ENHANCING DISTRIBUTIONAL ROBUSTNESS IN REGULARIZED PRINCIPAL COMPONENT ANALYSIS BY WASSERSTEIN DISTANCES*

LEI WANG†, XIN LIU‡, AND XIAOJUN CHEN§

Abstract. We consider the distributionally robust optimization (DRO) model of regularized principal component analysis (PCA) to account for uncertainty in the underlying probability distribution. The resulting formulation leads to a constrained min-max optimization problem, where the ambiguity set captures the distributional uncertainty by the type-2 Wasserstein distance. We prove that the inner maximization problem admits a closed-form optimal value. This explicit characterization equivalently reformulates the original DRO model into a minimization problem on the Stiefel manifold. To solve the resulting problem, we devise an efficient smoothing manifold proximal gradient algorithm. Our analysis establishes Riemannian gradient consistency and global convergence of our algorithm to a stationary point of the reformulated minimization problem. We also provide the iteration complexity $O(\epsilon^{-3})$ of our algorithm to achieve an ϵ-approximate stationary point. Finally, numerical experiments are conducted to validate the effectiveness and scalability of our algorithm, as well as to highlight the necessity and rationality of adopting the DRO model for PCA.

Key words. distributionally robust optimization, principal component analysis, Riemannian manifold, smoothing approximation, Wasserstein distance

MSC codes. 65K05, 90C17, 90C26, 90C30, 90C47

1. Introduction.

Let $\xi \in \mathbb{R}^d$ be a d-dimensional random vector governed by a probability distribution $\mathbb{P}_*$. In this paper, we consider the following distributionally robust optimization (DRO) model of principal component analysis (PCA),

$$\min_{X \in \mathcal{O}^{d,r}} \sup_{\mathbb{P} \in \mathscr{P}} \mathbb{E}_{\mathbb{P}} \left[\left\| \left(I_d - XX^\top \right) (\xi - \mathbb{E}_{\mathbb{P}}[\xi]) \right\|_{\mathrm{F}}^2 \right] + s(X). \tag{1.1}$$

Here, the feasible set $\mathcal{O}^{d,r} := \{X \in \mathbb{R}^{d \times r} \mid X^\top X = I_r\}$, commonly referred to as the Stiefel manifold [1, 7, 50], consists of all the $d \times r$ column-orthonormal matrices with $1 \leq r < d$. The ambiguity set $\mathscr{P}$ represents a collection of distributions that could plausibly contain $\mathbb{P}_*$ with high confidence. And $s : \mathbb{R}^{d \times r} \to \mathbb{R}$ is a convex and Lipschitz continuous function acting as a regularizer to promote certain desired structures of solutions in $\mathcal{O}^{d,r}$, such as sparsity [10, 48] or nonnegativity [13, 52].

In most practical scenarios, the underlying distribution $\mathbb{P}_*$ is unknown and can not be captured precisely, leaving us without the essential information required to solve the PCA problem exactly. Although the technique of sample average approximation provides a practical model, its solutions often suffer from poor out-of-sample performances when the sample size is limited [20]. This dilemma motivates us to

*Submitted to the editors DATE.

Funding: This work is supported by National Key R&D Program of China (2023YFA1009300), RGC grant JLFS/P-501/24 for the CAS AMSS-PolyU Joint Laboratory in Applied Mathematics, CAS-Croucher Funding Scheme for Joint Laboratories, Hong Kong Research Grant Council project PolyU15300024, and National Natural Science Foundation of China (12125108, 12021001, 12288201).

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (lei2wang@polyu.edu.hk).

‡State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, Beijing, China (liuxin@lsec.cc.ac.cn).

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (maxjchen@polyu.edu.hk).

investigate the DRO model (1.1) of PCA, which minimizes the worst case of the objective function across all distributions in $\mathscr{P}$.

In the realm of DRO [25, 36], there is a variety of ambiguity sets available, including those based on moment constraints [14, 17, 57], divergences [46, 56], and Wasserstein distances [20, 22], among others. For a thorough and insightful exposition of these concepts, we refer interested readers to a recent survey [36], which offers an in-depth exploration of DRO problems. Recently, Wasserstein DRO, where the discrepancy between probability measures is dictated by the Wasserstein distance, has garnered tremendous attentions across various domains [35, 36]. In a similar vein, we explore the use of the Wasserstein distance in constructing the ambiguity set $\mathscr{P}$ in problem (1.1). Let $\mathscr{Q}_p$ be the space of all probability distributions $\mathbb{P}$ supported on $\mathbb{R}^d$ with finite p-th moments, namely, $\mathbb{E}_{\mathbb{P}}(\|\xi\|^p) = \int_{\mathbb{R}^d} \|\xi\|^p \,\mathbb{P}(\mathrm{d}\xi) < \infty$. Below is the definition of the Wasserstein distance on $\mathscr{Q}_p$.

DEFINITION 1.1 ([34]). *The type-p Wasserstein distance $\mathbb{W}_p : \mathscr{Q}_p \times \mathscr{Q}_p \to \mathbb{R}_+$ between two probability distributions $\mathbb{P}_1 \in \mathscr{Q}_p$ and $\mathbb{P}_2 \in \mathscr{Q}_p$ is defined as*

$$\mathbb{W}_p(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\mathbb{Q} \in \mathscr{J}(\mathbb{P}_1, \mathbb{P}_2)} \left(\mathbb{E}_{(\xi_1, \xi_2) \sim \mathbb{Q}} \left[\|\xi_1 - \xi_2\|_p^p \right] \right)^{1/p},$$

where $\|\cdot\|_p$ represents the ℓ_p norm on $\mathbb{R}^d$, and $\mathscr{J}(\mathbb{P}_1, \mathbb{P}_2)$ is the set containing all the joint distributions of ξ_1 and ξ_2 with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

Throughout this paper, we are primarily interested in the case where $p \in [1, 2]$, which is of significant importance both theoretically and practically [21]. The corresponding Wasserstein DRO model of PCA can be formulated as

$$(\mathrm{P_{mM}}) \qquad \begin{array}{ll} \min\limits_{X \in \mathcal{O}^{d,r}} \sup\limits_{\mathbb{P} \in \mathscr{Q}_p} & \mathbb{E}_{\mathbb{P}} \left[\left\| \left(I_d - XX^\top \right) (\xi - \mathbb{E}_{\mathbb{P}}[\xi]) \right\|_{\mathrm{F}}^2 \right] + s(X) \\ \text{s.t.} & \mathbb{P} \in \mathscr{B}_p(\mathbb{P}_\circ, \rho) := \{ \mathbb{P} \in \mathscr{Q}_p \mid \mathbb{W}_p(\mathbb{P}, \mathbb{P}_\circ) \le \rho \}, \end{array}$$

where $\mathbb{P}_\circ \in \mathscr{Q}_p$ is a nominal distribution conceived as being close to $\mathbb{P}_*$ and $\rho \ge 0$ is a radius of the Wasserstein ball. It is noteworthy that the performance guarantees of the ambiguity set $\mathscr{B}_p(\mathbb{P}_\circ, \rho)$ defined in $(\mathrm{P_{mM}})$ can be inherited from the theoretical results in Wasserstein DRO [4, 5, 20, 21].

By exploiting the structure of the inner maximization problem, we find that the DRO model $(\mathrm{P_{mM}})$ is well-posed only when $p = 2$, as its optimal value becomes positive infinity for $1 \le p < 2$. Moreover, the optimal value of the inner maximization problem can be computed in an explicit form for the case $p = 2$. This leads us to an equivalent reformulation of the DRO model $(\mathrm{P_{mM}})$ with $p = 2$ as follows,

$$(\mathrm{P_m}) \qquad \min_{X \in \mathcal{O}^{d,r}} \ \mathrm{tr}\left(\left(I_d - XX^\top \right) \Sigma_\circ \right) + s(X) + 2\rho \left\| \left(I_d - XX^\top \right) \Sigma_\circ^{1/2} \right\|_{\mathrm{F}}$$

where $\Sigma_\circ = \mathbb{E}_{\mathbb{P}_\circ}[(\xi - \mathbb{E}_{\mathbb{P}_\circ}[\xi])(\xi - \mathbb{E}_{\mathbb{P}_\circ}[\xi])^\top] \in \mathbb{R}^{d \times d}$ is the covariance matrix of ξ under the nominal distribution $\mathbb{P}_\circ$. It is evident that problem $(\mathrm{P_m})$ is significantly easier to solve than problem $(\mathrm{P_{mM}})$ in its original min-max form. Therefore, we turn our attention to developing efficient algorithms for tackling problem $(\mathrm{P_m})$.

There have been substantial advancements in nonsmooth optimization on Riemannian manifolds in recent decades, with the majority of efforts directed toward locally Lipschitz continuous functions. During this period, a diverse range of algorithms has emerged, including subgradient-oriented approaches [29, 30, 38], proximal point methods [3, 9], primal-dual frameworks [37, 59], proximal gradient algorithms

[10, 31, 54], smoothing approaches [2, 49, 51], and so on. However, due to the presence of two possibly nonsmooth terms, none of the existing algorithms can be deployed to solve problem ($\mathrm{P_m}$) efficiently. In particular, the last term in problem ($\mathrm{P_m}$) poses a formidable challenge, as its proximal operator lacks a closed-form solution and entails costly computations. Since it can be represented by a composition of a smooth mapping and a convex function, we may resort to the proximal linear algorithm [54] for handling this issue. However, this approach requires calculating the Jacobian of $(I_d - XX^\top)\Sigma_\circ^{1/2}$, which involves the square root of $\Sigma_\circ$. And the presence of another nonsmooth term renders the subproblem particularly difficult to tackle. The Riemannian subgradient method proposed in [38] is capable of directly solving problem ($\mathrm{P_m}$). Nevertheless, owing to its reliance solely on subgradient information, it suffers from slow convergence with an iteration complexity of $O(\epsilon^{-4})$. Consequently, solving problem ($\mathrm{P_m}$) remains a challenging endeavor.

Within the scope of this work, proximal gradient methods share the closest theoretical and methodological connection with the present study. This class of algorithms is initially proposed in [10] on the Stiefel manifold, laying the foundation for a plethora of subsequent advancements [11]. Later on, Huang and Wei [31] extend this framework to general Riemannian manifolds. The crux of these approaches involves solving a proximal gradient subproblem on the tangent space. Another related line of work is the dynamic smoothing algorithm introduced by Beck and Rosset [2], which aims to solve a class of composite optimization problems based on Moreau envelopes. Although this smoothing-based perspective is related, our problem structure and algorithmic treatment are different.

In this paper, we present a comprehensive analysis of the DRO model ($\mathrm{P_{mM}}$) for PCA. The main contributions are summarized as follows.

• Our investigation reveals that, the optimal value of the inner maximization problem in ($\mathrm{P_{mM}}$) diverges to infinity when $1 \le p < 2$, whereas for $p = 2$, it admits a closed-form expression. Accordingly, we shift our attention to the case $p = 2$ in the subsequent analysis. This explicit characterization facilitates an equivalent reformulation of the original DRO model ($\mathrm{P_{mM}}$) into a minimization form ($\mathrm{P_m}$). A particularly intriguing discovery is that, for the classical PCA problem without regularizers, its distributionally robust counterpart is equivalent to the nominal model, uncovering an unexpected equivalence in this specific context.

• We design a novel smoothing method to solve a class of composite optimization problems on the Stiefel manifold. Our theoretical analysis elucidates the Riemannian gradient consistency for the smoothing function and establishes the global convergence of our algorithm to a stationary point. Furthermore, we provide the iteration complexity of $O(\epsilon^{-3})$ required to achieve an ϵ-approximate stationary point.

• Last but not least, preliminary experimental results present the numerical performance of the proposed algorithm, underscoring its practical viability. Moreover, these results validate the necessity and rationality of adopting the DRO model for PCA, demonstrating its superiority in addressing distributional uncertainty and enhancing solution robustness on three real-world datasets.

The rest of this paper proceeds as follows. Section 2 draws into some preliminaries of Riemannian optimization. In Section 3, we make a profound study on the DRO model of PCA and derive its equivalent reformulation. Section 4 discusses the stationarity condition and develops the smoothing algorithm. Its convergence analysis and iteration complexity are provided in Section 5. Numerical results are presented in Section 6. Finally, concluding remarks are given in Section 7.

2. Preliminaries. In this section, we introduce the notations and concepts used throughout this paper.

2.1. Basic Notations. We use $\mathbb{R}$ and $\mathbb{N}$ to denote the sets of real and natural numbers, respectively. And the notations $\mathbb{R}_+$ and $\mathbb{R}_{++}$ represent the sets of nonnegative and positive real numbers, respectively. The Euclidean inner product of two matrices Y_1, Y_2 with the same size is defined as $\langle Y_1, Y_2\rangle = \mathrm{tr}(Y_1^\top Y_2)$, where $\mathrm{tr}(B)$ stands for the trace of a square matrix B. And the notation $I_r \in \mathbb{R}^{r\times r}$ represents the $r \times r$ identity matrix. The Frobenius norm of a given matrix C is denoted by $\|C\|_{\mathrm{F}}$. We denote by $\mathbb{S}_+^d$ and $\mathbb{S}_{++}^d$ the spaces of symmetric positive semidefinite matrices and symmetric positive definite matrices in $\mathbb{R}^{d\times d}$, respectively. The notation $B^{1/2}$ stands for the square root of a symmetric positive semidefinite matrix B.

2.2. Riemannian Gradients and Clarke Subdifferentials. Let $\mathcal{M}$ be a complete submanifold embedded in $\mathbb{R}^{d\times r}$. For each point $X \in \mathcal{M}$, the tangent space to $\mathcal{M}$ at X is referred to as $\mathcal{T}_X\mathcal{M}$. In this paper, we consider the Riemannian metric $\langle\cdot,\cdot\rangle_X$ on $\mathcal{T}_X\mathcal{M}$ that is induced from the Euclidean inner product $\langle\cdot,\cdot\rangle$, i.e., $\langle V_1, V_2\rangle_X = \langle V_1, V_2\rangle = \mathrm{tr}(V_1^\top V_2)$ for any $V_1, V_2 \in \mathcal{T}_X\mathcal{M}$. The tangent bundle of $\mathcal{M}$ is denoted by $\mathcal{T}\mathcal{M} = \{(X, V) \mid X \in \mathcal{M}, V \in \mathcal{T}_X\mathcal{M}\}$, that is, the disjoint union of the tangent spaces of $\mathcal{M}$. Additionally, we use the notation $\mathrm{Proj}_{\mathcal{T}_X\mathcal{M}}(\cdot)$ to represent the orthogonal projection operator onto $\mathcal{T}_X\mathcal{M}$. In the context of the Stiefel manifold, the tangent space at $X \in \mathcal{O}^{d,r}$ can be described as $\mathcal{T}_X\mathcal{O}^{d,r} = \{D \in \mathbb{R}^{d\times r} \mid X^\top D + D^\top X = 0\}$, and its orthogonal projection operator is given by $\mathrm{Proj}_{\mathcal{T}_X\mathcal{O}^{d,r}}(V) = V - X(X^\top V + V^\top X)/2$ for any $V \in \mathbb{R}^{d\times r}$.

For a smooth function f, the Riemannian gradient at $X \in \mathcal{M}$, denoted by $\mathrm{grad}\, f(X)$, is defined as the unique element of $\mathcal{T}_X\mathcal{M}$ satisfying

$$\langle \mathrm{grad}\, f(X), V\rangle = \mathrm{D}f(X)[V], \quad \forall V \in \mathcal{T}_X\mathcal{M},$$

where $\mathrm{D}f(X)[V]$ is the directional derivative of f along the direction V at the point X. Since f is defined on an embedded submanifold in the Euclidean space, its Riemannian gradient can be computed by $\mathrm{grad}\, f(X) = \mathrm{Proj}_{\mathcal{T}_X\mathcal{M}}(\nabla f(X))$.

For a locally Lipschitz continuous function on the manifold, the Riemannian Clarke subdifferential has been intensively studied and used in the literature [58, 30, 29], which is a natural extension of the Clarke subdifferential [16, 44] in the Euclidean space. Throughout this paper, we adopt the following definition for the Riemannian Clarke subdifferential.

