $\mathbf{y}_j\in\mathbb{R}^m$ and $\boldsymbol{\xi}_j\in\partial g(\mathbf{y}_j)$ such that

$$\begin{aligned}\|\mathcal{P}_{T_{\mathbf{X}_j}\mathcal{M}}(\nabla f(\mathbf{X}_j)+\mathbf{A}^T\boldsymbol{\xi}_j)\|^2 &\leq s(j), &(4.19)\\ \|\mathbf{y}_j-\mathbf{A}\mathbf{X}_j\| &\leq \frac{DL_g}{\left\lfloor\frac{j}{2}\right\rfloor^v}. &(4.20)\end{aligned}$$

where $s:\mathbb{R}_{++}\to\mathbb{R}_{++}$ is a function satisfying that $s(t)\to 0$ as $t\to\infty$. Since $\mathbf{X}_j\overset{j\in K}{\to}\mathbf{X}$, it follows by (4.20) that $\mathbf{y}_j\overset{j\in K}{\longrightarrow}\mathbf{A}\mathbf{X}$. Since $\boldsymbol{\xi}_j\in\partial g(\mathbf{y}_j)$ and $\{\mathbf{y}_j\}_{j\in K}$ is a compact set (being a convergent sequence), it follows by [4, Theorem 3.16] that $\{\boldsymbol{\xi}_j\}_{j\in K}$ is bounded. Let $\{\boldsymbol{\xi}_j\}_{j\in T}$ ($T\subseteq K$) be a convergent subsequence of $\{\boldsymbol{\xi}_j\}_{j\in K}$ and denote its limit by $\boldsymbol{\xi}$. Then taking the limit $j\overset{j\in T}{\longrightarrow}\infty$ in the relation $\boldsymbol{\xi}_j\in\partial g(\mathbf{y}_j)$, we obtain that $\boldsymbol{\xi}\in\partial g(\mathbf{A}\mathbf{X})$. By Lemma A.1 in appendix A, we have that

$$\mathcal{P}_{T_{\mathbf{X}_j}\mathcal{M}}\left(\nabla f(\mathbf{x}_j)+\boldsymbol{\xi}_j\right)\overset{j\in T}{\longrightarrow}\mathcal{P}_{T_{\mathbf{X}}\mathcal{M}}\left(\nabla f(\mathbf{X})+\boldsymbol{\xi}\right).$$

The above along with (4.19) yields the relation $\mathcal{P}_{T_{\mathbf{X}}\mathcal{M}}\left(\nabla f(\mathbf{X})+\boldsymbol{\xi}\right)=\mathbf{0}$, which combined with the fact that $\boldsymbol{\xi}\in\partial g(\mathbf{A}\mathbf{X})$ implies that $\mathbf{X}$ is a stationary point (Definition 3.1). □

## 4.3 A Projection-Based Smooth Paths Set on the Stiefel Manifold

In this section we consider a projection-based paths set on the Stiefel manifold that obeys the smoothness conditions. We prove that the paths set is indeed smooth and calculate its smoothness constants, as given in Definition 3.3.
Consider the Stiefel manifold

$$\mathcal{M}=\{\mathbf{x}\in\mathbb{R}^{d\times p}\mid\mathbf{x}^T\mathbf{x}=\mathbf{I}_p\},$$

where $d\geq p$ are two fixed natural numbers. Since we deal with a matrix-space, the norm in $\mathbb{R}^{d\times p}$ is the Frobenius norm. A useful fact, proved in [2, Example 3.5.2] (and in many other places) is the following characterization of the tangent space of the Stiefel manifold:

$$T_{\mathbf{x}}\mathcal{M}=\{\mathbf{v}\in\mathbb{R}^{d\times p}\mid\mathbf{v}^T\mathbf{x}+\mathbf{x}^T\mathbf{v}=\mathbf{0}\}. \tag{4.21}$$

In [3, Proposition 5] it is shown that the following is a retraction on $\mathcal{M}$:

$$R_{\mathbf{x}}(\mathbf{v})=\mathcal{P}_{\mathcal{M}}(\mathbf{x}+\mathbf{v}),\ \mathbf{x}\in\mathcal{M},\mathbf{v}\in T_{\mathbf{x}}\mathcal{M}. \tag{4.22}$$

Thus, we are led to consider the following paths set:

$$\gamma_{\mathbf{x},\mathbf{v}}(t)=R_{\mathbf{x}}(t\mathbf{v})=P_{\mathcal{M}}(\mathbf{x}+t\mathbf{v}),\ \mathbf{x}\in\mathcal{M},\mathbf{v}\in T_{\mathbf{x}}\mathcal{M},\|\mathbf{v}\|=1. \tag{4.23}$$

The retraction (4.22) has the following useful representation (see for example [8])

$$R_{\mathbf{x}}(\mathbf{v}) = (\mathbf{x}+\mathbf{v})(\mathbf{I}_p + \mathbf{v}^T\mathbf{v})^{-1/2} \quad \mathbf{x} \in \mathcal{M}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}. \tag{4.24}$$

It is well known that the orthogonal projection of full column rank matrices onto the Stiefel manifold can be expressed in terms of their singular value decomposition[2]. The result is stated in the next lemma.

