On Relatively Smooth Optimization over Riemannian Manifolds

Chang He * Jiaxiang Li † Bo Jiang ‡ Shiqian Ma § Shuzhong Zhang ¶

Abstract

We study optimization over Riemannian embedded submanifolds, where the objective function is relatively smooth in the ambient Euclidean space. Such problems have broad applications but are still largely unexplored. We introduce two Riemannian first-order methods, namely the retraction-based and projection-based Riemannian Bregman gradient methods, by incorporating the Bregman distance into the update steps. The retraction-based method can handle nonsmooth optimization; at each iteration, the update direction is generated by solving a convex optimization subproblem constrained to the tangent space. We show that when the reference function is of the quartic form $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, the constraint subproblem admits a closed-form solution. The projection-based approach can be applied to smooth Riemannian optimization, which solves an unconstrained subproblem in the ambient Euclidean space. Both methods are shown to achieve an iteration complexity of $\mathcal{O}(1/\epsilon^2)$ for finding an ϵ-approximate Riemannian stationary point. When the manifold is compact, we further develop stochastic variants and establish a sample complexity of $\mathcal{O}(1/\epsilon^4)$. Numerical experiments on the nonlinear eigenvalue problem and low-rank quadratic sensing problem demonstrate the advantages of the proposed methods.

1 Introduction

In this work, we study the following constrained composite optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & F(x) = f(x) + g(x) \\ \text{s.t.} \quad & x \in \mathcal{M} \subseteq \mathbb{R}^n, \end{aligned} \tag{1.1}$$

where $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and may be nonconvex, and $g : \mathbb{R}^n \to \mathbb{R}$ is a convex, continuous (possibly nonsmooth) function. Here, convexity and smoothness are interpreted as the function is being considered in the ambient Euclidean space. Moreover, the objective function $F(\cdot)$ is bounded below on $\mathcal{M}$, i.e., $F^* = \inf_{x\in\mathcal{M}} F(x) > -\infty$. Such a constrained optimization problem has attracted considerable attention due to its numerous applications, including principal component analysis, low-rank matrix completion, and dictionary learning (Absil et al., 2009; Vandereycken, 2013; Boumal and Absil, 2015; Sun et al., 2015; Cherian and Sra, 2016; Liu et al., 2019; Boumal, 2023; Mishra et al., 2019; Li et al., 2024). By exploiting the geometric structure of the manifold, such as through a suitable retraction,

*School of Information Management and Engineering, Shanghai University of Finance and Economics; Department of Industrial and System Engineering, University of Minnesota. ischanghe@gmail.com

†Department of Electrical and Computer Engineering, University of Minnesota. li003755@umn.edu

‡School of Information Management and Engineering, Shanghai University of Finance and Economics. isyebojiang@gmail.com

§Department of Computational Applied Mathematics and Operations Research, Rice University. shiqian.ma@rice.edu

¶Department of Industrial and System Engineering, University of Minnesota. zhangs@umn.edu

manifold optimization problems can be tackled as unconstrained problems, often resulting in stronger convergence guarantees.

To measure the efficiency of Riemannian optimization methods, one typically considers the iteration complexity, which refers to the number of iterations needed to obtain an approximate solution. When optimizing over a Riemannian embedded submanifold $\mathcal{M}$, most theoretical analyses rely on a Riemannian version of the descent property (e.g., Property A3 in Boumal et al. (2019)), closely analogous to the Euclidean case. Consequently, the same iteration complexity bounds as in their unconstrained Euclidean counterparts typically hold. This descent property is usually established by combining Euclidean gradient Lipschitz continuity with retraction inequalities. However, standard Lipschitz gradient continuity can be sometimes restrictive. Even the simple polynomial $f(x) = x^4$ lacks a Lipschitz continuous gradient, and the widely used log-barrier function in interior-point methods fails to satisfy this condition either (Nesterov and Nemirovskii, 1994; Hinder and Ye, 2024; Jiang et al., 2024). Although it is often possible to argue that the iterates remain within a compact subset, the resulting Lipschitz constant can become too large that the corresponding complexity bound offers limited practical insight. This motivates us to develop a more general framework for Riemannian optimization, extending beyond the standard Lipschitz gradient assumption.

In this paper, we adopt the notion of *relative smoothness*[1], which is defined relative to a reference function (Bauschke et al., 2017; Lu et al., 2018). The formal definition is as follows:

Definition 1.1 (Relative smoothness)**.** *Given a differentiable convex function h, referred to as the reference function, function f is said to be L-smooth relative to h if, for all $x, y \in \mathbb{R}^d$,*

$$f(y) \;\leq\; f(x) + \langle \nabla f(x), y - x \rangle + L \cdot D_h(y, x),$$

where $D_h(y, x)$ is the Bregman distance induced by h, defined as

$$D_h(y, x) \;\triangleq\; h(y) - h(x) - \langle \nabla h(x), y - x \rangle.$$

Clearly, choosing $h(x) = \frac{1}{2}\|x\|^2$ recovers the standard notion of gradient Lipschitz continuity. In certain applications, carefully selecting the reference function h can yield a more accurate local approximation (Bolte et al., 2018). Below, we briefly highlight several representative applications for which the objective function is relatively smooth and the feasible region is a Riemannian manifold.

1.1 Motivating examples

Polynomial optimization over the Stiefel manifold. Constrained polynomial optimization is a widely studied class of problems, capturing applications in signal processing, machine learning, and control (Li et al., 2012). A prototypical instance over the Stiefel manifold is the nonlinear eigenvalue problems, which arises in electronic-structure calculations, such as Kohn–Sham and Hartree–Fock energy-minimization model (Cai et al., 2018; Yang et al., 2009; Liu et al., 2014). In particular, discretising a one-dimensional Kohn–Sham equation leads to

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times p}} \quad & f(X) = \frac{1}{2}\operatorname{Tr}(X^\top L X) + \frac{\beta}{4}\rho_X^\top L^\dagger \rho_X \\ \text{s.t.} \quad & X^\top X = I_p, \end{aligned} \tag{1.2}$$

where $\rho_X \triangleq \operatorname{diag}(XX^\top)$ collects the orbital densities, $L \in \mathbb{R}^{m \times m}$ is the tridiagonal matrix with 2 on the main diagonal and -1 on the first sub- and super-diagonals, $L^\dagger$ denotes its pseudo-inverse, and

[1] Some authors, e.g., Takahashi and Takeda (2024), refer to this property as the "adaptable smoothness property".

$\beta > 0$ is a parameter. Because the Hessian of f grows as a polynomial in $\|X\|$, we can invoke the result in Lu et al. (2018): if $\|\nabla^2 f(X)\| \leq p_r(\|X\|)$ for a univariate polynomial p_r of degree r, then f is relatively smooth with respect to $h(X) = \frac{1}{r+2}\|X\|^{r+2} + \frac{1}{2}\|X\|^2$. Specifically, when $p = 1$, the Stiefel manifold reduces to the sphere, yielding the classical polynomial optimization with sphere constraints.

Low-rank quadratic sensing problem. The quadratic sensing problem is a fundamental optimization problem arising in statistical models (Section 4 in Chi et al. (2018)). It appears in various applications, such as covariance sketching for streaming data; see, e.g., Chen et al. (2015); Cai and Zhang (2015). In this problem, we have access to N measurements of a rank-r matrix $\Sigma = X_* X_*^\top$ with $X_* \in \mathbb{R}^{m \times r}$:

$$c_j = \|X_*^\top y_j\|^2 = y_j^\top \Sigma y_j, j = 1, \dots, N,$$

where $y_j \in \mathbb{R}^m$ are known design vectors. The goal is to reconstruct X_* from these quadratic measures. Mathematically, this task can be formulated as the following optimization problem:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times r}} \quad & f(X) = \frac{1}{2} \sum_{j=1}^{N} \left(\|X^\top y_j\|^2 - c_j \right)^2 \\ \text{s.t.} \quad & \operatorname{rank}(X) = r. \end{aligned} \tag{1.3}$$

Note that f is a fourth-degree polynomial in the entries of X. Consequently, the objective function is relatively smooth with respect to the reference function $h(X) = \frac{1}{4}\|X\|^4 + \frac{1}{2}\|X\|^2$.

Low-rank Poisson matrix completion. Low-rank Poisson matrix completion (Cao and Xie, 2015; McRae and Davenport, 2021) seeks to recover a rank-r matrix $X \in \mathbb{R}^{m \times p}$ from partial nonnegative integer observations $\{Y_{ij}\}_{(i,j) \in \Omega}$. The observations are Poisson counts of the observed matrix entries:

$$Y_{ij} \sim \text{Poisson}\left(X_{ij}\right), \quad (i,j) \in \Omega,$$

where $\Omega \subseteq \{1, \dots, m\} \times \{1, \dots, p\}$ is the set of observed entries. We recover the matrix X via the maximum likelihood formulation; specifically, we minimize the negative log-likelihood:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times p}} \quad & f(X) = \sum_{(i,j) \in \Omega} \left(X_{ij} - Y_{ij} \log\left(X_{ij}\right) \right) \\ \text{s.t.} \quad & \operatorname{rank}(X) = r, \ X_{ij} > 0, \ \forall (i,j) \in \Omega, \end{aligned}$$

which is naturally posed on the embedded submanifold defined as $\mathcal{M} = \{X \in \mathbb{R}^{m \times p} : \operatorname{rank}(X) = r, X_{ij} > 0, \ \forall (i,j) \in \Omega\}$. Notice that the Euclidean Hessian $\nabla^2 f(X)$ has entries Y_{ij}/X_{ij}^2 for $(i,j) \in \Omega$, which blow up as $X_{ij} \to 0^+$, and thus f does not admit a global Lipschitz constant for its gradient. To identify the relative smoothness, we instead choose the logarithmic barrier function $h(X) = -\sum_{(i,j) \in \Omega} \log\left(X_{ij}\right)$ as the reference function. Then, f is relatively smooth with respect to h.

1.2 Related works

Optimization over Riemannian manifolds. For smooth Riemannian optimization problems, i.e., $g \equiv 0$ in (1.1), first-order methods are popular choices, such as Riemannian gradient descent and its variants, e.g., Riemannian conjugate gradient methods (Sato, 2022). When Hessian information is accessible, second-order methods such as Riemannian trust-region method (Absil et al., 2007) and cubic regularized Riemannian Newton method (Agarwal et al., 2021; Zhang and Zhang, 2018) can be applied, offering better convergence performances. When an efficient projection onto the manifold is available, Ding et al. (2024) proposed a projection-based framework specifically tailored for compact matrix manifolds. Additionally, stochastic variants of Riemannian gradient descent have been developed to improve scalability in large-scale settings (Bonnabel, 2013; Han and Gao, 2021; Hosseini and Sra, 2020; Zhang et al., 2016).

For composite problems involving a nonsmooth convex term g, Chen et al. (2020) proposed a Riemannian proximal gradient method over the Stiefel manifold, naturally extending the Euclidean proximal gradient framework to the manifold setting. Huang and Wei (2022, 2023) further generalized this approach to arbitrary Riemannian manifolds and introduced accelerated and inexact variants. Newton-type methods for nonsmooth composite Riemannian optimization problems have also been studied extensively; see, e.g., Grohs and Hosseini (2016), Hu et al. (2024), Si et al. (2024), Wang and Yang (2023), and Wang and Yang (2024). Recently, there have also been works extending Riemannian optimization methods to more complex settings, including minimax, bilevel, and zeroth-order optimization (Han et al., 2023a,b, 2024; He et al., 2024; Li et al., 2023; Li and Ma, 2025; Zhang et al., 2023).

Relatively smooth optimization. The concept of relative smoothness was initially proposed by Bauschke et al. (2017); Lu et al. (2018) in the context of convex optimization, relaxing the standard assumption of global gradient Lipschitz continuity. This notion motivates Bregman-type gradient methods, in which the traditional Euclidean distance is replaced by the Bregman divergence, thereby enabling them to accommodate a broader class of optimization problems. For instance, objective functions arising in the Poisson inverse problem and the D-optimal design problem have been shown to be relatively smooth with respect to the reference function $h(x) = -\sum_{i=1}^{n} \log(x_i)$. More applications can be found in Mukkamala (2022).

Variants of Bregman-type gradient methods have also been well studied. Hanzely et al. (2021); Hanzely and Richtárik (2021) proposed accelerated Bregman proximal gradient methods for relatively smooth convex optimization, and Takahashi and Takeda (2024) developed an inexact version, which employs a new formulation that approximates the Bregman distance, making the subproblem easier to solve. Stochastic Bregman gradient methods have been studied in Fatkhullin and He (2024); Ding et al. (2025); Dragomir et al. (2021), demonstrating that relative smoothness can be applied to deep learning and differentially private optimization. For compact convex constraint sets, Frank–Wolfe methods under relative smoothness have been recently introduced and analyzed theoretically in Takahashi and Takeda (2025); Vyguzov and Stonyakin (2024).

1.3 Main contributions

In this paper, we extend the methodology of relatively smooth minimization to the setting of Riemannian optimization and introduce two Riemannian Bregman gradient methods. In Section 3, we propose the retraction-based Riemannian Bregman gradient method (Algorithm 1) for nonsmooth optimization, which generalizes the update formulation of ManPG (Chen et al., 2020) by employing the Bregman distance. Consequently, at each iteration, we solve a convex optimization subproblem involving the reference function and subject to a tangent space constraint. We show that for the quartic reference function $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, the subproblem admits a closed-form solution (Proposition 3.1). Moreover, when the manifold is a sphere and the reference function is chosen to be either the log-barrier function or the entropy function, the subproblem simplifies to solving a one-dimensional nonlinear equation (Proposition 3.2). By viewing the subproblem as a parametric optimization problem, we prove convergence to a stationary point on smooth, complete Riemannian embedded submanifolds and establish an iteration complexity of $\mathcal{O}(1/\epsilon^2)$ for finding an ϵ-approximate Riemannian stationary point of problem (1.1) (Theorem 3.1).

In Section 4, we develop the projection-based Riemannian Bregman gradient method (Algorithm 2) for smooth Riemannian optimization ($g \equiv 0$ in (1.1)). In this case, the subproblem becomes an easier unconstrained convex optimization problem. After obtaining the update direction, we project directly onto the manifold with an appropriate stepsize. By using projection-related inequalities (Lemmas 4.2 and 4.3), we similarly establish convergence and iteration complexity (Theorem 4.1). Interestingly, we find that the projection-based method generates the same update direction as the retraction-based method when using the quartic reference function $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$ over fixed-rank manifolds (Proposition 4.1). In Section 5, for compact submanifolds $\mathcal{M}$, we further develop corresponding stochastic variants and establish their sample complexity guarantees; see Theorems 5.1 and 5.2, respectively. Numerical results on the nonlinear eigenvalue problem (1.2) and low-rank quadratic sensing problem (1.3) in Section 6 demonstrate the efficiency of our Riemannian Bregman gradient methods.

2 Preliminaries

Throughout this paper, we use lowercase letters (e.g., x, y, z) to denote vectors and uppercase letters (e.g., X, Y, Z) to denote matrices. Unless otherwise specified, we use lowercase symbols in the main text. A function is said to be C^k if it is k-times continuously differentiable; in particular, a C^∞ function is said to be smooth. In this section, we provide a brief introduction to optimization over Riemannian manifolds. For more details, we refer the interested reader to textbooks (Absil et al., 2009; Boumal, 2023). We first introduce the notion of a differentiable submanifold via the implicit-function theorem:

Definition 2.1 (Differentiable submanifolds)**.** *A subset $\mathcal{M} \subseteq \mathbb{R}^n$ is called a d-dimensional C^k embedded submanifold, $k \geq 1$, if for every $x \in \mathcal{M}$ there exist an open neighbourhood $\mathcal{U}_x \subseteq \mathbb{R}^n$ and a C^k map $\phi_x : \mathcal{U}_x \to \mathbb{R}^{n-d}$ such that $\mathcal{U}_x \cap \mathcal{M} = \{y \in \mathcal{U}_x : \phi_x(y) = 0\}$, and $\operatorname{rank}(\mathrm{J}\,\phi_x(y)) = n - d$ for all $y \in \mathcal{U}_x \cap \mathcal{M}$, where $\mathrm{J}\,\phi_x$ denotes the Jacobian matrix of ϕ_x.*

For a submanifold $\mathcal{M} \subseteq \mathbb{R}^n$, the tangent space of $\mathcal{M}$ at point $x \in \mathcal{M}$, denoted by $\mathcal{T}_x\mathcal{M}$, can be characterized as a linear subspace of $\mathbb{R}^n$ given by:

Definition 2.2 (Tangent space)**.** *Given a submanifold $\mathcal{M} \subseteq \mathbb{R}^n$, the tangent space of $\mathcal{M}$ at x is defined as*

$$\mathcal{T}_x\mathcal{M} = \left\{\gamma'(0) : \gamma \text{ is a smooth curve with } \gamma(0) = x,\ \gamma([-\delta, \delta]) \subseteq \mathcal{M},\ \text{for some } \delta > 0\right\}$$

Consequently, the tangent bundle is defined as $\mathcal{TM} = \{(x, \xi) : x \in \mathcal{M}, \xi \in \mathcal{T}_x\mathcal{M}\}$. The normal space of $\mathcal{M}$ at x, denoted by $\mathcal{N}_x\mathcal{M}$, is the orthogonal complement to the tangent space $\mathcal{T}_x\mathcal{M}$.

For example, one commonly encountered submanifold is the Stiefel manifold, defined as $\mathrm{St}(m, p) = \{X \in \mathbb{R}^{m\times p} : X^\top X - I_p = 0\}$, and the tangent space to $\mathrm{St}(m, p)$ at X is $\mathcal{T}_X \mathrm{St}(m, p) = \{V \in \mathbb{R}^{m\times p} : X^\top V + V^\top X = 0\}$. By the implicit function theorem, a useful result is that the tangent space can be expressed in terms of the Jacobian of some equations.