DEFINITION 2.1 ([30]). *Suppose that $f : \mathcal{M} \to \mathbb{R}$ is a locally Lipschitz continuous function. Let $\Omega_{\mathrm{R}}(f) = \{X \in \mathcal{M} \mid f \text{ is differentiable at } X\}$. Then the Riemannian Clarke subdifferential of f at $X \in \mathcal{M}$ is defined as*

$$\partial_{\mathrm{R}} f(X) = \mathrm{conv}\left\{D \in \mathcal{T}_X\mathcal{M} \mid \mathrm{grad}\, f(X_t) \to D, \Omega_{\mathrm{R}}(f) \ni X_t \to X\right\}.$$

2.3. Retractions. In contrast to the Euclidean setting, the point $X + V$ does not lie in the manifold in general for $X \in \mathcal{M}$ and $V \in \mathcal{T}_X\mathcal{M}$, due to the absence of a linear structure in $\mathcal{M}$. The interplay between $\mathcal{M}$ and $\mathcal{T}_X\mathcal{M}$ is typically carried out via the exponential mappings, which are usually computationally intensive to evaluate in practice. As an alternative, the concept of retraction, a first-order approximation of the exponential mapping, is proposed in the literature [1, 7] to alleviate the heavy computational burden.

DEFINITION 2.2 ([1]). *A retraction on a manifold $\mathcal{M}$ is a smooth mapping $\mathfrak{R} : \mathcal{T}\mathcal{M} \to \mathcal{M}$, and the restriction of $\mathfrak{R}$ to $\mathcal{T}_X\mathcal{M}$, denoted by $\mathfrak{R}_X$, satisfies the following two properties for any $X \in \mathcal{M}$.*

(i) It holds that $\mathfrak{R}_X(0_X) = X$, where 0_X is the zero vector in $\mathcal{T}_X\mathcal{M}$.

(ii) The differential of $\mathfrak{R}_X$ at 0_X is an identity map on $\mathcal{T}_X\mathcal{M}$.

By leveraging the retraction $\mathfrak{R}_X(V)$, we can obtain a point by moving away from $X \in \mathcal{M}$ along the direction $V \in \mathcal{T}_X\mathcal{M}$, while remaining on the manifold. In this way, it defines an update rule to preserve the feasibility. Following the proof of Lemma 2.7 in [8], we know that the retraction satisfies the following properties.

LEMMA 2.3 ([8]). *Let $\mathcal{M}$ be a compact embedded submanifold of a Euclidean space. Then there exist two constants $M_1 > 0$ and $M_2 > 0$ such that the following two relationships hold,*

$$\begin{cases} \|\mathfrak{R}_X(V) - X\|_{\mathrm{F}} \le M_1 \|V\|_{\mathrm{F}}, \\ \|\mathfrak{R}_X(V) - (X+V)\|_{\mathrm{F}} \le M_2 \|V\|_{\mathrm{F}}^2, \end{cases}$$

for any $X \in \mathcal{M}$ and $V \in \mathcal{T}_X\mathcal{M}$.

There are various practical realizations of retractions on the Stiefel manifold, such as QR factorization, polar decomposition and Cayley transformation. We refer interested readers to [1, 7, 55] for more details.

3. Model Analysis. The DRO model ($\mathrm{P_{mM}}$) of PCA is investigated in this section. We focus on the inner maximization problem in ($\mathrm{P_{mM}}$) as follows,

$$\begin{aligned} \varphi(X) &:= \sup_{\mathbb{P}\in\mathscr{B}_p(\mathbb{P}_\circ,\rho)} \mathbb{E}_{\mathbb{P}}\left[\left\|\left(I_d - XX^\top\right)(\xi - \mathbb{E}_{\mathbb{P}}[\xi])\right\|_{\mathrm{F}}^2\right] \\ &= \sup_{\mathbb{P}\in\mathscr{B}_p(\mathbb{P}_\circ,\rho)} \left\{\mathbb{E}_{\mathbb{P}}\left[\mathrm{tr}\left(\left(I_d - XX^\top\right)\left(\xi\xi^\top - \mathbb{E}_{\mathbb{P}}[\xi]\,\mathbb{E}_{\mathbb{P}}[\xi]^\top\right)\right)\right]\right\}. \end{aligned} \tag{3.1}$$

The objective function of (3.1) is quadratic in the underlying distribution $\mathbb{P}$ rather than linear, and hence, existing reformulations of Wasserstein DRO problems [6, 15, 20, 22, 60] are not applicable. To navigate this challenge, we introduce an auxiliary variable $\eta \in \mathbb{R}^d$ and propose the following splitting formulation of (3.1),

$$\begin{aligned} \sup_{\eta\in\mathscr{M}_p(\mathbb{P}_\circ,\rho)} \sup_{\mathbb{P}\in\mathscr{Q}_p} \; & \mathbb{E}_{\mathbb{P}}\left[\mathrm{tr}\left(\left(I_d - XX^\top\right)\xi\xi^\top\right)\right] - \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right) \\ \text{s.t.} \; & \mathbb{W}_p(\mathbb{P},\mathbb{P}_\circ) \le \rho, \quad \mathbb{E}_{\mathbb{P}}[\xi] = \eta, \end{aligned} \tag{3.2}$$

where $\mathscr{M}_p(\mathbb{P}_\circ,\rho) := \{\eta = \mathbb{E}_{\mathbb{P}}[\xi] \mid \mathbb{W}_p(\mathbb{P},\mathbb{P}_\circ) \le \rho\}$ collects all mean vectors of distributions in the ambiguity set $\mathscr{B}_p(\mathbb{P}_\circ,\rho)$. The constraint $\eta \in \mathscr{M}_p(\mathbb{P}_\circ,\rho)$ is imposed to avoid an empty feasible region and to ensure the well-posedness of problem (3.2). For $\tau \ge 0$, $\eta \in \mathscr{M}_p(\mathbb{P}_\circ,\rho)$, and $X \in \mathcal{O}^{d,r}$, we define

$$\psi(\tau \mid \eta, X) := \sup_{\mathbb{P}\in\mathscr{Q}_p} \left\{\mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] \;\middle|\; \mathbb{W}_p^p(\mathbb{P},\mathbb{P}_\circ) \le \tau, \mathbb{E}_{\mathbb{P}}[\xi] = \eta\right\}, \tag{3.3}$$

where $\omega_X(\xi) := \mathrm{tr}\left(\left(I_d - XX^\top\right)\xi\xi^\top\right)$. Then it holds that

$$\varphi(X) = \sup_{\eta\in\mathscr{M}_p(\mathbb{P}_\circ,\rho)} \left\{\psi(\rho^p \mid \eta, X) - \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right)\right\}. \tag{3.4}$$

Based on the preceding constructions, the objective function of the outer minimization problem in the Wasserstein DRO model ($\mathrm{P_{mM}}$) can be expressed as $\varphi(X) + s(X)$.

3.1. Dual Representation. In this subsection, we aim to derive the dual representation of $\psi(\tau \mid \eta, X)$ defined in (3.3). This part of analysis follows the idea of Zhang et al. [60] based on the Legendre transform [43]. We generalize existing results by handling an additional equality constraint $\mathbb{E}_{\mathbb{P}}[\xi] = \eta$. Although our focus is primarily on PCA, the techniques we propose can be naturally extended to more general settings.

The following lemma reveals some useful properties of the function $\psi(\cdot \mid \eta, X)$ for fixed $\eta \in \mathscr{M}_p(\mathbb{P}_\circ, \rho)$ and $X \in \mathcal{O}^{d,r}$.

LEMMA 3.1. *For fixed $\eta \in \mathscr{M}_p(\mathbb{P}_\circ, \rho)$ and $X \in \mathcal{O}^{d,r}$, the function $\psi(\cdot \mid \eta, X)$ is bounded from below, monotonically increasing, and concave on $\mathbb{R}_+$.*

Proof. The proof follows along the same lines as that of [60, Lemma 1] and is therefore omitted for brevity. ∎

For a function $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, we denote by $h^\diamond : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ its Legendre transform $h^\diamond(\lambda) := \sup_{\tau \in \mathbb{R}} \{\lambda\tau - h(\tau)\}$. Then the dual representation of $\psi(\tau \mid \eta, X)$ can be constructed by resorting to the Legendre transform.

THEOREM 3.2. *For any $p \in [1, 2]$ and $\tau > 0$, it holds that*

$$\psi(\tau \mid \eta, X) = \inf_{\lambda \geq 0, \varsigma \in \mathbb{R}^d} \left\{ \lambda\tau + \varsigma^\top \eta + \mathbb{E}_{\xi_\circ \sim \mathbb{P}_\circ} \left[\sup_{\xi \in \mathbb{R}^d} \bar{\omega}_X(\xi, \xi_\circ) \right] \right\}, \tag{3.5}$$

where $\bar{\omega}_X(\xi, \xi_\circ) := \omega_X(\xi) - \lambda \|\xi - \xi_\circ\|_p^p - \varsigma^\top \xi$.

Proof. For notational simplicity, we use $\psi(\cdot)$ in place of $\psi(\cdot \mid \eta, X)$ throughout this proof. Let $\lambda \in \mathbb{R}$. If $\lambda < 0$, we have $(-\psi)^\diamond(-\lambda) = \sup_{\tau \geq 0} \{-\lambda\tau + \psi(\tau)\} \geq \sup_{\tau \geq 0} \{-\lambda\tau + \psi(0)\} = +\infty$. Then our focus is on the case where $\lambda \geq 0$. Taking the Legendre transform of $-\psi$ leads to that

$$\begin{aligned}
(-\psi)^\diamond(-\lambda) &= \sup_{\tau \geq 0} \{-\lambda\tau + \psi(\tau)\} \\
&= \sup_{\tau \geq 0} \sup_{\mathbb{P} \in \mathscr{Q}_p} \left\{ \mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] - \lambda\tau \;\middle|\; \mathbb{W}_p^p(\mathbb{P}, \mathbb{P}_\circ) \leq \tau, \mathbb{E}_{\mathbb{P}}[\xi] = \eta \right\} \\
&= \sup_{\mathbb{P} \in \mathscr{Q}_p} \sup_{\tau \geq 0} \left\{ \mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] - \lambda\tau \;\middle|\; \mathbb{W}_p^p(\mathbb{P}, \mathbb{P}_\circ) \leq \tau, \mathbb{E}_{\mathbb{P}}[\xi] = \eta \right\} \\
&= \sup_{\mathbb{P} \in \mathscr{Q}_p} \left\{ \mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] - \lambda \mathbb{W}_p^p(\mathbb{P}, \mathbb{P}_\circ) \;\middle|\; \mathbb{E}_{\mathbb{P}}[\xi] = \eta \right\}.
\end{aligned}$$

According to Definition 1.1, it follows that

$$\begin{aligned}
(-\psi)^\diamond(-\lambda) &= \sup_{\mathbb{P} \in \mathscr{Q}_p} \left\{ \mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] - \lambda \inf_{\mathbb{Q} \in \bar{\mathscr{J}}(\mathbb{P}, \mathbb{P}_\circ)} \mathbb{E}_{(\xi, \xi_\circ) \sim \mathbb{Q}} \left[\|\xi - \xi_\circ\|_p^p \right] \;\middle|\; \mathbb{E}_{\mathbb{P}}[\xi] = \eta \right\} \\
&= \sup_{\mathbb{P} \in \mathscr{Q}_p, \mathbb{Q} \in \bar{\mathscr{J}}(\mathbb{P}, \mathbb{P}_\circ)} \left\{ \mathbb{E}_{\mathbb{P}}[\omega_X(\xi)] - \lambda \mathbb{E}_{(\xi, \xi_\circ) \sim \mathbb{Q}} \left[\|\xi - \xi_\circ\|_p^p \right] \;\middle|\; \mathbb{E}_{\mathbb{P}}[\xi] = \eta \right\} \\
&= \sup_{\mathbb{Q} \in \bar{\mathscr{J}}(\mathbb{P}_\circ)} \left\{ \mathbb{E}_{(\xi, \xi_\circ) \sim \mathbb{Q}} \left[\omega_X(\xi) - \lambda \|\xi - \xi_\circ\|_p^p \right] \;\middle|\; \mathbb{E}_{(\xi, \xi_\circ) \sim \mathbb{Q}}[\xi] = \eta \right\},
\end{aligned} \tag{3.6}$$

where $\bar{\mathscr{J}}(\mathbb{P}_\circ)$ stands for the set containing all the joint distributions of ξ and $\xi_\circ$ with second marginal $\mathbb{P}_\circ$. It is clear that both the objective function and the constraint of the last problem in (3.6) are linear with respect to $\mathbb{Q}$. Following the discussions in [45, Section 3] and [57, Section 2], we interpret it as a special instance of the problem of

moments. As a direct consequence of [57, Proposition 2.1], the Slater condition holds, and hence, the strong duality [45, Proposition 3.4] prevails. Hence, we can proceed to show that

$$\begin{aligned}(-\psi)^\diamond(-\lambda) &= \sup_{\mathbb{Q}\in\bar{\mathscr{F}}(\mathbb{P}_\circ)} \left\{ \mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}\left[\omega_X(\xi) - \lambda\left\|\xi-\xi_\circ\right\|_p^p\right] \;\middle|\; \mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}[\xi] = \eta \right\} \\ &= \inf_{\varsigma\in\mathbb{R}^d} \left\{ \varsigma^\top\eta + \sup_{\mathbb{Q}\in\bar{\mathscr{F}}(\mathbb{P}_\circ)} \mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}\left[\omega_X(\xi) - \lambda\left\|\xi-\xi_\circ\right\|_p^p - \varsigma^\top\xi\right] \right\} \\ &= \inf_{\varsigma\in\mathbb{R}^d} \left\{ \varsigma^\top\eta + \sup_{\mathbb{Q}\in\bar{\mathscr{F}}(\mathbb{P}_\circ)} \mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}\left[\bar{\omega}_X(\xi,\xi_\circ)\right] \right\},\end{aligned}$$

where $\varsigma \in \mathbb{R}^d$ is the Lagrangian multiplier associated with the equality constraint $\mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}[\xi] = \eta$. Lemma 3.1 illustrates that ψ is bounded from below, monotonically increasing, and concave on $\mathbb{R}_+$. Hence, either $\psi(\tau) < +\infty$ for all $\tau \geq 0$ or $\psi(\tau) = +\infty$ for all $\tau > 0$. In the former case, by invoking the result of [43, Theorem 12.2], we can obtain that

$$\begin{aligned}\psi(\tau) &= -(-\psi)^{\diamond\diamond}(\tau) = -\sup_{\lambda\in\mathbb{R}}\left\{-\lambda\tau - (-\psi)^\diamond(-\lambda)\right\} = \inf_{\lambda\geq 0}\left\{\lambda\tau + (-\psi)^\diamond(-\lambda)\right\} \\ &= \inf_{\lambda\geq 0,\varsigma\in\mathbb{R}^d} \left\{ \lambda\tau + \varsigma^\top\eta + \sup_{\mathbb{Q}\in\bar{\mathscr{F}}(\mathbb{P}_\circ)} \mathbb{E}_{(\xi,\xi_\circ)\sim\mathbb{Q}}\left[\bar{\omega}_X(\xi,\xi_\circ)\right] \right\}, \qquad (3.7)\end{aligned}$$

for all $\tau > 0$. In the latter case, it holds that $(-\psi)^\diamond(-\lambda) = +\infty$ for all $\lambda \geq 0$, which indicates that the relationship (3.7) is also valid. According to [60, Proposition 2], the function $\bar{\omega}_X$ satisfies the interchangeability principle. Then it follows that

$$\psi(\tau) = \inf_{\lambda\geq 0,\varsigma\in\mathbb{R}^d} \left\{ \lambda\tau + \varsigma^\top\eta + \mathbb{E}_{\xi_\circ\sim\mathbb{P}_\circ}\left[\sup_{\xi\in\mathbb{R}^d} \bar{\omega}_X(\xi,\xi_\circ)\right] \right\}.$$

The proof is completed. ∎

Theorem 3.2 also implies a remarkable result that the optimal value of the inner maximization problem in the DRO model ($\mathrm{P_{mM}}$) is always infinite for any $p \in [1, 2)$ and $\rho > 0$.