**Lemma 4.5.** Let $\mathbf{w} \in \mathbb{R}^{d\times p}$ be a matrix of rank $p$. Let $\mathbf{w} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^T$ be the singular value decomposition of $\mathbf{w}$. Then $\mathcal{P}_{\mathcal{M}}(\mathbf{w})$ is single-valued and equals[3] $\mathbf{L}\mathbf{I}_{d\times p}\mathbf{R}^T$.

We can now prove property A of smooth paths set (Definition 3.3) for the paths set given in (4.23).

**Theorem 4.3.** Let $\mathbf{x} \in \mathcal{M}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ such that $\|\mathbf{v}\|_F = 1$. Then for any $t, s \geq 0$,

$$\|\gamma_{\mathbf{x},\mathbf{v}}(t) - \gamma_{\mathbf{x},\mathbf{v}}(s)\| \leq |t-s|.$$

*Proof.* We use the following notation in this proof: the columns of a given matrix $\mathbf{u} \in \mathbb{R}^{d\times p}$ are denoted by $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(p)}$. Denote

$$\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^{d\times p} \mid \mathbf{x}^T\mathbf{x} \preceq \mathbf{I}_p\}.$$

Consider the singular value decomposition $\mathbf{x} + t\mathbf{v} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^T$ ($\mathbf{L} \in \mathbb{R}^{d\times d}, \mathbf{R} \in \mathbb{R}^{p\times p}$ orthogonal and $\mathbf{\Sigma} \in \mathbb{R}^{d\times p}$ diagonal). Then,

$$\mathcal{P}_{\mathcal{H}}(\mathbf{x}+t\mathbf{v}) = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{H}} \|\mathbf{x}+t\mathbf{v}-\mathbf{z}\|_F = \mathbf{L}\left(\operatorname*{argmin}_{\mathbf{w}\in\mathcal{H}} \|\mathbf{\Sigma}-\mathbf{w}\|_F\right)\mathbf{R}^T. \tag{4.25}$$

Note that $(\mathbf{x}+t\mathbf{v})^T(\mathbf{x}+t\mathbf{v}) = \mathbf{x}^T\mathbf{x} + t(\mathbf{x}^T\mathbf{v} + \mathbf{v}^T\mathbf{x}) + t^2\mathbf{v}^T\mathbf{v} = \mathbf{I}_p + t^2\mathbf{v}^T\mathbf{v} \succeq \mathbf{I}_p$, and hence all the singular values of $\mathbf{x}+t\mathbf{v}$ are greater or equal to 1, meaning that the diagonal elements of $\mathbf{\Sigma}$ are greater or equal to 1. Consider the set of $d\times p$ matrices that satisfy that the norm of its columns are at most 1: $\mathcal{R} = \{\mathbf{x} \in \mathbb{R}^{d\times p} : \|\mathbf{x}^{(i)}\|_2 \leq 1 \text{ for all } i = 1, ..., p\}$. Obviously, $\mathcal{H} \subseteq \mathcal{R}$. Consider the problem

$$\min_{\mathbf{w}\in\mathcal{R}} \|\mathbf{\Sigma}-\mathbf{w}\|_F^2 \equiv \min_{\mathbf{w}:\|\mathbf{w}^{(i)}\|_2\leq 1} \sum_{i=1}^{p} \|\Sigma_{i,i}\mathbf{e}_i - \mathbf{w}^{(i)}\|_2^2. \tag{4.26}$$

By the separability of the above problem, the optimal $\mathbf{w}^{(i)}$ is the optimal solution of $\min_{\mathbf{w}^{(i)}:\|\mathbf{w}^{(i)}\|_2\leq 1} \|\Sigma_{i,i}\mathbf{e}_i - \mathbf{w}^{(i)}\|_2$, which is $\mathbf{w}^{(i)} = \mathbf{e}_i$ by the fact that $\Sigma_{ii} \geq 1$. Thus, the optimal solution of (4.26) is $\mathbf{w} = \mathbf{I}_{d\times p}$. Since $\mathbf{I}_{d\times p} \in \mathcal{H}$ and $\mathcal{H} \subseteq \mathcal{R}$, it follows that $\mathbf{I}_{d\times p}$

[2]The result is a slight variation of [3, Proposition 7].