Corollary 2.1. *Let $\mathcal{M}$ be a d-dimensional embedded C^k submanifold of $\mathbb{R}^n$. Given $x \in \mathcal{M}$, for any $y \in \mathcal{U}_x \cap \mathcal{M}$, it holds that $\mathcal{T}_y\mathcal{M} = \ker(\mathrm{J}\,\phi_x(y)) = \{v \in \mathbb{R}^n : \mathrm{J}\,\phi_x(y)v = 0\}$, where $\phi_x(\cdot)$ is defined in Definition 2.1.*

Definition 2.3 (Riemannian (embedded) manifold)**.** *Let $\mathcal{M}$ be a differentiable submanifold of $\mathbb{R}^n$. We say $\mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^n$, if for any $x \in \mathcal{M}$ the tangent space $\mathcal{T}_x\mathcal{M}$ is endowed with a smooth inner product mapping $\langle \cdot, \cdot \rangle_x : \mathcal{T}\mathcal{M} \times \mathcal{T}\mathcal{M} \times \mathcal{M} \to \mathbb{R}$; that is, for any $\eta, \xi \in \mathcal{T}_x\mathcal{M}$, $\langle \xi, \eta \rangle_x$ forms an inner product on $\mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M}$. Denote the induced norm $\|\eta\|_x = \sqrt{\langle \eta, \eta \rangle_x}$ for any $\eta \in \mathcal{T}_x\mathcal{M}$.*

Further, we say $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, if for any $x \in \mathcal{M}$, the tangent space $\mathcal{T}_x\mathcal{M}$ is endowed with the Euclidean inner product; that is, for any $\eta, \xi \in \mathcal{T}_x\mathcal{M}$, $\langle \xi, \eta \rangle_x \triangleq \langle \xi, \eta \rangle$, where the latter is the standard Euclidean inner product. Hence, the norm $\| \cdot \|_x$ is also the same as the standard ℓ_2-norm or the Frobenius norm in the matrix case.

Without loss of generality, we omit the subscript x in the inner product $\langle \cdot, \cdot \rangle$ and the norm $\| \cdot \|$ since our focus is Riemannian embedded submanifolds. For any point $x \in \mathbb{R}^n$ and a nonempty subset $\mathcal{S} \subseteq \mathbb{R}^n$, we denote by $\mathcal{P}_{\mathcal{S}}(x)$ the projection of x onto $\mathcal{S}$ if it exists. We use $\overline{\mathcal{S}}$ to denote the closure of the set $\mathcal{S}$. The open Euclidean ball of radius $r > 0$ centered at $x \in \mathbb{R}^n$ is denoted by $\mathbb{B}(x, r) \triangleq \{y \in \mathbb{R}^n : \|y - x\| < r\}$. In this setting, the Riemannian gradient is defined as the projection of the Euclidean gradient onto the tangent space of the manifold.

Definition 2.4 (Riemannian gradient)**.** *Let f be a continuously differentiable function on $\mathbb{R}^n$. The Riemannian gradient* $\operatorname{grad} f(x)$ *of f with respect to a submanifold $\mathcal{M}$ is a tangent vector in $\mathcal{T}_x\mathcal{M}$ defined by*

$$\operatorname{grad} f(x) = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}(\nabla f(x)),$$

where $\mathcal{P}_{\mathcal{T}_x\mathcal{M}}$ is the orthogonal projection onto the tangent space $\mathcal{T}_x\mathcal{M}$. We say x is a Riemannian stationary point of the differentiable function f if it satisfies $\operatorname{grad} f(x) = 0$.

For a convex function g, its Euclidean subgradient at point x is denoted by $\partial g(x)$. Similarly, the Riemannian subgradient is defined as $\hat{\partial} g(x) = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}(\partial g(x))$. From Chen et al. (2020), the optimality condition for problem (1.1) is given as follows:

Definition 2.5 (Optimality condition)**.** *A point $x \in \mathcal{M}$ is called a Riemannian stationary point of problem* (1.1) *if it satisfies* $0 \in \operatorname{grad} f(x) + \mathcal{P}_{\mathcal{T}_x\mathcal{M}}(\partial g(x))$.

A key ingredient in Riemannian optimization is the notion of a retraction, which is a first-order approximation of the exponential mapping and is often more amenable for computation. Its formal definition is given below.

Definition 2.6 (Retraction)**.** *A retraction on a manifold $\mathcal{M}$ is a smooth mapping Retr from the tangent bundle $\mathcal{T}\mathcal{M}$ to $\mathcal{M}$ with the following properties. Let $\operatorname{Retr}(x, \cdot) : \mathcal{T}_x\mathcal{M} \to \mathcal{M}$ denote the restriction of Retr to $\mathcal{T}_x\mathcal{M}$.*

1. $\operatorname{Retr}(x, 0_x) = x$, *where 0_x is the zero vector in $\mathcal{T}_x\mathcal{M}$;*
2. *The differential of* $\operatorname{Retr}(x, \cdot)$ *at 0_x, i.e.,* $\operatorname{D}\operatorname{Retr}(x, 0_x)$, *is the identity map.*

When the manifold is complete, the domain of the retraction is the entire tangent bundle. By the smoothness of the retraction, for any $(x, v) \in \mathcal{T}\mathcal{M}$, there exist constants $M_1^{\mathrm{R}}(x, v), M_2^{\mathrm{R}}(x, v) \geq 0$ such that

$$\begin{aligned} \|\operatorname{Retr}(x, v) - x\| &\leq M_1^{\mathrm{R}}(x, v)\|v\|, \\ \|\operatorname{Retr}(x, v) - (x + v)\| &\leq M_2^{\mathrm{R}}(x, v)\|v\|^2, \end{aligned} \tag{2.1}$$

where $M_1^{\mathrm{R}}(x, v) = \max_{\xi \in \overline{\mathbb{B}}(x, \|v\|)} \|\operatorname{D}\operatorname{Retr}(x, \xi)\|$, and $M_2^{\mathrm{R}}(x, v) = \max_{\xi \in \overline{\mathbb{B}}(x, \|v\|)} \|\operatorname{D}^2 \operatorname{Retr}(x, \xi)\|$. These inequalities follow directly from Lemma 4 in Boumal et al. (2019). However, these two constants are no longer uniform since we do not restrict our analysis to compact manifolds. Specifically, $M_1^{\mathrm{R}}(x,v)$ and $M_2^{\mathrm{R}}(x,v)$ depend on both the current point x and tangent vector v. We close this section by stating the following assumptions used throughout the paper. These conditions are standard in Riemannian optimization (see, e.g., Chen et al. (2020); Zhang and Sra (2016)).

Assumption 2.1. *In problem (1.1), the Riemannian embedded submanifold $\mathcal{M}$ is C^∞ and complete. The smooth part f is L-smooth relative to a reference function h, where h is continuously differentiable and λ-strongly convex with $\lambda > 0$. The nonsmooth term g is L_g-Lipschitz continuous. The sublevel set of function F at some point $\widetilde{x}$ is compact, i.e., $\mathcal{L}(\widetilde{x}) \triangleq \{x \in \mathcal{M} : F(x) \leq F(\widetilde{x})\}$ is compact.*

3 Retraction-Based Riemannian Bregman Gradient Method

Most optimization methods over Riemannian manifolds share a common update scheme: they first solve a subproblem in the tangent space of the current iterate, which returns a suitable descent direction; then use a retraction along this direction with an appropriate stepsize to obtain the next iterate. This idea is natural and allows Riemannian optimization methods to mimic their Euclidean counterparts. In this section, we develop a Riemannian Bregman gradient method by following this update paradigm.

Given a point $x \in \mathcal{M}$, we use the Bregman distance induced by a reference function h to guide the update direction. Specifically, we solve the following subproblem at point x in the tangent space:

$$v^*(x) \;=\; \operatorname*{argmin}_{v \in \mathcal{T}_x\mathcal{M}} \langle \operatorname{grad} f(x), v\rangle + \gamma D_h(x+v, x) + g(x+v), \tag{3.1}$$

where $\gamma > 0$ can be viewed as the stepsize. If $h(x) = \frac{1}{2}\|x\|^2$, i.e., the Euclidean squared norm, and $\mathcal{M}$ is the Stiefel manifold, then the update rule in (3.1) reduces to the ManPG proposed in Chen et al. (2020). According to Theorem 4.1 in Yang et al. (2014), the first-order optimality condition for the subproblem (3.1) is characterized by

$$0 \;\in\; \operatorname{grad} f(x) + \gamma \cdot \mathcal{P}_{\mathcal{T}_x\mathcal{M}}\left(\nabla h(x + v^*(x)) - \nabla h(x)\right) + \mathcal{P}_{\mathcal{T}_x\mathcal{M}}\left(\partial g(x + v^*(x))\right).$$

Thus, if $v^*(x) = 0$, the condition reduces to $0 \in \operatorname{grad} f(x) + \mathcal{P}_{\mathcal{T}_x\mathcal{M}}(\partial g(x))$, which is exactly the optimality condition of problem (1.1). Hence, the magnitude of direction $v^*(x)$ can be viewed as a stationary measure. Since the subproblem is restricted to the tangent space, we have $\langle \operatorname{grad} f(x), v\rangle = \langle \nabla f(x), v\rangle,\ v \in \mathcal{T}_x\mathcal{M}$, by the definition of the Riemannian gradient. It is therefore unnecessary to compute the Riemannian gradient $\operatorname{grad} f(x)$; instead, the Euclidean gradient $\nabla f(x)$ suffices. After obtaining the direction $v^*(x)$, one can choose a suitable stepsize via backtracking linesearch with a shrinkage parameter. Combining the above components, we summarize our retraction-based method for solving (1.1) in Algorithm 1. During the iterations of this algorithm, we terminate once the norm of the update direction $\|v_t\|$ (Line 3 in Algorithm 1) becomes small. Specifically, we define the ϵ-approximate stationary point as follows.

Definition 3.1. *Given accuracy $\epsilon > 0$, we say x_t is an ϵ-approximate Riemannian stationary point of problem (1.1) whenever $\|v_t\| \leq \epsilon$, where v_t is defined in (3.2).*

Notice that the next iterate x_{t+1} lies on the manifold $\mathcal{M}$, whereas the subproblem (3.2) is approximately solved in the tangent space $\mathcal{T}_{x_t}\mathcal{M}$, this induces an approximation error. Let $x_t^+ \triangleq x_t + \alpha_t v_t$ denote the intermediate point in the tangent space. The following lemma provides an upper bound on the discrepancy between $D_h(x_{t+1}, x_t)$ and $D_h(x_t^+, x_t)$. As the analysis is localized around the iterate x_t, we introduce the radius $r(x,v) \triangleq M_1^{\mathrm{R}}(x,v)\|v\|$, where $M_1^{\mathrm{R}}(x,v)$ is the constant from inequality (2.1).

Algorithm 1: Retraction-Based Riemannian Bregman Gradient Method

Input: initial point $x_0 \in \mathcal{M}$, $\gamma_t \geq L$, $\rho \in (0,1)$

For $t = 0, 1, \ldots$ **do**

Obtain v_t by solving the subproblem

$$v_t \;=\; \operatorname*{argmin}_{v \in \mathcal{T}_{x_t}\mathcal{M}} \langle \nabla f(x_t), v\rangle + \gamma_t D_h(x_t + v, x_t) + g(x_t + v) \tag{3.2}$$

Set the initial stepsize $\alpha_t = 1$

While $F\left(\mathrm{Retr}(x_t, \alpha_t v_t)\right) - F(x_t) \;>\; -\frac{\gamma_t \lambda \alpha_t}{4}\|v_t\|^2$ **do**

$\alpha_t := \rho\alpha_t$

end While

Update $x_{t+1} = \mathrm{Retr}(x_t, \alpha_t v_t)$

Lemma 3.1. *Suppose Assumption 2.1 holds. Let v_t be the solution to (3.2). For any $\alpha_t \in (0,1)$, it holds that $D_h(x_t^+, x_t) \leq \alpha_t D_h(x_t + v_t, x_t)$ and*

$$D_h(x_{t+1}, x_t) - D_h(x_t^+, x_t) \leq 2G_h(x_t, r(x_t, v_t)) M_2^{\mathrm{R}}(x_t, v_t)\|\alpha_t v_t\|^2,$$

where $G_h(x_t, r(x_t, v_t))$ is defined in (3.3).

Proof. From the definition of the Bregman distance and the convexity of h, we obtain

$$\begin{aligned}
D_h(x_t^+, x_t) &= h(x_t + \alpha_t v_t) - h(x_t) - \langle \nabla h(x_t), \alpha_t v_t\rangle \\
&= h\left(\alpha_t(x_t + v_t) + (1-\alpha_t)x_t\right) - h(x_t) - \langle \nabla h(x_t), \alpha_t v_t\rangle \\
&\leq \alpha_t h(x_t + v_t) + (1-\alpha_t)h(x_t) - h(x_t) - \langle \nabla h(x_t), \alpha_t v_t\rangle \\
&= \alpha_t h(x_t + v_t) - \alpha_t h(x_t) - \langle \nabla h(x_t), \alpha_t v_t\rangle \\
&= \alpha_t D_h(x_t + v_t, x_t).
\end{aligned}$$

Next, we bound the error between $D_h(x_{t+1}, x_t)$ and $D_h(x_t^+, x_t)$. By the retraction inequality (2.1), it follows

$$\|x_{t+1} - x_t\| = \|\mathrm{Retr}(x_t, \alpha_t v_t) - x_t\| \leq \alpha_t M_1^{\mathrm{R}}(x_t, v_t)\|v_t\| \leq M_1^{\mathrm{R}}(x_t, v_t)\|v_t\|,$$

which implies that $x_{t+1} \in \overline{\mathbb{B}}(x_t, r(x_t, v_t))$. Let

$$G_h(x, r(x,v)) \triangleq \max_{y \in \overline{\mathbb{B}}(x, r(x,v))} \|\nabla h(y)\| \tag{3.3}$$

denote the maximum gradient norm of the reference function h over the Euclidean ball centered at x with radius $r(x,v)$. Then we have

$$\begin{aligned}
&D_h(x_{t+1}, x_t) - D_h(x_t^+, x_t) \\
&= h(x_{t+1}) - h(x_t^+) - \langle \nabla h(x_t), x_{t+1} - x_t^+\rangle \\
&\leq \langle \nabla h(x_{t+1}), x_{t+1} - x_t^+\rangle + \|\nabla h(x_t)\| \cdot \|x_{t+1} - x_t^+\| \\
&\leq 2G_h(x_t, r(x_t, v_t))\|x_{t+1} - x_t^+\|^2 \\
&\leq 2G_h(x_t, r(x_t, v_t)) M_2^{\mathrm{R}}(x_t, v_t)\|\alpha_t v_t\|^2
\end{aligned}$$

where the last inequality comes from (2.1). □

We now present the per-iteration descent lemma, which ensures a sufficient decrease in the function value for a small enough stepsize.

Lemma 3.2. *Suppose Assumption 2.1 holds. For any $\gamma_t \geq L$, there exists a constant $\alpha'_t > 0$ such that for any $0 < \alpha_t \leq \min\{1, \alpha'_t\}$, the next iterate x_{t+1} in Algorithm 1 satisfies*

$$F(x_{t+1}) - F(x_t) \;\leq\; -\frac{\gamma_t \lambda \alpha_t}{4}\|v_t\|^2,$$

where α'_t is defined in (3.4).

Proof. By the relatively L-smooth of f and $\gamma_t \geq L$, it holds that

$$\begin{aligned}
&f(x_{t+1}) - f(x_t)\\
&\leq \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \gamma_t D_h(x_{t+1}, x_t)\\
&= \langle \nabla f(x_t), x_{t+1} - x_t^+ + x_t^+ - x_t\rangle + \gamma_t D_h(x_t^+, x_t) + \gamma_t D_h(x_{t+1}, x_t) - \gamma_t D_h(x_t^+, x_t)\\
&\leq \langle \nabla f(x_t), x_{t+1} - x_t^+ + x_t^+ - x_t\rangle + \gamma_t \alpha_t D_h(x_t + v_t, x_t) + 2\gamma_t G_h(x_t, r(x_t, v_t)) M_2^{\mathrm{R}}(x_t, v_t)\|\alpha_t v_t\|^2.
\end{aligned}$$

We use Lemma 3.1 in the last inequality. For the inner product term, using the retraction property (2.1) yields

$$\begin{aligned}
&\langle \nabla f(x_t), x_{t+1} - x_t^+ + x_t^+ - x_t\rangle\\
&= \langle \nabla f(x_t), x_{t+1} - x_t^+\rangle + \alpha_t \langle \nabla f(x_t), v_t\rangle\\
&\leq M_2^{\mathrm{R}}(x_t, v_t)\|\nabla f(x_t)\| \cdot \|\alpha_t v_t\|^2 + \alpha_t \langle \nabla f(x_t), v_t\rangle.
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&f(x_{t+1}) - f(x_t)\\
&\leq \alpha_t \left(\langle \nabla f(x_t), v_t\rangle + \gamma_t D_h(x_t + v_t, x_t)\right) + \left(\|\nabla f(x_t)\| + 2\gamma_t G_h(x_t, r(x_t, v_t))\right) M_2^{\mathrm{R}}(x_t, v_t)\|\alpha_t v_t\|^2.
\end{aligned}$$

Since $\mathcal{M}$ is embedded in the Euclidean space $\mathbb{R}^n$, the tangent space $\mathcal{T}_{x_t}\mathcal{M}$ is closed and convex. Due to the optimality condition of constrained optimization, it follows

$$\langle \nabla f(x_t) + \gamma_t \nabla h(x_t + v_t) - \gamma_t \nabla h(x_t) + s_t, v - v_t\rangle \geq 0,\ \forall v \in \mathcal{T}_{x_t}\mathcal{M},$$

where $s_t \in \partial g(x_t + v_t)$. Specifically, choose v to be the zero vector in $\mathcal{T}_x\mathcal{M}$; this yields

$$\langle \nabla f(x_t) - \gamma_t \nabla h(x_t), v_t\rangle \leq \langle \gamma_t \nabla h(x_t + v_t), -v_t\rangle - \langle s_t, v_t\rangle.$$

