A RETRACTION-FREE METHOD FOR NONSMOOTH MINIMAX OPTIMIZATION OVER A COMPACT MANIFOLD

A PREPRINT

Necdet Serhat Aybat[*]
Department of Industrial and Manufacturing Engineering
Penn State University
University Park, PA, USA
nsa10@psu.edu

Jiang Hu[*]
Yau Mathematical Sciences Center
Tsinghua University
Beijing 100084, China
hujiangopt@mail.com

Zhanwang Deng[†]
Academy for Advanced Interdisciplinary Studies
Peking University
Beijing 100871, China
dzw_opt2022@stu.pku.edu.cn

December 9, 2025

ABSTRACT

We study the minimax optimization problem over a compact submanifold $\mathcal{M}$, i.e., $\min_{x\in\mathcal{M}} \max_y f_r(x,y) := f(x,y) - h(y)$, where f is continuously differentiable in (x,y), h is a closed, weakly-convex (possibly non-smooth) function and we assume that the regularized coupling function $-f_r(x,\cdot)$ is either μ-PL for some $\mu > 0$ or concave ($\mu = 0$) for any fixed x in the vicinity of $\mathcal{M}$. To address the nonconvexity due to the manifold constraint, we reformulate the problem using an *exact penalty* for the constraint $x \in \mathcal{M}$ and enforcing a convex constraint $x \in X$ for some $X \supset \mathcal{M}$ onto which projections can be computed efficiently. Building upon this new formulation for the manifold minimax problem in question, a single-loop smoothed manifold gradient descent-ascent (`sm-MGDA`) algorithm is proposed. Theoretically, any limit point of `sm-MGDA` sequence is a stationary point of the manifold minimax problem and `sm-MGDA` can generate an $\mathcal{O}(\epsilon)$-stationary point of the original problem with $\mathcal{O}(\kappa/\epsilon^2)$ and $\tilde{\mathcal{O}}(l^4/\epsilon^4)$ complexity for $\mu > 0$ and $\mu = 0$ scenarios, respectively, where $\kappa = l/\mu$ is the condition number and l denotes the Lipschitz constant of the gradient corresponding to the penalized problem over $X \times \operatorname{dom} h$. Moreover, for the $\mu = 0$ setting, through adopting Tikhonov regularization of the dual, one can improve the complexity to $\mathcal{O}(l^2/\epsilon^3)$ at the expense of asymptotic stationarity. The key component, common in the analysis of all cases, is to connect ϵ-stationary points between the penalized problem and the original problem by showing that the constraint $x \in X$ becomes inactive and the penalty term tends to 0 along any convergent subsequence. To our knowledge, `sm-MGDA` is the first retraction-free algorithm for minimax problems over compact submanifolds, and this is a very desirable algorithmic property since through avoiding retractions, one can get away with matrix orthogonalization subroutines required for computing retractions to manifolds arising in practice, which are not GPU friendly. Experiments on quadratic minimax problems, robust deep neural network training, and superquantile-based learning demonstrate clear advantages over state-of-the-art algorithms that rely on retraction operation in each iteration.

[*]These authors contributed equally to this work.
[†]The student coauthor contributed to the numerical experiments.

1 Introduction

Due to the broad applications in machine learning, minimax optimization has attracted significant attention arising in the context of generative adversarial networks (GANs) [17], robust deep neural network training [18, 28, 37], superquantile-based learning [10, 13], and reinforcement learning [14, 16]. Minimax problems involving manifold constraints naturally arise in applications such as robust geometry-aware PCA [21] and those involving subspace robust Wasserstein distance optimization [33, 40]. Furthermore, adding manifold constraints, such as orthogonality, on the model parameters has been found effective in accelerating the training [3, 11, 27], preventing the gradient sequence from diminishing or exploding, and improving generalization [12].

In this paper, we consider the following minimax optimization problem over a *compact submanifold* of a normed vector space $(\mathcal{X}, \|\cdot\|)$ with $\mathcal{X} := \mathbb{R}^{d_1 \times r}$ and $\mathcal{Y} := \mathbb{R}^{d_2}$:

$$\min_{x \in \mathcal{M}} \max_{y \in \mathcal{Y}} f_r(x, y) := f(x, y) - h(y), \tag{1}$$

where $\mathcal{M} := \{x \in \mathbb{R}^{d_1 \times r} : c(x) = 0\}$ for some smooth function $c : \mathbb{R}^{d_1 \times r} \to \mathbb{R}^p$, $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed, (possibly nonsmooth) ζ-weakly convex function[3] for some $\zeta \geq 0$ with a compact, convex domain $Y := \{y \in \mathcal{Y} : h(y) < \infty\}$, and f is continuously differentiable on some open set containing $\bar{X} \times Y$ where $\bar{X} \supset \mathcal{M}$ is some compact set that will be defined later. In the rest, we used $f_r(\cdot, \cdot)$ to denote $f(\cdot, \cdot)$ regularized with $h(\cdot)$.

We study (1) assuming that one of the following conditions holds: **(i)** $-f_r(x, \cdot)$ is μ-PL, i.e., it satisfies Polyak-Lojasiewicz condition for some $\mu > 0$, uniformly for all $x \in \bar{X}$ (see Definition 3), or **(ii)** $f_r(x, \cdot)$ is concave for all $x \in \bar{X}$ (abusing the notation, we abbreviate this scenario by $\mu = 0$). Throughout the paper we mainly focus on the setting where $\mathcal{M} := \{x \in \mathbb{R}^{d_1 \times r} : x^\top x = I_r\}$ is the Stiefel manifold, i.e., $c(x) = x^\top x - I_r$, and later in section 3.4 we discuss how our results can be extended to a more general class of smooth defining functions $c(\cdot)$.

Algorithms for solving the problem (1) have been extensively studied in the literature [7, 20, 23, 24, 25, 47, 48]. Among these works, [24, 47, 48] are the only works we are aware of studying a manifold minimax problem with a smooth f *without* assuming geodesic convexity, i.e., when f is not geodesically convex in x for some fixed $y \in Y$. This distinction is important, as by the Hopf-Rinow theorem, any geodesically convex function defined on a *compact* Riemannian manifold must be constant [31, Theorem 6.13]. Consequently, the results in [7, 20, 23, 25] are not directly applicable to problem (1) when $f(\cdot, y)$ is not constant for fixed $y \in Y$.

In [47, 48], the authors consider nonconvex-linear minimax problems on Riemannian manifolds, and under the linearity assumption in the dual, they establish $\mathcal{O}(1/\epsilon^3)$ iteration complexity. Neither RADA method in [47] nor ARPGDA in [48] is retraction-free, they require computing a retraction onto the manifold at each iteration – ARPGDA is a single-loop method and RADA requires inexactly solving smooth-strongly concave minimax subproblems over the manifold at each iteration. On the other hand, [24] considers nonconvex-strongly concave minimax problems, and it establishes $\mathcal{O}(1/\epsilon^2)$ complexity for the proposed RGDA method; that said, RGDA is also not retraction-free and it requires computing a retraction onto $\mathcal{M}$ at each iteration. For the main use case considered in this paper, the Stiefel manifold [1], the retraction always involves some expensive linear algebra operation, such as matrix inversion, exponential or square-root, which quickly become expensive as the dimension of the matrices grows (especially for the square Stiefel manifold where $d_1 = r$). Therefore, in this paper we will investigate the following natural question:

Can one design a retraction-free single-loop first-order method with cheap per-iteration complexity to efficiently compute stationary points of nonconvex-PL and nonconvex-concave minimax problems on compact submanifolds?

Our main contributions are listed below.

Retraction-free smoothed manifold GDA algorithm. We reformulate the manifold-constrained minimax problem using an *exact penalty* for the manifold constraint; moreover, we also introduce a norm-ball constraint $x \in X$ to ensure that f is smooth over the compact set $X \times Y$ with a global Lipschitz constant for ∇f –let $l > 0$ denote this constant. We propose a smoothed manifold gradient descent-ascent (`sm-MGDA`) method that operates entirely in Euclidean space, without requiring retractions or projections onto the manifold. This eliminates matrix orthogonalization and leads to a GPU-friendly implementation. Numerical experiments on quadratic minimax examples, robust deep neural network training, and superquantile-based learning demonstrate clear advantages over state-of-the-art methods. To our knowledge, `sm-MGDA` is the first retraction-free algorithm for minimax problems over compact submanifolds.

Convergence guarantee for both merely concave and PL settings. We establish that `sm-MGDA` finds an $\mathcal{O}(\epsilon)$-stationary point with $\tilde{\mathcal{O}}(l^4/\epsilon^4)$ complexity in the merely concave case and $\mathcal{O}(\kappa/\epsilon^2)$ in the PL case, where $\kappa = l/\mu$. The key technical ingredients in our analysis are to relate stationarity in the penalized problem to that of the original

[3] We say a function h is ζ-weakly convex if $h(\cdot) + \frac{\zeta}{2}\|\cdot\|$ is convex for some $\zeta \geq 0$.

problem by exploiting the exact penalty property of the penalty function adopted in our reformulation, and to show that the bound constraint $x \in X$ eventually becomes inactive under the assumption that the primal function is lower bounded. Compared to [24], our method removes the requirement for strong concavity and can handle the settings where $f_r(x,\cdot)$ is merely concave, or $-f_r(x,\cdot) = h(\cdot) - f(x,\cdot)$ is μ-PL with a possibly non-smooth weakly convex regularizer $h(\cdot)$. To our knowledge, `sm-MGDA` is the first algorithm with provable convergence guarantees for solving (1) in the merely concave and non-smooth μ-PL settings without resorting to retraction operations. Moreover, we also discuss how to relax the compactness assumption on the dual domain Y if we strengthen the μ-PL assumption to strong concavity with modulus $\mu > 0$ –in short, we call it μ-concave. We establish that in all of these setting, any limit point of `sm-MGDA` sequence is a stationary point of the manifold minimax problem. Finally, for the $\mu = 0$ setting, we also show that through adopting Tikhonov regularization of the dual, one can improve the complexity to $\mathcal{O}(l^2/\epsilon^3)$ for computing an $\mathcal{O}(\epsilon)$-stationary point at the expense of losing the asymptotic stationarity of the iterate sequence.

Notation. For a vector $x \in \mathbb{R}^d$, $\|x\|$ denotes the Euclidean norm; for a matrix $x \in \mathbb{R}^{d\times r}$, $\|x\|$ and $\|x\|_2$ denote the Frobenius and spectral norms, respectively. For a closed convex set X, let $\delta_X(\cdot)$ denote its indicator function, defined as $\delta_X(x) = 0$ if $x \in X$ and $+\infty$ otherwise. The projection of a point x onto X is denoted by $\mathcal{P}_X(x) := \arg\min_{y\in X}\|x-y\|^2$, and the distance from x to X is defined as $\operatorname{dist}(x,X) := \|x - \mathcal{P}_X(x)\|$. For any square matrix $x \in \mathbb{R}^{n\times n}$, we define the symmetrization operator as $\operatorname{sym}(x) := \frac{1}{2}(x + x^\top)$; moreover, $\operatorname{diag}(x) \in \mathbb{R}^n$ denotes the vector of diagonal elements of x. Given a proper, closed function h, we use $\partial h(x)$ to denote the Fréchet subdifferential[4] at x, i.e., $\partial h(x) := \{s : \liminf_{y\to x} \frac{h(y)-h(x)-\langle s, y-x\rangle}{\|y-x\|} \geq 0\}$. When $-h(\cdot)$ is weakly convex, we use $-\partial h(x)$ to denote the Fréchet subdifferential of $-h$ at x, i.e., $-\partial h(x) := \partial(-h(x))$. Finally, for a point $x \in \mathcal{M}$, $T_x\mathcal{M}$ denotes the tangent space of the manifold $\mathcal{M}$ at x. For a differentiable function h, we write $\nabla h(x)$ and $\operatorname{grad} h(x)$ for its Euclidean and Riemannian gradients at x, respectively. Let $g : \mathbb{R}^n \to \mathbb{R}^p$ be a differentiable map at x, then we use $\mathbf{J}_g(x) \in \mathbb{R}^{p\times n}$ to denote the Jacobian matrix at x, and for any given $u \in \mathbb{R}^p$ we use $\nabla g(x)[u] := \mathbf{J}_g(x)^\top u$.

Table 1: **Comparison of convergence guarantees and assumptions.** In the column **"Single-loop"**, we indicate whether the method consists of a single-loop iteration or not. In **"GNC $f(\cdot,y)$"** column, we indicate whether *geodesic nonconvexity* of $f(\cdot,y)$ over $\mathcal{M}$ is allowed when y is fixed; in **"Nonlinear $f(x,\cdot)$"** column we indicate whether nonlinearity of $f(x,\cot)$ is allowed when $x \in \mathcal{M}$ is fixed. In the two columns about complexity, $\mu = 0$ and $\mu > 0$ correspond to the cases where $-f_r(x,\cdot)$ are convex and μ-PL, respectively. Finally, in the column **"RF"** we state whether the method is *retraction free* or not. The stationarity metrics used across different works are as follows: [24] considers $\min_{x\in\mathcal{M}}\max_{y\in Y} f(x,y)$ and adopts the criterion $\|\operatorname{grad} F(x_\epsilon)\| \leq \epsilon$, where $F(x) := \max_{y\in Y} f(x,y)$ for $x \in \mathcal{X}$; [25] considers $\min_{x\in\mathcal{M}}\max_{y\in\mathcal{N}} f(x,y)$ where $\mathcal{M}$ and $\mathcal{N}$ are Riemannian manifolds, and the metric adopted for ϵ-stationarity is $\sqrt{\|x_\epsilon - x^*\|^2 + \|y_\epsilon - y^*\|^2} \leq \epsilon$, where (x^*, y^*) denotes the unique saddle point under the assumption on f satisfying geodesic strong convexity–geodesic strong concavity; [47] considers (1) such that $f(x,y) = g(x) + \langle \mathcal{A}(x), y\rangle$ and $h(\cdot)$ is a closed convex function, and adopts two different criteria for measuring ϵ-stationarity, i.e., ϵ-RGS: $\max\left\{\|\operatorname{grad}_x f(x_\epsilon, y_\epsilon)\|, \frac{1}{\gamma}\|y_\epsilon - \operatorname{prox}_{\gamma h}(y_\epsilon + \gamma\mathcal{A}(x_\epsilon))\|\right\} \leq \epsilon$, and ϵ-ROS: $\max\left\{\operatorname{dist}\left(0, \operatorname{grad} g(x_\epsilon) + \mathcal{P}_{T_{x_\epsilon}\mathcal{M}}\left(\nabla\mathcal{A}(x_\epsilon)^\top \partial h^*(p_\epsilon)\right)\right), \|p_\epsilon - \mathcal{A}(x_\epsilon)\|\right\} \leq \epsilon$, where h^* is the Fenchel conjugate of h; and [48] considers $\min_{x\in\mathcal{M}}\max_{y\in Y} f(x,y)$ such that $f(x,\cdot)$ is linear for every $x \in \mathcal{M}$ fixed. For our `sm-MGDA` algorithm we measure stationarity using the criterion $\max\left\{\|\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon)\|, \operatorname{dist}\Big(0, -\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)\Big)\right\} \leq \epsilon$.