Remark 3.3. Suppose that $p \in [1, 2)$ and $\rho > 0$. Then we have

$$\sup_{\xi\in\mathbb{R}^d} \bar{\omega}_X(\xi,\xi_\circ) = \sup_{\xi\in\mathbb{R}^d} \left\{ \mathrm{tr}\left(\left(I_d - XX^\top\right)\xi\xi^\top\right) - \lambda\left\|\xi-\xi_\circ\right\|_p^p - \varsigma^\top\xi \right\} = +\infty,$$

which, together with Theorem 3.2, implies that $\psi(\rho^p \mid \eta, X) = +\infty$. From the relationship (3.4), it can be deduced that $\varphi(X) = +\infty$ for all $X \in \mathcal{O}^{d,r}$.

3.2. Equivalent Reformulation for $p = 2$. This subsection is devoted to deriving the equivalent reformulation ($\mathrm{P_m}$) of the DRO model ($\mathrm{P_{mM}}$) for the specific situation where $p = 2$. To this end, we establish that the optimal value $\varphi(X)$ of problem (3.1) admits a closed-form formulation.

The following lemma first shows that the supremum $\sup\{\bar{\omega}_X(\xi,\xi_\circ) \mid \xi \in \mathbb{R}^d\}$ in the dual representation (3.5) of $\psi(\tau \mid \eta, X)$ can be explicitly computed.

LEMMA 3.4. *Suppose that $p = 2$ and $\rho > 0$. If $\lambda > 1$, it holds that*

$$\mathbb{E}_{\xi_\circ \sim \mathbb{P}_\circ}\left[\sup_{\xi \in \mathbb{R}^d} \bar{\omega}_X(\xi, \xi_\circ)\right] = \frac{\lambda}{\lambda - 1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}_\circ}\left[\xi_\circ \xi_\circ^\top\right]\right) + \theta_X(\lambda, \varsigma) - \varsigma^\top \eta,$$

where the function θ_X is defined as

$$\theta_X(\lambda, \varsigma) = \frac{1}{4\lambda(\lambda - 1)}\varsigma^\top\left(\lambda I_d - XX^\top\right)\varsigma + \varsigma^\top\left(\eta - \frac{1}{\lambda - 1}\left(\lambda I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}_\circ}\left[\xi_\circ\right]\right).$$

Moreover, if $\lambda \in [0, 1]$, we have

$$\mathbb{E}_{\xi_\circ \sim \mathbb{P}_\circ}\left[\sup_{\xi \in \mathbb{R}^d} \bar{\omega}_X(\xi, \xi_\circ)\right] = +\infty.$$

Proof. Straightforward calculations yield that

$$\begin{aligned}\bar{\omega}_X(\xi, \xi_\circ) &= \omega_X(\xi) - \lambda\left\|\xi - \xi_\circ\right\|_2^2 - \varsigma^\top \xi \\ &= -\xi^\top\left(\left(\lambda - 1\right) I_d + XX^\top\right)\xi + \left(2\lambda\xi_\circ - \varsigma\right)^\top \xi - \lambda\xi_\circ^\top \xi_\circ,\end{aligned}$$

which is a quadratic function with respect to $\xi \in \mathbb{R}^d$ for fixed $\xi_\circ \in \mathbb{R}^d$. We move on to investigate the following three cases.

Case I: $\lambda \in [0, 1)$. Since $r < d$, the matrix $(\lambda - 1)I_d + XX^\top$ has at least one negative eigenvalue $\lambda - 1 < 0$ associated with an eigenvector $z \in \mathbb{R}^d$ with $z^\top z = 1$. Then for any $t \in \mathbb{R}$, we have

$$\bar{\omega}_X(tz, \xi_\circ) = t^2(1 - \lambda) + t\left(2\lambda\xi_\circ - \varsigma\right)^\top z - \lambda\xi_\circ^\top \xi_\circ.$$

As a result of $\lambda \in [0, 1)$, it holds that $\sup_{t \in \mathbb{R}} \bar{\omega}_X(tz, \xi_\circ) = +\infty$, which further implies that

$$\sup_{\xi \in \mathbb{R}^d} \bar{\omega}_X(\xi, \xi_\circ) = +\infty.$$

Case II: $\lambda = 1$. It is clear that there exists $z \in \mathbb{R}^d$ with $z^\top z = 1$ such that $X^\top z = 0$ as $r < d$. Thus, it holds for all $t \in \mathbb{R}$ that

$$\bar{\omega}_X(tz, \xi_\circ) = t\left(2\xi_\circ - \varsigma\right)^\top z - \xi_\circ^\top \xi_\circ.$$

By simple manipulations, we arrive at

$$\sup_{t \in \mathbb{R}} \bar{\omega}_X(tz, \xi_\circ) = \begin{cases} -\xi_\circ^\top \xi_\circ, & \text{if } \left(2\xi_\circ - \varsigma\right)^\top z = 0, \\ +\infty, & \text{otherwise.}\end{cases}$$

Consequently, it can be readily verified that

$$\mathbb{E}_{\xi_\circ \sim \mathbb{P}_\circ}\left[\sup_{\xi \in \mathbb{R}^d} \bar{\omega}_X(\xi, \xi_\circ)\right] \geq \mathbb{E}_{\xi_\circ \sim \mathbb{P}_\circ}\left[\sup_{t \in \mathbb{R}} \bar{\omega}_X(tz, \xi_\circ)\right] = +\infty.$$

Case III: $\lambda > 1$. In this case, the matrix $(\lambda - 1)I_d + XX^\top$ is positive definite. Then $\bar{\omega}_X(\xi, \xi_\circ)$ is strictly concave with respect to $\xi \in \mathbb{R}^d$ for fixed $\xi_\circ \in \mathbb{R}^d$. Hence, we can

proceed to show that

$$\sup_{\xi\in\mathbb{R}^d} \bar{\omega}_X(\xi,\xi_\circ) = \left(\lambda\xi_\circ - \frac{1}{2}\varsigma\right)^\top \left((\lambda-1)\, I_d + XX^\top\right)^{-1} \left(\lambda\xi_\circ - \frac{1}{2}\varsigma\right) - \lambda\xi_\circ^\top\xi_\circ,$$

where the supremum is attained at $\xi = ((\lambda-1)I_d + XX^\top)^{-1}(\lambda\xi_\circ - \varsigma/2)$. Moreover, according to the Sherman–Morrison–Woodbury formula [28, page 329], we have

$$\left((\lambda-1)I_d + XX^\top\right)^{-1} = \frac{1}{\lambda-1} I_d - \frac{1}{\lambda(\lambda-1)} XX^\top.$$

Then a straightforward verification reveals that

$$\begin{aligned}\sup_{\xi\in\mathbb{R}^d} \bar{\omega}_X(\xi,\xi_\circ) = {} & \frac{\lambda}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\xi_\circ\xi_\circ^\top\right) - \frac{1}{\lambda-1}\varsigma^\top\left(\lambda I_d - XX^\top\right)\xi_\circ \\ & + \frac{1}{4\lambda(\lambda-1)}\varsigma^\top\left(\lambda I_d - XX^\top\right)\varsigma,\end{aligned}$$

which completes the proof. ∎

Leveraging the result of the previous lemma, we proceed to prove that $\psi(\rho^2 \mid \eta, X)$ can be expressed in an explicit form. Recall that $\Sigma_\circ = \mathbb{E}_{\mathbb{P}_\circ}[(\xi - \mathbb{E}_{\mathbb{P}_\circ}[\xi])(\xi - \mathbb{E}_{\mathbb{P}_\circ}[\xi])^\top]$ is the covariance matrix of ξ under the nominal distribution $\mathbb{P}_\circ$.

LEMMA 3.5. *Suppose that $p = 2$ and $\rho > 0$. Then, for any $\eta \in \mathscr{M}_2(\mathbb{P}_\circ, \rho)$ and $X \in \mathcal{O}^{d,r}$, it holds that*

$$\begin{aligned}\psi(\rho^2 \mid \eta, X) = {} & \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \left(\rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2\right)^{1/2}\right)^2 \\ & + \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right).\end{aligned}$$

Proof. Based on Lemma 3.4, we can restrict our discussion to the case where $\lambda > 1$. Moreover, it follows from Theorem 3.2 that

$$\psi(\rho^2 \mid \eta, X) = \inf_{\lambda>1}\left\{\lambda\rho^2 + \frac{\lambda}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}_\circ}\left[\xi_\circ\xi_\circ^\top\right]\right) + \inf_{\varsigma\in\mathbb{R}^d}\theta_X(\lambda,\varsigma)\right\}.$$

It is clear that $\theta_X(\lambda,\varsigma)$ is a quadratic function with respect to $\varsigma \in \mathbb{R}^d$ for fixed $\lambda > 1$. Since $\lambda I_d - XX^\top$ is positive definite, it holds that

$$\begin{aligned}\inf_{\varsigma\in\mathbb{R}^d}\theta_X(\lambda,\varsigma) = {} & -\frac{\lambda}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\,\mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]^\top\right) \\ & + \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right) - \lambda\,\|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2,\end{aligned}$$

where the infimum is attained at $\varsigma = 2\lambda\mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ] - 2\left((\lambda-1)\, I_d + XX^\top\right)\eta$. By simple calculations, we can obtain that

$$\begin{aligned}& \lambda\rho^2 + \frac{\lambda}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}_\circ}\left[\xi_\circ\xi_\circ^\top\right]\right) + \inf_{\varsigma\in\mathbb{R}^d}\theta_X(\lambda,\varsigma) \\ = {} & (\lambda-1)\left(\rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2\right) + \frac{1}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right) \\ & + \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right) + \rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2 + \mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right).\end{aligned}$$

For any $\eta \in \mathscr{M}_2(\mathbb{P}_\circ, \rho)$, there exists $\mathbb{P} \in \mathscr{B}_2(\mathbb{P}_\circ, \rho)$ such that $\eta = \mathbb{E}_{\mathbb{P}}[\xi]$. It follows from [23, Theorem 2.1] that $\|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2 \leq \mathbb{W}_2(\mathbb{P}, \mathbb{P}_\circ) \leq \rho$. Then it can be readily verified that

$$\begin{aligned}\psi(\rho^2 \mid \eta, X) = & \inf_{\lambda>1}\left\{(\lambda-1)\left(\rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2\right) + \frac{1}{\lambda-1}\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right\} \\ & + \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right) + \rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2 + \mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right) \\ = & \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \left(\rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2\right)^{1/2}\right)^2 \\ & + \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right).\end{aligned}$$

We complete the proof. ∎

We are now in a position to derive an explicit expression for $\varphi(X)$ based on the relationship (3.4), as established in the following theorem.

THEOREM 3.6. *For any $X \in \mathcal{O}^{d,r}$, $\mathbb{P}_\circ \in \mathscr{Q}_2$, and $\rho \geq 0$, the optimal value of problem* (3.1) *with $p = 2$ has the following explicit expression,*

$$\varphi(X) = \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho\right)^2. \tag{3.8}$$

Proof. We first consider the case where $\rho = 0$. Then the feasible region $\mathscr{B}_2(\mathbb{P}_\circ, 0)$ of problem (3.1) collapses to the singleton $\{\mathbb{P}_\circ\}$. As a result, we have

$$\varphi(X) = \mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right),$$

which indicates that the relationship (3.8) holds for $\rho = 0$.

Next, our focus is on the case where $\rho > 0$. As a direct consequence of Lemma 3.5, we can proceed to show that

$$\begin{aligned}\varphi(X) = & \sup_{\eta \in \mathscr{M}_2(\mathbb{P}_\circ, \rho)} \left\{\psi(\rho^2 \mid \eta, X) - \mathrm{tr}\left(\left(I_d - XX^\top\right)\eta\eta^\top\right)\right\} \\ = & \sup_{\eta \in \mathscr{M}_2(\mathbb{P}_\circ, \rho)} \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \left(\rho^2 - \|\eta - \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]\|_2^2\right)^{1/2}\right)^2 \\ = & \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho\right)^2,\end{aligned}$$

where the supremum is attained at $\eta = \mathbb{E}_{\mathbb{P}_\circ}[\xi_\circ]$. The proof is completed. ∎

By expanding the square term in (3.8), it can be obtained that

$$\varphi(X) = \mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right) + 2\rho\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho^2.$$

For any $X \in \mathcal{O}^{d,r}$, we have

$$\begin{aligned}\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} = & \left(\mathrm{tr}\left(\Sigma_\circ^{1/2}\left(I_d - XX^\top\right)\left(I_d - XX^\top\right)\Sigma_\circ^{1/2}\right)\right)^{1/2} \\ = & \left\|\left(I_d - XX^\top\right)\Sigma_\circ^{1/2}\right\|_{\mathrm{F}}.\end{aligned}$$

Based on this representation and Theorem 3.6, the DRO model ($\mathrm{P_{mM}}$) with $p = 2$ can be equivalently reformulated as problem ($\mathrm{P_m}$).

Remark 3.7. The equivalent reformulation we derive for the case $p = 2$ in Theorem 3.6 is similar to that presented in [47, Proposition 3.3] with a moment-based ambiguity set. The distinction of the present work is threefold. First, our ambiguity set is defined directly through the type-2 Wasserstein distance over probability measures, and the reformulation is derived from the original min-max problem without reducing the ambiguity description to moment information. Second, the nominal distribution $\mathbb{P}_\circ$ is allowed to be a general distribution in $\mathscr{Q}_2$, rather than being restricted to a discrete empirical distribution in [47]. Third, the splitting argument in (3.2) handles the nonlinear dependence of the objective on $\mathbb{P}$ and also clarifies why the model becomes ill-posed for $p < 2$. Finally, we note that the proposed splitting technique can be seamlessly generalized to broader settings beyond PCA.

We end this section by demonstrating that, the solutions of classical PCA without regularizers inherently possess robustness against data perturbations, as characterized by the type-2 Wasserstein distance.

COROLLARY 3.8. *Suppose that $p = 2$ and $s(X) = 0$ for all $X \in \mathcal{O}^{d,r}$. Then for any $\Sigma_\circ \in \mathbb{S}^d_+$ and $\rho \geq 0$, the global minimizers of problem* ($\mathrm{P_{mM}}$) *coincide with those of the nominal model* $\min_{X\in\mathcal{O}^{d,r}} \operatorname{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)$ *for PCA.*

Proof. According to Theorem 3.6, we know that the DRO model ($\mathrm{P_{mM}}$) is equivalent to the problem of minimizing φ over $\mathcal{O}^{d,r}$. Let $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous function defined by $\kappa(t) = (\sqrt{t} + \rho)^2$. Then κ is monotonically increasing on $\mathbb{R}_+$ for all $\rho \geq 0$. Consequently, the nonnegativity of $\operatorname{tr}((I_d - XX^\top)\Sigma_\circ)$ results in that

$$\operatorname*{arg\,min}_{X\in\mathcal{O}^{d,r}} \varphi(X) = \operatorname*{arg\,min}_{X\in\mathcal{O}^{d,r}} \kappa\left(\operatorname{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right) = \operatorname*{arg\,min}_{X\in\mathcal{O}^{d,r}} \operatorname{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right),$$

which completes the proof. ∎

4. Algorithm Design. The purpose of this section is to devise an efficient algorithm to solve the equivalent reformulation ($\mathrm{P_m}$) of the DRO model ($\mathrm{P_{mM}}$) with $p = 2$ based on the smoothing approximation technique. We consider a broader class of nonsmooth optimization problems of the following form,

$$\min_{X\in\mathcal{O}^{d,r}} \quad f(X) := u(X) + s(X) + w(X), \tag{4.1}$$

where u, s, and w are appropriate functions satisfying the following conditions.

(i) The function $u : \mathbb{R}^{d\times r} \to \mathbb{R}$ is continuously differentiable and its gradient ∇u is Lipschitz continuous with the corresponding Lipschitz constant $L_u > 0$.