Hence, we have

$$\begin{aligned}
&\alpha_t \langle \nabla f(x_t), v_t\rangle + \gamma_t D_h(x_t^+, x_t)\\
&\leq \alpha_t \left[\langle \nabla f(x_t), v_t\rangle + \gamma_t D_h(x_t + v_t, x_t)\right]\\
&= \alpha_t \left[\langle \nabla f(x_t), v_t\rangle + \gamma_t h(x_t + v_t) - \gamma_t h(x_t) - \gamma_t \langle \nabla h(x_t), v_t\rangle\right]\\
&= \alpha_t \left[\langle \nabla f(x_t) - \gamma_t \nabla h(x_t), v_t\rangle + \gamma_t h(x_t + v_t) - \gamma_t h(x_t)\right]\\
&\leq \alpha_t \left[\langle \gamma_t \nabla h(x_t + v_t), -v_t\rangle + \gamma_t h(x_t + v_t) - \gamma_t h(x_t)\right] - \alpha_t \langle s_t, v_t\rangle\\
&= -\alpha_t \gamma_t \left[h(x_t) - h(x_t + v_t) - \langle \nabla h(x_t + v_t), -v_t\rangle\right] - \alpha_t \langle s_t, v_t\rangle\\
&\leq -\frac{\alpha_t \gamma_t \lambda}{2}\|v_t\|^2 - \alpha_t \langle s_t, v_t\rangle.
\end{aligned}$$

Therefore, the descent property of smooth part can be established as follows:

$$f(x_{t+1}) - f(x_t) \;\leq\; -\frac{\alpha_t \gamma_t \lambda}{2}\|v_t\|^2 + \left(\|\nabla f(x_t)\| + 2\gamma_t G_h(x_t, r(x_t, v_t))\right) M_2^{\mathrm{R}}(x_t, v_t)\|\alpha_t v_t\|^2 - \alpha_t \langle s_t, v_t\rangle$$

$$= \left((\|\nabla f(x_t)\| + 2\gamma_t G_h(x_t, r(x_t, v_t)))\, M_2^{\mathrm{R}}(x_t, v_t) - \frac{\gamma_t \lambda}{2\alpha_t} \right) \|\alpha_t v_t\|^2 - \alpha_t \langle s_t, v_t \rangle.$$

As for the nonsmooth part g, we have

$$\begin{aligned} g(x_{t+1}) - g(x_t) &= g(x_{t+1}) - g(x_t^+) + g(x_t^+) - g(x_t) \\ &\leq L_g \|x_{t+1} - x_t^+\| + \alpha_t \left(g(x_t + v_t) - g(x_t) \right) \\ &\leq L_g M_2^{\mathrm{R}}(x_t, v_t) \|\alpha_t v_t\|^2 + \alpha_t \langle s_t, v_t \rangle, \end{aligned}$$

where we use the L_g-Lipschitz continuity of g and $g(x_t^+) = g(\alpha_t(x_t + v_t) + (1 - \alpha_t)x_t) \leq \alpha_t g(x_t + v_t) + (1 - \alpha_t) g(x_t)$ in the second inequality, and the last inequality holds due to the convexity of g. Combining the decrease of smooth part and nonsmooth part yields

$$F(x_{t+1}) - F(x_t) \leq \left((\|\nabla f(x_t)\| + 2\gamma_t G_h(x_t, r(x_t, v_t)) + L_g)\, M_2^{\mathrm{R}}(x_t, v_t) - \frac{\gamma_t \lambda}{2\alpha_t} \right) \|\alpha_t v_t\|^2.$$

By setting

$$\alpha_t' \triangleq \frac{\gamma_t \lambda}{4(\|\nabla f(x_t)\| + 2\gamma_t G_h(x_t, r(x_t, v_t)) + L_g) M_2^{\mathrm{R}}(x_t, v_t)}, \tag{3.4}$$

we conclude that for any $0 < \alpha_t \leq \min\{1, \alpha_t'\}$,

$$F(x_{t+1}) - F(x_t) \leq -\frac{\gamma_t \lambda \alpha_t}{4} \|v_t\|^2.$$

Thus, the proof is completed. □

The above lemma ensures that the while loop (Line 5) in Algorithm 1 is well-defined and terminates in a finite number of steps. It also guarantees that $x_{t+1} \in \mathcal{L}(\widetilde{x})$ whenever $x_t \in \mathcal{L}(\widetilde{x})$. However, the above descent lemma is a local result: since we do not assume the manifold $\mathcal{M}$ to be compact, constants such as the stepsize α_t depends on the current iterate x_t and the update direction v_t. If the stepsize sequence $\{\alpha_t\}_{t \geq 0}$ admits a uniform lower bound across iterations, we readily obtain a convergence result of Algorithm 1 from above lemma. To establish such a uniform lower bound, we view the subproblem (3.1) as a parametric optimization problem, and then show that the solution $v^*(x)$ is a continuous function. The proof requires standard concepts from variational analysis, which are provided in the Appendix (see Definitions 7.1 and 7.2).

Lemma 3.3. *Suppose $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is jointly continuous in (x, v), and for each fixed x, the function $v \mapsto \varphi(x, v)$ is λ-strongly convex with $\lambda > 0$. Let $S\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a continuous set-valued map such that, for every x, $S(x)$ is a linear subspace of $\mathbb{R}^n$. Then the minimizer $v^*(x) = \arg\min_{v \in S(x)} \varphi(x, v)$ defines a continuous function $v^* : \mathbb{R}^n \to \mathbb{R}^n$.*

Proof. Since $\varphi(x, \cdot)$ is λ–strongly convex for any fixed x, its restriction to $S(x)$ remains λ–strongly convex and hence coercive. Consequently, the minimum $v^*(x)$ on the subspace $S(x)$ exists and, by strong convexity, is unique. Notice that $\varphi(x, v^*(x)) \leq \varphi(x, 0)$, and $\varphi(x, v^*(x)) \geq \varphi(x, 0) + \langle s_0(x), v^*(x) \rangle + \frac{\lambda}{2}\|v^*(x)\|^2$, where $s_0(x) \in \partial_v \varphi(x, 0)$. Combining these, we get $0 \geq \langle s_0(x), v^*(x) \rangle + \frac{\lambda}{2}\|v^*(x)\|^2 \geq -\|s_0(x)\| \cdot \|v^*(x)\| + \frac{\lambda}{2}\|v^*(x)\|^2$. Suppose $v^*(x) \neq 0$, then we have $\|v^*(x)\| \leq 2\|s_0(x)\|/\lambda$; otherwise $v^*(x) = 0$, $\|v^*(x)\| \leq 2\|s_0(x)\|/\lambda$ stills holds.

Fix x, we claim that $s_0(\cdot)$ is locally bounded around x. Otherwise, there exists a sequence $x_k \to x$ and $s_0(x_k) \in \partial_v \varphi(x_k, 0)$ such that $\|s_0(x_k)\| \to \infty$. Let $\zeta(x_k) \triangleq s_0(x_k)/\|s_0(x_k)\|^2$. Clearly, $\langle \zeta(x_k), s_0(x_k) \rangle = 1$ and $\|\zeta(x_k)\| = 1/\|s_0(x_k)\| \to 0$. By the strong convexity of $\varphi(x_k, \cdot)$, it follows

$\varphi(\zeta(x_k), x_k) \geq \varphi(x_k, 0) + \langle s_0(x_k), \zeta(x_k)\rangle + \frac{\lambda}{2}\|\zeta(x_k)\|^2 \geq \varphi(x_k, 0) + 1$. Let $k \to \infty$. Since φ is jointly continuous, we have $\varphi(x, 0) \geq \varphi(x, 0) + 1$, which is a contradiction.

Choose a sequence $x_k \to x$ and set $v_k \triangleq v^*(x_k)$. By the above, the sequence $\{v_k\}$ is bounded, so it admits a convergent subsequence $v_{k_j} \to v'$. Since $v_{k_j} \in S(x_{k_j})$ and S is outer semicontinuous, it follows that $v' \in S(x)$. For any $w \in S(x)$, by the inner semicontinuous property of S, there exist a sequence $w_j \in S(x_{k_j})$ with $w_j \to w$. By continuity of φ, we have $\varphi(x_{k_j}, v_{k_j}) \to \varphi(x, v')$ and $\varphi(x_{k_j}, w_j) \to \varphi(x, w)$. Since v_{k_j} is the unique minimizer over $S(x_{k_j})$, we have $\varphi(x_{k_j}, v_{k_j}) \leq \varphi(x_{k_j}, w_j)$ for each j, hence $\varphi(x, v') \leq \varphi(x, w)$. Thus, v' is the unique minimizer of $\varphi(x, \cdot)$ over $S(x)$, i.e., $v' = v^*(x)$. Therefore, any convergent subsequence of $\{v_k\}$ converges to $v^*(x)$. Since all such subsequences have the same limit, by contradiction we could conclude that $v_k \to v^*(x)$ as $k \to \infty$, i.e., $v^*(x)$ is continuous in x. □

For the subproblem (3.1), we choose $\varphi(x, v) = \langle \operatorname{grad} f(x), v\rangle + \gamma D_h(x + v, x) + g(x + v)$, $\gamma > 0$. By Corollary 2.1, for any $x \in \mathcal{M}$, there exists a smooth mapping ϕ_x defined on an open neighborhood $\mathcal{U}_x$ of x, satisfying $\operatorname{rank}(\mathrm{J}\,\phi_x(y)) = n - d$, and $\mathcal{T}_y\mathcal{M} = \ker(\mathrm{J}\,\phi_x(y))$ for any $y \in \mathcal{U}_x$. Clearly, the tangent space mapping $\mathcal{T}_y\mathcal{M} = \ker(\mathrm{J}\,\phi_x(y))$ is a continuous set-valued map on $\mathcal{U}_x$. Additionally, φ is jointly continuous in (x, v), and λ-strongly convex for each fixed x. Thus, by applying the lemma above, we conclude that the solution mapping $v^*(x)$ is continuous, which can be used to establish the following Theorem.

Theorem 3.1. *Suppose Assumption 2.1 holds. Set the initial point $x_0 = \widetilde{x}$. Then every limit point of the sequence $\{x_t\}_{t\geq 0}$ generated by Algorithm 1 with $\gamma_t = L$ satisfies the optimality condition of problem (1.1). Moreover, for any given accuracy $\epsilon > 0$, after at most $\mathcal{O}(\epsilon^{-2})$ iterations, Algorithm 1 with $\gamma_t = L$ returns a direction v_t satisfying $\|v_t\| \leq \epsilon$.*

Proof. We first argue that the previously defined constants $M_1^{\mathrm{R}}(x, v)$, $M_2^{\mathrm{R}}(x, v)$, and $G_h(x, r(x, v))$ are continuous in x. Recall that

$$M_1^{\mathrm{R}}(x, v^*(x)) \;=\; \max_{\xi \in \overline{\mathbb{B}}(x, \|v^*(x)\|)} \|\mathrm{D}\,\mathrm{Retr}(x, \xi)\|.$$

Since $v^*(x)$ is continuous, then the set-valued map $\overline{\mathbb{B}}(x, \|v^*(x)\|)$ varies continuously with x. Clearly, for each x, $\overline{\mathbb{B}}(x, \|v^*(x)\|)$ is non-empty and compact. Hence, due to the smoothness of the retraction, it follows from Berge's Maximum Theorem (cf. Theorem 7.1) that $M_1^{\mathrm{R}}(x, v^*(x))$ is continuous in x. By the same argument, $M_2^{\mathrm{R}}(x, v^*(x))$ is also continuous in x. As a consequence, the radius $r(x, v^*(x)) = M_1^{\mathrm{R}}(x, v^*(x))\|v^*(x)\|$ is continuous. By applying Berge's Maximum Theorem again, and noting that $G_h(x, r(x, v^*(x)))$ defined in (3.3) is the maximum value over compact-valued continuous set-valued mapping, we conclude its continuity in x.

Now we proceed to show the convergence. Without loss of generality, assume that $v_t \neq 0$ for all $t \geq 0$; otherwise, x_t is already a stationary point. By Lemma 3.2, we know the Algorithm 1 is monotone, and the iterates $\{x_t\}_{t\geq 0}$ remain within the sublevel set $\mathcal{L}(\widetilde{x})$ when $x_0 = \widetilde{x}$. Since $\mathcal{L}(\widetilde{x})$ is compact, it follows that the sequence $\{v_t\}_{t\geq 0}$ is bounded. As a result, the sequences $\{M_1^{\mathrm{R}}(x_t, v_t)\}_{t\geq 0}$, $\{M_2^{\mathrm{R}}(x_t, v_t)\}_{t\geq 0}$, and $\{G_h(x_t, r(x_t, v_t))\}_{t\geq 0}$ are bounded. Hence, there exist constants $M_2^{\mathrm{R}} > 0$ and $G_h > 0$ such that $M_2^{\mathrm{R}}(x_t, v_t) \leq M_2^{\mathrm{R}}$ and $G_h(x_t, r(x_t, v_t)) \leq G_h$ for all $t \geq 0$. Due to the backtracking linesearch, it holds that $\alpha_t \geq \rho\alpha_t'$. Then we can find a constant $\alpha' > 0$ such that for all $t \geq 0$,

$$\begin{aligned} \alpha_t \;&\geq\; \frac{\rho\gamma_t\lambda}{4\left(\|\nabla f(x_t)\| + 2G_h(x_t, r(x_t, v_t))\gamma_t + L_g\right) M_2^{\mathrm{R}}(x_t, v_t)} \\ &\geq\; \frac{\rho L\lambda}{4\left(G_f + 2G_h L + L_g\right) M_2^{\mathrm{R}}} \;\triangleq\; \alpha', \end{aligned}$$

where G_f is the upper bound for sequence $\{\|\nabla f(x_t)\|\}_{t\geq 0}$. Since F is lower bounded on $\mathcal{M}$, the decrease property established in Lemma 3.2 implies that $\lim_{t\to\infty}\|v_t\| = 0$. Therefore, the sequence $\{x_t\}_{t\geq 0}$ converges to a stationary point of problem (1.1). Moreover, the compactness of $\mathcal{L}(x_0)$ yields that $\{x_t\}_{t\geq 0}$ admits at least one limit point.

Finally, we analyze the iteration complexity. Suppose that Algorithm 1 with $\gamma_t = L$ does not terminate after T iterations; i.e., $\|v_t\| > \epsilon/L$ for all $k = 0, 1, \ldots, T-1$, then it follows

$$F(x_{t+1}) - F(x_t) \;\leq\; -\frac{L\lambda\alpha'}{4}\|v_t\|^2 \;\leq\; -\frac{\lambda\alpha'}{4L}\epsilon^2.$$

Notice that $F^* - F(x_0) \leq F(x_t) - F(x_0) = \sum_{t=0}^{T-1}[F(x_{t+1}) - F(x_t)]$, and we conclude $T = \mathcal{O}(\epsilon^{-2})$. The proof is completed. $\square$

Remark 3.1. *When the reference function is chosen as $h(x) = \frac{1}{2}\|x\|^2$ and $\mathcal{M}$ is the Stiefel manifold, the iteration complexity result in Theorem 3.1 recovers the result established in Chen et al. (2020). Our analysis generalizes their result by incorporating a Bregman distance framework and allowing for general Riemannian embedded submanifolds. Moreover, we only assume that the sublevel set is bounded, which is a mild and commonly used condition in complexity analysis.*

The subproblem (3.2) in Algorithm 1 is a strongly convex minimization problem over a linear subspace. It can be efficiently solved via projected gradient descent or the regularized semismooth Newton method proposed in Chen et al. (2020). In the smooth setting, i.e., $g \equiv 0$ in (1.1), by substituting the definition of the Bregman distance, the subproblem (3.2) reduces to

$$\min_{v\in\mathcal{T}_{x_t}\mathcal{M}} \langle c_t, v\rangle + h(x_t + v), \tag{3.5}$$

where $c_t \triangleq \nabla f(x_t)/\gamma_t - \nabla h(x_t)$. Next, we prove that, with the quartic reference function $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, the update direction v_t admits an explicit closed-form expression on any Riemannian embedded submanifold. Besides, when the manifold is a sphere $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ and the reference function is chosen to be either the log-barrier function or the entropy function, the corresponding subproblem reduces to solving a one-dimensional nonlinear equation, which is similar to its Euclidean counterpart (see Eq. (18) in Lu et al. (2018)). Hence, one can use the bisection method, Newton's method, or any other suitable scalar root-finding method to efficiently compute the solution.