Algorithm	**Single-loop**	**GNC $f(\cdot,y)$**	**Nonlinear $f(x,\cdot)$**	**Complexity ($\mu = 0$)**	**Complexity ($\mu > 0$)**	**RF**
RGDA [24]	✓	✓	✓	N/A	$\mathcal{O}(\epsilon^{-2})$	✗
RCEG [25]	✓	✗	✓	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\log(\epsilon^{-1}))$	✗
RADA [47]	✗	✓	✗	$\mathcal{O}(\epsilon^{-3})$	N/A	✗
ARPGDA [48]	✓	✓	✗	$\mathcal{O}(\epsilon^{-3})$	N/A	✗
`sm-MGDA`	✓	✓	✓	$\mathcal{O}(\epsilon^{-3})$ or $\mathcal{O}(\epsilon^{-4})^\diamond$	$\mathcal{O}(\epsilon^{-2})$	✓

$^\diamond$ These two complexity results arise under different algorithmic settings and analysis frameworks. The $\mathcal{O}(\epsilon^{-3})$ bound for computing an ϵ-stationary point is achieved by introducing a Tikhonov regularization in the y-variable, i.e., given an arbitrary $\bar{y} \in Y$, solving the nonconvex–strongly concave approximate problem $\min_{x\in\mathcal{M}}\max_{y\in\mathcal{Y}} f_r(x,y) - \mu\|y - \bar{y}\|^2$ for $\mu = \mathcal{O}(\epsilon)$. In contrast, the $\mathcal{O}(\epsilon^{-4})$ bound corresponds to directly solving the original problem, with additional algorithmic parameters introduced to guarantee asymptotic convergence in terms of any limit point of the iterate sequence being a stationary point of the manifold minimax problem. Although the former bound is better for small $\epsilon > 0$, the generated iterates converge only to a solution of the approximate problem rather than the original one in (1).

[4] When h is proper closed convex, ∂h concides with the convex subdifferential, and if h is differentiable at x, $\partial h(x) = \{\nabla h(x)\}$.

2 Methodology & Smoothed Manifold GDA (sm-MGDA) Algorithm

Recently, the constraint dissolving method for Riemannian optimization has been proposed in [44] to cast a manifold-constrained optimization problem into an *unconstrained* one. Inspired by this methodology, which is designed for manifold optimization (in the primal sense), given a compact smooth submanifold $\mathcal{M} := \{x \in \mathbb{R}^{d_1 \times r} : c(x) = 0\}$ for some smooth $c : \mathbb{R}^{d\times r} \to \mathbb{R}^p$, we can rewrite the minimax problem in (1) as

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \tilde{f}_r(x, y) := \tilde{f}(x, y) - h(y), \quad \text{where} \quad \tilde{f}(x, y) := f(A(x), y) + \frac{\rho}{4}\|c(x)\|^2, \tag{2}$$

where $A : \mathcal{X} \to \mathcal{X}$ is the constraint dissolving operator, and $\rho > 0$ is a large enough penalty parameter–the assumptions on $c(\cdot)$ and $A(\cdot)$ are similar to [44, Assumptions 1.1 and 1.2] and will be specified later in section 3. The partial gradient $\nabla_x \tilde{f}(x, y)$ can be computed using chain rule as follows:

$$\nabla_x \tilde{f}(x, y) = \nabla A(x)[\nabla_x f(A(x), y)] + \frac{\rho}{2}\nabla c(x)[c(x)], \tag{3}$$

where $\nabla A(x)[u] := \mathbf{J}_A(x)^\top u$ for $x, u \in \mathbb{R}^{d_1}$ with $\mathbf{J}_A(x) \in \mathbb{R}^{d_1\times d_1}$ denoting the Jacobian of $A(\cdot)$ at x, and $\nabla c(x)[u] := \mathbf{J}_c(x)^\top u$ for $x \in \mathbb{R}^{d_1}$ and $u \in \mathbb{R}^p$ with $\mathbf{J}_c(x) \in \mathbb{R}^{p\times d_1}$ denoting the Jacobian of $c(\cdot)$ at x.

For example, when $\mathcal{M}$ is the Stiefel manifold, i.e., $\mathcal{M} = \{x \in \mathbb{R}^{d_1\times r} : x^\top x = I_r\}$, we can set $A(x) = x(\frac{3}{2}I_r - \frac{1}{2}x^\top x)$ and $c(x) = x^\top x - I_r$, for which we have $\nabla A(x)[u] := u(\frac{3}{2}I_r - \frac{1}{2}x^\top x) - x\mathrm{sym}(x^\top u)$ and $\nabla c(x)[u] = 2xu$.

We first state our assumptions on the manifold minimax problem in (1).

Assumption 1. ***(i)*** *Let $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed, ζ-weakly convex function with a closed domain $Y := \mathrm{dom}\, h = \{y \in \mathcal{Y} : h(y) < \infty\}$, and h is locally Lipschitz on its domain.* ***(ii)*** *Suppose Y is a bounded set.*

Remark 1. *Assumption 1 implies that there exists a constant $l_h > 0$ such that h is Lipschitz continuous over the compact set Y, i.e., $|h(y_1) - h(y_2)| \leq l_h\|y_1 - y_2\|$ for any $y_1, y_2 \in Y$.*

Definition 1. *Suppose $\mathcal{M} \subset \mathcal{X}$ is a compact submanifold. Define $\bar{X} := \{A(x) : \|x\| \leq C\} \subset \mathcal{X} = \mathbb{R}^{d_1\times r}$ for some $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$.*

Assumption 2. *Suppose that $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is differentiable on an open set containing $\bar{X} \times Y$, where $\bar{X} \subset \mathcal{X}$ is given in Definition 1 and $Y = \mathrm{dom}\, h \subset \mathcal{Y}$, and that there exist $L_{xx}, L_{xy}, L_{yx}, L_{yy} \geq 0$ such that*

$$\begin{aligned}
&\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \leq L_{xx}\|x_1 - x_2\| + L_{xy}\|y_1 - y_2\|, \\
&\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \leq L_{yx}\|x_1 - x_2\| + L_{yy}\|y_1 - y_2\|, \quad x_1, x_2 \in \bar{X}, \quad y_1, y_2 \in Y.
\end{aligned}$$

Let $L := \max\{L_{xx}, L_{xy}, L_{yx}, L_{yy}\}$.

Definition 2. *$F : \mathcal{X} \to \mathbb{R}$ denotes the primal function, i.e., $F(x) := \max_{y\in\mathcal{Y}} f_r(x, y)$, and let $F^* := \min_{x\in\mathcal{M}} F(x)$.*

In the rest, we analyze the convergence properties of the proposed sm-MGDA algorithm applied to (2) under different assumptions on f_r defined in (1). First, we consider the scenario where $-f_r(x, \cdot)$ satisfies the PL property for fixed x.

The PL property is usually given for smooth functions [26, Theorem 2] and its extension for composite functions $p_s + p_c$ where p_s is smooth and p_c is a proper, closed convex function, is given in [26, Eq. (12)]. On the other hand, the definition we adopted from [32] is more general and applies to a more general class of possibly nonsmooth ζ-weakly convex functions.

Definition 3 ([32])**.** *Let $\varphi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a proper, closed, ζ-weakly convex function, and let $S = \arg\min_x \varphi(x)$. Suppose $S \neq \emptyset$ and let $\varphi^* = \varphi(x^*)$ for some $x^* \in S$. $\varphi(\cdot)$ satisfies PL inequality with constant $\mu > 0$, i.e., we say $\varphi(\cdot)$ is μ-PL, if $2\mu(\varphi(x) - \varphi^*) \leq \mathrm{dist}^2(0, \partial\varphi(x))$ for all $x \in \mathrm{dom}\,\varphi$.*

For particular examples of nonsmooth ζ-weakly convex functions that satisfy PL condition, see [32, Sec. 3.2].

Consider f_r defined in (1). We assume that $-f_r(x, \cdot)$ satisfies one of the following conditions uniformly for all $x \in \bar{X}$: *(i)* μ-PL (see assumption 3), *(ii)* μ-strongly convex, and *(iii)* merely convex, i.e., $\mu = 0$.

Assumption 3. *There exists $\mu > 0$ such that $-f_r(x, \cdot)$ is μ-PL for all $x \in \bar{X} \subset \mathcal{X}$, i.e.,*

$$\max_{w\in\mathcal{Y}} f_r(x, w) - f_r(x, y) \leq \frac{1}{2\mu}\,\mathrm{dist}^2\Big(0, -\nabla_y f(x, y) + \partial h(y)\Big), \quad \forall x \in \bar{X},\ y \in Y.$$

Remark 2. *The solution set $\arg\max_{w\in\mathcal{Y}} f_r(x, w)$ is nonempty from the compactness of Y. By [32, Theorem 3.1], the assumption on $-f_r(x, \cdot)$ satisfying μ-PL condition is weaker than assuming $f_r(x, \cdot)$ is strongly concave with modulus $\mu > 0$; indeed, it is equivalent to the error–bound condition and further implies quadratic growth of $-f_r(x, \cdot)$. Examples of nonconvex–PL problems include $\min_x \max_y f(x, By)$ where $f(\cdot, \cdot)$ is nonconvex–strongly concave and B is an arbitrary matrix, as well as generative adversarial imitation learning for the linear–quadratic regulator [6].*

Assumption 4. *There exists $\mu > 0$ such that $f_r(x,\cdot)$ is μ-strongly concave for all $x \in \bar{X} \subset \mathcal{X}$, i.e., for any $y \in Y$ and $g \in \partial h(y)$, it holds that $f_r(x,y) + \langle \nabla_y f(x,y) - g,\ w - y\rangle - \frac{\mu}{2}\|w-y\|^2 \geq f_r(x,w)$ for all $x \in \bar{X}$, and $w \in Y$.*

Assumption 5. *$h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is merely convex, i.e., $\zeta = 0$ in assumption 1; moreover, $f(x,\cdot)$ is merely concave for all $x \in \bar{X} \subset \mathcal{X}$, i.e., for any $y \in Y$ fixed, $f(x,y) + \langle \nabla_y f(x,y),\ w - y\rangle \geq f(x,w)$ for all $x \in \bar{X}$ and $w \in Y$.*

Remark 3. *One can replace assumption 2 with the assumption that f is twice continuously differentiable on an open set containing $\bar{X} \times Y$, where $\bar{X}$ is given in Definition 1.*

Given $(x^*, y^*) \in \mathcal{X} \times Y$, the results in [44] imply that for sufficiently large but fixed $\rho > 0$, if $\operatorname{dist}(x^*, \mathcal{M}) \leq \delta$ for some small enough $\delta > 0$, then

$$\nabla_x \tilde{f}(x^*, y^*) = 0 \quad \Leftrightarrow \quad x^* \in \mathcal{M}, \quad \operatorname{grad}_x f(x^*, y^*) = 0, \tag{4}$$

where $\operatorname{grad}_x f(x^*, y^*)$ denotes the partial Riemannian gradient of f with respect to x over $\mathcal{M}$. For any $(\bar{x}, \bar{y}) \in \mathcal{M} \times Y$, the partial Riemannian gradient of f with respect to x can be computed using

$$\operatorname{grad}_x f(\bar{x}, \bar{y}) = \nabla_x f(\bar{x}, \bar{y}) - \bar{x}\operatorname{sym}(\bar{x}^\top \nabla_x f(\bar{x}, \bar{y})). \tag{5}$$

For details on the first-order optimality condition in manifold optimization, we refer to [2, 5, 22].

Definition 4. *$(x^*, y^*) \in \mathcal{M} \times Y$ is a stationary point for the minimax problem in* (1) *if*

$$\operatorname{grad}_x f(x^*, y^*) = 0, \quad 0 \in -\nabla_y f(x^*, y^*) + \partial h(y^*); \tag{6}$$

and $(x_\epsilon, y_\epsilon) \in \mathcal{M} \times Y$ is an ϵ-stationary point, if $\|\operatorname{grad}_x f(x_\epsilon, y_\epsilon)\| \leq \epsilon$ and $\operatorname{dist}\Big(0, -\nabla_y f(x_\epsilon, y_\epsilon) + \partial h(y_\epsilon)\Big) \leq \epsilon$.

Note that because of the penalty term $P(x) := \frac{\rho}{4}\|c(x)\|^2$ used within $\tilde{f}$, $\nabla\tilde{f}$ may not be Lipschitz on $\mathcal{X} \times \mathcal{Y}$ while ∇f is. For example, when $\mathcal{M}$ is the Stiefel manifold, we set $c(x) = x^\top x - I_r$; hence, the corresponding penalty term $P(\cdot)$ is quartic and $\nabla P(x) = \rho x(x^\top x - I_r)$ is not Lipschitz on $\mathcal{X} = \mathbb{R}^{d_1 \times r}$. However, since we are interested in stationary points $(x^*, y^*) \in \mathcal{M} \times Y$ as in (6) with $\mathcal{M}$ being a compact submanifold of $\mathbb{R}^{d_1 \times r}$, instead of (2), we will consider the following reformulation:

$$\min_{x \in X \subset \mathcal{X}} \max_{y \in \mathcal{Y}} \tilde{f}_r(x, y), \tag{7}$$

where $X := \{x \in \mathbb{R}^{d_1 \times r} : \|x\| \leq C\}$ for some $C > \frac{1}{2} + \sup_{x \in \mathcal{M}}\|x\|$ –for this reformulation, trivially $\nabla\tilde{f}$ is Lipschitz on the new domain $X \times Y$ that is compact, and $-\tilde{f}_r(x,\cdot)$ is μ-PL (or convex) for every fixed $x \in X$ since $-f_r(x,\cdot)$ is assumed to be μ-PL (or convex) for all $x \in \bar{X}$. It will be shown later that as long as the constant $C > 0$ is sufficiently large, i.e., $C > \frac{1}{2} + \sup_{x \in \mathcal{M}}\|x\|$, the choice does not affect the set of limit points of the proposed algorithm. Let $l(C)$ denote the Lipschitz constant of $\nabla\tilde{f}$ on $X \times Y$ –for simplicity of notation, we will suppress its dependence on C in the rest and use l instead.

In the rest of this paper, for notational convenience we focus on the Stiefel manifold to derive our results. That said, as mentioned in the introduction, we discuss their extensions to some other important compact manifolds with $c \in \mathcal{C}^2$; more specifically, the *oblique manifold* and the *generalized Stiefel manifold* will be discussed in section 3.4.

Lemma 1. *Let $\mathcal{M} \subset \mathcal{X}$ be the Steifel manifold, i.e., $c(x) = x^\top x - I_r$, and let $X = \{x \in \mathbb{R}^{d_1 \times r} : \|x\| \leq C\}$ for $C > 0$ as in Definition 1. Under Assumption 2, the function $\tilde{f}$ defined in* (2) *with $A(x) = x(\frac{3}{2}I_r - \frac{1}{2}x^\top x)$ is differentiable on an open set containing $X \times Y$, and there exist constants $l_{xx}, l_{xy}, l_{yx}, l_{yy} \geq 0$ such that $l_{yy} = L_{yy}$, $l_{yx} = \mathcal{O}(C^2 L_{yx})$, $l_{xy} = \mathcal{O}(C^2 L_{xy})$ and $l_{xx} = \mathcal{O}(C^2\rho + C^4 L_{xx})$ and that for all $x_1, x_2 \in X$ and $y_1, y_2 \in Y$, it holds that*

$$\begin{aligned}
\left\|\nabla_x \tilde{f}\,(x_1, y_1) - \nabla_x \tilde{f}\,(x_2, y_2)\right\| &\leq l_{xx}\,\|x_1 - x_2\| + l_{xy}\,\|y_1 - y_2\|\,, \\
\left\|\nabla_y \tilde{f}\,(x_1, y_1) - \nabla_y \tilde{f}\,(x_2, y_2)\right\| &\leq l_{yx}\,\|x_1 - x_2\| + l_{yy}\,\|y_1 - y_2\|\,.
\end{aligned}$$

Clearly, $\nabla\tilde{f}$ is Lipschitz on $X \times Y$ with constant $l := \max\{l_{xx}, l_{xy}, l_{yx}, l_{yy}\}$.

Proof. See Section 6.1. □

Remark 4. *If the condition $\|x\| \leq C$ in the definition of X is replaced by $\|x\|_2 \leq C$, the expressions for the Lipschitz constants $l_{xx}, l_{xy}, l_{yx}, l_{yy}$ in Lemma 1 remain the same.*

Remark 5. *The Lipschitz constants of $\tilde{f}$ calculated in the proof of Lemma 1 are specific to the case where $\mathcal{M}$ is the Stiefel manifold, i.e., $c(x) = x^\top x - I_r$. That said, they can be easily extended to other compact smooth submanifolds $\mathcal{M} = \{x \in \mathcal{X} :\ c(x) = 0\}$ over $X \times Y$ whenever the defining function $c(\cdot)$ is twice continuous differentiable.*

2.1 Smoothed AGDA (sm-MGDA) for solving problem (1)

The smoothed gradient descent ascent (sm-GDA) methods [51, 52] perform single-loop updates on the primal and dual variables, together with an additional update on a smoothing variable. These methods are designed for problems with a nonconvex–concave or nonconvex–PL (strongly concave) structure assuming Lipschitz continuous gradients, and under such conditions, they achieve better iteration complexity than the standard gradient descent ascent methods while computatinal cost per iteration being the same.