[3]$\mathbf{I}_{d\times p}$ is the $d \times p$ matrix defined by $(\mathbf{I}_{d\times p})_{ij} = 0$ if $i \neq j$ and 1 otherwise.

is an optimal solution of $\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{\Sigma}-\mathbf{w}\|_F^2$. Combining this with (4.25) and Lemma 4.5, we conclude that

$$\mathcal{P}_{\mathcal{H}}(\mathbf{x}+t\mathbf{v}) = \mathbf{L}\mathbf{I}_{d\times p}\mathbf{R}^T = \mathcal{P}_{\mathcal{M}}(\mathbf{x}+t\mathbf{v}).$$

By the non-expansiveness property of the orthogonal projection operator onto convex sets (see (2.3)), we conclude that for any $s,t\geq 0$,

$$\|\mathcal{P}_{\mathcal{M}}(\mathbf{x}+t\mathbf{v})-\mathcal{P}_{\mathcal{M}}(\mathbf{x}+s\mathbf{v})\|_F = \|\mathcal{P}_{\mathcal{H}}(\mathbf{x}+t\mathbf{v})-\mathcal{P}_{\mathcal{H}}(\mathbf{x}+s\mathbf{v})\|_F \leq \|\mathbf{x}+t\mathbf{v}-(\mathbf{x}+s\mathbf{v})\|_F = |s-t|,$$

which is the desired result. □

Property B of smooth paths set is established next.

**Theorem 4.4.** Let $\mathbf{x}\in\mathcal{M}, \mathbf{v}\in T_{\mathbf{x}}\mathcal{M}$ such that $\|\mathbf{v}\|_F = 1$. Then for any $t\geq 0$,

$$\|\gamma'_{\mathbf{x},\mathbf{v}}(t) - \gamma'_{\mathbf{x},\mathbf{v}}(0)\|_F \leq (\sqrt{p}+1)t.$$

*Proof.* By (4.22),

$$\gamma_{\mathbf{x},\mathbf{v}}(t) = (\mathbf{x}+t\mathbf{v})(\mathbf{I}_p + t^2\mathbf{v}^T\mathbf{v})^{-1/2}. \tag{4.27}$$

Denote $S(t) = (\mathbf{I}_p + t^2\mathbf{v}^T\mathbf{v})^{-1/2}$. By calculus of matrix functions we have

$$\begin{aligned}\|\gamma'_{\mathbf{x},\mathbf{v}}(t) - \gamma'_{\mathbf{x},\mathbf{v}}(0)\|_F &= \|(\mathbf{x}+t\mathbf{v})S'(t) + \mathbf{v}S(t) - \mathbf{v}\|_F \\ &\leq \|\mathbf{x}+t\mathbf{v}\|_F \cdot \|S'(t)\|_2 + \|\mathbf{v}\|_F \cdot \|S(t)-\mathbf{I}_p\|_2. \end{aligned}\tag{4.28}$$

Using the equality $\mathbf{x}^T\mathbf{v} + \mathbf{v}^T\mathbf{x} = \mathbf{0}$ we get

$$\|\mathbf{x}+t\mathbf{v}\|_F^2 = \mathrm{Tr}(\mathbf{x}^T\mathbf{x} + t^2\mathbf{v}^T\mathbf{v}) = \|\mathbf{x}\|_F^2 + t^2\|\mathbf{v}\|_F^2 = p + t^2. \tag{4.29}$$

Since $\mathbf{v}^T\mathbf{v}$ is positive semidefinite, it has a spectral decomposition of the form $\mathbf{v}^T\mathbf{v} = \mathbf{P}^T\,\mathrm{diag}(\boldsymbol{\sigma})\mathbf{P}$ where $\mathbf{P}\in\mathbb{R}^{p\times p}$ is orthogonal and $\boldsymbol{\sigma}\in\mathbb{R}^p_+$. Since $\|\mathbf{v}\|_F = 1$ we also have that $\sigma_i \leq 1$ for all $i = 1,2,\ldots,p$. Hence,

$$\|S'(t)\|_2 = \left\|\mathbf{P}^T\,\mathrm{diag}\left(\frac{-t\sigma_i}{(1+t^2\sigma_i)^{3/2}}\right)\mathbf{P}\right\|_2 = \max_{i=1,\ldots,p}\left|\frac{t\sigma_i}{(1+t^2\sigma_i)^{3/2}}\right| = \frac{\alpha t}{(1+\alpha t^2)^{3/2}} \tag{4.30}$$

for some $\alpha\in(0,1]$. Combining (4.29) and (4.30) yields

$$\|\mathbf{x}+t\mathbf{v}\|_F \cdot \|S'(t)\|_2 = \frac{\alpha\sqrt{p+t^2}}{(1+\alpha t^2)^{3/2}}t = \sqrt{\alpha\frac{p+t^2}{\frac{1}{\alpha}+t^2}}\cdot\frac{1}{1+\alpha t^2}t. \tag{4.31}$$