Proposition 3.1. *Suppose the nonsmooth term $g \equiv 0$, and consider the reference function $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$. Then the solution to subproblem (3.2) admits the closed form $v_t = -\theta_t \cdot \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(\nabla f(x_t)/\gamma_t - \nabla h(x_t)) - \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t)$, where $\theta_t > 0$ is the unique positive solution to equation (3.6).*

Proof. First notice that the subproblem solution v_t satisfies $\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t + \nabla h(x_t + v_t)) = 0$ due to the optimality condition of (3.5). When $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$, it implies that

$$\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t) + (\|x_t + v_t\|^2 + 1)\left(\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t) + v_t\right) = 0.$$

If $\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t) = 0$, then $v_t = -\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t)$; otherwise, there exists a constant $\theta_t > 0$ such that $-\theta_t \cdot \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t) = \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t) + v_t$, i.e., $v_t = -\theta_t \cdot \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t) - \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t)$. Hence, we obtain $\theta_t\left(\|x_t - \theta_t \cdot \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t) - \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(x_t)\|^2 + 1\right) = 1$. It follows that θ_t satisfies

$$\|\mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(c_t)\|^2\theta^3 + \left(\|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_t)\|^2 + 1\right)\theta - 1 = 0. \tag{3.6}$$

Clearly, θ_t is the unique positive solution of above equation. By Cardano's formula, θ_t can be expressed in a closed form. $\square$

Proposition 3.2. *Suppose the nonsmooth term $g \equiv 0$, and consider the sphere $\mathcal{M} = \mathbb{S}^{n-1}$. If the reference function is chosen as $h(x) = -\sum_{i=1}^n \log x_i$ or $h(x) = \sum_{i=1}^n x_i \log x_i$, then solving the subproblem (3.2) reduces to solving a one-dimensional nonlinear equation.*

Proof. Recall that the tangent space of the sphere at x_t is $\mathcal{T}_{x_t}\mathbb{S}^{n-1} = \{v \in \mathbb{R}^n : x_t^\top v = 0\}$. Hence the subproblem (3.5) becomes

$$\begin{aligned} \min_{v \in \mathbb{R}^n} \quad & \langle c_t, v \rangle + h(x_t + v) \\ \text{s.t.} \quad & x_t^\top v = 0. \end{aligned}$$

By associating the Lagrange multiplier $\lambda \in \mathbb{R}$, the Lagrangian function is $\mathcal{L}(v, \lambda) = \langle c_t, v \rangle + h(x_t + v) + \lambda\, x_t^\top v$. The KKT conditions are

$$c_t + \nabla h(x_t + v) + \lambda x_t = 0, \quad x_t^\top v = 0.$$

For the log-barrier reference function $h(x) = -\sum_{i=1}^n \log x_i$, the first KKT condition gives $c_t + \lambda x_t = (x_t + v)^{\odot -1}$, where the notation "$\odot -1$" denotes the element-wise inverse. Thus, $v = (c_t + \lambda x_t)^{\odot -1} - x_t$. Substituting into $x_t^\top v = 0$ yields the scalar equation

$$\sum_{i=1}^n \frac{x_{t,i}}{c_{t,i} + \lambda x_{t,i}} - 1 = 0,$$

which is strictly decreasing and therefore has a unique root λ^*. With this root, the subproblem solution is $v_t = (c_t + \lambda^* x_t)^{\odot -1} - x_t$. Similarly, for the entropy reference function $h(x) = \sum_{i=1}^n x_i \log x_i$, the first KKT relation becomes $c_t + \log(x_t + v) + \mathbf{1}_n + \lambda x_t = 0$, where $\exp(\cdot)$ is element-wise exponential, and $\mathbf{1}_n$ is the all-ones vector in $\mathbb{R}^n$. Plugging this into $x_t^\top v = 0$ gives

$$\sum_{i=1}^n x_{t,i} \exp\left(-c_{t,i} - \lambda x_{t,i} - 1\right) - 1 = 0.$$

Again, it is a strictly decreasing function with a unique root λ^*. The corresponding direction is $v_t = \exp\left(-c_t - \lambda^* x_t - \mathbf{1}_n\right) - x_t$. $\square$

4 Projection-Based Riemannian Bregman Gradient Method

As an alternative to the retraction-based approach described above, the classical projection method can also solve problem (1.1) efficiently (Hu et al., 2024; Zhang et al., 2024; Ding et al., 2024). For certain special submanifolds, it is possible to directly project onto the manifold; that is, one can easily compute $\mathcal{P}_{\mathcal{M}}(x + v)$, where $x \in \mathcal{M}$ and $v \in \mathbb{R}^n$. It is well known that for the Stiefel manifold, the projection can be computed via polar decomposition, and similar easily computable projections exist for the Grassmannian and fixed-rank manifold cases (Absil and Malick, 2012; Ding et al., 2024). The main advantage of the projection-based approach is that the update direction can be computed in the full ambient Euclidean space without being restricted to the tangent space. This simplification removes the tangent-space constraint, thus avoiding constrained subproblems such as

(3.2) in Algorithm 1. Consequently, at each iteration, determining the update direction reduces to solving an Euclidean unconstrained optimization problem.

In this section, we develop an efficient projection-based Bregman gradient method for smooth Riemannian optimization problems ($g \equiv 0$ in problem (1.1)). At iteration t, we solve the following unconstrained subproblem:

$$v_t \ = \ \operatorname*{argmin}_{v \in \mathbb{R}^n} \ \langle \operatorname{grad} f(x_t), v \rangle + \gamma_t D_h(x_t + v, x_t). \tag{4.1}$$

By the first-order optimality condition of (4.1), $v_t = 0$ if and only if $\operatorname{grad} f(x_t) = 0$. Hence $\|v_t\|$ also serves as a valid stationarity measure in the projection-based framework, analogous to the retraction-based case.

Definition 4.1. *Given accuracy $\epsilon > 0$, we say x_t is an ϵ-approximate Riemannian stationary point of problem (1.1) with $g \equiv 0$ whenever $\|v_t\| \leq \epsilon$, where v_t is defined in (4.1).*

Since the tangent-space constraint is removed, we decompose v_t into its tangential and normal components: $v_t = v_t^{\mathcal{T}} + v_t^{\mathcal{N}}$, where $v_t^{\mathcal{T}} = \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(v_t)$ and $v_t^{\mathcal{N}} = \mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(v_t)$. As v_t may have a large component in the normal space, we introduce a correction normal vector $u_t \in \mathcal{N}_{x_t}\mathcal{M}$ in the update step. This correction prevents the projection from introducing large deviations due to the normal component. Because u_t is chosen after computing v_t, we control its size via $\|u_t\| \leq \tau \|v_t\|$ for some parameter $\tau \geq 0$. For example, one may simply select $u_t = -v_t^{\mathcal{N}}$ at every iteration, and then $\tau = 1$. Finally, we summarize the projection-based Riemannian Bregman gradient method in Algorithm 2.

Algorithm 2: Projection-Based Riemannian Bregman Gradient Method

---

1: **Input:** initial point $x_0 \in \mathcal{M}$, $\gamma_t \geq L$, $\tau \geq 0$, $\rho \in (0, 1)$
2: **For** $t = 0, 1, \ldots$ **do**
3:     Obtain v_t by solving the subproblem (4.1)
4:     Choose the correction normal vector $u_t \in \mathcal{N}_{x_t}\mathcal{M}$ satisfying $\|u_t\| \leq \tau \|v_t\|$
5:     Set the initial stepsize $\alpha_t = 1$
6:     **While** $F\left(\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t))\right) - F(x_t) \ > \ -\frac{\gamma_t \lambda \alpha_t}{4}\|v_t\|^2$ **do**
7:         $\alpha_t := \rho \alpha_t$
8:     **end While**
9:     Update $x_{t+1} = \mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t))$

Before delving into the theoretical analysis of Algorithm 2, we first provide some properties of the projection operator onto a differentiable submanifold. The following result comes from Lemma 4 in Absil and Malick (2012) and Lemma 5.2 in Ding et al. (2024).

Lemma 4.1. *Let $\mathcal{M} \subseteq \mathbb{R}^n$ be a submanifold of class C^k with $k \geq 2$. Given any $x \in \mathcal{M}$, there exists $\varrho(x) > 0$ such that $\mathcal{P}_{\mathcal{M}}(y)$ uniquely exists for all $y \in \mathbb{B}(x, \varrho(x))$. Moreover, $\mathcal{P}_{\mathcal{M}}(y)$ is of class C^{k-1} for $y \in \mathbb{B}(x, \varrho(x))$, and its differential at x satisfies $\mathrm{D}\mathcal{P}_{\mathcal{M}}(x) = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}$. Additionally, for any $y \in \mathcal{M} \cap \mathbb{B}(x, \varrho(x))$ and $w \in \mathcal{N}_y\mathcal{M}$ satisfying $y + w \in \mathbb{B}(x, \varrho(x))$, we have $\mathcal{P}_{\mathcal{M}}(y + w) = y$.*

However, the above properties hold only locally, as the projection radius depends on the point x. Hence, the projection radius sequence $\{\varrho(x_t)\}_{t \geq 0}$ may converge to zero, where the iterates x_t, $t \geq 0$ are generated by Algorithm 2. To prevent such pathological behavior, we choose the initial point $x_0 = \widetilde{x}$. Therefore, as we will demonstrate later, the iterates generated by Algorithm 2 remain within $\mathcal{L}(\widetilde{x})$ due to the backtracking linesearch. Moreover, for any $x \in \mathcal{L}(\widetilde{x})$, we can derive the following projection inequalities.

Lemma 4.2. *Suppose Assumption 2.1 holds. For any $x \in \mathcal{L}(\widetilde{x})$, there exists a constant $\varrho > 0$ such that for any vectors $v \in \mathcal{T}_x\mathcal{M}$, $u \in \mathcal{N}_x\mathcal{M}$ satisfying $\|v + u\| \le \varrho/2$, we have*

$$\begin{aligned} \|\mathcal{P}_{\mathcal{M}}(x+v+u) - x\| &\le M_1^{\mathcal{P}}\|v\|, \\ \|\mathcal{P}_{\mathcal{M}}(x+v+u) - x - v\| &\le M_2^{\mathcal{P}}\|v\|^2 + M_3^{\mathcal{P}}\|v\|\|u\|, \end{aligned}$$

for some positive constants $M_1^{\mathcal{P}}, M_2^{\mathcal{P}}, M_3^{\mathcal{P}} > 0$.

Proof. First notice that $\mathcal{L}(\widetilde{x}) \subseteq \cup_{z\in\mathcal{L}(\widetilde{x})}\mathbb{B}(z, \varrho(z)/2)$, where $\varrho(z)$ is given in Lemma 4.1. Because $\mathcal{L}(\widetilde{x})$ is compact, by Lebesgue covering lemma, the open cover admits a finite sub-cover: there exist finite points $z_1, \ldots, z_m \in \mathcal{L}(\widetilde{x})$ such that $\mathcal{L}(\widetilde{x}) \subseteq \cup_{i=1}^m \mathbb{B}\left(z_i, \varrho(z_i)/2\right)$. Define $\varrho \triangleq \min_{1\le i\le m} \varrho(z_i)/2 > 0$. Then given $x \in \mathcal{L}(\widetilde{x})$, there exists a sub-cover such that $x \in \mathbb{B}\left(z_i, \varrho(z_i)/2\right)$ for some z_i. Hence, for any $y \in \mathbb{B}(x, \varrho)$, $\|y - z_i\| \le \|y - x\| + \|x - z_i\| \le \varrho + \varrho(z_i)/2 \le \varrho(z_i)$, which says $y \in \mathbb{B}(z_i, \varrho(z_i))$. Consequently, by Lemma 4.1, $\mathcal{P}_{\mathcal{M}}(y)$ uniquely exists and is of class C^∞. Its differential at x satisfies $\mathrm{D}\mathcal{P}_{\mathcal{M}}(x) = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}$. Then both $\mathcal{P}_{\mathcal{M}}(\cdot)$ and $\mathrm{D}\mathcal{P}_{\mathcal{M}}(\cdot)$ are Lipschitz continuous on $\mathbb{B}(x, \varrho)$, with Lipschitz constants $L_{\mathcal{P}_{\mathcal{M}}}(x)$ and $L_{\mathrm{D}\mathcal{P}_{\mathcal{M}}}(x)$, respectively. Besides, for any $w \in \mathcal{N}_x\mathcal{M}$ satisfying $\|w\| \le \varrho$, we also have $\|x + w - z_i\| \le \|w\| + \|x - z_i\| \le \varrho(z_i)$. It follows $\mathcal{P}_{\mathcal{M}}(x + w) = x$.

Now we prove two inequalities. Since $\|v+u\| \le \varrho/2$, then $x+v+u \in \mathbb{B}(x, \varrho)$ and $\mathcal{P}_{\mathcal{M}}(x+u) = x$. It holds that

$$\|\mathcal{P}_{\mathcal{M}}(x+v+u) - x\| \;=\; \|\mathcal{P}_{\mathcal{M}}(x+v+u) - \mathcal{P}_{\mathcal{M}}(x+u)\| \;\le\; L_{\mathcal{P}_{\mathcal{M}}}(x)\|v\| \;\le\; M_1^{\mathcal{P}}\|v\|,$$

which gives the first inequality with $M_1^{\mathcal{P}} \triangleq \max_{x\in\mathcal{L}(\widetilde{x})} L_{\mathcal{P}_{\mathcal{M}}}(x)$. As for the second inequality, consider the first-order Taylor expansion of $\mathcal{P}_{\mathcal{M}}$ at $x + u$. It follows

$$\|\mathcal{P}_{\mathcal{M}}(x+u+v) - \mathcal{P}_{\mathcal{M}}(x+u) - \mathrm{D}\mathcal{P}_{\mathcal{M}}(x+u)[v]\| \;\le\; \frac{L_{\mathrm{D}\mathcal{P}_{\mathcal{M}}}(x)}{2}\|v\|^2.$$

Besides, since $\mathrm{D}\mathcal{P}_{\mathcal{M}}(x) = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}$, we have $\mathrm{D}\mathcal{P}_{\mathcal{M}}(x)[v] = \mathcal{P}_{\mathcal{T}_x\mathcal{M}}(v) = v$. Using the Lipschitz continuity of $\mathrm{D}\mathcal{P}_{\mathcal{M}}$, it yields

$$\|\mathrm{D}\mathcal{P}_{\mathcal{M}}(x+u)[v] - v\| \;=\; \|\mathrm{D}\mathcal{P}_{\mathcal{M}}(x+u)[v] - \mathrm{D}\mathcal{P}_{\mathcal{M}}(x)[v]\| \;\le L_{\mathrm{D}\mathcal{P}_{\mathcal{M}}}(x)\|v\|\|u\|.$$

Recalling that $\mathcal{P}_{\mathcal{M}}(x+u) = x$. Let $M_2^{\mathcal{P}} \triangleq \max_{x\in\mathcal{L}(\widetilde{x})} L_{\mathrm{D}\mathcal{P}_{\mathcal{M}}}(x)/2$, and $M_3^{\mathcal{P}} \triangleq 2M_2^{\mathcal{P}}$. By combining the above two equations, we conclude

$$\|\mathcal{P}_{\mathcal{M}}(x+u+v) - x - v\| \;\le\; M_2^{\mathcal{P}}\|v\|^2 + M_3^{\mathcal{P}}\|v\|\|u\|.$$

The proof is completed. □

Remark 4.1. *The above result follows from Lemma 5.10 in Ding et al. (2024) (We restate it in the Appendix), which was originally proved for compact submanifolds. On a compact submanifold, one can always guarantee that the projection $\mathcal{P}_{\mathcal{M}}(x + v + u)$ remains on the manifold, so only a bound on the normal component is needed. In our setting, however, we work on a compact subset $\mathcal{L}(\widetilde{x})$ of a (potentially non-compact) differentiable submanifold, and $\mathcal{P}_{\mathcal{M}}(x + v + u)$ may not remain in $\mathcal{L}(\widetilde{x})$. Therefore, in order to prove the projection inequalities, we must control the magnitude of both tangent and normal vectors.*

Since the update direction in Algorithm 2 generally contains components in the normal space, it is necessary to control the magnitude of its normal component. The following lemma proves that $\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(x - y)\| = \mathcal{O}(\|x - y\|^2)$.

Lemma 4.3. *Suppose Assumption 2.1 holds. For any $x \in \mathcal{L}(\widetilde{x})$, we have $\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(x-y)\| \leq M_4^{\mathcal{P}}\|x-y\|^2$ for some $M_4^{\mathcal{P}} > 0$.*

Proof. We first argue that there exists a constant $\nu > 0$ such that for any $y \in \overline{\mathbb{B}}(x,\nu) \cap \mathcal{M}$, $\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(x-y)\| = \mathcal{O}(\|x-y\|^2)$ holds. By Definition 2.1 and Corollary 2.1, for any $z \in \mathcal{L}(\widetilde{x})$, there exists an open neighborhood $\mathcal{U}_z$ of z and a C^k map $\phi_z : \mathcal{U}_z \to \mathbb{R}^{n-d}$ such that $\mathcal{U}_z \cap \mathcal{M} = \{y \in \mathcal{U}_z : \phi_z(y) = 0\}$. Besides, for any $y \in \mathcal{U}_z \cap \mathcal{M}$, we have $\operatorname{rank}(\mathrm{J}\,\phi_z(y)) = n-d$, and $\mathcal{T}_y\mathcal{M} = \ker(\mathrm{J}\,\phi_z(y))$. Since $\mathcal{U}_z$ is open, we can choose a ball centered at z with radius $\nu(z) > 0$ such that $\overline{\mathbb{B}}(z,\nu(z)) \subseteq \mathcal{U}_z$. Clearly, $\cup_{z\in\mathcal{L}(\widetilde{x})}\mathbb{B}(z,\nu(z)/2)$ forms an open cover of $\mathcal{L}(\widetilde{x})$. Since $\mathcal{L}(\widetilde{x})$ is compact, then there exists a finite sub-cover such that $\mathcal{L}(\widetilde{x}) \subseteq \cup_{i=1}^m \mathbb{B}(z_i,\nu(z_i)/2)$.

Now we choose $\nu \triangleq \min_{i=1,\ldots,m}\nu(z_i)/2$. Hence for any $x \in \mathcal{L}(\widetilde{x})$ and $y \in \overline{\mathbb{B}}(x,\nu) \cap \mathcal{M}$, there exists some i such that $x, y \in \overline{\mathbb{B}}(z_i,\nu(z_i)) \subseteq \mathcal{U}_{z_i}$. Consequently, $\phi_{z_i}(y) = \phi_{z_i}(x) = 0$, and $\mathcal{T}_x\mathcal{M} = \ker(\mathrm{J}\,\phi_{z_i}(x))$. Then it holds that $\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(y-x) = \mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\,\mathrm{J}\,\phi_{z_i}(x)(y-x)$. Let $L_{\mathrm{J}\,\phi_{z_i}}$ be the Lipschitz constant of $\mathrm{J}\,\phi_{z_i}(\cdot)$ on the closed ball $\overline{\mathbb{B}}(z_i,\nu(z_i))$. We obtain

$$
\begin{aligned}
\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(y-x)\| &= \|\mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\,\mathrm{J}\,\phi_{z_i}(x)(y-x)\| \\
&\leq \|\mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\| \cdot \|\mathrm{J}\,\phi_{z_i}(x)(y-x)\| \\
&= \|\mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\| \cdot \|\phi_{z_i}(y) - \phi_{z_i}(x) - \mathrm{J}\,\phi_{z_i}(x)(y-x)\| \\
&\leq \|\mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\| \cdot \frac{L_{\mathrm{J}\,\phi_{z_i}}}{2}\|y-x\|^2.
\end{aligned}
$$

By choosing $M_{\mathrm{J}\,\phi} \triangleq \max_{i=1,\ldots,m} L_{\mathrm{J}\,\phi_{z_i}} \max_{x\in\overline{\mathbb{B}}(z_i,\nu(z_i))}\|\mathrm{J}\,\phi_{z_i}(x)^\top(\mathrm{J}\,\phi_{z_i}(x)\,\mathrm{J}\,\phi_{z_i}(x)^\top)^{-1}\|$, the inequality $\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(x-y)\| = \mathcal{O}(\|x-y\|^2)$ holds. For those $y \in \mathcal{M}$ satisfying $\|y-x\| > \nu$, it follows

$$
\|\mathcal{P}_{\mathcal{N}_x\mathcal{M}}(y-x)\| \;\leq\; \|x-y\| \;\leq\; \frac{\|x-y\|^2}{\nu}.
$$

Set $M_4^{\mathcal{P}} = \max\{M_{\mathrm{J}\,\phi}, 1/\nu\}$. The proof is completed. □

Compared with Algorithm 1, here we need an additional assumption to control growth of the gradient of the reference function h due to the existence of normal vectors.