It is known that every geodesically convex function over a compact manifold is a constant function [4]. This indicates that for any fixed $y \in Y$ and $z \in X$, $f(x,y;z) \triangleq f(x,y) + \frac{p}{2}\|x-z\|^2$ is not geodesically convex in x over $\mathcal{M}$ for any $p > 0$. Although we may use a retraction operator to design a smoothed descent-ascent type method as in [51, 52], its convergence analysis would be very intricate because for any given $z \in X$, the duality gap between $\min_{x\in\mathcal{M}} \max_{y\in Y} f(x,y;z)$ and $\max_{y\in Y} \min_{x\in\mathcal{M}} f(x,y;z)$ is not necessarily *zero* due to the lack of (geodesic) convexity. This motivates us to use more advanced tools to tackle the manifold constraint which leads to a nonconvex minimax problem.

Specifically, we consider the penalized problem in (7), which employs an exact penalty function $\tilde{f}_r(\cdot, y)$ for any $y \in Y$ to eliminate the manifold constraint $x \in \mathcal{M}$. By introducing a norm ball constraint $x \in X$, we avoid potential issues that would arise from the gradient of the penalty term $\frac{\rho}{4}\|c(x)\|^2$ not being Lipschitz continuous over $\mathcal{X}$. In this setup, we study the convergence behavior of smoothed gradient descent ascent iterate sequence [51, 52] to a stationary point of the original problem when sm-GDA is applied on the penalty problem in (7). Indeed, we incorporate the geometry of the manifold through using an exact penalty function, and we refer this particular implementation of sm-GDA framework on (7) as sm-MGDA.

Let $p \in \mathbb{R}_+$ such that $p > l$, e.g., $p = 2l$, and define

$$\hat{f}(x,y;z) := \tilde{f}(x,y) + \frac{p}{2}\|x-z\|^2, \qquad \forall\, x \in X,\ y \in Y,\ z \in \mathcal{X},$$

where $\frac{p}{2}\|x-z\|^2$ serves as a regularizer for dual smoothing, inspired by the Moreau-Yosida regularization, a.k.a. the Nesterov's smoothing [39]. Our proposed method, sm-MGDA, is presented in Algorithm 1, which consists of alternating x, y updates: one projected gradient descent in x using $\nabla_x \hat{f}(x_t, y_t; z_t)$, one proximal gradient ascent in y using $\nabla_y \hat{f}(x_{t+1}, y_t; z_t)$, and an averaging step in z to get z_{t+1}. On the other hand, for the case $\mu = 0$, rather than employing dual smoothing, we set $p = 0$ and incorporate an additional regularization term on y for primal smoothing, following [36, 49]. Indeed, for the scenario $\mu = 0$, rather than computing an ascent step based on $\tilde{f}(x_{t+1}, \cdot)$ for the y-update, we consider $\tilde{f}_{\theta_t}(x_{t+1}, \cdot)$ where $\tilde{f}_\theta(x,y) = \tilde{f}(x,y) - \frac{\theta}{2}\|y\|^2$ for $x \in X$ and $y \in Y$, i.e., we add a regularization term $-\frac{\theta_t}{2}\|y\|^2$ to $\tilde{f}$, and use $\nabla_y \tilde{f}_{\theta_t}(x_{t+1}, y_t)$ rather than $\nabla_y \hat{f}(x_{t+1}, y_t; z_t) = \nabla_y \tilde{f}(x_{t+1}, y_t)$, where the parameter sequence $\{\theta_t\} \subset \mathbb{R}_+$ is diminishing to 0. In our theoretical analysis, we show that the limit points of sm-MGDA iterate sequence are stationary points of the original problem (1) in the sense of Definition 4 under appropriate choices of the parameter $C > 0$ appearing in the definition of X and the penalty parameter ρ in $\tilde{f}$.

Remark 6. *One can avoid computing/estimating the Lipschitz constant l by adopting the line-search strategy of [54], which employs a carefully designed nonmonotone stepsize-search criterion and requires at most 3 backtracking steps per iteration. This type of extension would be helpful in practice as it can exploit local curvature through estimating local Lipschitz constants, which would lead to larger steps.*

Algorithm 1 Smoothed MGDA (sm-MGDA)

Input: $(x_0, y_0, z_0) \in \mathcal{M} \times Y \times \mathcal{M}$, $\{\theta_t\}, \{\tau_{1,t}\} \subset \mathbb{R}_+$, $\tau_2, p, \rho, C > 0$, $\beta \in (0,1]$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  $g_t \leftarrow \nabla A(x_t)[\nabla_x f(A(x_t), y_t)] + \frac{\rho}{2}\nabla c(x_t)[c(x_t)]$
  $x_{t+1} \leftarrow \mathcal{P}_X\left(x_t - \tau_{1,t}\Big(g_t + p(x_t - z_t)\Big)\right)$
  $y_{t+1} \leftarrow \mathrm{prox}_{\tau_2 h}\Big(y_t + \tau_2\Big(\nabla_y f(A(x_{t+1}), y_t) - \theta_t y_t\Big)\Big)$
  $z_{t+1} = z_t + \beta(x_{t+1} - z_t)$
end for
Output: $\{(x_t, y_t)\}_{t=0}^{T-1}$.

2.2 ϵ-stationary points of the penalty problem

Following [52, Definition 3.1], we call $(x_\epsilon, y_\epsilon) \in X \times Y$ an ϵ-stationary point of (7) if

$$\operatorname{dist}\Big(0, \nabla_x \tilde{f}(x_\epsilon, y_\epsilon) + \partial \delta_X(x_\epsilon)\Big) \leq \epsilon, \;\; \operatorname{dist}\Big(0, -\nabla_y \tilde{f}(x_\epsilon, y_\epsilon) + \partial h(y_\epsilon)\Big) \leq \epsilon. \tag{8}$$

Given $(x_\epsilon, y_\epsilon) \in X \times Y$ such that $\|x_\epsilon\| < C$, then $0 \in \nabla_x \tilde{f}(x_\epsilon, y_\epsilon) + \partial \delta_X(x_\epsilon)$ if and only if $\nabla_x \tilde{f}(x_\epsilon, y_\epsilon) = 0$; hence, if (8) holds with $\epsilon = 0$ for some $(x_\epsilon, y_\epsilon) \in X \times Y$ such that $\|x_\epsilon\| < C$, then

$$\nabla_x \tilde{f}(x_\epsilon, y_\epsilon) = 0, \quad 0 \in -\nabla_y \tilde{f}(x_\epsilon, y_\epsilon) + \partial h(y_\epsilon). \tag{9}$$

Thus, it follows from (4) that for $\rho > 0$ large enough and $C > \delta + \sup_{x \in \mathcal{M}} \|x\|$, there exists $\delta > 0$ such that if $\operatorname{dist}(x_\epsilon, \mathcal{M}) \leq \delta$ and (8) holds with $\epsilon = 0$, then $(x_\epsilon, y_\epsilon) \in \mathcal{M} \times Y$ and it is a stationary point for the original manifold constrained minimax problem in (1), i.e., (6) holds with $(x^*, y^*) = (x_\epsilon, y_\epsilon)$.

In practice, since the algorithm runs for only a finite number of iterations, we are also interested in how the ϵ-stationary point of the penalized problem relates to that of the original manifold-constrained problem. The following lemma describes this connection for the Steifel manifold.

Lemma 2. *Let $\mathcal{M} \subset \mathbb{R}^{d_1 \times r}$ be the Steifel manifold. For any given $\epsilon > 0$, let $(x_\epsilon, y_\epsilon) \in \mathcal{X} \times Y$ be such that $\|\nabla_x \tilde{f}(x_\epsilon, y_\epsilon)\| \leq \epsilon$, $\operatorname{dist}(0, -\nabla_y \tilde{f}(x_\epsilon, y_\epsilon) + \partial h(y_\epsilon)) \leq \epsilon$ and $\operatorname{dist}(x_\epsilon, \mathcal{M}) \leq \frac{1}{2}$, then*

$$\|x_\epsilon - \mathcal{P}_{\mathcal{M}}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon, \quad \|\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon)\| \leq \epsilon + \frac{11}{\rho}\Big(L_{xx} + L_x(y_\epsilon)\Big)\epsilon,$$
$$\operatorname{dist}\Big(0, -\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)\Big) \leq \Big(1 + \frac{3}{\rho} L_{yx}\Big)\epsilon,$$

whenever $\rho \geq 36 L_x(y_\epsilon)$, where $L_x(y) := \max\{\|\nabla_x f(x, y)\| : \|x\|_2 \leq 1\}$ is defined for $y \in Y$. Thus, $(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) \in \mathcal{M} \times Y$ is an $\mathcal{O}(\epsilon)$-stationary point for the minimax problem in (1) *in terms of Definition 4.*

Proof. See Section 6.2. □

For any given $\epsilon > 0$, in order to compute $(x_\epsilon, y_\epsilon) \in X \times Y$ satisfying the hypothesis of Lemma 2 we will employ `sm-MGDA` on the penalty problem in (7) with a properly chosen parameter $C > 0$, i.e., the choice of $C > 0$ is not arbitrary and it should be chosen carefully depending on δ in order to exclude the possibility of computing an $(x_\epsilon, y_\epsilon) \in X \times Y$ as in (8) with $\|x_\epsilon\| = C$. Indeed, Lemma 2 shows that when $\mathcal{M} \subset \mathbb{R}^{d_1 \times r}$ is the Steifel manifold, any $C > \frac{1}{2} + r$ works as $\delta = 1/2$ and $\sup_{x \in \mathcal{M}} \|x\| = \sqrt{r}$.

3 Convergence analysis

In this section, we present the convergence guarantees of our `sm-MGDA` algorithm under three different scenarios: We assume that $-f_r(x, \cdot)$ satisfies one of the following conditions uniformly for all $x \in \bar{X}$: *(i)* μ-PL (see assumption 3), *(ii)* μ-strongly convex, and *(iii)* merely convex, i.e., $\mu = 0$.

Definition 5. *Let $-\infty < \bar{F} := \min_{x \in X} F(A(x))$ where $F(\cdot) = \max_{y \in \mathcal{Y}} f_r(\cdot, y)$.*

3.1 Complexity for nonconvex-PL problems

In this section, we establish the convergence guarantees for computing an ϵ-stationary point of the manifold minimax problem in (1), and we also provide asymptotic convergence results. To achieve this goal, we first extend the analysis of [51] to handle weakly convex (possibly non-smooth) regularizer $h(\cdot)$ –the method proposed in [51] can only handle *smooth* problems of the form $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$ without any manifold constraint, where f is a smooth function satisfying assumption 2 on the vector space $\mathcal{X} \times \mathcal{Y}$ such that $-f(x, \cdot)$ is μ-PL for all $x \in \mathcal{X}$.

We next provide a convergence rate result for `sm-MGDA` under Assumptions 1,2 and 3. For this setting there is no clear relation between l and μ; hence, we define the modified condition number $\bar{\kappa} := \max\{1, \kappa\}$ where $\kappa := l/\mu$. Furthermore, we also define some other important quantities arising in our analysis.

Definition 6. *For any $z \in \mathcal{X}$, let $x^*(z) := \arg\min_{x \in X} \Phi(x; z)$ and $Y^*(z) := \arg\max_{y \in \mathcal{Y}} \Psi_r(y; z)$, where $\Phi(x; z) := \max_{y \in \mathcal{Y}} \hat{f}_r(x, y; z)$ defined for any $x \in X$, $\Psi_r(y; z) := \Psi(y; z) - h(y)$ and $\Psi(y; z) := \min_{x \in X} \hat{f}(x, y; z)$ defined for any $y \in Y$. Finally, $P(z) := \min_{x \in X} \Phi(x; z)$ for $z \in \mathcal{X}$.*

Theorem 1. *Suppose Assumptions 1,2 and 3 hold. For any given $C > 0$, let $X = \{x \in \mathcal{X} : \|x\| \leq C\}$, and let $\{x_t, y_t, z_t\}_{t\geq 0}$ be the sm-MGDA iterate sequence generated by Algorithm 1, initialized from an arbitrary $(x_0, y_0, z_0) \in X \times Y \times \mathcal{X}$, using the following parameters: $\tau_{1,t} = \tau_1$ and $\theta_t = 0$ for all $t \geq 0$ for some $\tau_1 \in (0, \frac{1}{3l}]$, $\tau_2 = \frac{1}{16}(\frac{3}{\tau_1} + \zeta)^{-1}$, $p = 2l$ and $\beta = \alpha \min\{\mu, l\}\tau_2$ for some $\alpha \in (0, 1/2306)$, where the constant $l > 0$ is defined in Lemma 1. Then, for any $T \geq 1$, it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\|G_t^x\|^2 + \bar{\kappa}\,\|G_t^y\|^2\Big) \leq \frac{O(1)\bar{\kappa}}{T}\Big(\frac{1}{\tau_1} + \zeta\Big)\Big(P(z_0) - \bar{F} + \Delta_0\Big), \tag{10}$$

where $\Delta_0 := \hat{f}_r(x_0, y_0; z_0) + P(z_0) - 2\Psi_r(y_0; z_0)$, and for all $t \geq 0$, G_t^x, G_t^y are defined as

$$\begin{aligned} G_{t+1}^x &:= \frac{x_t - x_{t+1}}{\tau_1} + \nabla_x \tilde{f}(x_{t+1}, y_{t+1}) - \nabla_x \tilde{f}(x_t, y_t) + p(z_t - x_t), \\ G_{t+1}^y &:= \frac{y_{t+1} - y_t}{\tau_2} + \nabla_y \tilde{f}(x_{t+1}, y_{t+1}) - \nabla \tilde{f}(x_{t+1}, y_t). \end{aligned} \tag{11}$$

Furthermore, $G_t^x \in \nabla_x \tilde{f}(x_t, y_t) + \partial\delta_X(x_t)$ and $G_t^y \in \nabla_y \tilde{f}(x_t, y_t) - \partial h(y_t)$ for all $t \geq 1$, which also implies that

$$\min_{1\leq t\leq T} \max\Big\{\mathrm{dist}\Big(0, \nabla_x \tilde{f}(x_t, y_t) + \partial\delta_X(x_t)\Big),\ \sqrt{\bar{\kappa}}\mathrm{dist}\Big(0, -\nabla_y \tilde{f}(x_t, y_t) + \partial h(y_t)\Big)\Big\} = \mathcal{O}\Big(\sqrt{\bar{\kappa}/T}\Big).$$

Finally, $\Delta_0 \geq 0$ can be bounded as $\Delta_0 \leq 2\,\mathrm{gap}_{\hat{f}_r}(x_0, y_0; z_0)$, where $\mathrm{gap}_{\hat{f}_r}(x_0, y_0; z_0) := \Phi(x_0; z_0) - \Psi_r(y_0; z_0)$.

Proof. See section 7. □

Remark 7. *If one chooses $x_0 = z_0$ and $y_0 \in Y^*(z_0)$ for any $z_0 \in \mathcal{M}$, then $\hat{f}_r(x_0, y_0; z_0) = f_r(z_0, y_0) \leq F(z_0)$. Moreover, choosing $y_0 \in Y^*(z_0)$ implies that $P(z_0) + \Delta_0 = \hat{f}_r(x_0, y_0; z_0)$ since $P(z_0) = \Psi_r(y_0; z_0)$ –see Lemma 7. Therefore, $P(z_0) + \Delta_0 - \bar{F} \leq F(z_0) - \bar{F}$.*

On the other hand, if one chooses $x_0 = x^*(z_0) \in X$ and $y_0 \in Y^*(z_0)$ for some $z_0 \in \mathcal{M}$, then one has $\Delta_0 = 0$ –see Lemma 7. Moreover, since $z_0 \in \mathcal{M}$, we have $P(z_0) = \min_{x\in X} \Phi(x; z_0) \leq \Phi(z_0; z_0) = F(z_0)$; therefore, $P(z_0) + \Delta_0 - \bar{F} \leq F(z_0) - \bar{F}$ as well.