If $p\leq 1/\alpha$, then $\frac{p+t^2}{1/\alpha+t^2}\leq 1$. Otherwise, if $p > 1/\alpha$, then

$$\sqrt{\alpha\frac{p+t^2}{1/\alpha+t^2}} < \sqrt{\alpha\frac{p}{1/\alpha}} = \alpha\sqrt{p} \leq \sqrt{p}, \tag{4.32}$$

where the first inequality follows from simple rearrangement of terms. Plugging (4.32) and the bound $\frac{1}{1+\alpha t^2} \leq 1$ into (4.31) yields

$$\|\mathbf{x} + t\mathbf{v}\|_F \cdot \|S'(t)\|_2 \leq \sqrt{p}t. \tag{4.33}$$

Now, the second term of (4.28) can be bounded as follows:

$$\|S(t)-\mathbf{I}_p\|_2 = \left\|\operatorname{diag}\left(\frac{1}{\sqrt{1+\sigma_i t^2}} - 1\right)\right\|_2 = \max_i \left(1 - \frac{1}{\sqrt{1+\sigma_i t^2}}\right) \leq 1 - \frac{1}{\sqrt{1+t^2}}, \tag{4.34}$$

where the inequality holds since $\sigma_i \in [0,1]$. Note that

$$1 - \frac{1}{\sqrt{1+t^2}} = \frac{(\sqrt{1+t^2}+1)(\sqrt{1+t^2}-1)}{\sqrt{1+t^2}(1+\sqrt{1+t^2})} = \frac{t^2}{\sqrt{1+t^2}+1+t^2} \leq t \cdot \frac{t}{1+t+t^2} \leq t.$$

Plugging this into (4.34) gives

$$\|S(t) - \mathbf{I}_p\|_2 \leq t. \tag{4.35}$$

Plugging (4.35) and (4.33) into (4.28) gives the desired result. □

We summarize Theorems 4.3 and 4.4 in the following result.

**Theorem 4.5.** The paths set defined in (4.23) forms a $(\sqrt{p}+1, 1)$-smooth paths set.

# 5 Numerical Results

## 5.1 SPCA

In this subsection we consider the following known formulation of the sparse principal component analysis (SPCA) problem:

$$\min_{\mathbf{V}\in\mathbb{R}^{d\times p}} \left\{\|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2 + \lambda\|\mathbf{V}\|_{1,1} : \mathbf{V}^T\mathbf{V} = \mathbf{I}_p\right\}, \tag{5.1}$$

where $\|\cdot\|_{1,1}$ denotes the sum of the absolute values of all the entries of the input matrix and $\lambda > 0$ is a trade-off parameter for the regularization factor. $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a data matrix, where $n$ is the number of samples and $d$ is the dimension of each sample. Problem (5.1) fits the general model (1.1) with $f(\mathbf{V}) = \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^T\|_F^2, g(\mathbf{V}) = \lambda\|\mathbf{V}\|_{1,1}, \mathbf{A}$ represents the identity mapping and $\mathcal{M}$ is the Stiefel manifold. Using a simple algebraic expansion, we get that using $f(\mathbf{V}) = -\operatorname{Tr}(\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V})$ will result in the same optimization problem and thus we will use this formulation instead.

We will compare the empirical results of the DSGM method proposed in this paper with the Riemmanian subgradient method on a set of synthesized examples of SPCA problems. The Riemmanian subgradient method for nonconvex objective functions was proposed and analyzed in [25]. The general update step of the method is

$$\mathbf{x}_{k+1} = R_{\mathbf{x}_k}\left(-t_k \mathcal{P}_{T_{\mathbf{x}_k}}(\boldsymbol{\xi}_k)\right),$$

where $R$ is a retraction, $t_k$ is the stepsize on iteration $k$ and $\boldsymbol{\xi}_k$ is some subgradient of the objective function at point $\mathbf{x}_k$. Here, we use the same retraction for the Riemmanian subgradient method as we use for DSGM:

$$R_{\mathbf{x}}(\boldsymbol{\xi}) = \mathcal{P}_{\mathcal{M}}(\mathbf{x} + \boldsymbol{\xi}).$$

This is also the retraction that is used in [25]. Note that the Riemmanian subgradient method only requires the tuning of the stepsize whereas in DSGM, one needs to tune both the stepsize and the smoothing factors $\mu_k$.

The Riemmanian subgradient method is a very natural baseline for our algorithm for a number of reasons. First, as proved in [25], it has convergence results in very broad settings, like those derived for DSGM. Second, both methods are very simple and easy to implement, as well as do not require any complex adjustments when performed on a new problem. In particular, they do not require the solution of complicated optimization problems at each iteration.

To show that the two methods share a similar computational complexity per iteration, we go over the computation process of a single step in each of the methods and find their computational complexity. We consider DSGM first. Denote by $\mathbf{x}_k \in \mathcal{M}$ the $k$th iterate point generated by the algorithm. To find the next iterate $\mathbf{x}_{k+1}$, DSGM performs the following operations:

1. Calculate $\nabla f(\mathbf{x}_k) = -2\mathbf{X}^T\mathbf{X}\mathbf{x}_k$. The computation of the product $\mathbf{X}^T\mathbf{X}$ is done in a preprocess, so we are left with a single multiplication of a $d \times d$ matrix with a $d \times p$ matrix that amounts to $O(p \cdot d^2)$ operations.