Assumption 4.1. *The reference function h is twice continuously differentiable.*

The above requirement is naturally satisfied by many widely-used reference functions, such as $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$ and $h(x) = -\sum_{i=1}^n \log(x_i)$. Now we move to the theoretical analysis of Algorithm 2. In the following analysis, we assume the update direction $v_t \neq 0$ at iteration t; otherwise, x_t is already a stationary point. Suppose the current iterate $x_t \in \mathcal{L}(\widetilde{x})$. By Lemma 4.2, if we choose the stepsize such that $\alpha_t \leq \min\{1, \varrho/(2\|v_t+u_t\|)\}$, we have $\|x_{t+1}-x_t\| = \|\mathcal{P}_{\mathcal{M}}(x_t+\alpha_t(v_t+u_t))-x_t\| \leq \alpha_t M_1^{\mathcal{P}}\|v_t^{\mathcal{T}}\| \leq M_1^{\mathcal{P}}\varrho$. Therefore, we define the maximum gradient norm of the reference function over the ball of radius $M_1^{\mathcal{P}}\varrho$ around x_t as $G_h(x_t, M_1^{\mathcal{P}}\varrho) \triangleq \max_{x\in\overline{\mathbb{B}}(x_t,M_1^{\mathcal{P}}\varrho)}\|\nabla h(x)\|$. Correspondingly, for analysis purposes, we define the maximum Hessian norm of the reference function over the ball $\overline{\mathbb{B}}(x_t, M_1^{\mathcal{P}}\varrho)$ as $H_h(x_t, M_1^{\mathcal{P}}\varrho) \triangleq \max_{x\in\overline{\mathbb{B}}(x_t,M_1^{\mathcal{P}}\varrho)}\|\nabla^2 h(x)\|$. Then, under these assumptions, we can rigorously establish that the deviation between the Bregman distances evaluated at v_t and v_t+u_t in the update step is bounded.

Lemma 4.4. *Suppose Assumptions 2.1 and 4.1 hold. Fix an iterate $x_t \in \mathcal{L}(\widetilde{x})$. For any $\alpha_t \leq \min\{1, \varrho/(2\|v_t+u_t\|)\}$, it holds that*

$$
D_h(x_{t+1}, x_t) - D_h(x_t+\alpha_t v_t, x_t) \;\leq\; \Psi_1(x_t)\|\alpha_t v_t\|^2,
$$

where $\Psi_1(x_t) \triangleq 2G_h(x_t, M_1^{\mathcal{P}}\varrho)(M_2^{\mathcal{P}} + M_3^{\mathcal{P}})(1+\tau) + H_h(x_t, M_1^{\mathcal{P}}\varrho)$.

Proof. First notice that

$$
\begin{aligned}
& D_h(x_{t+1}, x_t) - D_h(x_t + \alpha_t v_t, x_t) \\
= \; & D_h(x_{t+1}, x_t) - D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) + D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) - D_h(x_t + \alpha_t v_t, x_t),
\end{aligned}
$$

where we use the Bregman divergence between x_t and $x_t + \alpha_t v_t^{\mathcal{T}}$ as an intermediate term. On the one hand, due to the convexity of h, we have

$$
\begin{aligned}
& D_h(x_{t+1}, x_t) - D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) \\
= \; & h(x_{t+1}) - h(x_t + \alpha_t v_t^{\mathcal{T}}) - \langle \nabla h(x_t), x_{t+1} - x_t - \alpha_t v_t^{\mathcal{T}} \rangle \\
\leq \; & \langle \nabla h(x_{t+1}) - \nabla h(x_t), x_{t+1} - x_t - \alpha_t v_t^{\mathcal{T}} \rangle \\
\leq \; & \|\nabla h(x_{t+1}) - \nabla h(x_t)\| \cdot \|\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t)) - x_t - \alpha_t v_t^{\mathcal{T}}\| \\
\leq \; & 2G_h(x_t, M_1^{\mathcal{P}} \varrho) \|\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t)) - x_t - \alpha_t v_t^{\mathcal{T}}\|.
\end{aligned}
$$

Since the stepsize α_t satisfies $\alpha_t \leq \varrho/(2\|v_t + u_t\|)$, using the second inequality in Lemma 4.2 implies

$$
\|\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t)) - x_t - \alpha_t v_t^{\mathcal{T}}\| \;\leq\; M_2^{\mathcal{P}} \|\alpha_t v_t^{\mathcal{T}}\|^2 + M_3^{\mathcal{P}} \|\alpha_t v_t^{\mathcal{T}}\| \|\alpha_t(v_t^{\mathcal{N}} + u_t)\|.
$$

Note that $\|v_t^{\mathcal{T}}\| \leq \|v_t\|$, $\|v_t^{\mathcal{N}} + u_t\| \leq \|v_t\| + \|u_t\|$, and $\|u_t\| \leq \tau \|v_t\|$. Hence, we obtain $D_h(x_{t+1}, x_t) - D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) \leq 2G_h(x_t, M_1^{\mathcal{P}} \varrho)(M_2^{\mathcal{P}} + M_3^{\mathcal{P}})(1 + \tau)\|\alpha_t v_t\|^2$. On the other hand, we also have

$$
\begin{aligned}
& D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) - D_h(x_t + \alpha_t v_t, x_t) \\
= \; & h(x_t + \alpha_t v_t^{\mathcal{T}}) - h(x_t + \alpha_t v_t) - \alpha_t \langle \nabla h(x_t), v_t^{\mathcal{T}} - v_t \rangle \\
\leq \; & \alpha_t \langle \nabla h(x_t + \alpha_t v_t^{\mathcal{T}}) - \nabla h(x_t), -v_t^{\mathcal{N}} \rangle \\
\leq \; & \alpha_t \|\nabla h(x_t + \alpha_t v_t^{\mathcal{T}}) - \nabla h(x_t)\| \cdot \|v_t\|.
\end{aligned}
$$

Since h is twice continuously differentiable, Newton-Leibniz formula yields that

$$
\|\nabla h(x_t + \alpha_t v_t^{\mathcal{T}}) - \nabla h(x_t)\| \;=\; \|\int_0^1 \nabla^2 h(x_t + \alpha_t v_t^{\mathcal{T}} \cdot t) \alpha_t v_t^{\mathcal{T}} \, \mathrm{d}\, t\| \;\leq\; H_h(x_t, M_1^{\mathcal{P}} \varrho) \cdot \alpha_t \|v_t^{\mathcal{T}}\|.
$$

Thus we obtain $D_h(x_t + \alpha_t v_t^{\mathcal{T}}, x_t) - D_h(x_t^+, x_t)) \leq H_h(x_t, M_1^{\mathcal{P}} \varrho) \cdot \|\alpha_t v_t\|^2$. Combining the above inequalities yields

$$
\begin{aligned}
& D_h(x_{t+1}, x_t) - D_h(x_t + \alpha_t v_t, x_t) \\
\leq \; & \left(2G_h(x_t, M_1^{\mathcal{P}} \varrho)(M_2^{\mathcal{P}} + M_3^{\mathcal{P}})(1 + \tau) + H_h(x_t, M_1^{\mathcal{P}} \varrho)\right) \|\alpha_t v_t\|^2,
\end{aligned}
$$

which completes the proof. □

Lemma 4.5. *Suppose Assumptions 2.1 and 4.1 hold. Fix an iterate $x_t \in \mathcal{L}(\widetilde{x})$. For any $\gamma_t \geq L$ and $\alpha_t \in (0, 1)$, there exists a constant $\alpha_t' > 0$ such that for any $0 < \alpha_t \leq \alpha_t'$, the next iterate x_{t+1} in Algorithm 2 satisfies*

$$
F(x_{t+1}) - F(x_t) \;\leq\; -\frac{\gamma_t \lambda \alpha_t}{4} \|v_t\|^2,
$$

where α_t' is defined in (4.2).

Proof. By relative smoothness of f, it follows

$$\begin{aligned} f(x_{t+1}) - f(x_t) \ &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \gamma_t D_h(x_t + \alpha_t v_t, x_t) \\ &\quad + \gamma_t D_h(x_{t+1}, x_t) - \gamma_t D_h(x_t + \alpha_t v_t, x_t). \end{aligned}$$

Choosing the stepsize α_t such that $\alpha_t \leq \varrho/(2\|v_t + u_t\|)$, and applying Lemma 4.4, it yields

$$f(x_{t+1}) - f(x_t) \ \leq \ \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \gamma_t D_h(x_t + \alpha_t v_t, x_t) + \gamma_t \Psi_1(x_t) \|\alpha_t v_t\|^2.$$

The inner product term can be upper bounded as

$$\begin{aligned} &\langle \nabla f(x_t), x_{t+1} - x_t \rangle \\ = \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t \rangle + \langle \mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(\nabla f(x_t)), x_{t+1} - x_t \rangle \\ = \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t \rangle + \langle \mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(\nabla f(x_t)), \mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_{t+1} - x_t) \rangle \\ \leq \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t \rangle + \|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(\nabla f(x_t))\| \cdot \|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_{t+1} - x_t)\| \\ \leq \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t \rangle + \|\nabla f(x_t)\| \cdot \|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_{t+1} - x_t)\|. \end{aligned}$$

We now estimate $\|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_{t+1} - x_t)\|$. Recall that $\|x_{t+1} - x_t\| = \|\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t)) - x_t\| \leq \alpha_t M_1^{\mathcal{P}} \|v_t^{\mathcal{T}}\|$ when $\alpha_t \leq \varrho/(2\|v_t + u_t\|)$. By using Lemma 4.3, $\|\mathcal{P}_{\mathcal{N}_{x_t}\mathcal{M}}(x_{t+1} - x_t)\| \leq M_4^{\mathcal{P}} \|x_{t+1} - x_t\|^2 \leq (M_1^{\mathcal{P}})^2 M_4^{\mathcal{P}} \|\alpha_t v_t\|^2$. Let $x_t^+ \triangleq x_t + \alpha_t(v_t + u_t)$. Recall that the correction normal vector $u_t \in \mathcal{N}_{x_t}\mathcal{M}$. Hence, we obtain

$$\begin{aligned} &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t \rangle \\ = \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t^+ \rangle + \langle \operatorname{grad} f(x_t), x_t^+ - x_t \rangle \\ = \ &\langle \operatorname{grad} f(x_t), x_{t+1} - x_t - \alpha_t v_t^{\mathcal{T}} \rangle + \alpha_t \langle \operatorname{grad} f(x_t), v_t \rangle \\ \leq \ &\|\operatorname{grad} f(x_t)\| \cdot \|\mathcal{P}_{\mathcal{M}}(x_t + \alpha_t(v_t + u_t)) - x_t - \alpha_t v_t^{\mathcal{T}}\| + \alpha_t \langle \operatorname{grad} f(x_t), v_t \rangle \\ \leq \ &\|\nabla f(x_t)\| \cdot \left((M_2^{\mathcal{P}} + M_3^{\mathcal{P}})(1 + \tau) \|\alpha_t v_t\|^2 \right) + \alpha_t \langle \operatorname{grad} f(x_t), v_t \rangle, \end{aligned}$$

where we use the same argument in the proof of Lemma 4.4. Let $\Psi_2(x_t) \triangleq (1 + \tau)(M_2^{\mathcal{P}} + M_3^{\mathcal{P}}) \|\nabla f(x_t)\| + (M_1^{\mathcal{P}})^2 M_4^{\mathcal{P}} \|\nabla f(x_t)\|$. By Combining the above inequalities, we have

$$f(x_{t+1}) - f(x_t) \ \leq \ (\gamma_t \Psi_1(x_t) + \Psi_2(x_t)) \|\alpha_t v_t\|^2 + \alpha_t \langle \operatorname{grad} f(x_t), v_t \rangle + \gamma_t D_h(x_t + \alpha_t v_t, x_t).$$

To proceed, we use the first-order optimality condition of subproblem (4.1), which implies $\operatorname{grad} f(x_t) + \gamma_t \nabla h(x_t + v_t) - \gamma_t \nabla h(x_t) = 0$. From Lemma 3.1, we also have $D_h(x_t + \alpha_t v_t, x_t) \leq \alpha_t D_h(x_t + v_t, x_t)$ for any $\alpha_t \in (0, 1)$. We now obtain

$$\begin{aligned} &\alpha_t \langle \operatorname{grad} f(x_t), v_t \rangle + \gamma_t D_h(x_t + \alpha_t v_t, x_t) \\ = \ &\alpha_t \gamma_t \langle \nabla h(x_t) - \nabla h(x_t + v_t), v_t \rangle + \alpha_t \gamma_t D_h(x_t + v_t, x_t) \\ = \ &\alpha_t \gamma_t \left(h(x_t + v_t) - h(x_t) - \langle \nabla h(x_t + v_t), v_t \rangle \right) \\ \leq \ &- \frac{\alpha_t \gamma_t \lambda}{2} \|v_t\|^2, \end{aligned}$$

where the last inequality follows from the λ-strong convexity of h. Therefore, we obtain

$$f(x_{t+1}) - f(x_t) \ \leq \ (\gamma_t \Psi_1(x_t) + \Psi_2(x_t)) \|\alpha_t v_t\|^2 - \frac{\alpha_t \gamma_t \lambda}{2} \|v_t\|^2.$$

By setting

$$\alpha_t' = \min\left\{\frac{\varrho}{2\|v_t + u_t\|}, \frac{\gamma_t\lambda}{4\left(\gamma_t\Psi_1(x_t) + \Psi_2(x_t)\right)}\right\}, \tag{4.2}$$

we conclude that for any $0 < \alpha_t \leq \min\{1, \alpha_t'\}$,

$$F(x_{t+1}) - F(x_t) \leq -\frac{\gamma_t\lambda\alpha_t}{4}\|v_t\|^2.$$

The proof is completed. □

Similar to the retraction-based approach, we establish a per-iteration descent lemma for Algorithm 2, which ensures that the backtracking line-search procedure is well-defined. Moreover, it follows by induction that the entire sequence of iterates $\{x_t\}_{t\geq 0}$ remains within $\mathcal{L}(\widetilde{x})$ whenever $x_0 \in \mathcal{L}(\widetilde{x})$. Consequently, α_t' in the above lemma admits a strictly positive lower bound, which further implies the following convergence result.

Theorem 4.1. *Suppose Assumptions 2.1 and 4.1 hold. Set the initial point $x_0 = \widetilde{x}$. Then every limit point of the sequence $\{x_t\}_{t\geq 0}$ generated by Algorithm 2 with $\gamma_t = L$ satisfies the optimality condition of problem (1.1). Moreover, for any given accuracy $\epsilon > 0$, after at most $\mathcal{O}(\epsilon^{-2})$ iterations, Algorithm 2 with $\gamma_t = L$ returns a direction v_t satisfying $\|v_t\| \leq \epsilon$.*

Proof. By Lemma 3.3, we know that the minimizer of subproblem 4.1 is continuous with respect to x. Since all iterates belong to $\mathcal{L}(\widetilde{x})$, which is compact, then the sequence $\{v_t\}_{t\geq 0}$ is bounded by a uniform constant $\overline{v}$. Using an argument similar to the one used in the proof of Theorem 3.1, it can be shown that the sequences $\{\Psi_1(x_t)\}_{t\geq 0}$ and $\{\Psi_2(x_t)\}_{t\geq 0}$ are bounded. Hence, there exist constants $\Psi_1 > 0$ and $\Psi_2 > 0$ such that $\Psi_1(x_t) \leq \Psi_1$ and $\Psi_2(x_t) \leq \Psi_2$ for all $t \geq 0$. Due to the backtracking linesearch, it holds that $\alpha_t \geq \rho\alpha_t'$. Then we can find a constant $\alpha' > 0$ such that for all $t \geq 0$,

$$\begin{aligned}\alpha_t &\geq \min\left\{\frac{\rho\varrho}{2\|v_t + u_t\|}, \frac{\rho\gamma_t\lambda}{4\left(\gamma_t\Psi_1(x_t) + \Psi_2(x_t)\right)}\right\} \\ &\geq \min\left\{\frac{\rho\varrho}{2(1+\tau)\overline{v}}, \frac{\rho L\lambda}{4\left(L\Psi_1 + \Psi_2\right)}\right\} \triangleq \alpha'.\end{aligned}$$

By a similar argument in the proof of Theorem 3.1, we can conclude the sequence $\{x_t\}_{t\geq 0}$ converges to a stationary point of problem (1.1), and $\{x_t\}_{t\geq 0}$ admits at least one limit point. Besides, after at most $\mathcal{O}(\epsilon^{-2})$ iterations, the algorithm returns a direction v_t satisfying $\|v_t\| \leq \epsilon$. □

At the end of this section, we prove that for the fixed-rank manifold, the update direction generated by (4.1) coincides with the direction obtained in the retraction-based case when the reference function is chosen to be $h(X) = \frac{1}{4}\|X\|^4 + \frac{1}{2}\|X\|^2$.