(ii) The function $s : \mathbb{R}^{d\times r} \to \mathbb{R}$ is convex and Lipschitz continuous with the corresponding Lipschitz constant $L_s > 0$.

(iii) The function $w : \mathbb{R}^{d\times r} \to \mathbb{R}$ is of the form $w(X) = 2\rho\|(I_d - XX^\top)\Sigma_\circ^{1/2}\|_{\mathrm{F}}$ with $\Sigma_\circ \in \mathbb{S}^d_+$ and $\rho > 0$.

It is evident that model ($\mathrm{P_m}$) is a specific instance of problem (4.1) by identifying $u(X) = \operatorname{tr}((I_d - XX^\top)\Sigma_\circ)$. As a direct consequence of the continuity of f over the compact manifold $\mathcal{O}^{d,r}$, there always exists an optimal solution of problem (4.1).

Remark 4.1. The function w is continuously differentiable at every $X \in \mathcal{O}^{d,r}$ such that $\|(I_d - XX^\top)\Sigma_\circ^{1/2}\|_{\mathrm{F}} > 0$. In particular, when $\operatorname{rank}(\Sigma_\circ) > r$, this condition holds for any $X \in \mathcal{O}^{d,r}$, and hence w is smooth on the entire feasible set. Possible nonsmoothness occurs at zero-residual points satisfying $(I_d - XX^\top)\Sigma_\circ^{1/2} = 0$, which can arise when $\operatorname{rank}(\Sigma_\circ) \leq r$.

The case of $\operatorname{rank}(\Sigma_\circ) \le r$ is natural from the perspective of data-driven DRO [42]. As discussed in [20], the nominal distribution is often empirical and constructed from limited samples in practice. Therefore, a nominal covariance matrix with rank at most r does not imply that the true distribution is supported on an r-dimensional subspace. The Wasserstein ambiguity set is introduced to account for possible distributional perturbations, including mass or variance outside the observed nominal subspace. In this case, the consideration of regularized PCA remains meaningful, as it aims to select a loading matrix X with desirable structures among possibly many feasible or optimal solutions. Even in the regime where $\operatorname{rank}(\Sigma_\circ) > r$, the gradient of w contains a factor $\|(I_d - XX^\top)\Sigma_\circ^{1/2}\|_{\mathrm{F}}^{-1}$, which becomes numerically unstable when the denominator is close to zero. The smoothing algorithm developed in this section provides a unified and stable framework that remains applicable without imposing an additional rank condition on $\Sigma_\circ$. The underlying methodology can be seamlessly extended to a wider class of optimization problems, such as the regularized ℓ_1-norm PCA [40] and the DRO model of fair PCA [47].

4.1. Stationarity Condition. In this subsection, we establish the stationarity condition for local minimizers of problem (4.1). According to the discussions in [30, 58], a necessary condition for f to achieve a local minimum at X on $\mathcal{O}^{d,r}$ is that

$$0 \in \partial_{\mathrm{R}} f(X).$$

Since u is smooth, s is convex, and w is a composition of a smooth mapping and a convex function, they are all weakly convex and hence regular [18]. As a result, the objective function $f = u + s + w$ is also regular [16]. Then it follows from [58, Theorem 5.3] that $\partial_{\mathrm{R}} f(X) = \operatorname{grad} u(X) + \partial_{\mathrm{R}} s(X) + \partial_{\mathrm{R}} w(X)$ for any $X \in \mathcal{O}^{d,r}$.

Based on the above discussions, the stationarity condition of the optimization problem (4.1) can be stated as follows.

DEFINITION 4.2. *A point $X_* \in \mathcal{O}^{d,r}$ is called a stationary point of problem* (4.1) *if the inclusion $0 \in \operatorname{grad} u(X_*) + \partial_{\mathrm{R}} s(X_*) + \partial_{\mathrm{R}} w(X_*)$ holds.*

When w is smooth at X, the term $\partial_{\mathrm{R}} w(X)$ reduces to the singleton $\{\operatorname{grad} w(X)\}$. Thus, the Clarke subdifferential formulation provides a unified stationarity concept.

4.2. Smoothing Function. To address the challenges posed by the possibly nonsmooth term w, we propose to leverage the smoothing approximation technique [12]. Specifically, we construct the smoothing function of w as follows,

$$\tilde{w}(X,\mu) = \begin{cases} w(X), & \text{if } w(X) \ge \mu\rho, \\ \dfrac{w^2(X)}{2\mu\rho} + \dfrac{\mu\rho}{2}, & \text{if } w(X) < \mu\rho, \end{cases} \tag{4.2}$$

where $\mu > 0$ is a smoothing parameter. Interested readers can refer to [12] for more examples of smoothing functions.

When it is clear from the context, the Euclidean and Riemannian gradients of $\tilde{w}(X,\mu)$ with respect to X are simply denoted by $\nabla\tilde{w}(X,\mu)$ and $\operatorname{grad}\tilde{w}(X,\mu)$, respectively. The following proposition reveals that the smoothing function $\tilde{w}$ enjoys some favorable properties.

PROPOSITION 4.3. *The function $\tilde{w} : \mathbb{R}^{d\times r} \times \mathbb{R}_{++} \to \mathbb{R}$ constructed in* (4.2) *satisfies the following conditions.*

(i) For any $\mu > 0$, $\tilde{w}(\cdot,\mu)$ is continuously differentiable over $\mathbb{R}^{d\times r}$.

(ii) *For any* $X \in \mathbb{R}^{d\times r}$, *it holds that* $\lim_{X'\to X,\,\mu\downarrow 0} \tilde{w}(X',\mu) = w(X)$.
(iii) *For any* $X \in \mathbb{R}^{d\times r}$ *and* $\mu > 0$, *we have* $w(X) \le \tilde{w}(X,\mu) \le w(X) + \mu\rho/2$.
(iv) *There exists a constant* $M_w > 0$ *such that* $\|\nabla\tilde{w}(X,\mu)\|_{\mathrm{F}} \le M_w$ *for any* $X \in \mathcal{O}^{d,r}$ *and* $\mu > 0$.
(v) *There exists a constant* $L_w > 0$ *such that, for any* $\mu > 0$, $\nabla\tilde{w}(\cdot,\mu)$ *is Lipschitz continuous over* $\mathcal{O}^{d,r}$ *with the corresponding Lipschitz constant* $L_w\mu^{-1}$.

Proof. The proof is straightforward and is omitted here. ∎

We adopt the smoothing function $\tilde{w}$ given in (4.2) for two key reasons. First, the evaluation of both function values and gradients for $\tilde{w}$ circumvents the need to compute $\Sigma_\circ^{1/2}$. Second, this particular smoothing function satisfies Riemannian gradient consistency, which will be rigorously demonstrated in the next subsection. This property is crucial for guaranteeing the global convergence of our algorithm.

4.3. Riemannian Subdifferential. The algorithm proposed in this paper is based on the smoothing approximation technique. Thus, it is natural that the convergence result is closely tied to the specific smoothing function employed. Below is the definition of the Riemannian subdifferential associated with the smoothing function, which serves as a fundamental concept in our analysis.

DEFINITION 4.4. *The Riemannian subdifferential of* w *associated with the smoothing function* $\tilde{w}$ *at* $X \in \mathcal{O}^{d,r}$ *is defined as*

$$\partial_{\mathrm{R}}|_{\tilde{w}} w(X) = \left\{ G \in \mathcal{T}_X\mathcal{O}^{d,r} \mid \operatorname{grad}\tilde{w}(X_t,\mu_t) \to G, \mathcal{O}^{d,r} \ni X_t \to X, \mu_t \downarrow 0 \right\}.$$

The following theorem establishes the Riemannian gradient consistency of the smoothing function $\tilde{w}$. By bridging two subdifferentials, this property plays a pivotal role in showing the global convergence of the proposed algorithm to a stationary point of problem (4.1).

THEOREM 4.5. *For the smoothing function* $\tilde{w}$ *constructed in* (4.2), *it holds that*

$$\partial_{\mathrm{R}}|_{\tilde{w}} w(X) \subseteq \partial_{\mathrm{R}} w(X),$$

for any $X \in \mathcal{O}^{d,r}$.

Proof. We fix an arbitrary $X \in \mathcal{O}^{d,r}$. Let $G \in \partial_{\mathrm{R}}|_{\tilde{w}} w(X)$. Then there exist two sequences $\{X_t\} \subseteq \mathcal{O}^{d,r}$ and $\{\mu_t\} \subseteq \mathbb{R}_{++}$ with $X_t \to X$ and $\mu_t \downarrow 0$ as $t \to \infty$ such that

$$G = \lim_{\mathcal{O}^{d,r}\ni X_t\to X,\,\mu_t\downarrow 0} \operatorname{grad}\tilde{w}(X_t,\mu_t).$$

For convenience, we define the index set $\mathbb{T} := \{t \in \mathbb{N} \mid w(X_t) < \mu_t\rho\}$.

If $\mathbb{T}$ is a finite set, there exists $\bar{t} \in \mathbb{N}$ such that $w(X_t) \ge \mu_t\rho > 0$ for any $t \ge \bar{t}$. Hence, the function w is continuously differentiable near X_t and we have

$$\operatorname{grad}\tilde{w}(X_t,\mu_t) = \operatorname{grad} w(X_t),$$

for all $t \ge \bar{t}$. Then it can be obtained that

$$G = \lim_{\mathcal{O}^{d,r}\ni X_t\to X,\,\mu_t\downarrow 0} \operatorname{grad}\tilde{w}(X_t,\mu_t) = \lim_{\mathcal{O}^{d,r}\ni X_t\to X} \operatorname{grad} w(X_t),$$

which indicates that $G \in \partial_{\mathrm{R}} w(X)$.

Next, we consider the case where $\mathbb{T}$ is an infinite set. Then it is clear that $w(X) = 0$. Thus, the function w attains the global minimum at X, which further implies that $0 \in \partial_{\mathrm{R}} w(X)$. Let $\mathbb{T}' := \{t \in \mathbb{N} \mid w(X_t) = 0\}$ be a subset of $\mathbb{T}$. If $\mathbb{T}'$ is also an infinite set, we can obtain that

$$G = \lim_{\mathcal{O}^{d,r} \ni X_t \to X,\, \mu_t \downarrow 0,\, t \in \mathbb{T}'} \operatorname{grad} \tilde{w}(X_t, \mu_t).$$

Straightforward calculations yield that $\operatorname{grad} \tilde{w}(X_t, \mu_t) = 0 \in \partial_{\mathrm{R}} w(X_t)$ for any $t \in \mathbb{T}'$. Hence, the above relationship indicates that $G = 0 \in \partial_{\mathrm{R}} w(X)$. Now we assume that $\mathbb{T}'$ is a finite set. Passing to the subsequence of $\mathbb{T}$ and relabeling it if necessary, there exists $\hat{t} \in \mathbb{N}$ such that $0 < w(X_t) < \mu_t \rho$ for any $t \geq \hat{t}$. Moreover, the function w is continuously differentiable near X_t and we have

$$\operatorname{grad} \tilde{w}(X_t, \mu_t) = \tau_t \operatorname{grad} w(X_t), \tag{4.3}$$

where τ_t is a constant defined by

$$\tau_t = \frac{w(X_t)}{\mu_t \rho} \in (0, 1).$$

A straightforward verification reveals that

$$\|\operatorname{grad} w(X_t)\|_{\mathrm{F}} = 2\rho \frac{\left\|(I_d - X_t X_t^\top)\Sigma_\circ X_t\right\|_{\mathrm{F}}}{\left\|(I_d - X_t X_t^\top)\Sigma_\circ^{1/2}\right\|_{\mathrm{F}}} \leq 2\rho \left\|\Sigma_\circ^{1/2}\right\|_{\mathrm{F}},$$

for any $X_t \in \mathcal{O}^{d,r}$ satisfying $w(X_t) \neq 0$. Hence, the sequence $\{\operatorname{grad} w(X_t)\}_{t \geq \hat{t}}$ is bounded. By passing to a subsequence if necessary, we may assume without loss of generality that

$$H = \lim_{\mathcal{O}^{d,r} \ni X_t \to X} \operatorname{grad} w(X_t).$$

According to the definition of Riemannian Clarke subdifferentials, it holds that $H \in \partial_{\mathrm{R}} w(X)$. Since $\tau_t \in (0, 1)$ for any $t \geq \hat{t}$, we can assume without loss of generality that $\lim_{t \to \infty} \tau_t = \tau$ for a constant $\tau \in [0, 1]$. Consequently, it follows from the relationship (4.3) and the fact $0 \in \partial_{\mathrm{R}} w(X)$ that

$$\begin{aligned} G &= \lim_{\mathcal{O}^{d,r} \ni X_t \to X,\, \mu_t \downarrow 0} \operatorname{grad} \tilde{w}(X_t, \mu_t) = \lim_{\mathcal{O}^{d,r} \ni X_t \to X,\, \mu_t \downarrow 0} \tau_t \operatorname{grad} w(X_t) \\ &= \tau H = \tau H + (1 - \tau) 0 \in \operatorname{conv} \{\partial_{\mathrm{R}} w(X)\} = \partial_{\mathrm{R}} w(X). \end{aligned}$$

The proof is completed. ∎

Theorem 4.5 can be regarded as a Riemannian extension of the gradient consistency in Euclidean spaces [12]. It is worth noting, however, that on a Riemannian manifold one must additionally account for the projection onto tangent spaces as well as the use of Riemannian gradients.

4.4. Algorithm Development. Based on the smoothing function $\tilde{w}$, we can obtain the following approximation of the objective function f in problem (4.1),

$$\tilde{f}(X, \mu) := \tilde{g}(X, \mu) + s(X),$$

where $\tilde{g}$ is given by

$$\tilde{g}(X,\mu) := u(X) + \tilde{w}(X,\mu).$$

It is clear that the function $\tilde{f}(X,\mu)$ exhibits a composite structure. In particular, the first term $\tilde{g}(X,\mu)$ is smooth for fixed $\mu > 0$, whereas the second term $s(X)$ is possibly nonsmooth. This inherent structure naturally lends itself to the framework of the proximal gradient method on the manifold to minimize $\tilde{f}(\cdot,\mu)$ over $\mathcal{O}^{d,r}$.

Specifically, we intend to solve the following subproblem to find the descent direction $V_k \in \mathcal{T}_{X_k}\mathcal{O}^{d,r}$ at the k-th iteration,

$$V_k := \mathop{\arg\min}_{V \in \mathcal{T}_{X_k}\mathcal{O}^{d,r}} h_k(V) := \langle \nabla \tilde{g}(X_k,\mu_k), V\rangle + \frac{1}{2\mu_k}\|V\|_{\mathrm{F}}^2 + s(X_k+V), \tag{4.4}$$

where $X_k \in \mathcal{O}^{d,r}$ and $\mu_k > 0$ are the current iterate and smoothing parameter, respectively. The above subproblem involves minimizing a strongly convex function on the tangent space. Although V_k serves as a descent direction, the updated iterate $X_k + \alpha_k V_k$, for an arbitrary stepsize $\alpha_k > 0$, does not necessarily remain on $\mathcal{O}^{d,r}$. Consequently, we then perform a retraction to bring it back to $\mathcal{O}^{d,r}$.