2. Compute the descent direction computation $\boldsymbol{\xi}_k = \nabla f(\mathbf{x}_k) + \frac{1}{\mu_k}(\mathbf{x}_k - \text{prox}_{\mu_k g}(\mathbf{x}_k))$. This formula requires $O(d \cdot p)$ operations. Note that the proximal operator of the norm $\|\cdot\|_{1,1}$ is a component-wise soft-thresholding operator (see e.g., [4, Example 6.8]).

3. Find $\mathcal{P}_{T_{\mathbf{x}_k}\mathcal{M}}(\boldsymbol{\xi}_k) = (\mathbf{I}_d - \mathbf{x}_k\mathbf{x}_k^T)\boldsymbol{\xi}_k + \frac{1}{2}\mathbf{x}_k(\mathbf{x}_k^T\boldsymbol{\xi}_k - \boldsymbol{\xi}_k^T\mathbf{x}_k)$. The formula can be found in [2, Example 3.6.2], and its calculation requires $O(d \cdot p^2)$ operations.

4. Compute $\mathbf{x}_{k+1}$ via one retraction operation computation. Based on the presentation in 4.24, it follows the retraction evaluation requires $O(d \cdot p^2)$ operations for the sake of matrix multiplications and $O(p^3)$ operations in order to build the square root and inverse of a $p \times p$ matrix.

For RSG, all the steps are exactly the same, except for step 2, which is replaced by a calculation of a subgradient of the $\ell_1$ norm, which will not change the algorithm's complexity. Assuming that $d > p$, the overall complexity of a single step for both methods is $O(d^2p + p^3)$.

In their work, Li et al. recommend two optional choices for the stepsize $t_i$:

(a) $t_i = \alpha^i$ for some $\alpha \in (0,1)$. We refer to this method as **RSG_exp**.

(b) $t_i = \frac{D}{\sqrt{i}}$ for some $D > 0$. We refer to this method as **RSG_sqrt**.

We compare our results to those obtained by RSG_exp and RSG_sqrt. We used synthetic data that was generated in the following manner:

- Every synthesized case utilized two parameters $\sigma^2$ and $p$ where $p$ is the amount of principal components and $\sigma^2$ is the variance of the principal components. We take the dimension of the data to be $d = 1024$ in all of our experiments.
- We randomly generated a sparse matrix in the Stiefel manifold $\mathbf{W} \in \mathbb{R}^{d\times p}$, $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{p\times p}$, such that eighth of its entries are nonzero. We did it by partitioning $\mathbf{W}$ to block matrices and allowing only one eighth of them to be an orthogonal matrix. The rest of the block matrices are set to zero.
- A data point $\mathbf{x} \in \mathbb{R}^d$ was assumed to be generated using $\mathbf{x} = \mathbf{V}\boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_p)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_d)$.

We used the expectation of the matrix $\mathbf{X}^T\mathbf{X}$, which we denote by $\mathbf{R} = \sigma^2\mathbf{W}\mathbf{W}^T + \mathbf{I}_d$. For each covariance matrix $\mathbf{R}$ we solved the minimization problem

$$\min_{\mathbf{x}}\{-\operatorname{Tr}(\mathbf{x}^T\mathbf{R}\mathbf{x}) + \|\mathbf{x}\|_{1,1} : \mathbf{x}^T\mathbf{x} = \mathbf{I}_p\}. \tag{5.2}$$

We considered a total of 25 values for couples $(p, \sigma^2)$ by going over all the combinations of $p \in \{4, 16, 24, 32, 40\}$ and $\sigma^2 \in \{2, 4, 8, 16, 32\}$. For each couple $(p, \sigma^2)$, we performed the above synthesization procedure 50 times. Since all of the algorithms we use require some parameter tuning, we ran three algorithms in three settings each (overall nine runs) to solve (5.2) for each realization:

- **DSGM** algorithm with smoothing parameters $\frac{0.1}{i^\alpha}$ were $\alpha = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$. The stepsizes where chosen using the backtracking procedure.
- **RSG_sqrt** with stepsizes $\frac{D}{\sqrt{i}}$ were $D = 1, 0.1, 0.01$.
- **RSG_exp** with stepsize $\alpha^i$ where $\alpha = 0.7, 0.8, 0.9$.