Next, we argue that for both $C, \rho > 0$ sufficiently large, $\|x_t\| < C$ for all $t \geq 0$; thus, Theorem 1 implies that $\nabla f(x^*, y^*) = 0$ for any limit point (x^*, y^*) of the sm-MGDA iterate sequence $\{(x_t, y_t)\}$.

Theorem 2. *Under the premise of Theorem 1, suppose sm-MGDA is initialized from $(x_0, y_0, z_0) \in X \times Y \times \mathcal{X}$ such that $y_0 \in Y^*(z_0)$ and x_0 is set to either z_0 or $x^*(z_0)$ for some arbitrary $z_0 \in \mathcal{M}$. Then, for any $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$, there exists $\bar{\rho} > 0$ such that within $\mathcal{O}(\bar{\kappa}/\epsilon^2)$ gradient evaluations, sm-MGDA with given parameters $\rho \geq \bar{\rho}$ and $C > 0$ can generate $(x_\epsilon, y_\epsilon) \in X \times \mathcal{Y}$ such that $\|x_\epsilon - \mathcal{P}_{\mathcal{M}}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon$, $\|\mathrm{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon)\| = \mathcal{O}(\epsilon)$, and $\mathrm{dist}(0, -\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)) = \mathcal{O}(\epsilon)$.*

Moreover, any limit point (x^*, y^*) of the sm-MGDA sequence $\{(x_t, y_t)\}_{t\geq 0}$ is a stationary point for the minimax problem in (1), i.e., $x^* \in \mathcal{M}$, $\mathrm{grad}_x f(x^*, y^*) = 0$ and $0 \in -\nabla_y f(x^*, y^*) + \partial h(y^*)$.

Proof. See section 8. □

Remark 8. *For any given $z_0 \in \mathcal{M}$, let $x_0 = x^*(z_0) \in X$ and $y_0 \in Y^*(z_0)$. Then, for any $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$, choosing $\rho \geq 16\Big(F(z_0) - \bar{F} + (\bar{l}_y + l_h)D_Y\Big)$, where $\bar{l}_y := \max_{x\in X, y\in Y}\|\nabla_y f(A(x), y)\| < \infty$ and $D_Y := \sup_{y_1, y_2\in Y}\|y_1 - y_2\|$, implies that $\|c(x_t)\| \leq \frac{1}{2}$, which ensures $\mathrm{dist}(x_t, \mathcal{M}) \leq \frac{1}{2}$ for all $t \geq 0$.*

Remark 9. *The above result corresponds to the case where X is defined using the Frobenius norm. Instead, if X is defined by the spectral norm ball, the same iteration complexity bound continues to hold for any $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|_2 = \frac{3}{2}$. Since the expressions for the Lipschitz constants remain unchanged (see Remark 4), this smaller value of C (in contrast to $\frac{1}{2} + \sqrt{r}$) yields an improved iteration complexity $\mathcal{O}(\max\{L/\mu, 1\}/\epsilon^2)$ as $L \leq l$, where L is defined in Assumption 2. However, it should be noted that in this setting the projection operator $\mathcal{P}_X$ requires computing an SVD of an $n_1 \times r$ matrix at each iteration, which can be substantially more expensive than the projection onto the Frobenius norm ball, especially when $r \approx n_1$, e.g., $n_1 = r$.*

3.2 Complexity for nonconvex-strongly concave problems

In this section, we replace Assumption 3 with Assumption 4 which is clearly a stronger one; indeed, [32, Theorem 3.1] shows that Assumption 4 implies Assumption 3 –see also Remark 2. That said, while adopting a stronger condition of f_r, we now relax the compactness requirement on $Y = \operatorname{dom} h$, which is necessary for our analysis in the μ-PL case, i.e., we will study (1) under Assumptions 1.**(i)**, 2 and 4.

Definition 7. *Let $\Phi(x) := \max_{y\in\mathcal{Y}} \tilde{f}_r(x,y)$ and $\Phi^* = \min_{x\in X}\Phi(x)$. Note that $\Phi(x) = F(A(x)) + \frac{\rho}{4}\|c(x)\|^2$; hence, $\Phi^* \geq \bar{F}$. Moreover, define $r^*(x) \triangleq \arg\max_{y\in\mathcal{Y}} \tilde{f}_r(x,y) = \arg\max_{y\in\mathcal{Y}} f_r(A(x),y)$ for all $x \in X$, and let $\delta_t := \|y_t - r^*(x_t)\|^2$ for $t \geq 0$.*

First, we argue that when $f_r(x,\cdot)$ is strongly concave on Y for all $x \in \bar{X}$, the `sm-MGDA` iterate sequence is bounded. Note that for every $z \in \mathcal{X}$, due to strong concavity $Y^*(z)$ is a singleton, let $Y^*(z) = \{y^*(z)\}$.

Lemma 3. *Suppose Assumptions 1.**(i)**,2 and 4 hold. Given an arbitrary $C > 0$ such that $X \supset \mathcal{M}$, let $\{x_t, y_t, z_t\} \subset X \times \mathcal{Y} \times \mathcal{X}$ be generated by `sm-MGDA` using $\tau_1, \tau_2, p, \beta, \alpha$ as stated in Theorem 1. Then starting from any $(x_0, y_0, z_0) \in X \times \mathcal{Y} \times X$, $\{x_t, y_t, z_t\}_{t\geq 0}$ is a bounded sequence; indeed, $\max\{\|x_t\|, \|z_t\|\} \leq C$. If one initializes `sm-MGDA` from (x_0, y_0, z_0) as in Theorem 2, then $\|y_t - y_0\| \leq \sqrt{\frac{2}{\mu}(F(z_0) - \Phi^*)} + 2\kappa_{yx}\|z_t - z_0\|$ for all $t \geq 0$, where $\kappa_{yx} := l_{yx}/\mu$.*

Proof. See section 9.1. □

Next, we state a result that establishes $\{\|y_t - r^*(x_t)\|^2\} \to 0$ as $t \to \infty$.

Lemma 4. *For $t \geq 0$, $\delta_{t+1} \leq (1 - \tau_2\mu/2)\delta_t + \frac{(1-\tau_2\mu)(2-\tau_2\mu)}{\tau_2\mu}\frac{l_{yx}^2}{\mu^2}\|x_{t+1} - x_t\|^2$; moreover, $\sum_{t=0}^{\infty}\delta_t < \infty$, which implies that $\delta_t \to 0$ as $t \to \infty$.*

Proof. See section 9.2. □

Next, we argue that for both $C, \rho > 0$ sufficiently large, $\|x_t\| < C$ for all $t \geq 0$; thus, Theorem 1 implies that any limit point of $\{x_t, y_t\}$ is a stationary point for the minimax problem in (1) in terms of Definition 4.

Theorem 3. *Under Assumptions 1.**(i)**,2 and 4, there exists some $\bar{\rho} > 0$ such that for every $\rho > \bar{\rho}$ there is $\bar{T}_\rho \in \mathbb{Z}_+$, which is non-increasing in ρ, such that the results of Theorem 2 continue to hold for all $t \geq \bar{T}_\rho$.*

Proof. See section 9. □

3.3 Complexity for nonconvex-merely concave problems

Beyond the μ-PL and strong concavity settings for $f_r(x,\cdot)$, we can also establish an $\tilde{\mathcal{O}}(\epsilon^{-4})$ complexity of our `sm-MGDA` algorithm to compute an ϵ-stationary point of the original problem under mere concavity of $f_r(x,\cdot)$ for any $x \in \bar{X}$. This result follows from Theorem 3 and the potential function construction provided in [49, Theorem 3.2].

Theorem 4. *Suppose Assumptions 1,2 and 5 hold. For any $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$, there exists $\bar{\rho} > 0$ such that within $\tilde{\mathcal{O}}(l^4\epsilon^{-4})$ gradient evaluations, initialized from an arbitrary $(x_0, y_0, x_0) \in \mathcal{M} \times Y \times \mathcal{M}$ and using parameters $\rho > \bar{\rho}, \tau_2 \leq \frac{1}{10l_{yy}}, \tau_{1,t} = \frac{2}{2\tau_2 l^2(1+b\sqrt{t+1})-l}, b > \max\left\{\frac{2}{\tau_2 l} - 1, \frac{32\cdot 20^2}{19^2}\right\}, \theta_t = \frac{19}{20}\cdot\frac{1}{\tau_2}\cdot\frac{1}{(t+1)^{1/4}}, p = 0$, and $\beta = 1$, `sm-MGDA` can generate $(x_\epsilon, y_\epsilon) \in X \times \mathcal{Y}$ such that $\|x_\epsilon - \mathcal{P}_{\mathcal{M}}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon$, $\|\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon)\| = \mathcal{O}(\epsilon)$, and $\operatorname{dist}\Big(0, -\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)\Big) = \mathcal{O}(\epsilon)$. Moreover, for any limit point (x^*, y^*), it holds that $x^* \in \mathcal{M}$, $\operatorname{grad}_x f(x^*, y^*) = 0$ and $0 \in -\nabla_y f(x^*, y^*) + \partial h(y^*)$.*

Proof. See section 10 in the appendix. □

Remark 10. *The above theorem establishes the first complexity result for a retraction-free method to solve nonconvex-merely-concave (NCMC) minimax problems over compact submanifolds.*

Remark 11. *In addition to the $\mathcal{O}(l^4\epsilon^{-4})$ complexity result established in Theorem 4, we can also obtain an improved $\mathcal{O}(\epsilon^{-3})$ bound by applying `sm-MGDA` to the regularized problem:*

$$\min_{x\in\mathcal{M}}\max_{y\in\mathcal{Y}} f_r^\epsilon(x,y) := f(x,y) - h(y) - \frac{\epsilon}{4D_Y}\|y - y_0\|^2,$$

where $D_Y := \max_{y_1,y_2\in Y}\|y_1 - y_2\|$ *and* $y_0 \in Y$. $f_r^\epsilon(x,\cdot)$ *is strongly concave with modulus* $\mu = \dfrac{\epsilon}{2D_Y}$ *for any fixed* x. *Therefore, by Theorem 2,* `sm-MGDA` *yields a pair* (x_ϵ, y_ϵ) *such that* $\|x_\epsilon - \mathcal{P}_\mathcal{M}(x_\epsilon)\| \leq \dfrac{3}{\rho}\epsilon$, $\|\mathrm{grad}_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon)\| = \mathcal{O}(\epsilon)$, $\mathrm{dist}\left(0, -\nabla_y f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) + \dfrac{\epsilon}{2D_Y}(y_\epsilon - y_0) + \partial h(y_\epsilon)\right) = \mathcal{O}(\epsilon)$. *Since* $\bar{\kappa} = \kappa = l/\mu = \mathcal{O}(l\epsilon^{-1})$ *and* $D_Y \geq \|y_\epsilon - y_0\|$, *it follows that* $\mathrm{dist}(0, -\nabla_y f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)) = \mathcal{O}(\epsilon)$; *therefore,* (x_ϵ, y_ϵ) *is an* ϵ*-stationary point of the original NCMC minimax problem* $\min_{x\in\mathcal{M}}\max_{y\in\mathcal{Y}} f_r(x,y)$, *and it can be computed within* $\mathcal{O}(l^2\epsilon^{-3})$ *iterations.*

3.4 Extension to other compact submanifolds

In previous sections, for the theoretical analysis of `sm-MGDA`, we focus on the case where $\mathcal{M}$ is the Stiefel manifold. However, the concept of exact penalty functions introduced in [45] applies more broadly to all compact submanifolds satisfying some mild regularity conditions. In the following, we discuss how our algorithm and its guarantees can be extended to some other important compact submanifolds.

Oblique manifold: $\mathcal{M} = \{x \in \mathbb{R}^{d\times r} : \mathrm{diag}(x^\top x) = I_r\}$. In this case, we define

$$A(x) = 2x\left(I_r + \mathrm{diag}(x^\top x)\right)^{-1}, \quad c(x) = \mathrm{diag}(x^\top x) - I_r.$$

To ensure that the same convergence guarantees hold, it suffices to (i) validate Lemma 2, and (ii) ensure that small constraint violation $c(x)$ still implies that x is close to the manifold.

For (i), we can use [44, Theorem 1] to control the feasibility error and Riemannian gradient norm by gradient norm of the penalized objective function when x is sufficiently close to the manifold and ρ is chosen large enough.

For (ii), let $\bar{x}$ denote the projection of x onto $\mathcal{M}$, given by $\bar{x} = x\,\mathrm{diag}(x^\top x)^{-\frac{1}{2}}$. Then, $\|x - \bar{x}\| = \|x\left(I_r - \mathrm{diag}(x^\top x)^{-\frac{1}{2}}\right)\|$. If $\|c(x)\| \leq \epsilon$ with $\epsilon \leq \frac{1}{2}$, then we have $1 - \epsilon \leq \max_{1\leq i\leq r}\|x_i\| \leq 1 + \epsilon$, and

$$\|x - \bar{x}\| \leq \|x\|_2\left\|I_r - \mathrm{diag}(x^\top x)^{-\frac{1}{2}}\right\| \leq \frac{3}{2}\left\|I_r - \mathrm{diag}(x^\top x)^{-\frac{1}{2}}\right\| \leq \frac{3}{2}\sqrt{r\epsilon}.$$

Generalized Stiefel Manifold: $\mathcal{M} = \{x \in \mathbb{R}^{d\times r} : x^\top Bx = I_r\}$ where $B \in \mathbb{R}^{d\times d}$ is a positive definite matrix. In this case, we define

$$A(x) = x\left(\frac{3}{2}I_r - \frac{1}{2}x^\top Bx\right), \quad c(x) = x^\top Bx - I_r.$$

Following [44, Theorem 1], we can establish results analogous to Lemma 2, ensuring approximate stationarity under the exact penalty framework.

To show that small $c(x)$ implies small $\mathrm{dist}(x,\mathcal{M})$, consider the compact SVD of $B^{\frac{1}{2}}x$: $B^{\frac{1}{2}}x = usv^\top$. Define $\bar{x} = B^{-\frac{1}{2}}uv^\top$. Then, $\|x - \bar{x}\| = \left\|B^{-\frac{1}{2}}\left(usv^\top - uv^\top\right)\right\| \leq \|B^{-\frac{1}{2}}\|\|s - I_r\|$. Noting that $c(x) = x^\top Bx - I_r = vs^\top sv^\top - I_r = v(s^2 - I_r)v^\top$, we obtain $\|s^2 - I_r\| = \|c(x)\|$, which implies $\|s - I_r\| \leq \|c(x)\|$. Thus, we establish the desired result by having $\|x - \bar{x}\| \leq \|B^{-\frac{1}{2}}\|\,\|c(x)\|$.

4 Numerical experiments

In this section, we conduct experiments on three tasks, i.e., the quadratic nonconvex-merely concave (NCMC) problem, the quadratic nonconvex-strongly concave (NCSC) problem, and robust DNNs training over a Riemannian manifold. In experiments, we compare our `sm-MGDA` with MGDA, (i.e., $p = 0$ and $\beta = 0$ in Algorithm 1), RGDA [24, 35], and its stochastic version RSGDA [24] as the comparison baselines. We use constant primal step sizes for `sm-MGDA`, i.e., $\tau_{1,t} = \tau_1$ for all $t = 0, 1, \ldots$. For all figures in this section, we run each algorithm several times independently, and plot the geometric mean and standard deviation in solid lines and shaded regions, respectively.