In practice, subproblem (4.4) can be addressed effectively via a semi-smooth Newton method [10] or through a fixed-point method [39]. Below, we provide a concise exposition of the computational procedure underlying the latter approach. The KKT conditions of subproblem (4.4) can be expressed as

$$\begin{cases} 0 \in \nabla \tilde{g}(X_k,\mu_k) - X_k\Lambda + \dfrac{1}{\mu_k}V + \partial s(X_k+V), \\ 0 = X_k^\top V + V^\top X_k, \end{cases} \tag{4.5}$$

where $\Lambda \in \mathbb{R}^{r\times r}$ represents the Lagrangian multiplier associated with the linear constraint $V \in \mathcal{T}_{X_k}\mathcal{O}^{d,r}$. For a given Λ, the first relationship of (4.5) allows V to be characterized explicitly by

$$V(\Lambda) := \mathfrak{P}_{\mu_k,s}\left(X_k - \mu_k\left(\nabla\tilde{g}(X_k,\mu_k) - X_k\Lambda\right)\right) - X_k, \tag{4.6}$$

where the proximal mapping $\mathfrak{P}_{\mu,s}$ is defined as

$$\mathfrak{P}_{\mu,s}(Y) := \mathop{\arg\min}_{Z\in\mathbb{R}^{d\times r}} \; s(Z) + \frac{1}{2\mu}\|Z-Y\|_{\mathrm{F}}^2.$$

Substituting (4.6) into the second relationship of (4.5) reveals that Λ satisfies the nonlinear equation $E(\Lambda) = 0$ with $E(\Lambda) := X_k^\top V(\Lambda) + V(\Lambda)^\top X_k$. Building upon this insight, Liu et al. [39] propose the following fixed-point method for solving $E(\Lambda) = 0$ with respect to Λ,

$$\Lambda_{t+1} = \Lambda_t - \frac{1}{\mu_k}E(\Lambda_t).$$

In practice, the stopping criterion can be specified as $\|E(\Lambda_t)\|_{\mathrm{F}} \le \varepsilon$ for a prescribed tolerance $\varepsilon \in (0,1)$. He et al. [27] have shown that this fixed-point method—viewed as a special instance of their primal-dual hybrid gradient algorithm—enjoys an $O(1/t)$ convergence rate in the ergodic sense under mild conditions, which are satisfied in the present setting.

The complete framework of our approach for solving problem (4.1) is outlined in Algorithm 1, which is named *smoothing manifold proximal gradient* and abbreviated as SMPG. It is noteworthy that SMPG involves an Armijo line search procedure in (4.7) to determine the stepsize. As we will show later, this backtracking line search procedure is well-defined and guaranteed to terminate in a finite number of steps.

Algorithm 1: smoothing manifold proximal gradient (SMPG).

Input: $X_0 \in \mathcal{O}^{d,r}$, $\mu_0 > 0$, $\bar{\mu} \in [0, \mu_0]$, $\theta \in (0,1)$, and $\beta \in (0,1)$.
for $k = 0, 1, 2, \ldots$ **do**
  Solve subproblem (4.4) to obtain V_k.
  if $\|V_k\|_{\mathrm{F}} \le \bar{\mu}^2$ *and* $\mu_k \le \bar{\mu}$ **then**
    Return X_k.
  else
    Find $\alpha_k := \beta^{m_k}$ such that m_k is the smallest integer satisfying

$$\tilde{f}(\mathfrak{R}_{X_k}(\beta^{m_k} V_k), \mu_k) \le \tilde{f}(X_k, \mu_k) - \frac{\beta^{m_k}}{2\mu_k} \|V_k\|_{\mathrm{F}}^2 . \tag{4.7}$$

    Update $X_{k+1} := \mathfrak{R}_{X_k}(\alpha_k V_k)$.
    Set

$$\mu_{k+1} := \begin{cases} \mu_k, & \text{if } \|V_k\|_{\mathrm{F}} > \mu_k^2, \\ \theta \mu_k, & \text{if } \|V_k\|_{\mathrm{F}} \le \mu_k^2. \end{cases}$$

Output: X_k.

Remark 4.6. Among the four parameters introduced in SMPG, μ_0 denotes the initial smoothing parameter, θ governs the decay rate of the smoothing parameter, β controls the decay rate of the stepsize in the line search procedure, and $\bar{\mu}$ serves as the termination tolerance. In practice, $\bar{\mu}$ specifies the desired accuracy and therefore does not require tuning. Our numerical experiments presented in Section 6.2 indicate that SMPG is not particularly sensitive to the other three parameters.

Algorithm 1 bears a certain resemblance to the ManPG method proposed in [10]. The latter similarly solves a subproblem akin to (4.4) in the tangent space, followed by a retraction to preserve the feasibility. It should be emphasized, however, that in Algorithm 1 we further employ a smoothing approximation to handle the possibly nonsmooth term arising from the DRO formulation, together with an update scheme for the smoothing parameter. These features introduce additional challenges to the convergence analysis of Algorithm 1, which we shall elaborate on in the subsequent section.

5. Convergence Analysis. This section delves into a comprehensive convergence analysis of the proposed algorithm. Specifically, we establish that any accumulation point of the sequence generated by Algorithm 1 is a stationary point. And the iteration complexity of Algorithm 1 is provided to attain an approximate stationary point.

5.1. Descent Property. In the following lemma, we first prove that V_k, obtained by solving subproblem (4.4), serves as a descent direction in the tangent space $\mathcal{T}_{X_k}\mathcal{O}^{d,r}$ at the current iterate $X_k \in \mathcal{O}^{d,r}$.

LEMMA 5.1. *Suppose that $\{X_k\}$ is the sequence generated by Algorithm 1. Then it holds that*

$$h_k(0) - h_k(\alpha V_k) \geq \frac{\alpha(2-\alpha)}{2\mu_k} \|V_k\|_{\mathrm{F}}^2,$$

for any $\alpha \in [0,1]$.

Proof. Since $h_k(V)$ is strongly convex with the modulus μ_k^{-1}, we have

$$h_k(\bar{V}) \geq h_k(V) + \left\langle \partial h_k(V), \bar{V} - V \right\rangle + \frac{1}{2\mu_k} \left\|\bar{V} - V\right\|_{\mathrm{F}}^2, \tag{5.1}$$

for any $V, \bar{V} \in \mathbb{R}^{d\times r}$. In particular, for $V, \bar{V} \in \mathcal{T}_{X_k}\mathcal{O}^{d,r}$, it holds that

$$\left\langle \partial h_k(V), \bar{V} - V \right\rangle = \left\langle \mathrm{Proj}_{\mathcal{T}_{X_k}\mathcal{O}^{d,r}}\left(\partial h_k(V)\right), \bar{V} - V \right\rangle.$$

Moreover, it follows from the optimality condition of subproblem (4.4) that $0 \in \mathrm{Proj}_{\mathcal{T}_{X_k}\mathcal{O}^{d,r}}\left(\partial h_k(V_k)\right)$. Taking $V = V_k$ and $\bar{V} = 0$ in (5.1) yields that

$$h_k(0) \geq h_k(V_k) + \frac{1}{2\mu_k} \|V_k\|_{\mathrm{F}}^2,$$

which, after a suitable rearrangement, can be equivalently written as

$$s(X_k) \geq \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + \frac{1}{\mu_k} \|V_k\|_{\mathrm{F}}^2 + s(X_k + V_k).$$

According to the convexity of s, we have

$$\begin{aligned} s(X_k + \alpha V_k) - s(X_k) &= s((1-\alpha)X_k + \alpha(X_k + V_k)) - s(X_k) \\ &\leq \alpha\left(s(X_k + V_k) - s(X_k)\right). \end{aligned}$$

Collecting the above two relationships together results in that

$$\begin{aligned} h_k(\alpha V_k) - h_k(0) &= \alpha \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + \frac{\alpha^2}{2\mu_k} \|V_k\|_{\mathrm{F}}^2 + s(X_k + \alpha V_k) - s(X_k) \\ &\leq \alpha \left(\langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + \frac{\alpha}{2\mu_k} \|V_k\|_{\mathrm{F}}^2 + s(X_k + V_k) - s(X_k) \right) \\ &\leq \frac{\alpha(\alpha-2)}{2\mu_k} \|V_k\|_{\mathrm{F}}^2, \end{aligned}$$

which completes the proof. ∎

Based on Lemma 5.1, we can proceed to show that the line search procedure in Algorithm 1 is well-defined, ensuring that the stepsize α_k can be determined in a finite number of trials.

LEMMA 5.2. *Let $\{X_k\}$ be the sequence generated by Algorithm 1. Then there exists a constant $\bar{\alpha} \in (0,1]$ such that*

$$\tilde{f}(X_k, \mu_k) - \tilde{f}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) \geq \frac{\alpha}{2\mu_k} \|V_k\|_{\mathrm{F}}^2,$$

for any $\alpha \in (0, \bar{\alpha}]$.

Proof. According to the Lipschitz continuity of $\nabla \tilde{g}(X, \mu)$ for fixed $\mu > 0$, it follows that

$$\begin{aligned}\tilde{g}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) &\le \tilde{g}(X_k, \mu_k) + \langle \nabla \tilde{g}(X_k, \mu_k), \mathfrak{R}_{X_k}(\alpha V_k) - X_k \rangle \\ &\quad + \frac{L_u \mu_k + L_w}{2\mu_k} \left\| \mathfrak{R}_{X_k}(\alpha V_k) - X_k \right\|_{\mathrm{F}}^2 \\ &\le \tilde{g}(X_k, \mu_k) + \langle \nabla \tilde{g}(X_k, \mu_k), \mathfrak{R}_{X_k}(\alpha V_k) - (X_k + \alpha V_k) \rangle \\ &\quad + \alpha \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + \frac{L_u \mu_0 + L_w}{2\mu_k} \left\| \mathfrak{R}_{X_k}(\alpha V_k) - X_k \right\|_{\mathrm{F}}^2,\end{aligned}$$

where the last inequality holds due to the fact that $\mu_k \le \mu_0$. Since ∇u is continuous over the compact manifold $\mathcal{O}^{d,r}$, there exists a constant $M_u > 0$ such that $\|\nabla u(X)\|_{\mathrm{F}} \le M_u$ for any $X \in \mathcal{O}^{d,r}$. Hence, it holds that

$$\left\| \nabla \tilde{g}(X_k, \mu_k) \right\|_{\mathrm{F}} \le \left\| \nabla u(X_k) \right\|_{\mathrm{F}} + \left\| \nabla \tilde{w}(X_k, \mu_k) \right\|_{\mathrm{F}} \le \frac{(M_u + M_w)\mu_0}{\mu_k}.$$

By invoking Lemma 2.3, we can obtain that

$$\begin{aligned}&\langle \nabla \tilde{g}(X_k, \mu_k), \mathfrak{R}_{X_k}(\alpha V_k) - (X_k + \alpha V_k) \rangle \\ \le\ & \left\| \nabla \tilde{g}(X_k, \mu_k) \right\|_{\mathrm{F}} \left\| \mathfrak{R}_{X_k}(\alpha V_k) - (X_k + \alpha V_k) \right\|_{\mathrm{F}} \\ \le\ & \frac{\alpha^2 \mu_0 M_2 (M_u + M_w)}{\mu_k} \left\| V_k \right\|_{\mathrm{F}}^2,\end{aligned}$$

and

$$\left\| \mathfrak{R}_{X_k}(\alpha V_k) - X_k \right\|_{\mathrm{F}}^2 \le \alpha^2 M_1^2 \left\| V_k \right\|_{\mathrm{F}}^2.$$

Let $C = 2\mu_0 M_2 (M_u + M_w) + M_1^2 (L_u \mu_0 + L_w) > 0$ be a constant. Then it can be readily verified that

$$\tilde{g}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) - \tilde{g}(X_k, \mu_k) \le \alpha \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + \frac{\alpha^2 C}{2\mu_k} \left\| V_k \right\|_{\mathrm{F}}^2.$$

Moreover, according to the Lipschitz continuity of s, we have

$$\begin{aligned}s(\mathfrak{R}_{X_k}(\alpha V_k)) - s(X_k) &= s(\mathfrak{R}_{X_k}(\alpha V_k)) - s(X_k + \alpha V_k) + s(X_k + \alpha V_k) - s(X_k) \\ &\le L_s \left\| \mathfrak{R}_{X_k}(\alpha V_k) - (X_k + \alpha V_k) \right\|_{\mathrm{F}} + s(X_k + \alpha V_k) - s(X_k) \\ &\le \frac{\alpha^2 \mu_0 M_2 L_s}{\mu_k} \left\| V_k \right\|_{\mathrm{F}}^2 + s(X_k + \alpha V_k) - s(X_k),\end{aligned}$$

where the last inequality results from Lemma 2.3 and the fact that $\mu_k \le \mu_0$. Collecting the above two inequalities together yields that

$$\begin{aligned}&\tilde{f}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) - \tilde{f}(X_k, \mu_k) \\ &= \tilde{g}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) - \tilde{g}(X_k, \mu_k) + s(\mathfrak{R}_{X_k}(\alpha V_k)) - s(X_k) \\ &\le \alpha \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + s(X_k + \alpha V_k) - s(X_k) + \frac{\alpha^2 (C + 2\mu_0 M_2 L_s)}{2\mu_k} \left\| V_k \right\|_{\mathrm{F}}^2.\end{aligned} \tag{5.2}$$

As a direct consequence of Lemma 5.1, we can proceed to show that

$$\alpha \langle \nabla \tilde{g}(X_k, \mu_k), V_k \rangle + s(X_k + \alpha V_k) - s(X_k) \leq -\frac{\alpha}{\mu_k} \|V_k\|_{\mathrm{F}}^2 ,$$

which together with the relationship (5.2) implies that

$$\tilde{f}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) - \tilde{f}(X_k, \mu_k) \leq -\frac{\alpha}{2\mu_k} \left(2 - \alpha \left(C + 2\mu_0 M_2 L_s\right)\right) \|V_k\|_{\mathrm{F}}^2 .$$

Let $\bar{\alpha} = \min\{1, 1/(C + 2\mu_0 M_2 L_s)\} \in (0, 1]$. Then for any $\alpha \in (0, \bar{\alpha}]$, we can conclude that

$$\tilde{f}(\mathfrak{R}_{X_k}(\alpha V_k), \mu_k) - \tilde{f}(X_k, \mu_k) \leq -\frac{\alpha}{2\mu_k} \|V_k\|_{\mathrm{F}}^2 ,$$

as desired. The proof is completed. ∎

Lemma 5.2 guarantees that the line search procedure in (4.7) terminates in at most $\lceil \log_\beta \bar{\alpha} \rceil$ steps, which is independent of the smoothing parameter μ_k. Here, the notation $\lceil m \rceil$ represents the smallest integer greater than or equal to $m \in \mathbb{R}$.

5.2. Global Convergence. The following lemma lays the foundation for the global convergence analysis of Algorithm 1 with $\bar{\mu} = 0$, as established in Theorem 5.4.

LEMMA 5.3. *Let $\{X_k\}$ be the sequence generated by Algorithm 1 with $\bar{\mu} = 0$. Then $\mathbb{K} := \{k \in \mathbb{N} \mid \|V_k\|_{\mathrm{F}} \leq \mu_k^2\}$ is an infinite set.*

Proof. Suppose, on the contrary, that $\mathbb{K}$ is a finite set. Then there exists $\bar{k} \in \mathbb{N}$ such that

$$\mu_k = \mu_{\bar{k}} > 0, \text{ and } \|V_k\|_{\mathrm{F}} > \mu_{\bar{k}}^2 > 0, \tag{5.3}$$

for any $k \geq \bar{k}$. Thus, we have $X_{k+1} = \mathfrak{R}_{X_k}(\alpha_k V_k)$ for all $k \geq \bar{k}$, where the stepsize α_k is obtained by using the line search procedure in (4.7) with μ_k fixed as $\mu_{\bar{k}}$. From Lemma 5.2, we know that $\alpha_k \geq \bar{\alpha}\beta$ and

$$\tilde{f}(X_k, \mu_{\bar{k}}) - \tilde{f}(X_{k+1}, \mu_{\bar{k}}) \geq \frac{\alpha_k}{2\mu_{\bar{k}}} \|V_k\|_{\mathrm{F}}^2 \geq \frac{\bar{\alpha}\beta}{2\mu_{\bar{k}}} \|V_k\|_{\mathrm{F}}^2 ,$$

for any $k \geq \bar{k}$. This observation indicates that the sequence $\{\tilde{f}(X_k, \mu_{\bar{k}})\}_{k \geq \bar{k}}$ is monotonically decreasing. Moreover, as a direct consequence of Proposition 4.3, we can proceed to show that

$$f(X) \leq \tilde{f}(X, \mu) \leq f(X) + \frac{\mu\rho}{2},$$

for any $X \in \mathcal{O}^{d,r}$ and $\mu > 0$. In light of the continuity of f over the compact manifold $\mathcal{O}^{d,r}$, there exist two constants $\tilde{f}_{\min}$ and $\tilde{f}_{\max}$ such that

$$\tilde{f}_{\min} \leq \tilde{f}(X, \mu) \leq \tilde{f}_{\max}, \tag{5.4}$$

for any $X \in \mathcal{O}^{d,r}$ and $\mu \in (0, \mu_0]$. Hence, the sequence $\{\tilde{f}(X_k, \mu_{\bar{k}})\}_{k \geq \bar{k}}$ is convergent. Then we can obtain that

$$\lim_{k \to \infty} \|V_k\|_{\mathrm{F}}^2 \leq \frac{2\mu_{\bar{k}}}{\bar{\alpha}\beta} \lim_{k \to \infty} \left(\tilde{f}(X_k, \mu_{\bar{k}}) - \tilde{f}(X_{k+1}, \mu_{\bar{k}})\right) = 0,$$

which contradicts the second relationship in (5.3). Consequently, we can conclude that $\mathbb{K}$ is an infinite set. The proof is completed. ∎

Now we are in a position to establish the global convergence of Algorithm 1 to a stationary point of problem (4.1) under the setting $\bar{\mu} = 0$.