Proposition 4.1. *Suppose the nonsmooth term $g \equiv 0$, and consider the fixed-rank manifold $\mathcal{M}_r = \{X \in \mathbb{R}^{m\times p} : \mathrm{rank}(X) = r\}$, with $0 < r \leq \min\{m, p\}$. If the reference function is chosen as $h(X) = \frac{1}{4}\|X\|^4 + \frac{1}{2}\|X\|^2$, then the direction V_t generated by (4.1) is also the solution to (3.2).*

Proof. For clarity, let $V_t^{\mathcal{P}}$ denote the solution to (4.1), and let V_t^{R} denote the solution to (3.2). Define $Y_t \triangleq X_t + V_t^{\mathcal{P}}$. Then Y_t is the solution to the problem

$$Y_t = \operatorname*{argmin}_{Y\in\mathbb{R}^{m\times p}} \langle \mathrm{grad}\, f(X_t)/\gamma_t, Y\rangle + D_h(Y, X_t).$$

Let $C_t' \triangleq \operatorname{grad} f(X_t)/\gamma_t - \nabla h(X_t)$. By an argument similar to that in the proof of Proposition 3.1, we have $Y_t = -\theta_t C_t'$, where θ_t is the unique positive solution to the equation $\|C_t'\|^2\theta^3 + \theta - 1 = 0$. Thus, $V_t^{\mathcal{P}} = -\theta_t C_t' - X_t$. Let $X_t \triangleq U_t\Sigma_t V_t^\top$ be the SVD decomposition. By Proposition 2.1 in Vandereycken (2012), the tangent space at X_t is

$$\mathcal{T}_{X_t}\mathcal{M}_r = \{U_t M V_t^\top + U V_t^\top + U_t V^\top : M \in \mathbb{R}^{r\times r}, U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{p\times r}\}.$$

Clearly, $X_t \in \mathcal{T}_{X_t}\mathcal{M}_r$ when $M = \Sigma_t$, $U = 0$, and $V = 0$. Hence, $\mathcal{P}_{\mathcal{N}_{X_t}\mathcal{M}_r}(X_t) = 0$. Moreover, for the chosen reference function $h(X) = \frac{1}{4}\|X\|^4 + \frac{1}{2}\|X\|^2$, we have $\nabla h(X_t) = (\|X_t\|^2 + 1)X_t$. Notice that $\nabla h(X_t) = \mathcal{P}_{\mathcal{T}_{X_t}\mathcal{M}_r}(\nabla h(X_t))$, which implies that $C_t' = \mathcal{P}_{\mathcal{T}_{X_t}\mathcal{M}_r}(C_t)$, where C_t is defined in (3.5). Consequently, equation (3.6) reduces to $\|C_t'\|^2\theta^3 + \theta - 1 = 0$, and thus, θ_t is precisely the positive solution of (3.6). Therefore, by proposition 3.1, $V_t^{\mathrm{R}} = -\theta_t \cdot \mathcal{P}_{\mathcal{T}_{X_t}\mathcal{M}_r}(\nabla f(X_t)/\gamma_t - \nabla h(X_t)) - \mathcal{P}_{\mathcal{T}_{X_t}\mathcal{M}_r}(X_t) = -\theta_t C_t' - X_t = V_t^{\mathcal{P}}$, completing the proof. □

5 Extension to the Stochastic Setting

In this section, we show that both retraction-based and projection-based Bregman gradient methods can be extended to solve the Riemannian stochastic optimization settings. Specifically, we consider the composite optimization problem:

$$\min_{x\in\mathcal{M}} F(x) = f(x) + g(x), \text{ with } f(x) \triangleq \mathbb{E}_\pi[f(x,\pi)] \tag{5.1}$$

where $\mathbb{E}_\pi$ is the expectation with respect to the random variable π. We assume the access to the stochastic first-order oracle that returns stochastic gradients $\nabla f(x,\pi)$, which are unbiased estimators of the true gradient with bounded variance. Specifically, for all $x \in \mathbb{R}^n$, we have $\mathbb{E}_\pi[\nabla f(x,\pi)] = \nabla f(x)$, and $\mathbb{E}_\pi\left[\|\nabla f(x,\pi) - \nabla f(x)\|^2\right] \le \sigma^2$.

For the retraction-based approach, at each iteration, we replace the full gradient in the update step (3.2) of Algorithm 1 with a stochastic estimator of the Euclidean gradient. Specifically, we randomly sample a mini-batch $\mathcal{B}_t$ and define $\nabla f_{\mathcal{B}_t}(x_t) \triangleq \frac{1}{|\mathcal{B}_t|}\sum_{j\in\mathcal{B}_t}\nabla f(x_t, \pi_t^{(j)})$, where $\{\pi_t^{(j)}\}_{j\in\mathcal{B}_t}$ are i.i.d. samples drawn from the underlying distribution. The corresponding mini-batch Riemannian gradient follows $\operatorname{grad} f_{\mathcal{B}_t}(x_t) = \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(\nabla f_{\mathcal{B}_t}(x_t))$. We then solve the subproblem (5.2) using this stochastic gradient and update the iterate accordingly. The resulting procedure is summarized in Algorithm 3. Similarly, for smooth Riemannian optimization problems, ($g \equiv 0$ in problem (5.1)), the projection-based Riemannian Bregman gradient method can be generalized in a similar manner. This stochastic variant is presented in Algorithm 4. In the stochastic case, we set $u_t = 0$ for simplicity.

Algorithm 3: Stochastic Retraction-Based Riemannian Bregman Gradient Method

---

1: **Input:** initial point $x_0 \in \mathcal{M}$, $\gamma_t \ge L$, $\alpha_t > 0$

2: **For** $t = 0, 1, \ldots$ **do**

3: Obtain update direction ζ_t^{R} by solving the subproblem

$$\zeta_t^{\mathrm{R}} = \operatorname*{argmin}_{\zeta\in\mathcal{T}_{x_t}\mathcal{M}} \langle \nabla f_{\mathcal{B}_t}(x_t), \zeta\rangle + \gamma_t D_h(x_t + \zeta, x_t) + g(x_t + \zeta) \tag{5.2}$$

4: Update $x_{t+1} = \mathrm{Retr}(x_t, \alpha_t \zeta_t^{\mathrm{R}})$

Algorithm 4: Stochastic Projection-Based Riemannian Bregman Gradient Method

1: **Input:** initial point $x_0 \in \mathcal{M}$, $\gamma_t \geq L$, $\alpha_t > 0$

2: **For** $t = 0, 1, \ldots$ **do**

3: Obtain update direction $\zeta_t^{\mathcal{P}}$ by solving the subproblem

$$\zeta_t^{\mathcal{P}} = \underset{\zeta \in \mathbb{R}^n}{\operatorname{argmin}} \ \langle \operatorname{grad} f_{\mathcal{B}_t}(x_t), \zeta \rangle + \gamma_t D_h(x_t + \zeta, x_t) \tag{5.3}$$

4: Update $x_{t+1} = \mathcal{P}_{\mathcal{M}}(x_t + \alpha_t \zeta_t^{\mathcal{P}})$

To distinguish the update directions in the stochastic methods from those in the deterministic setting, we denote them by ζ_t^{R} and $\zeta_t^{\mathcal{P}}$, corresponding to the retraction-based and projection-based methods, respectively. In the deterministic setting, we denote the update directions by v_t^{R} and $v_t^{\mathcal{P}}$. Recall that in Theorems 3.1 and 4.1, the norms $\|v_t^{\mathrm{R}}\|$ and $\|v_t^{\mathcal{P}}\|$ are used as measures of approximate stationarity, that is, x is an ϵ-approximate Riemannian stationary point if $\|v_t^{\mathrm{R}}\| \leq \epsilon$ or $\|v_t^{\mathcal{P}}\| \leq \epsilon$. However, in the stochastic setting, these vectors cannot be directly computed because the true gradient $\nabla f(x)$ is not accessible. The following lemma establishes a relationship between $\|v_t^{\mathrm{R}}\|$, $\|v_t^{\mathcal{P}}\|$and $\|\zeta_t^{\mathrm{R}}\|$, $\|\zeta_t^{\mathcal{P}}\|$. This allows v_t^{R} and $v_t^{\mathcal{P}}$ to remain valid theoretical measures of stationarity, even though they cannot be evaluated.

Lemma 5.1. *Suppose Assumption 2.1 holds. Then it holds that*

$$\begin{aligned} \|v_t^{\mathrm{R}}\|^2 &\leq 2\mathbb{E}\left[\|\zeta_t^{\mathrm{R}}\|^2 \mid x_t\right] + 2\sigma^2/(\gamma_t^2 \lambda^2 |\mathcal{B}_t|), \\ \|v_t^{\mathcal{P}}\|^2 &\leq 2\mathbb{E}\left[\|\zeta_t^{\mathcal{P}}\|^2 \mid x_t\right] + 2\sigma^2/(\gamma_t^2 \lambda^2 |\mathcal{B}_t|), \end{aligned}$$

where the expectation is taken with respect to $\pi_t^{(1)}, \ldots, \pi_t^{(\mathcal{B}_t)}$.

Proof. To prove the first inequality, we first observe that

$$\|v_t^{\mathrm{R}}\|^2 \leq 2\|\zeta_t^{\mathrm{R}}\|^2 + 2\|v_t^{\mathrm{R}} - \zeta_t^{\mathrm{R}}\|^2. \tag{5.4}$$

It therefore suffices to bound the term $\|v_t^{\mathrm{R}} - \zeta_t^{\mathrm{R}}\|$. Recall $v_t^{\mathrm{R}}, \zeta_t^{\mathrm{R}}$ satisfy

$$\begin{aligned} v_t^{\mathrm{R}} &= \underset{v \in \mathcal{T}_{x_t}\mathcal{M}}{\operatorname{argmin}} \ \langle \nabla f(x_t), v \rangle + \gamma_t D_h(x_t + v, x_t) + g(x_t + v), \\ \zeta_t^{\mathrm{R}} &= \underset{\zeta \in \mathcal{T}_{x_t}\mathcal{M}}{\operatorname{argmin}} \ \langle \nabla f_{\mathcal{B}_t}(x_t), \zeta \rangle + \gamma_t D_h(x_t + \zeta, x_t) + g(x_t + \zeta). \end{aligned}$$

From the optimality condition of constrained optimization, there exist $s_t \in \partial g(x_t + v_t^{\mathrm{R}})$ and $s_t' \in \partial g(x_t + \zeta_t^{\mathrm{R}})$ such that

$$\langle \nabla f(x_t) + \gamma_t \nabla h(x_t + v_t^{\mathrm{R}}) - \gamma_t \nabla h(x_t) + s_t, v - v_t^{\mathrm{R}} \rangle \geq 0, \ \forall v \in \mathcal{T}_{x_t}\mathcal{M}, \tag{5.5a}$$

$$\langle \nabla f_{\mathcal{B}_t}(x_t) + \gamma_t \nabla h(x_t + \zeta_t^{\mathrm{R}}) - \gamma_t \nabla h(x_t) + s_t', \zeta - \zeta_t^{\mathrm{R}} \rangle \geq 0, \ \forall \zeta \in \mathcal{T}_{x_t}\mathcal{M}. \tag{5.5b}$$

By summing over the above two inequalities with $v = \zeta_t^{\mathrm{R}}$ in equation (5.5a) and $\zeta = v_t^{\mathrm{R}}$ in equation (5.5b), we obtain

$$\langle \nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}} \rangle \geq \gamma_t \langle \nabla h(x_t + \zeta_t^{\mathrm{R}}) - \nabla h(x_t + v_t^{\mathrm{R}}), \zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}} \rangle + \langle s_t' - s_t, \zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}} \rangle.$$

Due to the strong convexity of the reference function h, we have $\langle \nabla h(x_t + \zeta_t^{\mathrm{R}}) - \nabla h(x_t + v_t^{\mathrm{R}}), \zeta_t^{\mathrm{R}} - v_t\rangle \geq \lambda\|\zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}}\|^2$. Also, $\langle s_t' - s_t, \zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}}\rangle \geq 0$ since g is convex. We conclude $\|\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t)\| \cdot \|\zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}}\| \geq \gamma_t\lambda\|\zeta_t^{\mathrm{R}} - v_t^{\mathrm{R}}\|^2$. Substituting the above inequality into equation (5.4) yields

$$\|v_t^{\mathrm{R}}\|^2 \;\leq\; 2\|\zeta_t^{\mathrm{R}}\|^2 + 2\frac{\|\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t)\|^2}{\gamma_t^2\lambda^2}.$$

By the batching property, it follows $\mathbb{E}\left[\|\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t)\|^2 \mid x_t\right] \leq \sigma^2/|\mathcal{B}_t|$, where the expectation is taken over the mini-batch samples $\pi_t^{(1)}, \ldots, \pi_t^{(\mathcal{B}_t)}$. As a consequence, we obtain

$$\|v_t^{\mathrm{R}}\|^2 \;\leq\; 2\mathbb{E}\left[\|\zeta_t^{\mathrm{R}}\|^2 \mid x_t\right] + \frac{2\sigma^2}{\gamma_t^2\lambda^2|\mathcal{B}_t|}.$$

As for the second one, we can use a similar argument. By combining the optimality conditions of $\|v_t^{\mathcal{P}}\|$ and $\|\xi_t^{\mathcal{P}}\|$, we have $\|\operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t)\| = \gamma_t\|\nabla h(x_t + v_t^{\mathcal{P}}) - \nabla h(x_t + \zeta_t^{\mathcal{P}})\|$. Again, using the strong convexity of h yields $\gamma_t\lambda\|v_t^{\mathcal{P}} - \zeta_t^{\mathcal{P}}\| \leq \|\operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t)\|$. Since $\operatorname{grad} f_{\mathcal{B}_t}(x_t) = \mathcal{P}_{\mathcal{T}_{x_t}\mathcal{M}}(\nabla f_{\mathcal{B}_t}(x_t))$, the batching property still holds, and we can conclude

$$\|v_t^{\mathcal{P}}\|^2 \;\leq\; 2\mathbb{E}\left[\|\zeta_t^{\mathcal{P}}\|^2 \mid x_t\right] + \frac{2\sigma^2}{\gamma_t^2\lambda^2|\mathcal{B}_t|}.$$

The proof is completed. □

We now turn to establishing the sample complexity of the two stochastic methods. In the context of stochastic optimization, we assume that $\mathcal{M}$ is a compact Riemannian embedded submanifold. This assumption ensures that gradient-related quantities remain uniformly bounded throughout the analysis. Besides, we can also obtain uniform constants $M_1^{\mathrm{R}}, M_2^{\mathrm{R}} > 0$ in retraction inequalities (2.1) by Boumal et al. (2019).

Assumption 5.1. *The Riemannian embedded submanifold in problem* (5.1) *is compact. Accordingly, we define* $G_f \triangleq \max_{x\in\mathcal{M}} \|\nabla f(x)\|$, *and* $G_h \triangleq \max_{x\in\mathcal{M}} \|\nabla h(x)\|$.

Lemma 5.2. *Suppose Assumptions 2.1 and 5.1 hold. For any stepsize* $\gamma_t \geq L$ *and* $\alpha_t > 0$, *the iterate* x_{t+1} *generated by Algorithm 3 satisfies*

$$\mathbb{E}\left[F(x_{t+1}) - F(x_t) \mid x_t\right] \;\leq\; -\left(\frac{\gamma_t\lambda\alpha_t}{4} - (G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}}\alpha_t^2\right)\|\zeta_t^{\mathrm{R}}\|^2 + \frac{\sigma^2\alpha_t}{\gamma_t\lambda|\mathcal{B}_t|},$$

where the expectation is taken with respect to $\pi_t^{(1)}, \ldots, \pi_t^{(\mathcal{B}_t)}$.

Proof. Following a similar argument in the proof of Lemma 3.2 implies

$$f(x_{t+1}) - f(x_t) \;\leq\; \alpha_t\left(\langle\nabla f(x_t), \zeta_t^{\mathrm{R}}\rangle + \gamma_t D_h(x_t + \zeta_t^{\mathrm{R}}, x_t)\right) + (G_f + 2\gamma_t G_h)\, M_2^{\mathrm{R}}\|\alpha_t\zeta_t^{\mathrm{R}}\|^2.$$

It remains to upper the term

$$\begin{aligned}&\alpha_t\left(\langle\nabla f(x_t), \zeta_t^{\mathrm{R}}\rangle + \gamma_t D_h(x_t + \zeta_t^{\mathrm{R}}, x_t)\right)\\ =\;& \alpha_t\left(\langle\nabla f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathrm{R}}\rangle + \gamma_t D_h(x_t + \zeta_t^{\mathrm{R}}, x_t)\right) + \alpha_t\langle\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathrm{R}}\rangle.\end{aligned}$$

From the update of (5.2), we have

$$\langle\nabla f_{\mathcal{B}_t}(x_t) + \gamma_t\nabla h(x_t + \zeta_t^{\mathrm{R}}) - \gamma_t\nabla h(x_t) + s_t', v - \zeta_t^{\mathrm{R}}\rangle \geq 0,\;\; \forall\zeta \in \mathcal{T}_{x_t}\mathcal{M},$$

where $s'_t \in \partial g(x_t + \zeta_t^{\mathrm{R}})$. Specifically, choose ζ to be the zero vector in $\mathcal{T}_x\mathcal{M}$ and it yields $\langle \nabla f_{\mathcal{B}_t}(x_t) - \gamma_t \nabla h(x_t), \zeta_t^{\mathrm{R}} \rangle \leq \langle \gamma_t \nabla h(x_t + \zeta_t^{\mathrm{R}}), -\zeta_t^{\mathrm{R}} \rangle - \langle s'_t, \zeta_t^{\mathrm{R}} \rangle$. By using the strong convexity of h. we have

$$\alpha_t \left(\langle \nabla f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathrm{R}} \rangle + \gamma_t D_h(x_t + \zeta_t^{\mathrm{R}}, x_t) \right) \;\leq\; -\frac{\alpha_t \gamma_t \lambda}{2} \|\zeta_t^{\mathrm{R}}\|^2 - \alpha_t \langle s'_t, \zeta_t^{\mathrm{R}} \rangle.$$

Besides, we can apply Young's inequality to obtain

$$\alpha_t \langle \nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathrm{R}} \rangle \;\leq\; \frac{\alpha_t}{\gamma_t \lambda} \|\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t)\|^2 + \frac{\gamma_t \lambda \alpha_t}{4} \|\zeta_t^{\mathrm{R}}\|^2.$$

For the nonsmooth part g, it holds that

$$g(x_{t+1}) - g(x_t) \;\leq\; L_g M_2^{\mathrm{R}} \|\alpha_t \zeta_t^{\mathrm{R}}\|^2 + \alpha_t \langle s'_t, \zeta_t^{\mathrm{R}} \rangle.$$

Therefore, the descent property of F can be shown as

$$F(x_{t+1}) - F(x_t) \;\leq\; -\left(\frac{\gamma_t \lambda \alpha_t}{4} - (G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}} \alpha_t^2 \right) \|\zeta_t^{\mathrm{R}}\|^2 + \frac{\alpha_t}{\gamma_t \lambda} \|\nabla f(x_t) - \nabla f_{\mathcal{B}_t}(x_t)\|^2.$$

Finally, it remains to take the expectation

$$\mathbb{E}\left[F(x_{t+1}) - F(x_t) \mid x_t \right] \;\leq\; -\left(\frac{\gamma_t \lambda \alpha_t}{4} - (G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}} \alpha_t^2 \right) \|\zeta_t^{\mathrm{R}}\|^2 + \frac{\sigma^2 \alpha_t}{\gamma_t \lambda |\mathcal{B}_t|},$$

and the proof is completed. □

Theorem 5.1. *Suppose Assumptions 2.1 and 5.1 hold, and let $T \geq 1$ be the total number of iterations. Under the following parameter setting:*

$$\gamma_t = \gamma \geq L, \;\; \alpha_t = \alpha < \frac{\gamma \lambda}{8\,(G_f + 2\gamma G_h + L_g)\, M_2^{\mathrm{R}}}, \;\; |\mathcal{B}_t| = |\mathcal{B}|, \tag{5.6}$$

the sequence $\{x_t\}_{t=0}^{T-1}$ generated by Algorithm 3 satisfies

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|v_t^{\mathrm{R}}\|^2 \right] \;\leq\; \mathcal{O}\left(\frac{F(x_0) - F^*}{T} + \frac{\sigma^2}{|\mathcal{B}|} \right),$$

where the expectation is taken with respect to all the randomness.