4.1 Quadratic NCMC problem ($\mu = 0$)

4.1.1 Case 1: Vector variables

In the first scenario, we consider the following quadratic NCMC minimax problem:

$$\min_{x\in\mathcal{M}}\max_{y\in Y}\mathcal{L}(x,y) = \frac{1}{2}x^\top Qx + x^\top Ay, \tag{12}$$

where $Y = \{y \in \mathbb{R}^n : \|y\| \leq 1\}$, $\mathcal{M} = \{x \in \mathbb{R}^n : x^\top x = 1\}$ denotes the sphere manifold, and $A, Q \in \mathbb{R}^{n\times n}$ are symmetric random matrices. We set $n = 500$, and randomly generate A and Q such that $A = V\Lambda_A V^\top$ and $Q = V\Lambda_Q V^\top$, where Λ_A, Λ_Q are diagonal matrices, and $V \in \mathbb{R}^{500\times 500}$ is an orthogonal matrix, i.e., $VV^\top = V^\top V = I$. We set $\Lambda_Q = \frac{L\cdot\Lambda_Q^0}{\|\Lambda_Q^0\|_2}$ for $L \in \{5, 10\}$, where Λ_Q^0 is a random diagonal matrix with diagonal elements being sampled uniformly at random from the interval $[-1, 1]$ and $\|\Lambda_Q^0\|_2$ denotes the spectral norm of Λ_Q^0; the diagonal Λ_A is generated the same way, except we ensure that $\|\Lambda_A\|_2 = 1$. Thus, $\nabla\mathcal{L}$ is Lipschitz with constant $L = \max\{\|Q\|_2, \|A\|_2\}$. The primal loss function is $F(x) = \frac{1}{2}x^\top Qx + \|Ax\|$. Since x lies on the sphere, the minimum value of $F(x)$ is $F^* = \lambda_{\min}\left(\frac{1}{2}Q + A\right)$. We randomly regenerated 10 instances as described above and the results are shown in Figure 1 for the algorithms running on these instances. The relative gradient norm denotes $\min_{1\leq t\leq T}\left\{\|G_t^x\|^2 + \|G_t^y\|^2\right\}/(\|G_1^x\|^2 + \|G_1^y\|^2)$, where G_t^x and G_t^y are defined by (11). In the experiments, we set $\beta = 0.9$, $p = 2L$, $\tau_1 = \frac{1}{6L}$, $\tau_2 = \frac{\tau_1}{3}$, $\rho = 10$, and $C = 1000$ for `sm-MGDA`, and we use the same τ_1 and τ_2 for MGDA, i.e., $\tau_1 = \frac{1}{6L}$ and $\tau_2 = \frac{\tau_1}{3}$. For RGDA, we set $\tau_1 = \frac{1}{16(\mu_y+1)^2 L}$ and $\tau_2 = \frac{1}{L}$ according to [35, Theorem 17]. We observe that `sm-MGDA` is competitive against MGDA and RGDA. Furthermore, the manifold constraint violations show that the iterates x_t gradually land on the manifold and the norm constraint, $\|x\| \leq C$, is always inactive.

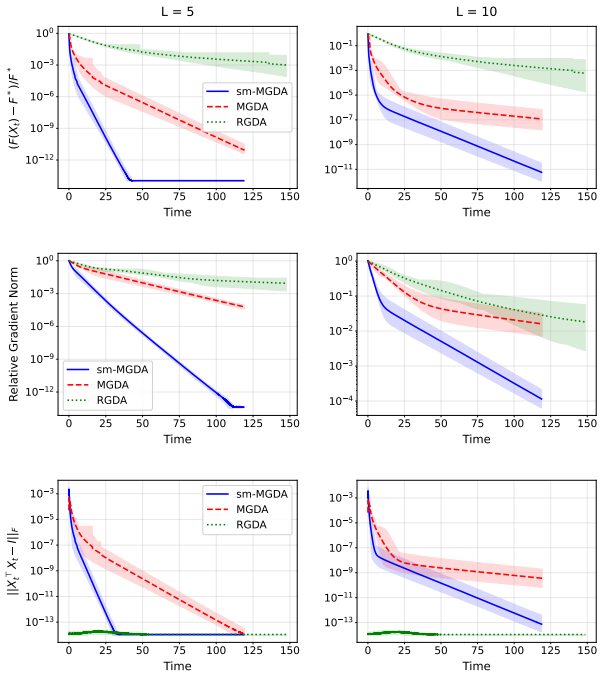

Figure 1: Comparison of `sm-MGDA` with MGDA and RGDA on (12) for $\mu = 0$ with vector variables $x \in \mathbb{R}^n$ with $n = 500$ for $L \in \{5, 10\}$.

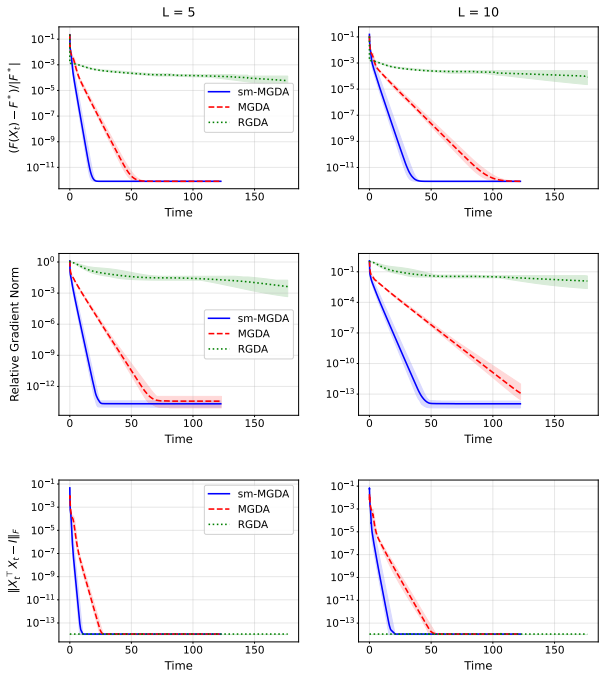

Figure 2: Comparison of `sm-MGDA` with MGDA and RGDA on (13) for $\mu = 0$ with matrix variables $x \in \mathrm{St}(500, 450)$ for for $L \in \{5, 10\}$.

4.1.2 Case 2: Matrix variables

In the second scenario, we consider the following quadratic NCMC minimax problem:

$$\min_{x\in\mathrm{St}(n,k)} \max_{y\in Y} \mathcal{L}(x, y) = \frac{1}{2}\,\mathrm{tr}(x^\top Qx) + \mathrm{tr}(x^\top Ay), \tag{13}$$

where $A, Q \in \mathbb{R}^{n\times n}$ are symmetric matrices, $\mathrm{St}(n, k) = \{x \in \mathbb{R}^{n\times k} : x^\top x = I_k\}$ is the Stiefel manifold, and $Y = \{y \in \mathbb{R}^n : \|y\| \leq 1\}$. We set $n = 500$ and $k = 0.9n = 450$. The matrices $A, Q \in \mathbb{R}^{500\times 500}$ are randomly generated such that $A = V\Lambda_A V^\top$ and $Q = V\Lambda_Q V^\top$, where $V \in \mathbb{R}^{500\times 500}$ is orthogonal and Λ_A, Λ_Q are diagonal. We scale $\Lambda_Q = \frac{L\,\Lambda_Q^0}{\|\Lambda_Q^0\|_2}$ for $L \in \{5, 10\}$, where the diagonal of Λ_Q^0 is sampled uniformly from $[-1, 1]$. Maximizing over $y \in Y$ yields the primal objective $F(x) = \frac{1}{2}\,\mathrm{tr}(x^\top Qx) + \|A^\top x\|_*$, where $\|\cdot\|_*$ denotes the nuclear norm. The parameters are chosen as $\beta = 0.9$, $p = 2L$, $\tau_1 = \frac{1}{6L}$, $\tau_2 = \frac{\tau_1}{3}$, $\rho = 10$, and $C = 1000$ for `sm-MGDA`, and we use the

same τ_1 and τ_2 for MGDA, i.e., $\tau_1 = \frac{1}{6L}$ and $\tau_2 = \frac{\tau_1}{3}$. For RGDA, we set $\tau_1 = \frac{1}{16(\mu_y+1)^2 L}$ and $\tau_2 = \frac{1}{L}$ according to [35, Theorem 17]. The results in Figure 2 show that `sm-MGDA` converges the fastest among MGDA and RGDA in terms of primal values and gradient norms. Furthermore, the manifold constraint violations, i.e., $\|x_t^\top x_t - I\|$, show that the iterates x_t gradually land on the manifold.

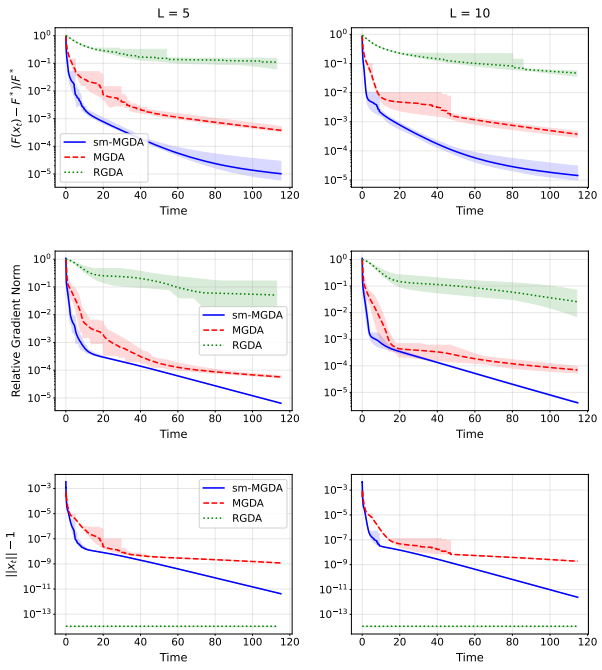

Figure 3: Comparison of `sm-MGDA` with MGDA and RGDA on (14) for $\mu = 1$ with vector variables $x \in \mathbb{R}^n$ with $n = 500$ for $L \in \{5, 10\}$.

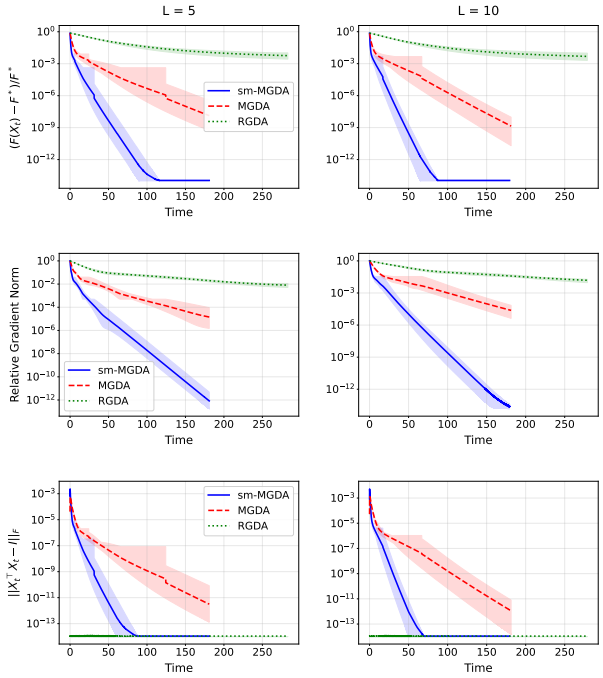

Figure 4: Comparison of `sm-MGDA` with MGDA and RGDA on (15) for $\mu = 1$ with matrix variables $x \in \mathrm{St}(500, 450)$ for for $L \in \{5, 10\}$.

4.2 Quadratic NCSC problem ($\mu > 0$)

4.2.1 Case 1: Vector variables

Again in the first case, we consider the following quadratic NCSC minimax problem:

$$\min_{x\in\mathcal{M}} \max_{y\in\mathbb{R}^n} \mathcal{L}(x,y) = \frac{1}{2}x^\top Q x + x^\top A y - \frac{\mu_y}{2}\|y\|^2, \tag{14}$$

where $\mathcal{M} = \{x \in \mathbb{R}^n : x^\top x = 1\}$ denotes the sphere manifold, $A, Q \in \mathbb{R}^{n\times n}$ are symmetric matrices and $\mu_y > 0$. This class of problems includes robust principal component analysis [25], PL game [9, 53], image processing [8], and robust regression [46]. Except for setting $\mu_y = 1$ and removing the constraint on y, we randomly generate both Q and A as described in Section 4.1 with $n = 500$. Given $x \in \mathbb{R}^n$, we can compute $y^*(x) = \frac{1}{\mu_y}Ax$. Consequently, the primal loss function is $F(x) = \frac{1}{2}x^\top \left(Q + \frac{1}{\mu_y}A^2\right)x$. Since $x \in \mathbb{R}^n$ lies in the sphere manifold, the minimum value of $F(\cdot)$ over $\mathcal{M}$ is $F^* = \frac{1}{2}\lambda_{\min}\left(Q + \frac{1}{\mu_y}A^2\right)$. We adopt the same hyperparameters from Section 4.1 for the three algorithms. The results are shown in Figure 2 for 10 randomly generated problems. Similar to Section 4.1, our `sm-MGDA` performs better than MGDA and RGDA.

4.2.2 Case 2: Matrix variables

In the second case, we consider the following quadratic NCSC minimax problem:

$$\min_{x\in\mathrm{St}(m,k)} \max_{y\in\mathbb{R}^{n\times k}} \mathcal{L}(x,y) = \frac{1}{2}\,\mathrm{tr}(x^\top Q x) + \mathrm{tr}(x^\top A y) - \frac{\mu_y}{2}\|y\|_F^2, \tag{15}$$

where $\mathrm{St}(n,k) = \{x \in \mathbb{R}^{n\times k} : x^\top x = I_k\}$ is the Stiefel manifold, $A, Q \in \mathbb{R}^{n\times n}$ symmetric matrices and $\mu_y > 0$. Except for setting $\mu_y = 1$ and removing the constraint on y, we randomly generate both Q and A as described in Section

4.1 with $n = 500$. We also set $k = 0.9n = 450$. Given $x \in \mathbb{R}^{n\times k}$, maximizing (15) over $y \in \mathbb{R}^{n\times k}$ yields $y^*(x) = \frac{1}{\mu_y} A^\top x$, and thus the primal objective is $F(x) = \frac{1}{2}\operatorname{tr}(x^\top Q x) + \frac{1}{2\mu_y}\|A^\top x\|_F^2 = \frac{1}{2}\operatorname{tr}\Big(x^\top\Big(Q + \frac{1}{\mu_y}AA^\top\Big)x\Big)$. Since $x \in \mathrm{St}(n,k)$, the minimum value is $F^* = \frac{1}{2}\sum_{i=1}^k \lambda_i\Big(Q + \frac{1}{\mu_y}AA^\top\Big)$. We adopt the same hyperparameters from Section 4.1 for the three algorithms. The results are shown in Figure 4 for 10 randomly generated problems. Similar to Section 4.1, our `sm-MGDA` algorithm exhibits the fastest convergence behavior in terms of relative gradient norms and primal function values.

4.3 Robust DNN training

We cast the original robust training against adversarial attacks problem into the following nonconvex-concave problem:

$$\min_{x\in\mathcal{M}}\max_{u\in\mathcal{U}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} u_j \ell\left(h\left(a_{ij}^K; x\right), b_i\right) - r(u), \quad \text{s.t.} \quad \mathcal{U} = \{u \in \mathbb{R}^C \mid u \geq 0,\ \|u\|_1 = 1\}, \tag{16}$$

where a_{ij}^K is the permuted sample after K iterations of Projected Gradient Descent (PGD) attack [29], and C is the number of classes for the dataset. Here $r(u)$ is a convex regularization term, e.g., $r(u) = \alpha\|u - 1/C\|^2$, where $\alpha \geq 0$ is a hyper-parameter. In the experiment, we use Stiefel manifold $\mathcal{M} = \mathrm{St}(r,d) = \left\{X \in \mathbb{R}^{d\times r} : X^\top X = I_r\right\}$ on parameters x of DNNs (convolution layers and linear layers). The DNN architecture used in our experiments is summarized in Table 2. To simplify the operation of the Retraction operator, the network structure calls `mctorch` package [38].

Table 2: The DNN architecture used in our experiments. C is the number of classes, and d is the number of channels for inputs.