THEOREM 5.4. *Suppose that $\{X_k\}$ is the sequence generated by Algorithm 1 with $\bar{\mu} = 0$. Then the sequence $\{X_k\}$ has at least one accumulation point. Moreover, any accumulation point is a stationary point of problem* (4.1).

Proof. For each $k \in \mathbb{K}$, we have $\mu_{k+1} = \theta\mu_k$ with $\theta \in (0, 1)$ being a decaying factor. According to Lemma 5.3, the index set $\mathbb{K}$ is infinite. Then it can be readily verified that

$$\lim_{\mathbb{K}\ni k\to\infty} \frac{1}{\mu_k} \|V_k\|_{\mathrm{F}} \leq \lim_{\mathbb{K}\ni k\to\infty} \mu_k = 0. \tag{5.5}$$

Since $\mathcal{O}^{d,r}$ is a compact manifold, the sequence $\{X_k\}$ is bounded. Then from the Bolzano-Weierstrass theorem, it can be deduced that the sequence $\{X_k\}$ has at least one accumulation point. Let X_* be an accumulation point of $\{X_k\}$. The completeness of $\mathcal{O}^{d,r}$ guarantees that $X_* \in \mathcal{O}^{d,r}$. By passing to a subsequence if necessary, we may assume without loss of generality that $\lim_{\mathbb{K}\ni k\to\infty} X_k = X_*$.

Next, by virtue of the optimality condition [10, 58] of subproblem (4.4), there exists $H_k \in \partial s(X_k + V_k)$ such that

$$\operatorname{grad} u(X_k) + \operatorname{grad} \tilde{w}(X_k, \mu_k) + \operatorname{Proj}_{\mathcal{T}_{X_k}\mathcal{O}^{d,r}}(H_k) + \frac{1}{\mu_k} V_k = 0, \tag{5.6}$$

for any $k \in \mathbb{K}$. It is clear that the sequence $\{\operatorname{grad} u(X_k)\}$ is convergent and

$$\lim_{\mathbb{K}\ni k\to\infty} \operatorname{grad} u(X_k) = \operatorname{grad} u(X_*).$$

According to the Lipschitz continuity of s, the sequence $\{H_k\}$ is bounded [16]. Without loss of generality, we assume that it is also convergent. Then there exists H_* such that $\lim_{\mathbb{K}\ni k\to\infty} H_k = H_*$. It follows from [43, Theorem 24.4] that $H_* \in \partial s(X_*)$. In addition, the boundedness of the sequence $\{\mu_k^{-1} V_k\}$ results from the fact that it is convergent in (5.5). As a result, the sequence $\{\operatorname{grad} \tilde{w}(X_k, \mu_k)\}$ is also bounded. Without loss of generality, we can assume that there exists G_* such that

$$\lim_{\mathbb{K}\ni k\to\infty} \operatorname{grad} \tilde{w}(X_k, \mu_k) = G_*.$$

According to the definition of the Riemannian subdifferential of w associated with $\tilde{w}$, it holds that $G_* \in \partial_{\mathrm{R}}|_{\tilde{w}} w(X_*)$. Then it follows from Theorem 4.5 that $G_* \in \partial_{\mathrm{R}} w(X_*)$.

Finally, upon taking $k \to \infty$ in the relationship (5.6), we can conclude that

$$0 = \operatorname{grad} u(X_*) + G_* + \operatorname{Proj}_{\mathcal{T}_{X_*}\mathcal{O}^{d,r}}(H_*) \in \operatorname{grad} u(X_*) + \partial_{\mathrm{R}} s(X_*) + \partial_{\mathrm{R}} w(X_*),$$

which indicates that X_* is a stationary point of problem (4.1). We complete the proof. ∎

5.3. Iteration Complexity. The final task is to derive the iteration complexity of Algorithm 1, a critical challenge that remains unresolved in existing works. The proof of Theorem 5.4 previously discussed leads to the insight that an accumulation point of $\{X_k\}$ is a stationary point of (4.1) if the following conditions are satisfied,

$$\begin{cases} \lim\limits_{k\to\infty} \operatorname{dist}\Big(0, \operatorname{grad} u(X_k) + \operatorname{grad} \tilde{w}(X_k, \mu_k) + \operatorname{Proj}_{\mathcal{T}_{X_k}\mathcal{O}^{d,r}}(\partial s(X_k + V_k))\Big) = 0, \\ \lim\limits_{k\to\infty} \|V_k\|_{\mathrm{F}} = 0, \quad \lim\limits_{k\to\infty} \mu_k = 0. \end{cases}$$

This observation motivates us to define the concept of ϵ-approximate stationarity for problem (4.1) as follows.

DEFINITION 5.5. *A point $X \in \mathcal{O}^{d,r}$ is called an ϵ-approximate stationary point of problem* (4.1) *if there exists $V \in \mathcal{T}_X\mathcal{O}^{d,r}$ with $\|V\|_{\mathrm{F}} \leq \epsilon$ and $\mu \in (0, \epsilon]$ such that*

$$\operatorname{dist}\left(0, \operatorname{grad} u(X) + \operatorname{grad} \tilde{w}(X, \mu) + \operatorname{Proj}_{\mathcal{T}_X\mathcal{O}^{d,r}}\left(\partial s(X+V)\right)\right) \leq \epsilon.$$

We show that Algorithm 1 is capable of identifying an ϵ-approximate stationary point of problem (4.1) under the setting $\bar{\mu} = \epsilon$.

LEMMA 5.6. *For any $\epsilon \in (0,1)$ and $\mu_0 \geq \epsilon$, Algorithm 1 with $\bar{\mu} = \epsilon$ will terminate at an ϵ-approximate stationary point of problem* (4.1).

Proof. We first define the constant

$$\iota_\epsilon = \lceil \log_\theta (\epsilon/\mu_0) \rceil + 1.$$

Let $\Bbbk_i$ be the i-th smallest element of $\mathbb{K}$. Then it holds that

$$\mu_{\Bbbk_{\iota_\epsilon}} = \theta^{\iota_\epsilon - 1}\mu_0 \leq \epsilon, \text{ and } \left\|V_{\Bbbk_{\iota_\epsilon}}\right\|_{\mathrm{F}} \leq \mu^2_{\Bbbk_{\iota_\epsilon}} \leq \epsilon^2,$$

which reveals that Algorithm 1 will terminate at the iterate $X_{\Bbbk_{\iota_\epsilon}}$.

The next step is to show that $X_{\Bbbk_{\iota_\epsilon}}$ is an ϵ-approximate stationary point of problem (4.1). According to the relationship (5.6), there exists $H_{\Bbbk_{\iota_\epsilon}} \in \partial s(X_{\Bbbk_{\iota_\epsilon}} + V_{\Bbbk_{\iota_\epsilon}})$ such that

$$-\frac{1}{\mu_{\Bbbk_{\iota_\epsilon}}} V_{\Bbbk_{\iota_\epsilon}} = \operatorname{grad} u(X_{\Bbbk_{\iota_\epsilon}}) + \operatorname{grad} \tilde{w}(X_{\Bbbk_{\iota_\epsilon}}, \mu_{\Bbbk_{\iota_\epsilon}}) + \operatorname{Proj}_{\mathcal{T}_{X_{\Bbbk_{\iota_\epsilon}}}\mathcal{O}^{d,r}}\left(H_{\Bbbk_{\iota_\epsilon}}\right),$$

which implies that

$$\begin{aligned} &\operatorname{dist}\left(0, \operatorname{grad} u(X_{\Bbbk_{\iota_\epsilon}}) + \operatorname{grad} \tilde{w}(X_{\Bbbk_{\iota_\epsilon}}, \mu_{\Bbbk_{\iota_\epsilon}}) + \operatorname{Proj}_{\mathcal{T}_{X_{\Bbbk_{\iota_\epsilon}}}\mathcal{O}^{d,r}}\left(\partial s(X_{\Bbbk_{\iota_\epsilon}} + V_{\Bbbk_{\iota_\epsilon}})\right)\right) \\ &\leq \frac{1}{\mu_{\Bbbk_{\iota_\epsilon}}}\left\|V_{\Bbbk_{\iota_\epsilon}}\right\|_{\mathrm{F}} \leq \mu_{\Bbbk_{\iota_\epsilon}} \leq \epsilon. \end{aligned}$$

Therefore, we conclude that $X_{\Bbbk_{\iota_\epsilon}}$ is an ϵ-approximate stationary point of problem (4.1), which completes the proof. ∎

The iteration complexity of Algorithm 1 is established in the following theorem for finding an ϵ-approximate stationary point.

THEOREM 5.7. *For any $\epsilon \in (0,1)$ and $\mu_0 \geq \epsilon$, Algorithm 1 with $\bar{\mu} = \epsilon$ will reach an ϵ-approximate stationary point of problem* (4.1) *after at most $O(\epsilon^{-3})$ iterations.*

Proof. According to Lemma 5.6, Algorithm 1 with $\bar{\mu} = \epsilon$ will terminate at $X_{\Bbbk_{\iota_\epsilon}}$, which is an ϵ-approximate stationary point of problem (4.1). We give an upper bound on the iteration number $\Bbbk_{\iota_\epsilon}$. For convenience, we denote $\Bbbk_0 = 0$. The iterations from $\Bbbk_i$ to $\Bbbk_{i+1}$ solve subproblem (4.4) with the smoothing parameter fixed as $\mu_{\Bbbk_{i+1}}$ for $0 \leq i \leq \iota_\epsilon - 1$. According to Lemma 5.2, we have

$$\tilde{f}(X_k, \mu_{\Bbbk_{i+1}}) - \tilde{f}(X_{k+1}, \mu_{\Bbbk_{i+1}}) \geq \frac{\alpha_k}{2\mu_{\Bbbk_{i+1}}}\|V_k\|^2_{\mathrm{F}} \geq \frac{\bar{\alpha}\beta}{2\mu_{\Bbbk_{i+1}}}\|V_k\|^2_{\mathrm{F}},$$

for $\Bbbk_i \le k \le \Bbbk_{i+1} - 1$. Then it can be readily verified that

$$\sum_{k=\Bbbk_i}^{\Bbbk_{i+1}-1} \|V_k\|_{\mathrm{F}}^2 \le \frac{2\mu_{\Bbbk_{i+1}}}{\bar{\alpha}\beta} \sum_{k=\Bbbk_i}^{\Bbbk_{i+1}-1} \left(\tilde{f}(X_k, \mu_{\Bbbk_{i+1}}) - \tilde{f}(X_{k+1}, \mu_{\Bbbk_{i+1}})\right)$$
$$= \frac{2\mu_{\Bbbk_{i+1}}}{\bar{\alpha}\beta} \left(\tilde{f}(X_{\Bbbk_i}, \mu_{\Bbbk_{i+1}}) - \tilde{f}(X_{\Bbbk_{i+1}}, \mu_{\Bbbk_{i+1}})\right) \le \frac{2\mu_{\Bbbk_{i+1}}}{\bar{\alpha}\beta} \left(\tilde{f}_{\max} - \tilde{f}_{\min}\right),$$

where the last inequality follows from (5.4). From the definition of $\Bbbk_{i+1}$, we know that $\|V_k\|_{\mathrm{F}} > \mu_{\Bbbk_{i+1}}^2$ for $\Bbbk_i \le k \le \Bbbk_{i+1} - 1$. Hence, it holds that

$$\sum_{k=\Bbbk_i}^{\Bbbk_{i+1}-1} \|V_k\|_{\mathrm{F}}^2 \ge (\Bbbk_{i+1} - \Bbbk_i)\, \mu_{\Bbbk_{i+1}}^4,$$

which together with the relationship $\mu_{\Bbbk_{i+1}} = \theta^i \mu_0$ implies that

$$\Bbbk_{i+1} - \Bbbk_i \le \frac{2(\tilde{f}_{\max} - \tilde{f}_{\min})}{\bar{\alpha}\beta\mu_{\Bbbk_{i+1}}^3} = \frac{2(\tilde{f}_{\max} - \tilde{f}_{\min})}{\bar{\alpha}\beta\mu_0^3\theta^{3i}}.$$

Finally, a straightforward verification reveals that

$$\Bbbk_{\iota_\epsilon} = \Bbbk_0 + \sum_{i=0}^{\iota_\epsilon - 1} (\Bbbk_{i+1} - \Bbbk_i) \le \frac{2(\tilde{f}_{\max} - \tilde{f}_{\min})}{\bar{\alpha}\beta\mu_0^3} \sum_{i=0}^{\iota_\epsilon - 1} \frac{1}{\theta^{3i}} \le \frac{2\theta^3(\tilde{f}_{\max} - \tilde{f}_{\min})}{\bar{\alpha}\beta(1-\theta^3)\epsilon^3}.$$

The proof is completed. ∎

Theorem 5.7 demonstrates that SMPG achieves an iteration complexity of $O(\epsilon^{-3})$. This result represents a clear improvement over the Riemannian subgradient method proposed in [38], which suffers from an inferior iteration complexity of $O(\epsilon^{-4})$.

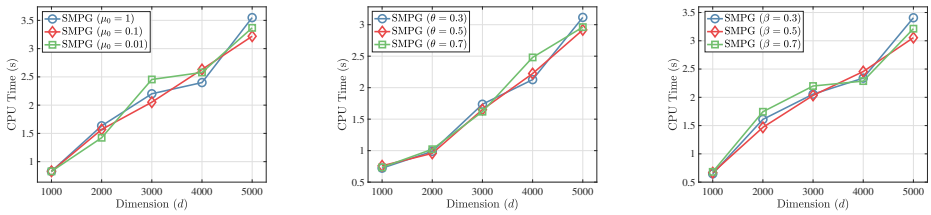
6. Numerical Experiments. Preliminary numerical results are presented in this section to provide additional insights into the performance guarantees of model ($\mathrm{P_{mM}}$) and Algorithm 1 (SMPG). All code is implemented in MATLAB R2018b on a workstation with dual Intel Xeon Gold 6242R CPU processors (at 3.10 GHz×20 × 2) and 510 GB of RAM under Ubuntu 20.04.

6.1. Experimental Setting. In the numerical experiments, we estimate an empirical distribution [20] to serve as the nominal distribution $\mathbb{P}_\circ$ in problem ($\mathrm{P_{mM}}$). Although the true distribution $\mathbb{P}_*$ remains inherently elusive, it is often partially observable through a finite collection of $n \in \mathbb{N}$ independent samples [20, 41], such as past realizations of the random vector ξ. Let the training dataset comprising these samples be denoted by $\hat{\Xi}_n := \{\hat{\xi}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then we construct the empirical distribution $\hat{\mathbb{P}}_n := \sum_{i=1}^n \delta_{\hat{\xi}_i}/n$, where $\delta_{\hat{\xi}_i}$ represents the Dirac distribution concentrating unit mass at $\hat{\xi}_i \in \mathbb{R}^d$. In fact, the empirical distribution $\hat{\mathbb{P}}_n$ can be interpreted as the uniform distribution over the finite samples in $\hat{\Xi}_n$.