Proof. From Lemma 5.2, rearranging terms gives

$$\left(\frac{\gamma_t \lambda \alpha_t}{4} - (G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}} \alpha_t^2 \right) \|\zeta_t^{\mathrm{R}}\|^2 \;\leq\; \mathbb{E}\left[F(x_t) - F(x_{t+1}) \mid x_t \right] + \frac{\sigma^2 \alpha_t^2}{2|\mathcal{B}_t|}$$

for any $\gamma_t \geq L$ and $\alpha_t > 0$. Notice that the choice $\alpha_t \leq \gamma_t \lambda / (8\,(G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}})$ guarantees that

$$\frac{\gamma_t \lambda \alpha_t}{4} - (G_f + 2\gamma_t G_h + L_g)\, M_2^{\mathrm{R}} \alpha_t^2 \;\geq\; \frac{\gamma_t \lambda \alpha_t}{8}.$$

Consequently, we obtain

$$\|\zeta_t^{\mathrm{R}}\|^2 \leq \frac{8\mathbb{E}\left[F(x_t) - F(x_{t+1}) \mid x_t\right]}{\gamma_t \lambda \alpha_t} + \frac{4\sigma^2 \alpha_t}{\gamma_t \lambda |\mathcal{B}_t|}.$$

By combining with Lemma 5.1, and substituting the parameter choice (5.6), it follows

$$\alpha_t \|v_t^{\mathrm{R}}\|^2 \;\leq\; \frac{16\mathbb{E}\left[F(x_t) - F(x_{t+1}) \mid x_t\right]}{\gamma_t \lambda \alpha_t} + \frac{8\sigma^2 \alpha_t}{\gamma_t \lambda |\mathcal{B}_t|} + \frac{2\sigma^2}{\gamma_t^2 \lambda^2 |\mathcal{B}_t|}.$$

summing over t from 0 to $T-1$ imply

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|v_t^{\mathrm{R}}\|^2\right] \;\leq\; \frac{16(F(x_0) - F(x^*))}{\gamma \lambda \alpha T} + \frac{8\sigma^2 \alpha}{\gamma \lambda |\mathcal{B}|} + \frac{2\sigma^2}{\gamma^2 \lambda^2 |\mathcal{B}|}$$

where the expectation is taken with respect to all the randomness. The proof is completed. □

Given an accuracy $\epsilon > 0$, choose a minibatch size $|\mathcal{B}| = \mathcal{O}(\epsilon^{-2})$. Then, after $T = \mathcal{O}(\epsilon^{-2})$ terations, an ϵ-approximate stationary point can be found in expectation, and the overall sample complexity is $\mathcal{O}(\epsilon^{-4})$. Now we establish the sample complexity bound of the stochastic projection-based method (Algorithm 4). As discussed in Remark 4.1, when $\mathcal{M}$ is compact, we can use the stronger Lemma 7.1 in place of Lemma 4.2, and hence no bound on the tangent component is required. Let $H_h \triangleq \max_{x\in\mathcal{M}} \|\nabla^2 h(x)\|$. Define $\Psi_1 \triangleq 2G_h(M_2^{\mathcal{P}} + M_3^{\mathcal{P}}) + H_h$ and $\Psi_2 \triangleq \left((M_2^{\mathcal{P}} + M_3^{\mathcal{P}}) + (M_1^{\mathcal{P}})^2 M_4^{\mathcal{P}}\right) G_f$, where $M_i^{\mathcal{P}}$ $(i = 1, 2, 3)$ are the constants in Lemma 7.1. Therefore, since $u_t = 0$ in Algorithm 4, we can conclude the following descent property

$$f(x_{t+1}) - f(x_t) \;\leq\; (\gamma_t \Psi_1 + \Psi_2)\,\|\alpha_t \zeta_t^{\mathcal{P}}\|^2 + \alpha_t \langle \operatorname{grad} f(x_t), \zeta_t^{\mathcal{P}}\rangle + \alpha_t \gamma_t D_h(x_t + \zeta_t^{\mathcal{P}}, x_t) \tag{5.7}$$

by using a similar argument in the proof of Theorem 4.1.

Theorem 5.2. *Suppose Assumptions 2.1, 4.1 and 5.1 hold. Let $T \geq 1$ be the total number of iterations. Under the following parameter setting:*

$$\gamma_t = \gamma \geq L,\;\; \alpha_t = \alpha \leq \frac{\gamma\lambda}{8(\gamma\Psi_1 + \Psi_2)},\;\; |\mathcal{B}_t| = |\mathcal{B}|, \tag{5.8}$$

the sequence $\{x_t\}_{t\geq 0}$ generated by Algorithm 4 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|v_t^{\mathcal{P}}\|^2\right] \;\leq\; \mathcal{O}\left(\frac{F(x_0) - F^*}{T} + \frac{\sigma^2}{|\mathcal{B}|}\right).$$

where the expectation is taken with respect to all the randomness.

Proof. Recall that $g \equiv 0$. From (5.7), it holds that

$$F(x_{t+1}) - F(x_t) \;\leq\; (\gamma_t \Psi_1 + \Psi_2)\,\|\alpha_t \zeta_t^{\mathcal{P}}\|^2 + \alpha_t \langle \operatorname{grad} f(x_t), \zeta_t^{\mathcal{P}}\rangle + \alpha_t \gamma_t D_h(x_t + \zeta_t^{\mathcal{P}}, x_t).$$

Then, we decompose $\operatorname{grad} f(x_t) = \operatorname{grad} f_{\mathcal{B}_t}(x_t) + \operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t)$. By using the optimality condition of the subproblem 5.3, we obtain $\operatorname{grad} f_{\mathcal{B}_t}(x_t) = \gamma_t \nabla h(x_t) - \gamma_t \nabla h(x_t + \zeta_t^{\mathcal{P}})$, which further implies

$$\alpha_t \langle \operatorname{grad} f(x_t), \zeta_t^{\mathcal{P}}\rangle + \alpha_t \gamma_t D_h(x_t + \zeta_t^{\mathcal{P}}, x_t)$$

$$
\begin{aligned}
&= \alpha_t \langle \gamma_t \nabla h(x_t) - \gamma_t \nabla h(x_t + \zeta_t^{\mathcal{P}}), \zeta_t^{\mathcal{P}} \rangle + \alpha_t \gamma_t D_h(x_t + \zeta_t^{\mathcal{P}}, x_t) + \alpha_t \langle \operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathcal{P}} \rangle \\
&= \alpha_t \gamma_t \left(h(x_t + \zeta_t^{\mathcal{P}}) - h(x_t) - \langle \nabla h(x_t + \zeta_t^{\mathcal{P}}), \zeta_t^{\mathcal{P}} \rangle \right) + \alpha_t \langle \operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t), \zeta_t^{\mathcal{P}} \rangle \\
&\leq -\frac{\alpha_t \gamma_t \lambda}{2} \|\zeta_t^{\mathcal{P}}\|^2 + \frac{\alpha_t}{\gamma_t \lambda} \| \operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t) \|^2 + \frac{\gamma_t \lambda \alpha_t}{4} \|\zeta_t^{\mathcal{P}}\|^2 \\
&= -\frac{\alpha_t \gamma_t \lambda}{4} \|\zeta_t^{\mathcal{P}}\|^2 + \frac{\alpha_t}{\gamma_t \lambda} \| \operatorname{grad} f(x_t) - \operatorname{grad} f_{\mathcal{B}_t}(x_t) \|^2.
\end{aligned}
$$

As a result, the descent property follows

$$
\mathbb{E}\left[F(x_{t+1}) - F(x_t) \mid x_t\right] \leq -\left(\frac{\gamma_t \lambda \alpha_t}{4} - (\gamma_t \Psi_1 + \Psi_2)\, \alpha_t^2 \right) \|\zeta_t^{\mathcal{P}}\|^2 + \frac{\sigma^2 \alpha_t^2}{2|\mathcal{B}_t|}.
$$

Notice that under the choice $\alpha_t \leq \gamma\lambda/(8(\gamma\Psi_1 + \Psi_2))$, we have

$$
\frac{\gamma_t \lambda \alpha_t}{4} - (\gamma \Psi_1 + \Psi_2) \alpha_t^2 \geq \frac{\gamma_t \lambda \alpha_t}{8}.
$$

Consequently, we obtain

$$
\|\zeta_t^{\mathcal{P}}\|^2 \leq \frac{8\mathbb{E}\left[F(x_t) - F(x_{t+1}) \mid x_t\right]}{\gamma_t \lambda \alpha_t} + \frac{4\sigma^2 \alpha_t}{\gamma_t \lambda |\mathcal{B}_t|}.
$$

By combining with Lemma 5.1 and substituting the parameter choice (5.6), we can conclude

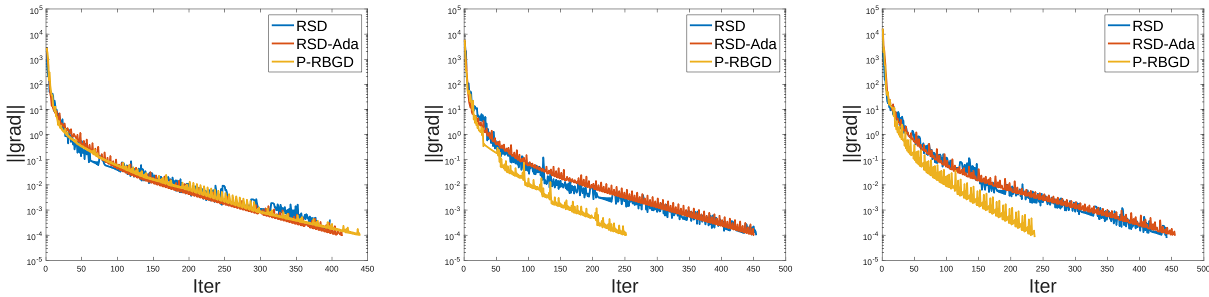
$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|v_t^{\mathcal{P}}\|^2\right] \leq \mathcal{O}\left(\frac{F(x_0) - F^*}{T} + \frac{\sigma^2}{|\mathcal{B}|} \right).
$$

The proof is completed. □

6 Numerical Experiments

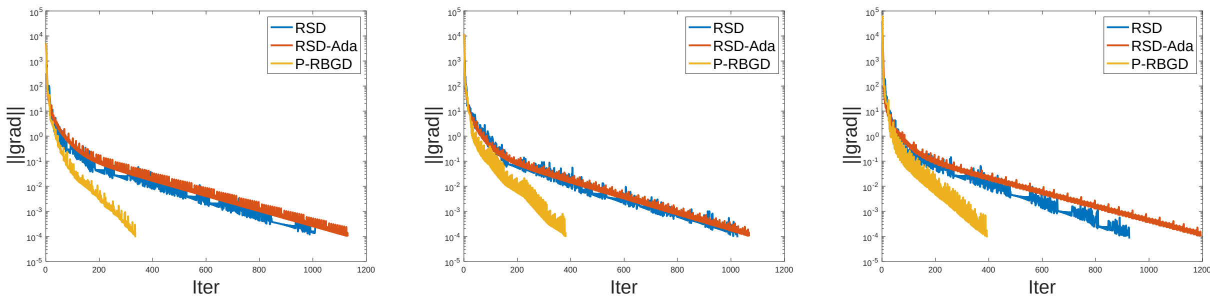
In this section, we numerically test our Riemannian Bregman gradient methods on the nonlinear eigenvalue problem (1.2) and the low-rank quadratic sensing problem (1.3). All numerical experiments reported here are performed on a platform equipped with an Apple M1 CPU and 8GB of RAM. We test Algorithm 1 (R-RBGD), Algorithm 2 without a corrective normal vector (P-RBGD), and Algorithm 2 with a corrective normal vector $u_t = -v_t^{\mathcal{N}}$ (P-BRGD-C). We compare our methods against the `steepestdescent` solver in Manopt (Boumal et al., 2014), employing both the default line search (RSD) and adaptive line search (RSD-Ada) strategies. For RSD and RSD-Ada, all parameters are kept at their default values provided by the package. The initial points are randomly generated.

We begin by testing the low-rank quadratic sensing problem (1.3). The data $\{(y_j, c_j)\}_{j=1}^N$ are randomly generated by MATLAB's `randn` function, and we set $N = 100$. By proposition 4.1, the update directions generated by R-RBGD and P-RBGD coincide. Hence, we only test P-RBGD, as the associated subproblem is computationally simpler. The parameters in the backtracking line search are set to an initial stepsize of 0.1 and a contraction factor of 0.5. All algorithms are terminated when the norm of the Riemannian gradient is less than 10^{-4}. We test all algorithms with various parameter combinations of m and r. The results are illustrated in Figures 1, 2, 3, 4. From these figures, we observe that our projection-based Riemannian Bregman gradient method outperforms RSD and RSD-Ada in most cases, requiring fewer iterations to achieve the specified accuracy.



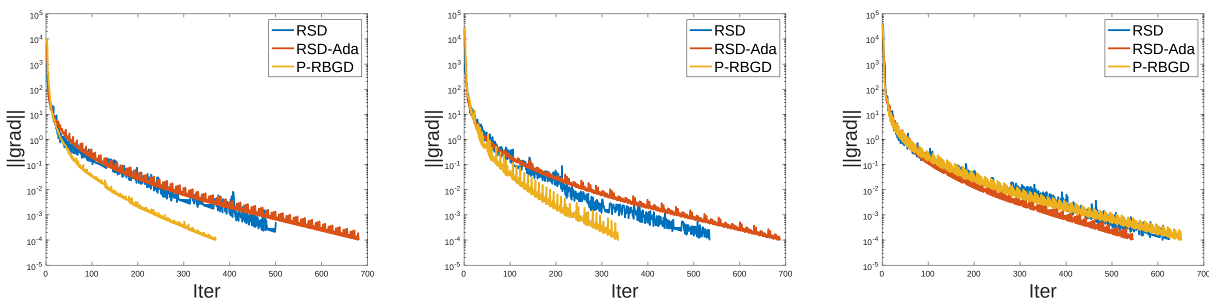
(a) $m = 500, r = 10$ (b) $m = 500, r = 20$ (c) $m = 500, r = 40$

Figure 1: The results of problem (1.3) with $m = 500$ and varying rank $r = 10, 20, 40$.



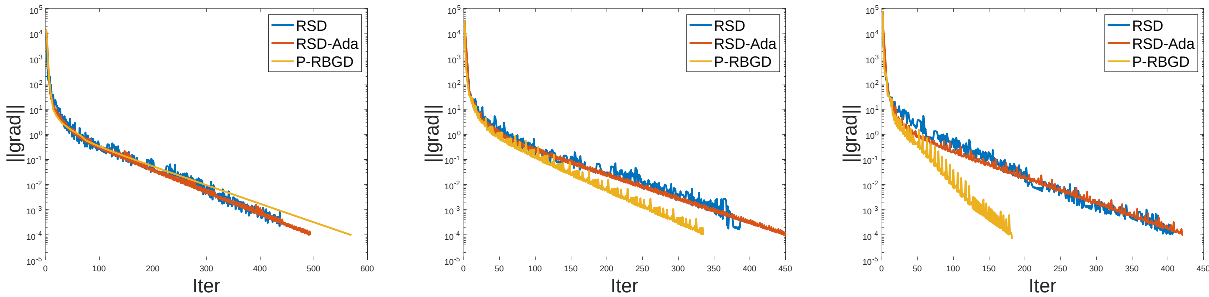
(a) $m = 1000, r = 10$ (b) $m = 1000, r = 20$ (c) $m = 1000, r = 40$

Figure 2: The results of problem (1.3) with $m = 1000$ and varying rank $r = 10, 20, 40$.



(a) $m = 2000, r = 10$ (b) $m = 2000, r = 20$ (c) $m = 2000, r = 40$

Figure 3: The results of problem (1.3) with $m = 2000$ and varying rank $r = 10, 20, 40$.



(a) $m = 4000, r = 10$ (b) $m = 4000, r = 20$ (c) $m = 4000, r = 40$

Figure 4: The results of problem (1.3) with $m = 4000$ and varying rank $r = 10, 20, 40$.