Layer	**Configuration**
Inputs	d channels
Conv	$d \to 32$, Batchnorm, ReLU
Conv	$32 \to 64$, Batchnorm, ReLU
Conv	$64 \to 128$, Batchnorm, ReLU
Max Pool	
Linear	$512 \to 256$, Batchnorm, ReLU
Linear	$256 \to 128$, Batchnorm, ReLU
Linear	$128 \to C$
Outputs	

We choose five datasets for this experiment: CIFAR10, CIFAR100, STL10, FashionMNIST, and MNIST. For the CIFAR100 dataset, we selected the data from the first 20 classes for training and testing. We set $K = 5$, corresponding to five attacks on the data. The numerical results against different attacks (i.e., PGD attack [29] and Fast Gradient Sign Method (FGSM) attack [18]) are shown in Table 3. We set $\tau_1 = \tau_2 = 10^{-3}$, $\beta = 0.9$, $p = 1$, $\rho = 10, C = 1000$ for `sm-MGDA` with the same τ_1 and τ_2 for MGDA and RSGDA. In `sm-MGDA`, we use an inexact form of g_t, given by $\nabla A(x_t)[\nabla_x f(x_t, y_t)] + \frac{\rho}{2}\nabla c(x_t)[c(x_t)]$, to simplify computation. This approximation quality is controlled by the small difference between x_t and $A(x_t)$ when x_t lies close to the manifold. The detailed training trajectories are listed in Figure 5. It is shown that RSGDA has worse performance compared with `sm-MGDA` and MGDA, while `sm-MGDA` has the best performance in the sense of gradient norm, primal loss, and test accuracy. This demonstrates the enhanced robustness and practical utility of `sm-MGDA` in robust DNN training. The results of the test accuracy of the compared algorithms under the attack are summarized in Table 3, which demonstrates that `sm-MGDA` has the highest test accuracy compared with MGDA and RSGDA.

The manifold error of the model with epoch on robust DNN training task is shown in Figure 6. It is shown that the error has a tendency to decrease with the number of epochs, indicating that the parameter falls on the manifold.

5 Conclusion

We propose a retraction-free smoothed manifold gradient descent-ascent (`sm-MGDA`) method for solving minimax problems over compact submanifolds through solving an exact penalized problem with an additional norm constraint. We establish convergence guarantees under nonsmooth PL condition for weakly convex functions, and more specifically

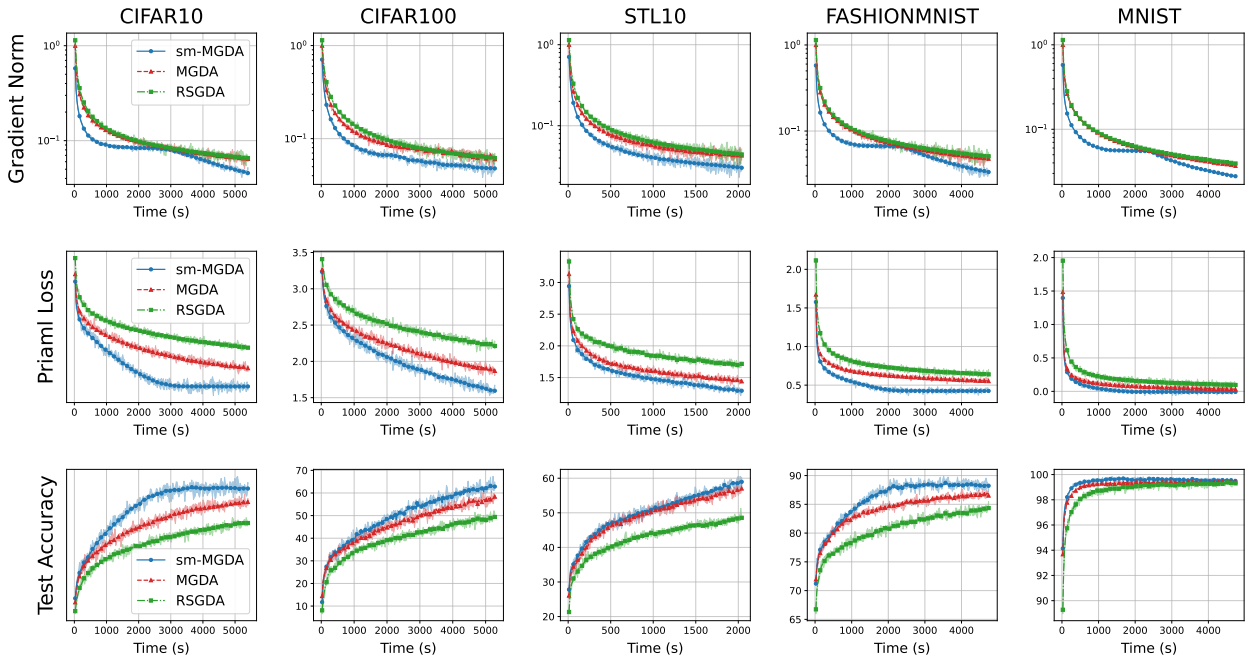

Figure 5: Gradient norm, primal loss, and test accuracy on robust DNN training problems over 3 runs.

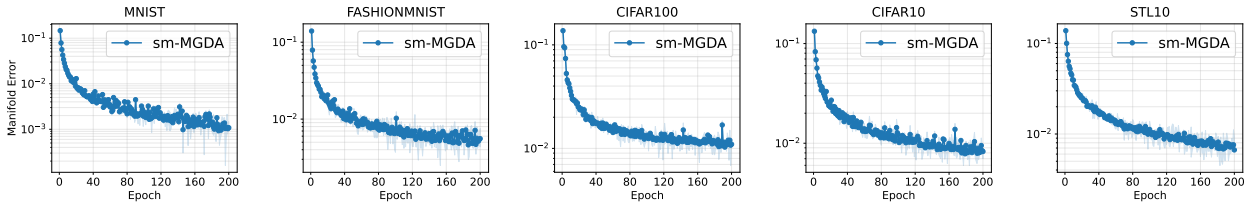

Figure 6: Manifold error of the model with epoch on robust DNN training task.

Table 3: Test accuracy against nature images and different attacks for (from top to bottom) **CIFRA10, CIFAR100, STL10, FashionMNIST, MNIST** datasets.

Methods	**Original Image**	**PGD^{40} L_∞**			**FGSM L_∞**		
		$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$
MGDA	75.06%	72.74%	70.48%	66.01%	72.76%	70.56%	66.41%
RSGDA	66.22%	64.04%	61.31%	56.19%	64.06%	61.44%	56.77%
sm-MGDA	**80.22%**	**77.73%**	**75.06%**	**69.11%**	**77.78%**	**75.29%**	**70.12%**
		$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$
MGDA	58.40%	55.15%	51.80%	45.50%	55.15%	52.10%	46.15%
RSGDA	47.20%	47.27%	43.10%	31.80%	49.98%	46.67%	37.40%
sm-MGDA	**62.35%**	**59.55%**	**57.00%**	**50.90%**	**59.55%**	**56.95%**	**50.90%**
		$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$
MGDA	57.04%	54.26%	52.18%	47.02%	54.26%	54.18%	46.19%
RSGDA	51.86%	47.90%	45.85%	34.90%	53.95%	51.10%	46.05%
sm-MGDA	**58.49%**	**55.85%**	**53.29%**	**47.88%**	**55.85%**	**53.41%**	**48.76%**
		$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.2$
MGDA	86.53%	80.70%	72.82%	49.08%	81.12%	77.70%	67.77%
RSGDA	84.72%	79.40%	70.52%	51.46%	78.41%	70.52%	61.15%
sm-MGDA	**87.73%**	**82.12%**	**75.11%**	**55.80%**	**82.85%**	**77.75%**	**67.87%**
		$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.4$
MGDA	99.29%	98.21%	95.39%	83.43%	98.53%	97.03%	92.41%
RSGDA	99.26%	98.11%	95.83%	84.76%	93.05%	91.26%	90.81%
sm-MGDA	**99.45%**	**98.60%**	**96.69%**	**87.42%**	**98.69%**	**97.62%**	**92.91%**

under both merely and strongly concave settings, and show that the algorithm asymptotically recovers stationarity of the original manifold minimax problem. Numerical results on robust training and risk-sensitive learning tasks demonstrate the practical advantages of our approach over existing methods.

Our work motivates several interesting directions. First, a natural next step is to extend our algorithm and analysis to general submanifolds defined by smooth equality constraints $\{x : c(x) = 0\}$. Finally, we also plan to study the stochastic version of `sm-MGDA`, which is essential for solving large-scale manifold minimax problems.

References

[1] Pierre Ablin and Gabriel Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In *International Conference on Artificial Intelligence and Statistics*, pages 5636–5657. PMLR, 2022.

[2] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[3] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pages 1120–1128. PMLR, 2016.

[4] Richard L Bishop and Barrett O'Neill. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969.

[5] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[6] Qi Cai, Mingyi Hong, Yongxin Chen, and Zhaoran Wang. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019.

[7] Yang Cai, Michael I Jordan, Tianyi Lin, Argyris Oikonomou, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. Curvature-independent last-iterate convergence for games on riemannian manifolds. *arXiv preprint arXiv:2306.16617*, 2023.

[8] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[9] Lesi Chen, Boyuan Yao, and Luo Luo. Faster stochastic algorithms for minimax optimization under polyak-lojasiewicz condition. *Advances in Neural Information Processing Systems*, 35:13921–13932, 2022.

[10] Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018.

[11] Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.

[13] Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. *Advances in Neural Information Processing Systems*, 33:1036–1047, 2020.

[14] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International conference on machine learning*, pages 1125–1134. PMLR, 2018.

[15] Kangkang Deng and Jiang Hu. Decentralized projected riemannian gradient method for smooth optimization on compact submanifolds. *arXiv preprint arXiv:2304.08241*, 2023.

[16] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International conference on machine learning*, pages 1049–1058. PMLR, 2017.

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] Andi Han and Junbin Gao. Riemannian stochastic recursive momentum method for non-convex optimization. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2505–2511. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[20] Andi Han, Bamdev Mishra, Pratik Jawanpuria, Pawan Kumar, and Junbin Gao. Riemannian hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3):1797–1827, 2023.

[21] Inbal Horev, Florian Yger, and Masashi Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *Asian Conference on Machine Learning*, pages 1–16. PMLR, 2016.

[22] Jiang Hu, Kangkang Deng, Jiayuan Wu, and Quanzheng Li. A projected semismooth newton method for a class of nonconvex composite programs with strong prox-regularity. *arXiv preprint arXiv:2303.05410*, 2023.

[23] Zihao Hu, Guanghui Wang, Xi Wang, Andre Wibisono, Jacob D Abernethy, and Molei Tao. Extragradient type methods for riemannian variational inequality problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2080–2088. PMLR, 2024.

[24] Feihu Huang and Shangqian Gao. Gradient descent ascent for minimax problems on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] Michael Jordan, Tianyi Lin, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. *Advances in Neural Information Processing Systems*, 35:6557–6574, 2022.

[26] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[27] Lingkai Kong, Yuqing Wang, and Molei Tao. Momentum stiefel optimizer, with applications to suitably-orthogonal attention, and optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[29] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[30] Yassine Laguel, Yasa Syed, Necdet Serhat Aybat, and Mert Gürbüzbalaban. High-probability complexity guarantees for nonconvex minimax problems. *arXiv preprint arXiv:2405.14130*, 2024.

[31] John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

[32] Feng-Yi Liao, Lijun Ding, and Yang Zheng. Error bounds, pl condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods. In *6th Annual Learning for Dynamics & Control Conference*, pages 993–1005. PMLR, 2024.

[33] Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.

[34] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pages 2738–2779. PMLR, 2020.

[35] Tianyi Lin, Chi Jin, and Michael I Jordan. Two-timescale gradient descent ascent algorithms for nonconvex minimax optimization. *Journal of Machine Learning Research*, 26(11):1–45, 2025.

[36] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[38] Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, and Bamdev Mishra. Mctorch, a manifold optimization library for deep learning. Technical report, arXiv preprint arXiv:1810.01811, 2018.

[39] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[40] François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

[41] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[42] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016.

[43] Nachuan Xiao and Xin Liu. Solving optimization problems over the stiefel manifold by smooth exact penalty functions. *Journal of Computational Mathematics*, 42(5):1246–1276, 2024.

[44] Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. Dissolving constraints for riemannian optimization. *Mathematics of Operations Research*, 49(1):366–397, 2024.

[45] Xiantao Xiao, Yongfeng Li, Zaiwen Wen, and Liwei Zhang. A regularized semi-smooth newton method with projection steps for composite convex programs. *Journal of Scientific Computing*, 76:364–389, 2018.

[46] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *Advances in neural information processing systems*, 21, 2008.

[47] Meng Xu, Bo Jiang, Ya-Feng Liu, and Anthony Man-Cho So. A riemannian alternating descent ascent algorithmic framework for nonconvex-linear minimax problems on riemannian manifolds. *arXiv preprint arXiv:2409.19588*, 2024.

[48] Meng Xu, Bo Jiang, Wenqiang Pu, Ya-Feng Liu, and Anthony Man-Cho So. An efficient alternating riemannian/projected gradient descent ascent algorithm for fair principal component analysis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7195–7199. IEEE, 2024.

[49] Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201(1):635–706, 2023.

[50] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in neural information processing systems*, 33:1153–1165, 2020.

[51] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

[52] Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

[53] Xuan Zhang, Gabriel Mancino-Ball, Necdet Serhat Aybat, and Yangyang Xu. Jointly improving the sample and communication complexities in decentralized stochastic minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20865–20873, 2024.

[54] Xuan Zhang, Qiushui Xu, and Necdet Serhat Aybat. Agda+: Proximal alternating gradient descent ascent method with a nonmonotone adaptive step-size search for nonconvex minimax problems. *arXiv preprint arXiv:2406.14371*, 2024.

6 Proofs

6.1 Proof of Lemma 1

Note that for any $(x, y) \in X \times Y$, we have $\nabla_y \tilde{f}(x, y) = \nabla_y f(A(x), y)$ and

$$\nabla_x \tilde{f}(x, y) = \nabla A(x)[\nabla_x f(A(x), y)] + \rho x(x^\top x - I_r),$$

where $\nabla A(x)[u] := u(\frac{3}{2}I_r - \frac{1}{2}x^\top x) - x\mathrm{sym}(x^\top u)$. Since $\|x\| \leq C$, assumption 2 implies that

$$\|\nabla A(x)\|_{\mathrm{op}} := \max_{u \neq 0} \frac{\|\nabla A(x)[u]\|}{\|u\|} \leq \frac{3}{2} + \frac{1}{2}\|x^\top x\| + \|x\|^2 \leq \frac{3}{2} + \frac{3}{2}\|x\|^2 \leq \frac{3}{2} + \frac{3}{2}C^2$$

for all $x \in X$, where we use $\|xy\| \leq \|x\|_2\|y\| \leq \|x\|\|y\|$.

Lipschitz Continuity of $\nabla_y \tilde{f}$. It directly follows from the expression of $\nabla_y \tilde{f}$ that we have $l_{yy} = L_{yy}$.

Next, we consider l_{yx}. Given $x_1, x_2 \in X$ and $y \in Y$, we have

$$\begin{aligned}
&\|\nabla_y \tilde{f}(x_1, y) - \nabla_y \tilde{f}(x_2, y)\| \\
&\quad = \|\nabla_y f(A(x_1), y) - \nabla_y f(A(x_2), y)\| \leq L_{yx}\|A(x_1) - A(x_2)\| \leq \frac{3}{2}L_{yx}(1 + C^2)\|x_1 - x_2\|,
\end{aligned}$$

where we used the uniform bound on $\|\nabla A(\cdot)\|_{\mathrm{op}}$. Thus, $l_{yx} = \frac{3}{2}(1 + C^2)L_{yx}$.