Based on the preceding constructions, the sample average approximation (SAA) model of PCA can be expressed as

$$\min_{X \in \mathcal{O}^{d,r}} \mathbb{E}_{\hat{\mathbb{P}}_n} \left[\left\| \left(I_d - XX^\top\right) \left(\xi - \mathbb{E}_{\hat{\mathbb{P}}_n}[\xi]\right) \right\|_2^2 \right] + s(X). \tag{6.1}$$

The above formulation, which does not account for uncertainty in the underlying distribution, has been extensively investigated in the literature [41, 48, 53]. In the



(a) Different μ_0 (b) Different θ (c) Different β

FIG. 1. *Numerical comparison of different algorithmic parameters for SMPG.*

subsequent experiments, we will engage in a performance comparison between the equivalent reformulation ($\mathrm{P_m}$) of the DRO model ($\mathrm{P_{mM}}$) and the SAA model (6.1).

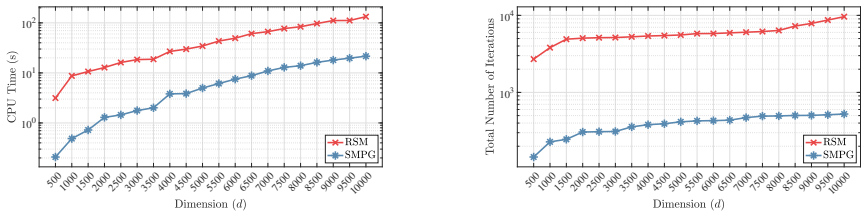
In addition, we focus on the ℓ_1-norm regularizer $s(X) = \gamma \|X\|_1$ with a parameter $\gamma > 0$ to control the amount of sparsity, where $\|X\|_1 := \sum_{i,j} |X_{i,j}|$ with $X_{i,j}$ being the (i,j)-th entry of X. Our empirical experiments reveal that the choice of regularizers does not affect the numerical results dramatically.

6.2. Implementation Details. The proposed algorithm SMPG involves three tuning parameters, including the initial smoothing parameter $\mu_0 > 0$, the decay rate $\theta \in (0,1)$ of the smoothing parameter, and the decay rate $\beta \in (0,1)$ of the stepsize in the line search procedure. This subsection investigates the performance of SMPG on problem ($\mathrm{P_m}$) under various choices of these parameters.

For our testing, we randomly generate a collection of matrices $\Sigma_\circ \in \mathbb{S}^d_+$ with dimension $d \in \{1000, 2000, 3000, 4000, 5000\}$. The other parameters in problem ($\mathrm{P_m}$) are set to $r = 10$, $\gamma = 0.1$, and $\rho = 1$. We then evaluate the behavior of SMPG under a broad range of parameter configurations, taking $\mu_0 \in \{0.01, 0.1, 1\}$, $\theta \in \{0.3, 0.5, 0.7\}$, and $\beta \in \{0.3, 0.5, 0.7\}$. The CPU times consumed by SMPG to attain the termination tolerance $\bar{\mu} = 10^{-3}$ are recorded and displayed in Figure 1. As evidenced therein, SMPG demonstrates a remarkable degree of stability across different parameter choices, with its performance exhibiting no pronounced sensitivity to the specific values selected. Accordingly, we fix $\mu_0 = 0.1$, $\theta = 0.5$, and $\beta = 0.5$ by default in the algorithm SMPG.

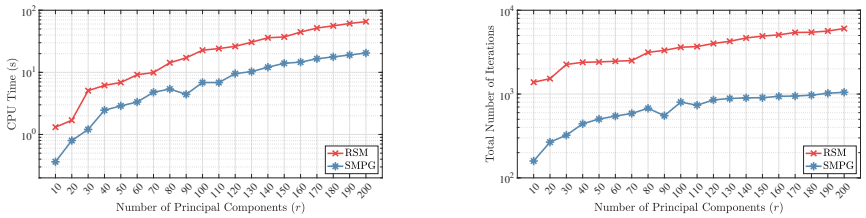
6.3. Performance of SMPG. The experiments conducted in this subsection are designed to demonstrate the effectiveness and efficiency of SMPG for solving problem ($\mathrm{P_m}$) in comparison with the Riemannian subgradient method (RSM) proposed in [38]. Building on the implications of [38, Theorem 2], we equip RSM with a stepsize scheme of the form $\sigma k^{-1/2}$ at the k-th iteration with $\sigma > 0$. For each test instance, the parameter σ is tuned via a manual grid search in $[1, 20]$ with an increment of 1 to identify the value that delivers the best empirical performance. We employ the fixed-point method proposed in [39] to solve the subproblem (4.4) of SMPG. The retraction operator is realized by the polar decomposition in both SMPG and RSM. We set the termination tolerance of SMPG to $\bar{\mu} = 10^{-3}$. And RSM is stopped once the function value at the current iterate does not exceed $1.0001 \times f_{\mathrm{SMPG}}$, where f_{SMPG} denotes the final function value returned by SMPG. The initial point for both algorithms is generated by the leading r eigenvectors of $\Sigma_\circ$ in problem ($\mathrm{P_m}$).

We begin by assessing the performance of SMPG and RSM on synthetic datasets. Specifically, the true distribution $\mathbb{P}_*$ is chosen as the normal distribution $\mathbb{N}(0, \Sigma_*)$,



(a) CPU Time (b) Total Number of Iterations

FIG. 2. *Numerical comparison between SMPG and RSM for different values of d with $r = d/100$.*



(a) CPU Time (b) Total Number of Iterations

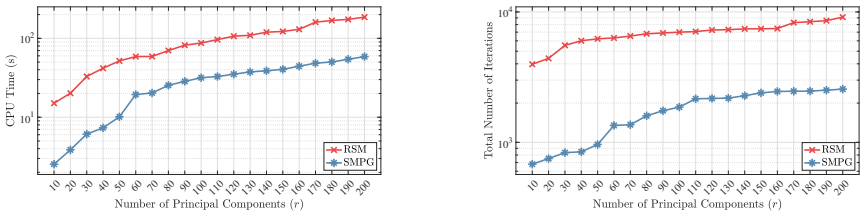
FIG. 3. *Numerical comparison between SMPG and RSM on MNIST for different values of r.*

where the true covariance matrix $\Sigma_* \in \mathbb{S}_+^d$ is obtained by projecting a randomly generated matrix onto $\mathbb{S}_+^d$. Then the nominal distribution $\mathbb{P}_\circ$ is constructed by the empirical distribution $\hat{\mathbb{P}}_n$ based on $n = 300$ samples drawn independently and identically from $\mathbb{N}(0, \Sigma_*)$. In this test, we vary the dimension d within $\{500j \mid j = 1, 2, \ldots, 20\}$ while fixing $r = d/100$, $\gamma = 0.05$, and $\rho = 1$ in problem $(\mathrm{P_m})$. Figure 2 comprises two subplots that depict CPU times and total numbers of iterations required by two algorithms. It can be observed that SMPG outperforms RSM in terms of computational efficiency. Although each iteration of RSM incurs a relatively low cost, its slow convergence necessitates a large number of iterations to reach a desirable accuracy. Therefore, the total CPU time of RSM is considerably higher than that of SMPG.

In the subsequent experiments, we then extend our evaluation to two real-world datasets, MNIST[1] and CIFAR-10[2], to further examine the practical behavior of SMPG and RSM. MNIST contains 60000 samples, each with $d = 784$ features; and CIFAR-10 consists of 50000 samples with $d = 3072$ features per sample. The nominal distribution $\mathbb{P}_\circ$ is constructed in the same manner as before, based on the empirical distribution $\hat{\mathbb{P}}_n$ derived from the first $n = 500$ samples. Other parameters in problem $(\mathrm{P_m})$ are set to $\gamma = 0.02$ and $\rho = 1$ with $r \in \{10j \mid j = 1, 2, \ldots, 20\}$. The corresponding numerical results are presented in Figure 3 and Figure 4 for MNIST and CIFAR-10, respectively. These evidences clearly show that the computational advantage of SMPG is not confined to synthetic scenarios, whose superiority remains pronounced when applied to real-world datasets.

6.4. Performance of DRO Model. The purpose of this subsection is to illustrate the rationality and necessity of adopting the DRO model for PCA. For conve-

[1]https://yann.lecun.com/exdb/mnist/
[2]https://www.cs.toronto.edu/~kriz/cifar.html



(a) CPU Time (b) Total Number of Iterations

FIG. 4. *Numerical comparison between SMPG and RSM on CIFAR-10 for different values of r.*

nience, the DRO model ($\mathrm{P_m}$) and the SAA model (6.1) are denoted by DRPCA and PCA-SAA, respectively. The performances of DRPCA and PCA-SAA are evaluated on three real-world datasets, including MNIST, CIFAR-10, and WHO-Mortality[3]. The last one includes 6080 samples, each characterized by $d = 24$ features.

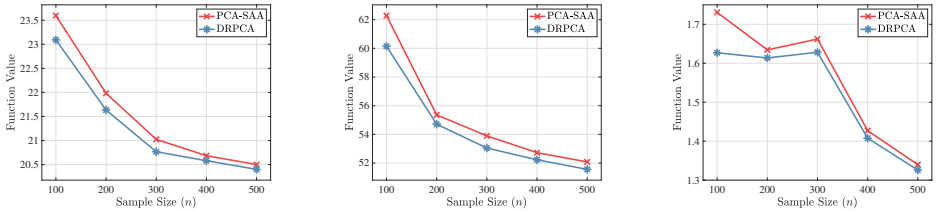
In the first experiment, we extract the first n samples of each dataset to construct the empirical distribution $\hat{\mathbb{P}}_n$ as the nominal distribution $\mathbb{P}_\circ$. Then SMPG and ManPG [10] are deployed to solve the DRPCA model ($\mathrm{P_m}$) and the PCA-SAA model (6.1), respectively. For our simulation in this case, we fix $r = 5$ and $\gamma = 0.02$. The radius ρ in problem ($\mathrm{P_m}$) is set to $5n^{-1/2}$ for each sample size n. To assess the quality of solutions in DRPCA and PCA-SAA, we calculate the out-of-sample performance [20] for the function value of PCA as follows,

$$f_*(X) = \mathbb{E}_{\mathbb{P}_*}\left[\left\|\left(I_d - XX^\top\right)\left(\xi - \mathbb{E}_{\mathbb{P}_*}[\xi]\right)\right\|_2^2\right], \tag{6.2}$$

which represents the reconstruction error of PCA with the true distribution $\mathbb{P}_*$ being the empirical distribution generated by all samples in the dataset. Furthermore, the sparsity levels of solutions, quantified by the percentage of zero entries, are recorded as well. When computing the sparsity level, we set an entry to zero when its absolute value is less than 10^{-5} (see [10]). Figure 5 and Figure 6 visualize the out-of-sample performances of DRPCA and PCA-SAA on three datasets, evaluated across sample sizes $n \in \{100, 200, 300, 400, 500\}$. As evidenced therein, the solutions of DRPCA consistently demonstrate superior performances compared to those of PCA-SAA in terms of function values. In addition, the solutions obtained from both models exhibit comparable sparsity levels, with their differences remaining within an acceptable margin. These numerical results highlight the rationality and necessity of adopting the DRO model for PCA.

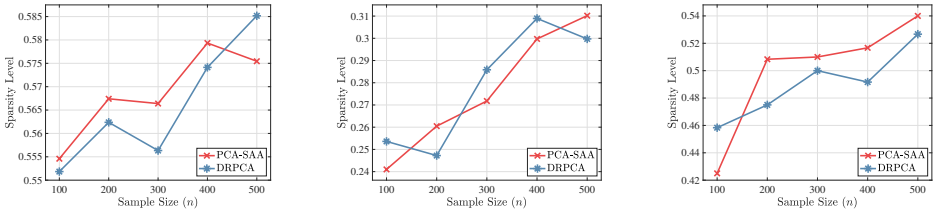
Due to the presence of noise and the variability of practical environments, real-world applications often face the risk of distributional corruption or the emergence of extreme scenarios. We then perform a stress test [19, 26] on DRPCA and PCA-SAA to simulate this situation. The robustness of solutions delivered by both models is examined under a contaminated distribution $\mathbb{P}_\nu = (1-\nu)\hat{\mathbb{P}}_n + \nu\bar{\mathbb{P}}$, where $\bar{\mathbb{P}}$ represents a contaminant of $\hat{\mathbb{P}}_n$ and $\nu \in [0, 1]$ encodes the contamination level. Similar to the setting of [32], we choose a Gaussian contaminant inspired by the extremal-covariance construction in Appendix A. Specifically, we set $\bar{\mathbb{P}} = \mathbb{N}(\zeta, \bar{\Sigma}_{\bar{X}})$ for a solution $\bar{X}$ of DRPCA, where $\zeta \in \mathbb{R}^d$ is a d-dimensional vector of all ones and $\bar{\Sigma}_{\bar{X}} \in \mathbb{S}_+^d$ is given in Theorem A.1. Next, the function values of PCA under $\mathbb{P}_\nu$ are evaluated for $\nu \in [0, 1]$,

[3]https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates



(a) MNIST (b) CIFAR-10 (c) WHO-Mortality

FIG. 5. *Numerical comparison of function values between DRPCA and PCA-SAA.*



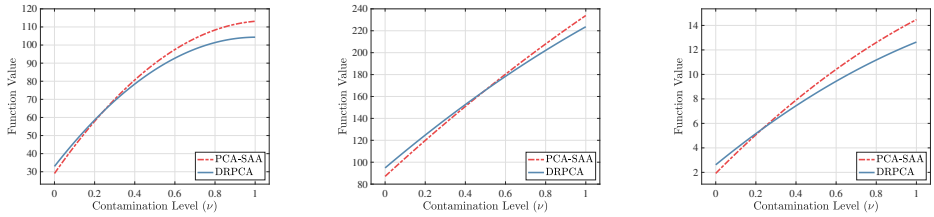
(a) MNIST (b) CIFAR-10 (c) WHO-Mortality

FIG. 6. *Numerical comparison of sparsity levels between DRPCA and PCA-SAA.*

obtained by substituting $\mathbb{P}_*$ with $\mathbb{P}_\nu$ in (6.2). In this experiment, we fix $r = 20$, $\gamma = 0.1$, $n = 100$, and $\rho = 10$. Figure 7 illustrates the numerical results of DRPCA and PCA-SAA on three datasets under the stress test. Beyond a certain level of contamination, the solutions of DRPCA begin to exhibit better performance than those of PCA-SAA. Moreover, the function values of PCA-SAA change more steeply than those of DRPCA, which suggests the enhanced robustness of DRPCA against contaminated data.

7. Concluding Remarks. The DRO model ($\mathrm{P_{mM}}$) of regularized PCA constitutes a constrained min-max optimization problem on a Riemannian manifold. When the ambiguity set is characterized by the type-2 Wasserstein distance, we equivalently reformulate it as a minimization problem ($\mathrm{P_m}$) by providing a closed-form expression for the optimal value of the inner maximization problem in ($\mathrm{P_{mM}}$). However, problem ($\mathrm{P_m}$) can hardly be solved efficiently by existing Riemannian optimization algorithms due to the involvement of two possibly nonsmooth terms in the objective function. To surmount this issue, we develop an efficient algorithm SMPG for problem ($\mathrm{P_m}$), which incorporates the smoothing approximation technique into the proximal gradient method on Riemannian manifolds. We rigorously demonstrate that SMPG achieves global convergence to a stationary point and further establish an iteration complexity. Preliminary numerical results are presented to validate the efficiency of SMPG and the effectiveness of our DRO model, illuminating their potential in addressing the challenges inherent in PCA under distributional uncertainty.

Acknowledgments. We sincerely express our gratitude to Prof. Man-Chung Yue and two anonymous reviewers for their insightful and constructive comments, which have substantially improved the quality, clarity, and presentation of this paper.