Next, we apply our methods to the nonlinear eigenvalue problem (1.2). The parameter β is set to 10. For these three algorithms, we use a backtracking line search with an initial stepsize of 0.5 and a contraction factor of 0.5. Similar to the previous experiment, all algorithms are terminated when the norm of the Riemannian gradient is smaller than 10^{-4}. Tables 1 and 2 report the performance of all compared solvers on problem (1.2) for various combinations of the parameters m and p. From these tables, we observe that our Riemannian Bregman gradient methods achieve function values comparable to those obtained by RSD and RSD-Ada, yet require fewer iterations (the function values differ only in the 8th decimal place). Due to the need to solve an auxiliary subproblem at each iteration, the CPU time is similar to that of RSD and RSD-Ada. However, for challenging instances (e.g., $m = 5000, p = 60$), RSD and RSD-Ada fail to satisfy the termination criterion because the stepsize becomes excessively small, causing premature termination.

7 Conclusions

In this paper, we developed two Riemannian Bregman gradient methods for solving relatively smooth optimization problems over Riemannian embedded submanifolds. The retraction-based method handles nonsmooth optimization by solving a convex subproblem constrained to the tangent space at each iteration. We identified particular reference functions, such as the quartic form, log-barrier, and entropy functions, for which the subproblem admits either closed-form or significantly simplified solutions. The projection-based approach, suitable for smooth optimization, involves solving an unconstrained subproblem in the ambient Euclidean space followed by a projection onto the manifold. Both methods achieve an iteration complexity of $\mathcal{O}(1/\epsilon^2)$ for finding an ϵ-approximate Riemannian stationary point. Additionally, for compact manifolds, we proposed stochastic variants with sample complexities of $\mathcal{O}(1/\epsilon^4)$. Numerical experiments demonstrated the effectiveness of proposed Riemannian Bregman gradient methods.

Table 1: The results of problem (1.2) with varying m.

Solver	$m=500,\ p=50$				$m=1000,\ p=50$			
	Fval	‖ grad ‖	Iter	Time	Fval	‖ grad ‖	Iter	Time
RSD	2.7674e+04	5.1152e-05	7566	36.2087	2.7674e+04	2.5639e-04	8873	68.4400
RSD-Ada	2.7674e+04	4.3667e-04	7650	40.2719	2.7674e+04	3.8527e-04	8988	70.3007
R-RBGD	2.7674e+04	9.6010e-05	4938	23.5305	2.7674e+04	9.8616e-05	5518	53.8013
P-RBGD	2.7674e+04	9.9854e-05	4864	40.7854	2.7674e+04	9.9820e-05	5503	74.4051
P-RBGD-C	2.7674e+04	9.7726e-05	4901	46.7791	2.7674e+04	9.8943e-05	5510	80.5897
Solver	$m=1500,\ p=50$				$m=2000,\ p=50$			
	Fval	‖ grad ‖	Iter	Time	Fval	‖ grad ‖	Iter	Time
RSD	2.7674e+04	9.1536e-04	7108	117.2364	2.7674e+04	2.6418e-04	8902	159.2698
RSD-Ada	2.7674e+04	4.7544e-04	6304	113.7995	2.7674e+04	4.3806e-04	9115	196.9259
R-RBGD	2.7674e+04	9.9155e-05	5160	102.2899	2.7674e+04	9.5463e-05	5964	192.0930
P-RBGD	2.7674e+04	9.9829e-05	5144	130.3309	2.7674e+04	9.8332e-05	5884	171.8766
P-RBGD-C	2.7674e+04	9.8907e-05	5322	133.2718	2.7674e+04	9.7909e-05	6019	177.7956
Solver	$m=2500,\ p=50$				$m=3000,\ p=50$			
	Fval	‖ grad ‖	Iter	Time	Fval	‖ grad ‖	Iter	Time
RSD	2.7674e+04	2.3527e-04	7840	206.9153	2.7674e+04	2.1955e-04	7732	221.2304
RSD-Ada	2.7674e+04	3.5709e-04	7238	232.8811	2.7674e+04	5.1092e-04	6015	204.8033
R-RBGD	2.7674e+04	9.9000e-05	5197	204.8740	2.7674e+04	9.8652e-05	5315	239.3315
P-RBGD	2.7674e+04	9.8332e-05	5246	179.6978	2.7674e+04	9.8969e-05	5265	213.2356
P-RBGD-C	2.7674e+04	9.7069e-05	5303	207.1347	2.7674e+04	9.8572e-05	5233	226.3744

Table 2: The results of problem (1.2) with varying p.

Solver	$m = 5000,\ p = 10$				$m = 5000,\ p = 20$			
	Fval	$\|\text{grad}\|$	Iter	Time	Fval	$\|\text{grad}\|$	Iter	Time
RSD	2.8429e+02	8.1462e-05	249	1.0864	1.9443e+03	9.9204e-05	1333	9.7795
RSD-Ada	2.8429e+02	9.8615e-05	307	1.2623	1.9443e+03	9.8288e-05	1401	10.5765
R-RBGD	2.8429e+02	9.3002e-05	317	1.3235	1.9443e+03	9.7837e-05	1282	12.6708
P-RBGD	2.8429e+02	9.3140e-05	315	1.1404	1.9443e+03	9.9858e-05	1274	11.4878
P-RBGD-C	2.8429e+02	9.5805e-05	314	1.3440	1.9443e+03	9.8893e-05	1267	12.6460
Solver	$m = 5000,\ p = 30$				$m = 5000,\ p = 40$			
	Fval	$\|\text{grad}\|$	Iter	Time	Fval	$\|\text{grad}\|$	Iter	Time
RSD	6.2293e+03	9.9957e-05	3328	77.0571	1.4389e+04	9.5968e-05	5728	171.1722
RSD-Ada	6.2293e+03	9.9447e-05	3228	84.4482	1.4389e+04	3.1914e-04	3872	126.8553
R-RBGD	6.2293e+03	9.9393e-05	1681	43.5144	1.4389e+04	9.8178e-05	5367	241.8630
P-RBGD	6.2293e+03	9.9030e-05	1679	38.0242	1.4389e+04	9.6651e-05	5356	218.3186
P-RBGD-C	6.2293e+03	9.9961e-05	1665	40.7001	1.4389e+04	9.8343e-05	5339	223.8977
Solver	$m = 5000,\ p = 50$				$m = 5000,\ p = 60$			
	Fval	$\|\text{grad}\|$	Iter	Time	Fval	$\|\text{grad}\|$	Iter	Time
RSD	2.7674e+04	2.7698e-04	8007	277.4160	4.7334e+04	1.3049e-03	10090	392.3834
RSD-Ada	2.7674e+04	4.3422e-04	7477	300.1134	4.7334e+04	6.6243e-04	10260	415.5020
R-RBGD	2.7674e+04	9.9805e-05	5326	328.1556	4.7334e+04	9.9850e-05	5128	353.4060
P-RBGD	2.7674e+04	9.9563e-05	5228	303.5396	4.7334e+04	9.9838e-05	5267	382.2308
P-RBGD-C	2.7674e+04	9.7795e-05	5423	337.2280	4.7334e+04	9.8905e-05	5300	411.1444

Appendix

In the appendix, we first introduce several concepts from variational analysis that are used in the proof of Lemma 3.3. Then, we restate the well-known Berge's Maximum Theorem, which is utilized in the proof of Theorem 3.1, and Lemma 5.10 from Ding et al. (2024), which provides inequalities related to projections onto a compact submanifold.

Definition 7.1 (Set-valued map)**.** *Let X and Y be topological spaces. A set-valued map from X to Y is a mapping $S : X \rightrightarrows Y$ that assigns to each $x \in X$ a subset $S(x) \subseteq Y$.*

Definition 7.2 (Continuity of set-valued maps)**.** *Let X and Y be metric spaces, and $S : X \rightrightarrows Y$ a set-valued map. We say S is outer semicontinuous at $x \in X$ if, whenever $x_k \to x$ and $y_k \to y$ with $y_k \in S(x_k)$, we have $y \in S(x)$; S is inner semicontinuous at $x \in X$ if, for any sequence $x_k \to x$ and any $y \in S(x)$, there exists a sequence $y_k \in S(x_k)$ with $y_k \to y$. S is continuous at x if it is both outer and inner semicontinuous at x.*

Theorem 7.1 (Berge's Maximum Theorem)**.** *Let $S\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a non-empty, compact-valued, continuous set-valued map, and let $\varphi\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be continuous. Then the maximum value function*

$$\Phi(x) \triangleq \max_{v \in S(x)} \varphi(x, v), \qquad x \in \mathbb{R}^n$$

is well-defined and continuous.

Lemma 7.1 (Restatement of Lemma 5.10 in Ding et al. (2024))**.** *Let $\mathcal{M} \subseteq \mathbb{R}^n$ be a compact submanifold of class C^3. Then there exists a constant $\varrho > 0$ such that for all $x \in \mathcal{M}, v \in \mathcal{T}_x\mathcal{M}$ and $u \in \mathcal{N}_x\mathcal{M}$ satisfying $\|u\| \leq \varrho/2$, we have*

$$\|\mathcal{P}_{\mathcal{M}}(x + v + u) - x\| \leq M_1^{\mathcal{P}} \|v\|,$$
$$\|\mathcal{P}_{\mathcal{M}}(x + v + u) - x - v\| \leq M_2^{\mathcal{P}} \|v\|^2 + M_3^{\mathcal{P}} \|v\| \|u\|,$$

for some positive constants $M_1^{\mathcal{P}}, M_2^{\mathcal{P}}, M_3^{\mathcal{P}} > 0$.

References

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. (Cited on pages 13 and 14.)

P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7:303–330, 2007. (Cited on page 4.)

P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. (Cited on pages 1 and 5.)

N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188:85–134, 2021. (Cited on page 4.)

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. (Cited on pages 2 and 4.)

J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018. (Cited on page 2.)

S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. (Cited on page 4.)

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. (Cited on pages 1 and 5.)

N. Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015. (Cited on page 1.)

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014. (Cited on page 25.)

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. (Cited on pages 2, 6, and 22.)

T. T. Cai and A. Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015. (Cited on page 3.)

Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li. On an eigenvector-dependent nonlinear eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1360–1382, 2018. (Cited on page 2.)

Y. Cao and Y. Xie. Poisson matrix completion. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1841–1845. IEEE, 2015. (Cited on page 3.)

S. Chen, S. Ma, A. Man-Cho So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020. (Cited on pages 4, 6, 7, and 12.)

Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015. (Cited on page 3.)

A. Cherian and S. Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems*, 28(12):2859–2871, 2016. (Cited on page 1.)

Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018. (Cited on page 3.)

K. Ding, J. Li, and K.-C. Toh. Nonconvex stochastic bregman proximal gradient method with application to deep learning. *Journal of Machine Learning Research*, 26(39):1–44, 2025. (Cited on page 4.)

W. Ding, J. Li, and S. Zhang. Convergence analysis of the transformed gradient projection algorithms on compact matrix manifolds. *arXiv preprint arXiv:2404.19392*, 2024. (Cited on pages 4, 13, 14, 15, and 30.)

R. A. Dragomir, M. Even, and H. Hendrikx. Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. In *International Conference on Machine Learning*, pages 2815–2825. PMLR, 2021. (Cited on page 4.)

I. Fatkhullin and N. He. Taming nonconvex stochastic mirror descent with general bregman divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3501. PMLR, 2024. (Cited on page 4.)

P. Grohs and S. Hosseini. Nonsmooth trust region algorithms for locally lipschitz functions on riemannian manifolds. *IMA Journal of Numerical Analysis*, 36(3):1167–1192, 2016. (Cited on page 4.)

A. Han and J. Gao. Improved variance reduction methods for riemannian non-convex optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7610–7623, 2021. (Cited on page 4.)

A. Han, B. Mishra, P. Jawanpuria, and J. Gao. Nonconvex-nonconcave min-max optimization on riemannian manifolds. *Transactions on Machine Learning Research*, 2023a. (Cited on page 4.)

A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3):1797–1827, 2023b. (Cited on page 4.)

A. Han, B. Mishra, P. K. Jawanpuria, and A. Takeda. A framework for bilevel optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 37:103829–103872, 2024. (Cited on page 4.)

F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *Computational Optimization and Applications*, 79:717–766, 2021. (Cited on page 4.)

F. Hanzely, P. Richtarik, and L. Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79:405–440, 2021. (Cited on page 4.)

C. He, Z. Pan, X. Wang, and B. Jiang. Riemannian accelerated zeroth-order algorithm: improved robustness and lower query complexity. *arXiv preprint arXiv:2405.05713*, 2024. (Cited on page 4.)

O. Hinder and Y. Ye. Worst-case iteration bounds for log barrier methods on problems with nonconvex constraints. *Mathematics of Operations Research*, 49(4):2402–2424, 2024. (Cited on page 2.)

R. Hosseini and S. Sra. Recent advances in stochastic riemannian optimization. *Handbook of Variational Methods for Nonlinear Geometric Data*, pages 527–554, 2020. (Cited on page 4.)

J. Hu, K. Deng, J. Wu, and Q. Li. A projected semismooth newton method for a class of nonconvex composite programs with strong prox-regularity. *Journal of Machine Learning Research*, 25(56): 1–32, 2024. (Cited on pages 4 and 13.)

W. Huang and K. Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194 (1):371–413, 2022. (Cited on page 4.)

W. Huang and K. Wei. An inexact riemannian proximal gradient method. *Computational Optimization and Applications*, 85(1):1–32, 2023. (Cited on page 4.)

X. Jiang, J. Li, M. Hong, and S. Zhang. Barrier function for bilevel optimization with coupled lower-level constraints: Formulation, approximation and algorithms. *arXiv preprint arXiv:2410.10670*, 2024. (Cited on page 2.)

J. Li and S. Ma. Riemannian bilevel optimization. *Journal of Machine Learning Research*, 26(18): 1–44, 2025. (Cited on page 4.)

J. Li, K. Balasubramanian, and S. Ma. Stochastic zeroth-order riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 48(2):1183–1211, 2023. (Cited on page 4.)

J. Li, S. Ma, and T. Srivastava. A riemannian alternating direction method of multipliers. *Mathematics of Operations Research*, 2024. (Cited on page 1.)

Z. Li, S. He, and S. Zhang. *Approximation methods for polynomial optimization: Models, Algorithms, and Applications*. Springer Science & Business Media, 2012. (Cited on page 2.)

H. Liu, A. M.-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1):215–262, 2019. (Cited on page 1.)

X. Liu, X. Wang, Z. Wen, and Y. Yuan. On the convergence of the self-consistent field iteration in kohn–sham density functional theory. *SIAM Journal on Matrix Analysis and Applications*, 35(2): 546–558, 2014. (Cited on page 2.)

H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. (Cited on pages 2, 3, 4, and 12.)

A. D. McRae and M. A. Davenport. Low-rank matrix completion and denoising under poisson noise. *Information and Inference: A Journal of the IMA*, 10(2):697–720, 2021. (Cited on page 3.)

B. Mishra, H. Kasai, P. Jawanpuria, and A. Saroop. A riemannian gossip approach to subspace learning on grassmann manifold. *Machine Learning*, 108:1783–1803, 2019. (Cited on page 1.)

M. C. Mukkamala. *Bregman proximal minimization algorithms, analysis and applications*. PhD thesis, Universität Tübingen, 2022. (Cited on page 4.)

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994. (Cited on page 2.)

H. Sato. Riemannian conjugate gradient methods: General framework and specific algorithms with convergence analyses. *SIAM Journal on Optimization*, 32(4):2690–2717, 2022. (Cited on page 3.)

W. Si, P.-A. Absil, W. Huang, R. Jiang, and S. Vary. A riemannian proximal newton method. *SIAM Journal on Optimization*, 34(1):654–681, 2024. (Cited on page 4.)

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 407–410. IEEE, 2015. (Cited on page 1.)

S. Takahashi and A. Takeda. Approximate bregman proximal gradient algorithm for relatively smooth nonconvex optimization. *Computational Optimization and Applications*, pages 1–30, 2024. (Cited on pages 2 and 4.)

S. Takahashi and A. Takeda. Approximate bregman proximal gradient algorithm for relatively smooth nonconvex optimization. *Computational Optimization and Applications*, 90(1):227–256, 2025. (Cited on page 4.)

B. Vandereycken. Low-rank matrix completion by riemannian optimization—extended version. *arXiv preprint arXiv:1209.3834*, 2012. (Cited on page 20.)

B. Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. (Cited on page 1.)

A. Vyguzov and F. Stonyakin. Adaptive variant of frank-wolfe method for relative smooth convex optimization problems. *arXiv preprint arXiv:2405.12948*, 2024. (Cited on page 4.)

Q. Wang and W. H. Yang. Proximal quasi-newton method for composite optimization over the stiefel manifold. *Journal of Scientific Computing*, 95(2):39, 2023. (Cited on page 4.)

Q. Wang and W. H. Yang. An adaptive regularized proximal newton-type methods for composite optimization over the stiefel manifold. *Computational Optimization and Applications*, 89(2): 419–457, 2024. (Cited on page 4.)

C. Yang, W. Gao, and J. C. Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM journal on matrix analysis and applications*, 30(4): 1773–1788, 2009. (Cited on page 2.)

W. H. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014. (Cited on page 7.)

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016. (Cited on page 7.)

H. Zhang, S. J Reddi, and S. Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016. (Cited on page 4.)

J. Zhang and S. Zhang. A cubic regularized newton's method over riemannian manifolds. *arXiv preprint arXiv:1805.05565*, 2018. (Cited on page 4.)

J. Zhang, J. Hu, A. M.-C. So, and M. Johansson. Nonconvex federated learning on compact smooth submanifolds with heterogeneous data. *Advances in Neural Information Processing Systems*, 37: 109817–109844, 2024. (Cited on page 13.)

P. Zhang, J. Zhang, and S. Sra. Sion's minimax theorem in geodesic metric spaces and a riemannian extragradient algorithm. *SIAM Journal on Optimization*, 33(4):2885–2908, 2023. (Cited on page 4.)