All the algorithms were initialized using the same random point on the Stiefel manifold. In all runs we used a code implemented in Python (Github link appears at the end of the paper) with the same stopping criterion - the run stopped after 2 seconds. We chose to stop the run after 2 seconds as we saw that both algorithms reach stagnation at that time frame. Running **RSG** for 2 seconds means running it for between 220 (for larger $p$) to 300 (for smaller $p$) iterations on average. **DSGM** ran for around 20% less iterations in the same time frame.

| | $\sigma^2 = 2$ | $\sigma^2 = 4$ | $\sigma^2 = 8$ | $\sigma^2 = 16$ | $\sigma^2 = 32$ |
|---|---|---|---|---|---|
| $p = 8$ | 68% | 2% | 86% | 0% | 0% |
| $p = 16$ | 68% | 0% | 8% | 52% | 58% |
| $p = 24$ | 96% | 20% | 38% | 100% | 88% |
| $p = 32$ | 94% | 4% | 6% | 100% | 100% |
| $p = 40$ | 84% | 2% | 0% | 100% | 100% |

Table 1: SPCA: The percentage of times that DSGM found the best performing point from all algorithms across all 50 realizations that were generated for each couple $(p, \sigma^2)$.

For each couple of values $(p, \sigma^2)$, we searched for the method (either RSG_exp, RSG_sqrt or DSGM) that resulted in the point with the best objective function value, for each of the 50 runs. RSG_sqrt never found the best point. Table 1 shows the percentage of the simulations for which DSGM (with one of the three used smoothing techniques) found the best performing point (meaning, the point with the smallest objective function value), over all the possible couples $(p, \sigma^2)$. We see that for some parameters DSGM is preferable and for others, RSG is better.

### 5.1.1 Comparison with ManPG

We also compared the DSGM method to the ManPG algorithm that was recently proposed by Chen et. al. in [10]. ManPG can also be used to solve the model (1.1). In ManPG, the descent direction at the $k$th iteration is computed by the formula

$$\mathbf{v}_k := \underset{\mathbf{v} \in T_{\mathbf{x}_k}\mathcal{M}}{\operatorname{argmin}} \left\{ \langle \nabla f(\mathbf{x}_k), \mathbf{v} \rangle + \frac{1}{2t}\|\mathbf{v}\|_F^2 + g(\mathbf{x}_k + \mathbf{v}) \right\}, \tag{5.3}$$

where $t > 0$ is a stepsize. In [10], $\mathcal{M}$ is assumed to be the Stiefel manifold, and show how (5.3) can be solved using a semismooth Newton method.

Using the MATLAB code provided by the authors of [10] for the generation of problems and for their solution using ManPG, we compare the performance of DSGM and ManPG. To make the comparison fair, we solved each problem instance with a code implementing DSGM that was written in MATLAB as well.

For every $d \in \{128, 256, 512, 768, 1024\}$ and $p \in \{4, 8, 16, 32\}$, we randomly generated 50 realizations of the data matrix $\mathbf{X}$ with $n = 50$. For each of these realizations, we ran the adaptive version of the method in [10] (ManPG-Ada) for solving (5.1), as recommended in [10]. We ran DSGM with smoothing parameters $\frac{1}{\sqrt{i}}$ and the stepsizes were chosen using the backtracking procedure. Each of the methods was run for $0.5s$, regardless of the choice of the parameters. Here also we observed that this time limit is enough for the

| | $d = 128$ | $d = 256$ | $d = 512$ | $d = 1024$ |
|---|---|---|---|---|
| $p = 4$ | 26% | 34% | 32% | 18% |
| $p = 8$ | 46% | 48% | 50% | 50% |
| $p = 16$ | 46% | 66% | 92% | 74% |
| $p = 32$ | 98% | 98% | 98% | 100% |

Table 2: SPCA: percentage of times that DSGM found a point that performs better than ManPG-Ada across all 50 realizations that were generated for each couple $(p, d)$.

algorithms to reach stagnation. The percentage of times that DSGM found the better performing point across the realizations of each problem setting is illustrated in 2.

We see that for larger values of $p$, DSGM performs better than ManPG-Ada.

## 5.2 Robust Subspace Recovery

Robust subspace recovery (RSR) rises when one has corrupted data in a high dimensional space. A review on the RSR problem is given in [16]. One of the mathematical formulations of the problem, as described in [16] and [25] is the following:

$$\min_{\mathbf{x}\in\mathbb{R}^{n\times r}} \left\{ \sum_{i=1}^{m} \|\mathbf{y}_i^T\mathbf{x}\|_2 : \mathbf{x}^T\mathbf{x} = \mathbf{I}_r \right\}, \tag{5.4}$$

where $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m \in \mathbb{R}^n$ are the data points and $n - r$ is the dimension of the subspace one tries to recover. The span of the columns of a solution of (5.4) is an approximation of its orthogonal complement. The objective function is convex and Lipschitz continuous and hence (5.4) can be formulated as (1.1) where the smooth part is zero, $g(\mathbf{x}) = \|\mathbf{x}\|_{2,1}$ and $\mathbf{A}$ is the $m \times n$ matrix that has $\mathbf{y}_i^T$ as its $i$-th row. The norm $\|\cdot\|_{2,1}$ stands for the sum of the $\ell_2$ norm of the rows of the input matrix.