(a) MNIST (b) CIFAR-10 (c) WHO-Mortality

FIG. 7. *Numerical comparison of stress tests between DRPCA and PCA-SAA.*

Appendix A. Extremal Distribution. In this appendix, we focus on the case where $p = 2$. For any $X \in \mathcal{O}^{d,r}$, we construct a candidate extremal distribution $\bar{\mathbb{P}}_X$ that achieves the worst-case function value in problem (3.1), namely,

$$\phi(X, \bar{\mathbb{P}}_X) = \varphi(X), \tag{A.1}$$

where $\phi(X, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\|(I_d - XX^\top)(\xi - \mathbb{E}_{\mathbb{P}}[\xi])\|_{\mathrm{F}}^2]$ denotes the objective function of problem (3.1). The closed-form expression of φ is provided in Theorem 3.6. Recall that $\Sigma_\circ \in \mathbb{S}_+^d$ is the covariance matrix of ξ under the nominal distribution $\mathbb{P}_\circ$. For convenience, we define $\mathcal{S}(\Sigma_\circ) = \left\{X \in \mathcal{O}^{d,r} \mid \left(I_d - XX^\top\right)\Sigma_\circ = 0\right\}$. Given that $\Sigma_\circ \in \mathbb{S}_+^d$ and $I_d - XX^\top \in \mathbb{S}_+^d$, it is immediately evident that $X \in \mathcal{S}(\Sigma_\circ)$ if and only if $\mathrm{tr}((I_d - XX^\top)\Sigma_\circ) = 0$.

THEOREM A.1. *Let $\bar{\mathbb{P}}_X$ be a probability distribution with the following covariance matrix,*

$$\bar{\Sigma}_X = \begin{cases} \bar{L}_X \Sigma_\circ \bar{L}_X, & \text{if } X \notin \mathcal{S}(\Sigma_\circ), \\ \Sigma_\circ + \dfrac{\rho^2}{d-r}(I_d - XX^\top), & \text{if } X \in \mathcal{S}(\Sigma_\circ), \end{cases} \tag{A.2}$$

where $\bar{L}_X = I_d + \rho\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{-1/2}\left(I_d - XX^\top\right) \in \mathbb{S}_{++}^d$. Then, for any $X \in \mathcal{O}^{d,r}$, the relationship (A.1) *holds.*

Proof. By simple manipulations, we can obtain that

$$\phi(X, \mathbb{P}) = \mathrm{tr}\left(\left(I_d - XX^\top\right)\mathbb{E}_{\mathbb{P}}\left[\left(\xi - \mathbb{E}_{\mathbb{P}}\left[\xi\right]\right)\left(\xi - \mathbb{E}_{\mathbb{P}}\left[\xi\right]\right)^\top\right]\right) = \mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma\right),$$

where $\Sigma \in \mathbb{S}_+^d$ is the covariance matrix of ξ under the distribution $\mathbb{P}$. Thus, the objective function of problem (3.1) only depends on the covariance matrix. We investigate the following two cases.

Case I: $X \notin \mathcal{S}(\Sigma_\circ)$. It can be readily verified that

$$\mathrm{tr}\left(\bar{\Sigma}_X\right) = \mathrm{tr}\left(\bar{L}_X^2 \Sigma_\circ\right) = \mathrm{tr}\left(\Sigma_\circ\right) + 2\rho\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho^2.$$

By virtue of the orthogonality of X, we have $\bar{L}_X X = X$. Hence, it holds that

$$\begin{aligned} \phi(X, \bar{\mathbb{P}}_X) &= \mathrm{tr}\left(\left(I_d - XX^\top\right)\bar{\Sigma}_X\right) = \mathrm{tr}\left(\bar{\Sigma}_X\right) - \mathrm{tr}\left(X^\top \bar{L}_X \Sigma_\circ \bar{L}_X X\right) \\ &= \mathrm{tr}\left(\Sigma_\circ\right) + 2\rho\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho^2 - \mathrm{tr}\left(X^\top \Sigma_\circ X\right) \\ &= \left(\left(\mathrm{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right)\right)^{1/2} + \rho\right)^2 = \varphi(X). \end{aligned}$$

Case II: $X \in \mathcal{S}(\Sigma_\circ)$. Straightforward calculations give rise to that

$$\begin{aligned}\phi(X, \bar{\mathbb{P}}_X) &= \operatorname{tr}\left(\left(I_d - XX^\top\right)\bar{\Sigma}_X\right) = \operatorname{tr}\left(\left(I_d - XX^\top\right)\Sigma_\circ\right) + \frac{\rho^2}{d-r}\operatorname{tr}\left(I_d - XX^\top\right) \\ &= \rho^2 = \varphi(X).\end{aligned}$$

We complete the proof. ∎

Remark A.2. The formulation in Theorem A.1 does not imply that the eigenvectors of $\bar{\Sigma}_X$ are sparse. In general, the sparsity pattern of these eigenvectors depends on the eigenspace of $\Sigma_\circ$, the loading matrix X, and the Wasserstein radius ρ. No sparsity guarantee should be expected. Sparse PCA seeks to construct a sparse approximation to the principal subspace, balancing explained variance and interpretability. In general, the solution of sparse PCA does not necessarily coincide with that of PCA. From a theoretical perspective, sparsity regularizers are useful to remediate certain inconsistency phenomena inherent in PCA [33].

Remark A.3. The pair $(X, \bar{\Sigma}_X)$ can not be interpreted as a Nash equilibrium in general. Intuitively, the construction of $\bar{\Sigma}_X$ increases the variance in the residual subspace spanned by $I_d - XX^\top$. As a result, the leading eigenspace of $\bar{\Sigma}_X$ may move away from X, rather than coincide with it. From the viewpoint of min-max optimization, the outer variable X in our DRO model (P_{mM}) is constrained to lie on the nonconvex Stiefel manifold. Moreover, the objective function, with respect to X, is neither convex when viewed as a function in the Euclidean space, nor geodesically convex when regarded as a function on the Stiefel manifold. Therefore, the standard convex-concave assumptions for the existence of a Nash equilibrium are not available.

Even though Theorem A.1 furnishes a probability distribution $\bar{\mathbb{P}}_X$ that attains the worst-case function value in problem (3.1), whether this construction falls within the feasible region $\mathscr{B}_2(\mathbb{P}_\circ, \rho)$ remains, at present, beyond our reach to confirm for general cases. This difficulty stems from the fact that computing the Wasserstein distance between two arbitrary distributions is typically intractable, which is shown to be #P-hard in [35]. Fortunately, the type-2 Wasserstein distance between two normal distributions admits a closed-form expression [23, 24]. Leveraging this result, we are able to identify a probability distribution that not only satisfies the relationship (A.1) but is also feasible for problem (3.1), when the nominal distribution is normal.

COROLLARY A.4. *Suppose that the nominal distribution $\mathbb{P}_\circ$ is the normal distribution $\mathbb{N}(\zeta_\circ, \Sigma_\circ)$ with $\zeta_\circ \in \mathbb{R}^d$ and $\Sigma_\circ \in \mathbb{S}_+^d$. Let $\bar{\Sigma}_X \in \mathbb{S}_+^d$ be the matrix constructed in* (A.2)*. Then the global maximum of problem* (3.1) *is attained at $\bar{\mathbb{P}}_X = \mathbb{N}(\zeta_\circ, \bar{\Sigma}_X)$. Namely, the relationship* (A.1) *holds with $\bar{\mathbb{P}}_X \in \mathscr{B}_2(\mathbb{P}_\circ, \rho)$.*

Proof. This is a direct consequence of Theorem A.1 and [23, Theorem 2.1]. The proof is omitted here. ∎

REFERENCES

[1] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, 2008.

[2] A. BECK AND I. ROSSET, *A dynamic smoothing technique for a class of nonsmooth optimization problems on manifolds*, SIAM J. Optim., 33 (2023), pp. 1473–1493.

[3] G. C. BENTO, O. P. FERREIRA, AND J. G. MELO, *Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds*, J. Optim. Theory Appl., 173 (2017), pp. 548–562.

[4] J. Blanchet and Y. Kang, *Sample out-of-sample inference based on Wasserstein distance*, Oper. Res., 69 (2021), pp. 985–1013.
[5] J. Blanchet, Y. Kang, and K. Murthy, *Robust Wasserstein profile inference and applications to machine learning*, J. Appl. Probab., 56 (2019), pp. 830–857.
[6] J. Blanchet and K. Murthy, *Quantifying distributional model risk via optimal transport*, Math. Oper. Res., 44 (2019), pp. 565–600.
[7] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*, Cambridge University Press, 2023.
[8] N. Boumal, P.-A. Absil, and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, IMA J. Numer. Anal., 39 (2018), pp. 1–33.
[9] S. Chen, Z. Deng, S. Ma, and A. M.-C. So, *Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning*, IEEE Trans. Signal Process., 69 (2021), pp. 4759–4773.
[10] S. Chen, S. Ma, A. M.-C. So, and T. Zhang, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM J. Optim., 30 (2020), pp. 210–239.
[11] S. Chen, S. Ma, A. M.-C. So, and T. Zhang, *Nonsmooth optimization over the Stiefel manifold and beyond: Proximal gradient method and recent variants*, SIAM Rev., 66 (2024), pp. 319–352.
[12] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99.
[13] X. Chen, Y. He, and Z. Zhang, *Tight error bounds for the sign-constrained Stiefel manifold*, SIAM J. Optim., 35 (2025), pp. 302–329.
[14] X. Chen, H. Sun, and H. Xu, *Discrete approximation of two-stage stochastic and distributionally robust linear complementarity problems*, Math. Program., 177 (2019), pp. 255–289.
[15] H. T. M. Chu, M. Lin, and K.-C. Toh, *Wasserstein distributionally robust optimization and its tractable regularization formulations*, arXiv:2402.03942, (2024).
[16] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.
[17] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Oper. Res., 58 (2010), pp. 595–612.
[18] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Math. Program., 178 (2019), pp. 503–558.
[19] J. Dupačová, *Stress testing via contamination*, in Coping with Uncertainty: Modeling and Policy Issues, Berlin, 2006, Springer, pp. 29–46.
[20] P. M. Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Math. Program., 171 (2018), pp. 115–166.
[21] R. Gao, *Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality*, Oper. Res., 71 (2023), pp. 2291–2306.
[22] R. Gao and A. Kleywegt, *Distributionally robust stochastic optimization with Wasserstein distance*, Math. Oper. Res., 48 (2023), pp. 603–655.
[23] M. Gelbrich, *On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces*, Math. Nachr., 147 (1990), pp. 185–203.
[24] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, Mich. Math. J., 31 (1984), pp. 231–240.
[25] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Oper. Res., 58 (2010), pp. 902–917.
[26] G. A. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler, *Distributionally robust multi-item newsvendor problems with multimodal demand distributions*, Math. Program., 152 (2015), pp. 1–32.
[27] B. He, Y. You, and X. Yuan, *On the convergence of primal-dual hybrid gradient algorithm*, SIAM J. Imaging Sci., 7 (2014), pp. 2526–2537.
[28] N. J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, 2008.
[29] S. Hosseini, W. Huang, and R. Yousefpour, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, SIAM J. Optim., 28 (2018), pp. 596–619.
[30] S. Hosseini and A. Uschmajew, *A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds*, SIAM J. Optim., 27 (2017), pp. 173–189.
[31] W. Huang and K. Wei, *Riemannian proximal gradient methods*, Math. Program., 194 (2022), pp. 371–413.
[32] J. Jiang, H. Sun, and X. Chen, *Data-driven distributionally robust multiproduct pricing problems under pure characteristics demand models*, SIAM J. Optim., 34 (2024), pp. 2917–

2942.
[33] I. M. Johnstone and A. Y. Lu, *On consistency and sparsity for principal components analysis in high dimensions*, J. Am. Stat. Assoc., 104 (2009), pp. 682–693.
[34] L. V. Kantorovich and S. Rubinshtein, *On a space of totally additive functions*, Vestn. St. Petersb. Univ.: Math., 13 (1958), pp. 52–59.
[35] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, *Wasserstein distributionally robust optimization: Theory and applications in machine learning*, in Operations Research & Management Science in the Age of Analytics, INFORMS, 2019, pp. 130–166.
[36] D. Kuhn, S. Shafiee, and W. Wiesemann, *Distributionally robust optimization*, Acta Numer., 34 (2025), pp. 579–804.
[37] R. Lai and S. Osher, *A splitting method for orthogonality constrained problems*, J. Sci. Comput., 58 (2014), pp. 431–449.
[38] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. M.-C. So, *Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods*, SIAM J. Optim., 31 (2021), pp. 1605–1634.
[39] X. Liu, N. Xiao, and Y.-X. Yuan, *A penalty-free infeasible approach for a class of nonsmooth optimization problems over the Stiefel manifold*, J. Sci. Comput., 99 (2024), p. 30.
[40] G.-F. Lu, J. Zou, Y. Wang, and Z. Wang, *L1-norm-based principal component analysis with adaptive regularization*, Pattern Recognit., 60 (2016), pp. 901–907.
[41] Z. Lu and Y. Zhang, *An augmented Lagrangian approach for sparse principal component analysis*, Math. Program., 135 (2012), pp. 149–193.
[42] V. A. Nguyen, D. Kuhn, and P. M. Esfahani, *Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator*, Oper. Res., 70 (2022), pp. 490–515.
[43] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.
[44] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer Science & Business Media, 2009.
[45] A. Shapiro, *On duality theory of conic linear problems*, in Semi-Infinite Programming: Recent Advances, M. Á. Goberna and M. A. López, eds., Springer, Boston, 2001, pp. 135–165.
[46] B. P. Van Parys, P. M. Esfahani, and D. Kuhn, *From data to decisions: Distributionally robust optimization is optimal*, Manag. Sci., 67 (2021), pp. 3387–3402.
[47] H. Vu, T. Tran, M.-C. Yue, and V. A. Nguyen, *Distributionally robust fair principal components via geodesic descents*, in Proc. 10th Int. Conf. Learn. Represent., 2022, pp. 1–23.
[48] L. Wang, L. Bao, and X. Liu, *A decentralized proximal gradient tracking algorithm for composite optimization on Riemannian manifolds*, J. Mach. Learn. Res., 26 (2025), pp. 1–37.
[49] L. Wang and X. Chen, *An adaptive smoothing algorithm for non-Lipschitz optimization on manifolds with complexity guarantees*, arXiv:2604.18325, (2026).
[50] L. Wang and X. Liu, *Decentralized optimization over the Stiefel manifold by an approximate augmented Lagrangian function*, IEEE Trans. Signal Process., 70 (2022), pp. 3029–3041.
[51] L. Wang and X. Liu, *Smoothing gradient tracking for decentralized optimization over the Stiefel manifold with non-smooth regularizers*, in Proc. 62nd IEEE Conf. Decis. Control., IEEE, 2023, pp. 126–132.
[52] L. Wang, X. Liu, and X. Chen, *A support-set algorithm for optimization problems with nonnegative and orthogonal constraints*, arXiv:2511.03443, (2025).
[53] L. Wang, X. Liu, and Y. Zhang, *Seeking consensus on subspaces in federated principal component analysis*, J. Optim. Theory Appl., 203 (2024), pp. 529–561.
[54] Z. Wang, B. Liu, S. Chen, S. Ma, L. Xue, and H. Zhao, *A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis*, INFORMS J. Optim., 4 (2022), pp. 200–214.
[55] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Math. Program., 142 (2013), pp. 397–434.
[56] W. Wiesemann, D. Kuhn, and B. Rustem, *Robust Markov decision processes*, Math. Oper. Res., 38 (2013), pp. 153–183.
[57] H. Xu, Y. Liu, and H. Sun, *Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods*, Math. Program., 169 (2018), pp. 489–529.
[58] W. H. Yang, L.-H. Zhang, and R. Song, *Optimality conditions for the nonlinear programming problems on Riemannian manifolds*, Pac. J. Optim., 10 (2014), pp. 415–434.
[59] J. Zhang, S. Ma, and S. Zhang, *Primal-dual optimization algorithms over Riemannian manifolds: An iteration complexity analysis*, Math. Program., 184 (2020), pp. 445–490.
[60] L. Zhang, J. Yang, and R. Gao, *A short and general duality proof for Wasserstein distributionally robust optimization*, Oper. Res., 73 (2025), pp. 2146–2155.