Lipschitz Continuity of $\nabla_x \tilde{f}$. From the expression of $\nabla_x \tilde{f}$, for any $x \in X$ and $y_1, y_2 \in Y$, we have

$$\begin{aligned}
\|\nabla_x \tilde{f}(x, y_1) - \nabla_x \tilde{f}(x, y_2)\| &= \|\nabla A(x)[\nabla_x f(A(x), y_1) - \nabla_x f(A(x), y_2)]\| \\
&\leq \|\nabla A(x)\|_{\mathrm{op}} L_{xy}\|y_1 - y_2\| \leq \frac{3}{2}(1 + C^2)L_{xy}\|y_1 - y_2\|,
\end{aligned}$$

which implies that $l_{xy} = \frac{3}{2}(1 + C^2)L_{xy}$.

Next, we consider l_{xx}. Define $s(x) := x(x^\top x - I_r)$. Then $\nabla s(x)[u] = u(x^\top x - I_r) + x(u^\top x + x^\top u)$, which implies that $\|\nabla s(x)\|_{\mathrm{op}} = \max_{u \neq 0} \frac{\|\nabla s(x)[u]\|}{\|u\|} \leq C^2 + 1 + 2C^2 = 3C^2 + 1$ for all $x \in X$. Moreover, let $\bar{l}_x := \max_{x \in X, y \in Y} \|\nabla_x f(A(x), y)\|$. Since $A(x) := x(\frac{3}{2}I_r - \frac{1}{2}x^\top x)$, we also have $\|A(x)\| \leq C\frac{1}{2}(3 + C^2)$ for all $x \in X$. Therefore, $\bar{l}_x \leq \max\{\|\nabla_x f(x, y)\| \colon \|x\| \leq C(\frac{3}{2} + \frac{1}{2}C^2),\ y \in Y\}$. For any $x_1, x_2 \in X$ and $y \in Y$, we have

$$\begin{aligned}
&\|\nabla_x \tilde{f}(x_1, y) - \nabla_x \tilde{f}(x_2, y)\| \\
&= \|\nabla A(x_1)[\nabla_x f(A(x_1), y)] - \nabla A(x_2)[\nabla_x f(A(x_2), y)]\| + \rho\|s(x_1) - s(x_2)\| \\
&\leq \|\Big(\nabla A(x_1) - \nabla A(x_2)\Big)[\nabla_x f(A(x_1), y)]\| + \|\nabla A(x_2)[\nabla_x f(A(x_1), y) - \nabla_x f(A(x_2), y)]\| \\
&\quad + \rho \max_{x \in X} \|\nabla s(x)\|_{\mathrm{op}}\|x_2 - x_1\| \\
&\leq 3C\bar{l}_x\|x_2 - x_1\| + \left(\frac{9}{4}(1 + C^2)^2\right) L_{xx}\|x_2 - x_1\| + \rho(3C^2 + 1)\|x_2 - x_1\| \\
&\leq \left(3C\bar{l}_x + \frac{9}{4}(1 + C^2)^2 L_{xx} + \rho(3C^2 + 1)\right) \|x_2 - x_1\|,
\end{aligned}$$

where we used the definition of $\bar{l}_x$ and

$$\begin{aligned}
\max_{\|x_1\| \leq C, \|x_2\| \leq C} \|\Big(\nabla A(x_1) - \nabla A(x_2)\Big)[u]\| &\leq \frac{1}{2}\|x_1^\top x_1 - x_2^\top x_2\|\|u\| + (\|x_1\| + \|x_2\|)\|x_1 - x_2\|\|u\| \\
&= \frac{1}{2}\|(x_1 - x_2)^\top x_1 + x_2^\top (x_1 - x_2)\|\|u\| + (\|x_1\| + \|x_2\|)\|x_1 - x_2\|\|u\| \\
&\leq 3C\|x_1 - x_2\|\|u\|.
\end{aligned}$$

Thus, we have $l_{xx} = 3C\bar{l}_x + \frac{9}{4}(1 + C^2)^2 L_{xx} + \rho(3C^2 + 1)$. □

6.2 Proof of Lemma 2

Let $x_\epsilon = usv^\top$ denote the compact SVD of x_ϵ and $\mathbb{R}^r_+ \ni \sigma = \mathrm{diag}(s)$. Since $\mathrm{dist}(x_\epsilon, \mathcal{M}) \leq \frac{1}{2}$, we have $\frac{1}{2} \leq \sigma_i \leq \frac{3}{2}$ for $i = 1, \ldots, r$, which also implies that $\frac{1}{2} \leq \|x_\epsilon\|_2 \leq \frac{3}{2}$. Since $\mathcal{M}$ is the Stiefel manifold and $A(x) = x(\frac{3}{2}I_r - \frac{1}{2}x^\top x)$ for $x \in X$, [43, Proposition 2.8] implies that

$$\nabla_x \tilde{f}(x, y) = G(x, y) + \rho x(x^\top x - I_r), \quad \forall\, (x, y) \in X \times Y,$$

where $G(x, y) := \nabla_x f(A(x), y)(\frac{3}{2}I_r - \frac{1}{2}x^\top x) - x\mathrm{sym}(x^\top \nabla_x f(A(x), y))$. Then, for (x_ϵ, y_ϵ) we get

$$\begin{aligned}
\|\nabla_x \tilde{f}(x_\epsilon, y_\epsilon)\|^2 &= \|G(x_\epsilon, y_\epsilon)\|^2 + \rho^2 \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 + 2\rho \left\langle G(x_\epsilon, y_\epsilon), x_\epsilon(x_\epsilon^\top x_\epsilon - I_r) \right\rangle \\
&= \|G(x_\epsilon, y_\epsilon)\|^2 + \rho^2 \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 - 3\rho \left\langle \mathrm{sym}(x_\epsilon^\top \nabla_x f(A(x_\epsilon), y_\epsilon)), (x_\epsilon^\top x_\epsilon - I_r)^2 \right\rangle \\
&\geq \|G(x_\epsilon, y_\epsilon)\|^2 + \rho^2 \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 - 3\rho \|x_\epsilon^\top \nabla_x f(A(x_\epsilon), y_\epsilon)\| \|x_\epsilon^\top x_\epsilon - I_r\|^2 \\
&\geq \|G(x_\epsilon, y_\epsilon)\|^2 + \rho^2 \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 - 3\rho \cdot \frac{3}{2}\tilde{L}_x \cdot 4 \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 \\
&= \|G(x_\epsilon, y_\epsilon)\|^2 + \rho(\rho - 18\tilde{l}_x) \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 \\
&\geq \|G(x_\epsilon, y_\epsilon)\|^2 + \frac{\rho^2}{2} \|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2,
\end{aligned} \tag{17}$$

where $\tilde{L}_x := L_x(y_\epsilon) = \max\{\|\nabla_x f(x, y_\epsilon)\| :\ \|x\|_2 \leq 1\}$. Here, the second equation follows from the definition of $G(\cdot, \cdot)$ and the fact that $\langle A, B\rangle = \langle \mathrm{sym}(A), B\rangle$ whenever B is symmetric, the first inequality is due to $\|\mathrm{sym}(A)\| \leq \|A\|$ and $\langle A, B\rangle \leq \|A\|\|B\|$. Next, we argue for the the second inequality. Note that

$$\|A(x_\epsilon)\|_2 = \frac{1}{2} \max_{i=1,\ldots,r} \sigma_i(3 - \sigma_i^2) \leq \frac{1}{2} \max\left\{ t(3 - t^2) :\ t \in \left[\frac{1}{2}, \frac{3}{2}\right] \right\} = 1.$$

Thus, $\|\nabla_x f(A(x_\epsilon), y_\epsilon)\| \leq \tilde{L}_x$. Consequently, the second inequality follows from $\|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\| \geq \frac{1}{2}\|x_\epsilon^\top x_\epsilon - I_r\|$ where we use $\|AB\| \geq \sigma_{\min}(A) \cdot \|B\|$, and $\sigma_{\min}(A) \geq 0$ denotes the minimum singular value of A. Finally, the last line uses $\rho \geq 36\tilde{L}_x$.

In addition, it holds $\mathcal{P}_\mathcal{M}(x_\epsilon) = uv^\top$, and using $\|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\| \geq \frac{1}{2}\|x_\epsilon^\top x_\epsilon - I_r\|$, we also get

$$\|x_\epsilon - \mathcal{P}_\mathcal{M}(x_\epsilon)\| = \|usv^\top - uv^\top\| = \|s - I_r\| \leq \|s^2 - I_r\| = \|x_\epsilon^\top x_\epsilon - I_r\| \leq 2\|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|.$$

Thus, we get $\|x_\epsilon - \mathcal{P}_\mathcal{M}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon$, which follows from using (17) together with

$$\|x_\epsilon - \mathcal{P}_\mathcal{M}(x_\epsilon)\|^2 \leq 4\|x_\epsilon(x_\epsilon^\top x_\epsilon - I_r)\|^2 \leq \frac{8}{\rho^2}\|\nabla_x \tilde{f}(x_\epsilon, y_\epsilon)\|^2 \leq \frac{9}{\rho^2}\epsilon^2. \tag{18}$$

Next, we provide a bound on $\|\mathrm{grad}_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon)\|$. Using the definition of $G(x_\epsilon, y_\epsilon)$ and the expression for the Riemannian gradient given in (5), we get

$$\begin{aligned}
&\|\mathrm{grad}_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) - G(x_\epsilon, y_\epsilon)\| \\
&= \|\nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) - \mathcal{P}_\mathcal{M}(x_\epsilon)\mathrm{sym}(\mathcal{P}_\mathcal{M}(x_\epsilon)^\top \nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon)) \\
&\quad - \nabla_x f(A(x_\epsilon), y_\epsilon)\Big(\frac{3}{2}I_r - \frac{1}{2}x_\epsilon^\top x_\epsilon\Big) - x_\epsilon \mathrm{sym}(x_\epsilon^\top \nabla_x f(A(x_\epsilon), y_\epsilon))\| \\
&\leq \frac{1}{2}\|\nabla_x f(A(x_\epsilon), y_\epsilon)\| \|x_\epsilon^\top x_\epsilon - I_r\| + L_{xx}\|\mathcal{P}_\mathcal{M}(x_\epsilon) - A(x_\epsilon)\| \\
&\quad + (\mathcal{P}_\mathcal{M}(x_\epsilon) - x_\epsilon)\mathrm{sym}\Big(\mathcal{P}_\mathcal{M}(x_\epsilon)^\top \nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon)\Big) \\
&\quad + x_\epsilon \mathrm{sym}\Big(\mathcal{P}_\mathcal{M}(x_\epsilon)^\top \nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) - x_\epsilon^\top \nabla_x f(A(x_\epsilon), y_\epsilon)\Big) \\
&\leq \frac{1}{2}\|\nabla_x f(A(x_\epsilon), y_\epsilon)\| \|x_\epsilon^\top x_\epsilon - I_r\| + L_{xx}\|\mathcal{P}_\mathcal{M}(x_\epsilon) - A(x_\epsilon)\| + \tilde{L}_x\|\mathcal{P}_\mathcal{M}(x_\epsilon) - x_\epsilon\| \\
&\quad + \|x_\epsilon\|_2\Big(\|\mathcal{P}_\mathcal{M}(x_\epsilon) - x_\epsilon\| \|\nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon))\| + \|x_\epsilon\|_2 \|\nabla_x f(\mathcal{P}_\mathcal{M}(x_\epsilon), y_\epsilon) - \nabla_x f(A(x_\epsilon), y_\epsilon)\|\Big) \\
&\leq \frac{1}{2}\|\nabla_x f(A(x_\epsilon), y_\epsilon)\| \|x_\epsilon^\top x_\epsilon - I_r\| + L_{xx}\|\mathcal{P}_\mathcal{M}(x_\epsilon) - A(x_\epsilon)\| + \tilde{L}_x\|\mathcal{P}_\mathcal{M}(x_\epsilon) - x_\epsilon\|
\end{aligned}$$

$$+\frac{3}{2}\tilde{L}_x\|\mathcal{P}_{\mathcal{M}}(x_\epsilon)-x_\epsilon\|+\frac{9}{4}L_{xx}\|\mathcal{P}_{\mathcal{M}}(x_\epsilon)-A(x_\epsilon)\|. \tag{19}$$

Furthermore, since $x_\epsilon = usv^\top$, we also have $A(x_\epsilon)=\frac{1}{2}us(3I_r-s^2)v^\top$; hence,

$$\begin{aligned}\|\mathcal{P}_{\mathcal{M}}(x_\epsilon)-A(x_\epsilon)\| &= \|u\Big(I_r-\frac{1}{2}s(3I_r-s^2)\Big)v^\top\|\\ &= \|(s-I_r)\Big(\frac{1}{2}(s^2+s)-I_r\Big)\|\leq \|s-I_r\| = \|\mathcal{P}_{\mathcal{M}}(x_\epsilon)-x_\epsilon\|,\end{aligned} \tag{20}$$

where the inequality follows from $\|\frac{1}{2}(s^2+s)-I_r\|_2\leq 1$, which is implied by $\frac{1}{2}\leq\sigma_i\leq\frac{3}{2}$ for $i=1,\ldots,r$. Moreover, also note that $\|x_\epsilon(x_\epsilon^\top x_\epsilon-I_r)\|\geq\frac{1}{2}\|x_\epsilon^\top x_\epsilon-I_r\|$ together with (18) gives us $\|x_\epsilon^\top x_\epsilon-I_r\|\leq\frac{3}{\rho}$. Finally, (17) implies that $\|G(x_\epsilon,y_\epsilon)\|\leq\epsilon$. Therefore, using (19), we get

$$\begin{aligned}\|\mathrm{grad}_x\, f(\mathcal{P}_{\mathcal{M}}(x_\epsilon),y_\epsilon)\| &\leq \|G(x_\epsilon,y_\epsilon)\|+\|\mathrm{grad}_x\, f(\mathcal{P}_{\mathcal{M}}(x_\epsilon),y_\epsilon)-G(x_\epsilon,y_\epsilon)\|\\ &\leq \epsilon+\frac{3}{\rho}\Big(\frac{7}{2}\tilde{L}_x+\frac{13}{4}L_{xx}\Big)\epsilon.\end{aligned}$$

Finally, the desired bound for $\mathrm{dist}\Big(0,-\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon),y_\epsilon)+\partial h(y_\epsilon)\Big)$ directly follows from the Lipschitz continuity of $\nabla_y f$ together with (18) and (20). □

7 Proof of Theorem 1

In the rest $\mathcal{X}$ and $\mathcal{Y}$ denote finite dimensional Euclidean vector spaces. Let us start with some necessary notations. Let

- $\tilde{f}_r(x,y):=\tilde{f}(x,y)-h(y)$ with $\tilde{f}(x,y):=f(A(x),y)+\frac{\rho}{4}\|c(x)\|^2$ for $x\in X$ and $y\in\mathcal{Y}$;
- $\Phi^*:=\min_{x\in X}\Phi(x)$ with $\Phi(x):=\max_{y\in\mathcal{Y}}\tilde{f}_r(x,y)$;
- $\hat{f}_r(x,y;z):=\hat{f}(x,y;z)-h(y)$ with $\hat{f}(x,y;z):=\tilde{f}(x,y)+\frac{p}{2}\|x-z\|^2$ for any $(x,y)\in X\times Y$ and $z\in\mathcal{X}$. Consider the auxiliary problem $\min_{x\in X}\max_{y\in\mathcal{Y}}\hat{f}_r(x,y;z)$. For any $y\in Y$, due to Lemma 1, $\tilde{f}(\cdot,y)$ is l-weakly convex over X; hence, $\hat{f}(\cdot,y;z)$ is $(p-l)$-strongly convex on X for any fixed $z\in\mathcal{X}$ and $y\in Y$.
- $\Phi(x;z):=\max_{y\in\mathcal{Y}}\hat{f}_r(x,y;z)$ for $x\in X$ and $z\in\mathcal{X}$, denotes the primal function of the auxiliary problem –note that $\Phi(x;z)=\Phi(x)+\frac{p}{2}\|x-z\|^2$ for any $x\in X$ and $z\in\mathcal{X}$;
- $\Psi_r(y;z):=\Psi(y;z)-h(y)$ with $\Psi(y;z):=\min_{x\in X}\hat{f}(x,y;z)$ for $y\in Y$ and $z\in\mathcal{X}$, denotes the dual function of the auxiliary problem;
- $P(z):=\min_{x\in X}\max_{y\in\mathcal{Y}}\hat{f}_r(x,y;z)$ is the optimal value for the auxiliary primal problem, i.e., $P(z)=\min_{x\in X}\Phi(x;z)$ for any fixed $z\in\mathcal{X}$;
- $x^*(y,z):=\arg\min_{x\in X}\hat{f}_r(x,y;z)=\arg\min_{x\in X}\hat{f}(x,y;z)$ for any $y\in Y$ and $z\in\mathcal{X}$;
- $x^*(z):=\arg\min_{x\in X}\Phi(x;z)$ is the optimal solution to the auxiliary primal problem for any fixed $z\in\mathcal{X}$;
- $Y^*(z):=\arg\max_{y\in\mathcal{Y}}\Psi_r(y;z)$ is the set of optimal solutions to the auxiliary dual problem for any fixed $z\in\mathcal{X}$;
- $y^+(z_t):=\mathrm{prox}_{\tau_2 h}(y_t+\tau_2\nabla_y\tilde{f}(x^*(y_t,z_t),y_t))$ for any $y_t\in Y$ and $z_t\in\mathcal{X}$, and it denotes a prox-gradient update in y corresponding to the dual function $\Psi_r(\cdot;z)$ since $\Psi_r(\cdot;z)=\Psi(\cdot;z)-h(\cdot)$ and $\nabla\Psi(y;z)=\nabla_y\tilde{f}(x^*(y,z),y)$;
- $V(x,y;z):=\hat{f}_r(x,y;z)-2\Psi_r(y;z)+2P(z)$ for $x\in X$, $y\in Y$ and $z\in\mathcal{X}$ denotes the potential function.