Similarly to [25], we simulated data in the following manner. We set the data dimension to be $n = 100$. For each simulated example, we chose randomly a subspace of dimension $n-r$. We randomly generated $I$ points from the unit sphere in the subspace and $5000-I$ points from the unit sphere in $\mathbb{R}^n$. We performed it for $r = 5, 10, 15, 20$ and for $I = 250, 500, 1000, 1500$ which is equivalent to $5\%, 10\%, 20\%, 30\%$ of inlier points. For each simulated problem we ran the DSGM method two times with smoothing parameter $\mu_j = \frac{0.1}{j^\alpha}$ for $\alpha \in \{1/2, 2/3\}$ at iteration $i$. We chose only those parameters as they were the best performing ones in SPCA, which is a similar problem. For RSG_exp we chose the stepsize to be equal to $0.1 \cdot 0.9^j$ in one run and $0.1 \cdot 0.75^j$ for another run. Here we have not used the RSG_sqrt since it was proved in [25] that problem (5.4) is sharp and thus exponential stepsize should have exponential convergence.

| | $r = 5$ | $r = 10$ | $r = 15$ | $r = 20$ |
|---|---|---|---|---|
| $I = 250$ | 96% | 100% | 100% | 100% |
| $I = 500$ | 92% | 100% | 100% | 100% |
| $I = 1000$ | 100% | 100% | 98% | 86% |
| $I = 1500$ | 68% | 48% | 34% | 8% |

Table 3: RSR: Percentage of times that the DSGM algorithm found a point with the best objective value among all the others.

For each couple of values $(I, r)$, we simulated 50 sets of data for 50 different problems. For each simulated problem, we checked which algorithm found the best value. Table 3 shows the percentage of the times that the best value was found by our algorithm, for each $(I, r)$. It can be seen that in the cases that the inlier data points are less significant, the DSGM method outperforms RSG_exp.

Our implementation of the methods and the simulations on python are available here: https://github.com/israelross/DSGM

# References


[1] P.-A Absil, Christopher Baker, and Kyle Gallivan, *Trust-region methods on Riemannian manifolds*, Foundations of Computational Mathematics **7** (2007), 303–330.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, Princeton, NJ, 2008.

[3] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM J. Optim. **22** (2012), no. 1, 135–158. MR 2902688

[4] A. Beck, *First-order methods in optimization*, SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2017.

[5] A. Beck and M. Teboulle, *Smoothing and first order methods: A unified framework*, SIAM Journal on Optimization **22** (2012), no. 2, 557–580.

[6] G. Bento, J. Cruz N., and P. Oliveira, *A new approach to the proximal point method: Convergence on general Riemannian manifolds*, Journal of Optimization Theory and Applications **168** (2016), 743–755.

[7] G. Bento, O. Ferreira, and J. Melo, *Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds*, Journal of Optimization Theory and Applications **173** (2017), 548–562.

[8] N. Boumal, *An introduction to optimization on smooth manifolds*, 2022, Preprint.

[9] R. Boţ and C. Hendrich, *A variable smoothing algorithm for solving convex optimization problems*, Top **23** (2015), 124–150.

[10] S. Chen, S. Ma, A. So, and T. Zhang, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM J. Optim. **30** (2020), 210–239.

[11] O. Ferreira and P. Oliveira, *Proximal point algorithm on Riemannian manifolds*, Optimization **51** (2002), 257–270.

[12] O. P. Ferreira and P. R. Oliveira, *Subgradient algorithm on Riemannian manifolds*, Journal of Optimization Theory and Applications **97** (1998), no. 1, 93–104.

[13] S. Hosseini, W. Huang, and R. Yousefpour, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, SIAM J. Optim. **28** (2016), 596–619.

[14] S. Hosseini and A. Uschmajew, *A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds*, SIAM J. Optim. **27** (2016), 173–189.

[15] W. Huang and K. Wei, *Riemannian proximal gradient methods*, Mathematical Programming (2021), Online First.

[16] G. Lerman and T. Maunu, *An overview of robust subspace recovery*, Proceedings of the IEEE **106** (2018), no. 8, 1380–1410.

[17] J-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France **93** (1965), 273–299.

[18] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming **103** (2005), no. 1, 127–152.

[19] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR 0274683

[20] Jonathan Siegel, *Accelerated optimization with orthogonality constraints*, Journal of Computational Mathematics **39** (2021), 207–226.

[21] M. Soltanalian and P. Stoica, *Designing unimodular codes via quadratic optimization*, IEEE Trans. Signal Process. **62** (2014), no. 5, 1221–1234. MR 3168147

[22] Q. Tran-Dinh, *Adaptive smoothing algorithms for nonsmooth composite convex minimization*, Computational Optimization and Applications **66** (2017), no. 3, 425–451.

[23] I. Waldspurger, A. d'Aspremont, and S. Mallat, *Phase recovery, MaxCut and complex semidefinite programming*, Math. Program. **149** (2015), no. 1-2, Ser. A, 47–81. MR 3300456

[24] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming **142** (2013), no. 1, 397–434.

[25] L. Xiao, C. Shixiang, D. Zengde, Q. Qing, Z. Zhihui, and M. C. S. Anthony, *Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods*, SIAM J. Optim. **31** (2021), 1605–1634.

[26] H. Zhang and S. Sra, *First-order methods for geodesically convex optimization*, 29th Annual Conference on Learning Theory, Proceedings of Machine Learning Research, vol. 49, PMLR, 2016, pp. 1617–1638.

[27] Leihong Zhang, Wei Hong Yang, and Ruyi Song, *Optimality conditions for the nonlinear programming problems on riemannian manifolds*, Pacific Journal of Optimization (2013).

# A A Lemma on Manifolds

Let $\mathcal{M} \subseteq \mathbb{R}^n$ be an embedded submanifold of dimension $d$. For every point $x \in \mathcal{M}$, denote $\phi_x = \mathcal{P}_{T_x\mathcal{M}}$. Using this notation we wish to prove the following lemma.

**Lemma A.1.** The correspondence $\phi : \mathcal{M} \to \mathbb{R}^{n\times n}$ that maps $x \in \mathcal{M}$ to $\phi_x$ is a continuous mapping on $\mathcal{M}$.

First we need to prove the following lemma.

**Lemma A.2.** Let $\mathcal{H} = \{L \in \mathbb{R}^{n\times d} \mid \operatorname{rk} L = d\}$. Define $\pi : \mathcal{H} \to \mathbb{R}^{n\times n}$ by $\pi(L) = \mathcal{P}_{\operatorname{Im} L}$ where $\mathcal{P}_{\operatorname{Im} L}$ is the orthogonal projection on the linear subspace $\operatorname{Im} L$. Then, $\pi$ is a continuous mapping.

*Proof.* Let $L \in \mathcal{H}$. We will show continuity of $\pi$ in an open neighbourhood of $L$. Assume, without loss of generality, that the upper $d \times d$ minor of $L$ is invertible.
Let

$$U = \left\{ R = \left[\begin{array}{c} R_d \\ \hline r_1 \\ \cdots \\ r_{n-d} \end{array}\right] ; R_d \in \mathbb{R}^{d\times d}, \forall i : r_i^T \in \mathbb{R}^d, \det(R_d) \neq 0 \right\}$$

be an open neighbourhood of $L$. For every $R \in U$ and for every $0 < i \leq n - d$, define $v_i(R) \in \mathbb{R}^n$ as the vector that has its $d$ first entries equal to $R_d^{-1} r_i$, its $i + d$ entry equals

$-1$ and the rest of the entries equal to zero. Note that $v_i(R)^T R = 0$ and that the set $\{v_i(R)\}_{i=1}^{n-d}$ is linearly independent for every $R \in U$.
We can now define a mapping $\psi : U \to \mathbb{R}^{n\times n}$ by:

$$\psi(R) = \Big( \; R \;\Big|\; v_1(R) \;\Big|\; \cdots \;\Big|\; v_{n-d}(R) \; \Big).$$

By the construction of $v_i$, this map is continuous. Denote by $\mathcal{S}$ the diagonal matrix with 1 on the first $d$ diagonal elements, and zero on the rest. Observing that for every $R \in U$,

$$\pi(R) = \psi(R)\mathcal{S}\psi(R)^{-1}$$

finishes the proof of the continuity of $\pi$. □

*Proof of Lemma A.1.* Let $x_0 \in \mathcal{M}$ be some point on the manifold. Let $U \subset \mathbb{R}^n, V \subset \mathbb{R}^d$ be open subsets such that there exists a smooth one-to-one map

$$h : V \to \mathbb{R}^n,$$

where $h(V) = U \cap \mathcal{M}$ and such that $x_0 \in U$. Moreover, assume that $\mathrm{rk}\,(Dh(x)) = d$ for every $x \in V$. This map exists because one can compose a the inverse of a chart of $\mathcal{M}$ with the inclusion map $\iota : \mathcal{M} \to \mathbb{R}^n$. Since the inclusion map is an immersion, this composition is smooth and has a differential with full rank in every point. By the definition of the tangent space, for every $x \in V$ we have that $\mathrm{Im}\,(Dh(x)) = T_x\mathcal{M}$.
Using the map

$$\pi : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times n};\ \pi(L) = \mathcal{P}_{ImL}$$

and the fact that

$$\phi\,|_{U\cap\mathcal{M}} = \pi \circ Dh \circ h^{-1},$$

we get that $\phi$ is continuous on a neighborhood of $x_0$. Since this can be shown for every point $x_0 \in \mathcal{M}$, the continuity of $\phi$ on $\mathcal{M}$ follows. □