Based on Assumptions 1 and 3, we first list some helpful results from [51, 52].

Lemma 5. *Let $\varphi:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}\cup\{\pm\infty\}$ with domain $\mathrm{dom}\,\varphi:=\{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m:\ -\infty<\varphi(x,y)<\infty\}$. Let $\mathcal{D}_x=\{x:\ \exists\, y \text{ s.t. } (x,y)\in\mathrm{dom}\,\varphi\}$ and $\mathcal{D}_y=\{y:\ \exists\, x \text{ s.t. } (x,y)\in\mathrm{dom}\,\varphi\}$. Suppose that*

(i) $\varphi(\cdot,y)$ is a proper, closed ζ_x-weakly convex function that is μ_x-PL uniformly for all $y\in\mathcal{D}_y$;

(ii) $-\varphi(x,\cdot)$ is a proper, closed ζ_y-weakly convex function that is μ_y-PL uniformly for all $x\in\mathcal{D}_x$.

If there exists $(x^*, y^*) \in \operatorname{dom}\varphi$ *such that* $0 \in \partial_x \varphi(x, y^*) \mid_{x=x^*}$ *and* $0 \in \partial_y - \varphi(x^*, y) \mid_{y=y^*}$*, then* (x^*, y^*) *is a saddle point, i.e.,* $\varphi(x^*, y) \leq \varphi(x^*, y^*) \leq \varphi(x, y^*)$ *for all* $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$*; hence, the minimax switch holds:* $\min_x \max_y \varphi(x, y) = \max_y \min_x \varphi(x, y)$.

Proof. The proof proceeds along the lines of [50, Lemma 2.1]. In particular, suppose (x^*, y^*) is a stationary point, i.e., $0 \in \partial_x \varphi(x, y^*) \mid_{x=x^*}$ and $0 \in \partial_y - \varphi(x^*, y) \mid_{y=y^*}$. From assumption 3, it follows that

$$\varphi(x^*, y^*) - \min_x \varphi(x, y^*) \leq \frac{1}{2\mu_x} \operatorname{dist}^2(0, \partial_x p(x^*, y^*)) = 0,$$

$$\max_y \varphi(x^*, y) - \varphi(x^*, y^*) \leq \frac{1}{2\mu_y} \operatorname{dist}^2(0, \partial_y - \varphi(x^*, y^*)) = 0;$$

hence, we can conclude that $\max_y \varphi(x^*, y) = \varphi(x^*, y^*) \leq \min_x \varphi(x, y^*)$. □

Lemma 6. *Under Assumptions 1, 2 and 3, let* $\Phi : X \to \mathbb{R}$ *such that* $\Phi(x) := \max_{y \in \mathcal{Y}} \tilde{f}_r(x, y)$ *for* $x \in X$*. Then* $\Phi(\cdot)$ *is differentiable on* X*; indeed, for all* $x \in X$*,*

$$\nabla \Phi(x) = \nabla_x \tilde{f}_r(x, r^*(x)),$$

for any $r^*(x) \in \arg\max_{y \in \mathcal{Y}} \tilde{f}_r(x, y)$*, and* $\nabla \Phi(\cdot)$ *is* $l(1 + 2\kappa)$*-Lipschitz on* X*.*

Moreover, for any fixed $z \in \mathcal{X}$*, let* $\Psi(y; z) = \min_{x \in X} \hat{f}(x, y; z)$ *for* $y \in Y$*. Then,* $\Psi(\cdot; z)$ *is differentiable*[5] *on* Y*; indeed, for all* $y \in Y$*,* $\nabla \Psi(y; z) = \nabla_y \tilde{f}(x^*(y, z), y)$ *where* $x^*(y, z) = \arg\min_{x \in X} \hat{f}(x, y; z)$*.*

Proof. Since we assume that for any $x \in X$, $-f_r(x, \cdot)$ is μ-PL, it holds that

$$2\mu\Big(\Phi(x) - f_r(x, y)\Big) \leq \operatorname{dist}^2\Big(0, -\partial_y f_r(x, y)\Big),$$

using the notational convention that $-\partial_y f_r(x, y) := \partial_y(-f_r(x, y))$. By [32, Theorem 3.1], any μ-PL function also satisfies quadratic growth condition, namely,

$$\frac{\mu}{4} \operatorname{dist}^2(y, R^*(x)) \leq \Phi(x) - f_r(x, y),$$

where $R^*(x) := \arg\max_{y \in \mathcal{Y}} f_r(x, y)$ is a closed set. For given $x_1, x_2 \in X$, it holds for any $y_1 \in R^*(x_1)$ that $0 \in -\nabla_y f(x_1, y_1) + \partial h(y_1)$; hence,

$$\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_1) \in -\nabla_y f(x_2, y_1) + \partial h(y_1). \tag{21}$$

Consider $\Delta := \Phi(x_2) - f_r(x_2, y_1) \geq 0$. By the μ-PL property of $-f_r(x_2, \cdot)$, it holds that

$$\Delta \leq \frac{1}{2\mu} \operatorname{dist}^2\Big(0, -\nabla_y f(x_2, y_1) + \partial h(y_1)\Big) \leq \frac{l^2}{2\mu} \|x_2 - x_1\|^2,$$

where the second inequality uses eq. (21) and the Lipschitz continuity of $\nabla_y f(\cdot, y_1)$. Furthermore, let $y_2 = \arg\min_{y \in R^*(x_2)} \|y - y_1\|$. Since $\operatorname{dist}(y_1, R^*(x_2)) = \|y_1 - y_2\|$, the quadratic growth property implies that $\Delta \geq \frac{\mu}{4} \|y_1 - y_2\|^2$; hence, combining both inequalities on Δ leads to

$$\|y_1 - y_2\|^2 \leq 2 \frac{l^2}{\mu^2} \|x_1 - x_2\|^2. \tag{22}$$

Now given $x \in X$ and $d \in \mathcal{X}$, let $y^* \in R^*(x)$. From eq. (22), there exists $y^*(\tau) \in \arg\max_{y \in \mathcal{Y}} f_r(x + \tau d, y)$ such that $\|y^*(\tau) - y^*\|^2 \leq 2\tau^2 \frac{l^2}{\mu^2} \|d\|^2$. Moreover, since h is ζ-weakly convex and $\nabla_y f(x, y^*) \in \partial h(y^*)$, we have

$$h(y^*) - h(y^*(\tau)) + \langle \nabla_y f(x, y^*), y^*(\tau) - y^* \rangle \leq \frac{\zeta}{2} \|y^*(\tau) - y^*\|^2. \tag{23}$$

Then, we can bound the primal value difference, $\Phi(x + \tau d) - \Phi(x)$, from above as follows:

$$\begin{aligned}
\Phi(x + \tau d) - \Phi(x) =& f_r(x + \tau d, y^*(\tau)) - f_r(x, y^*) \\
=& f(x + \tau d, y^*(\tau)) - f(x, y^*) + h(y^*) - h(y^*(\tau)) \\
=& \tau \nabla_x f(x, y^*)^\top d + \nabla_y f(x, y^*)^\top (y^*(\tau) - y^*) + h(y^*) - h(y^*(\tau)) + o(\tau) \\
\leq& \tau \nabla_x f(x, y^*)^\top d + \frac{\zeta}{2} \|y^*(\tau) - y^*\|^2 + o(\tau).
\end{aligned}$$

[5] The classic Danskin's theorem is stated for the case $\hat{f}(x, \cdot; z)$ is concave. In our case, as we only assume $\hat{f}(x, y; z)$ is smooth in y, for completeness we state the result for our setting.

This implies

$$\limsup_{\tau\downarrow 0}\frac{\Phi(x+\tau d)-\Phi(x)}{\tau}\leq \nabla_x f(x,y^*)^\top d+\lim_{\tau\downarrow 0}\Big(\frac{o(\tau)}{\tau}+\zeta\frac{l^2}{\mu^2}\|d\|^2\tau\Big)=\nabla_x f(x,y^*)^\top d.$$

On the other hand, we also have

$$\Phi(x+\tau d)\geq f_r(x+\tau d,y^*)=f_r(x,y^*)+\tau\nabla_x f(x,y^*)^\top d+o(\tau),$$

which implies that

$$\liminf_{\tau\downarrow 0}\frac{\Phi(x+\tau d)-\Phi(x)}{\tau}\geq \nabla_x f(x,y^*)^\top d.$$

Thus,

$$\lim_{\tau\downarrow 0}\frac{\Phi(x+\tau d)-\Phi(x)}{\tau}=\nabla_x f(x,y^*)^\top d. \tag{24}$$

Since (24) holds for all $d\in\mathcal{X}$, we have $\nabla\Phi(x)=\nabla_x f(x,y^*)$ for any $y^*\in R^*(x)$.

Next, to argue for the smoothness of $\Phi(\cdot)$, let $x_1,x_2\in X$ be some arbitrary points, consider any $y_1\in R^*(x_1)$ and $y_2\in\arg\min_{y\in R^*(x_2)}\|y-y_1\|$. Then, we complete the proof by using (22) to reach at the conclusion:

$$\|\nabla\Phi(x_2)-\nabla\Phi(x_1)\|=\|\nabla_x f(x_2,y_2)-\nabla_x f(x_1,y_1)\|\leq l(\|x_2-x_1\|+\|y_2-y_1\|)\leq l\left(1+2\frac{l}{\mu}\right)\|x_2-x_1\|.$$

Next, for any fixed $z\in\mathcal{X}$, consider $\Psi(y;z)=\min_{x\in X}\hat{f}(x,y;z)$ for $y\in Y$. Note that $\hat{f}(\cdot,y;z)$ is $(p-l)$-strongly convex for any $y\in Y$; hence, $\hat{f}(\cdot,y;z)$ is μ-PL and let $x^*(y,z)=\arg\min_{x\in X}\hat{f}(x,y;z)$ denote the unique optimal solution. Observing that $\Psi(y;z)=-\max_{x\in X}-\hat{f}(x,y;z)$, using the result from the first part, we can conclude that $\nabla\Psi(y;z)=\nabla_y\hat{f}(x^*(y,z),y;z)$ for all $y\in Y$, which gives us the desired result. □

Lemma 7. *For any given $z\in\mathcal{X}$, consider $\hat{f}_r(\cdot,\cdot;z)$ defined above. Under Assumptions 1, 2 and 3, it holds that $\min_{x\in X}\max_{y\in\mathcal{Y}}\hat{f}_r(x,y;z)=\max_{y\in\mathcal{Y}}\min_{x\in X}\hat{f}_r(x,y;z)$; hence, $P(z)=\min_{x\in X}\Phi(x;z)=\max_{y\in\mathcal{Y}}\Psi_r(y;z)$.*

Proof. Fix an arbitrary $z\in\mathcal{X}$, and define $\varphi(x,y):=\hat{f}_r(x,y;z)+\delta_X(x)$, where $\delta_X(\cdot)$ denotes the indicator function of the set X. For any $y\in Y$, $\varphi(\cdot,y)$ is strongly convex with modulus $p-l$, and according to [32, Theorem 3.1], the non-smooth function $\varphi(\cdot,y)$ is μ_x-PL with $\mu_x=p-l$. On the other hand, according to assumption 3, $\varphi(x,\cdot)$ is μ_y-PL with $\mu_y=\mu$.

Since $\hat{f}_r(\cdot,y;z)$ is strongly convex with modulus $p-l$ for any $y\in Y$, $\Phi(\cdot;z)$ is also strongly convex with modulus $p-l$ as it is the pointwise maximum of strongly convex functions. Therefore, $x^*(z)=\arg\min_{x\in X}\Phi(x;z)$ exists. Let $y^*(z)\in\arg\max_{y\in\mathcal{Y}}\hat{f}_r(x^*(z),y;z)$ –note that $Y=\operatorname{dom}h$ is compact and $x^*(z)\in X$ imply that $\hat{f}_r(x^*(z),\cdot;z)$ is continuous on Y; therefore, $y^*(z)$ exists. Thus, first-order optimality conditions and Lemma 6 imply that

$$0\in\nabla\Phi(x^*(z);z)+\partial\delta_X(x^*(z))=\nabla_x\hat{f}_r(x^*(z),y^*(z);z)+\partial\delta_X(x^*(z))=\partial_x\varphi(x^*(z),y^*(z)),$$
$$0\in-\partial_y\hat{f}_r(x^*(z),y^*(z);z)=-\partial_y\varphi(x^*(z),y^*(z));$$

hence, $(x^*(z),y^*(z))$ is a stationary point of $\min_{x\in X}\max_{y\in\mathcal{Y}}\varphi(x,y)$. Therefore, using Lemma 5 we can conclude that $(x^*(z),y^*(z))$ is a saddle point, which implies that $y^*(z)\in Y^*(z)$. This completes the proof. □

Lemma 8. *Suppose that Assumptions 1 and 3 hold, and p is chosen such that $p>l$. Then, we have*

$$\|x^*(y,z)-x^*(y,z')\|\leq\gamma_1\|z-z'\|,\quad\forall\,y\in Y,\,\forall\,z,z'\in X, \tag{25a}$$
$$\|x^*(z)-x^*(z')\|\leq\gamma_1\|z-z'\|,\quad\forall\,z,z'\in X, \tag{25b}$$
$$\|x^*(y,z)-x^*(y',z)\|\leq\gamma_2\|y-y'\|,\quad\forall\,z\in X,\,\forall\,y,y'\in Y, \tag{25c}$$
$$\|x_{t+1}-x^*(y_t,z_t)\|\leq\gamma_3\|x_{t+1}-x_t\|, \tag{25d}$$

where $\gamma_1:=\frac{p}{p-l}$, $\gamma_2:=\frac{p+l}{p-l}$, and $\gamma_3:=1+\frac{1}{\tau_1(p-l)}$.

Proof. For the proof of (25a) and (25b), see [30, Lemma 21] and [30, Lemma 23], respectively – these proofs simply extends to the constrained optimization setting here with $x\in X$. The inequality in (25c) follows from the proof of [34, Lemma B.2(c)] which can also handle $x\in X$ constraint. The inequality in (25d) is provided in [52, Lemma B.2] of which proof can be found in Appendix C at `https://arxiv.org/pdf/1812.10229`. □

We first establish the following lemma.

Lemma 9. *Under Assumptions 1, 2 and 3, for all $t \in \mathbb{Z}_+$, it holds that*

$$\left\|x^*\left(y^+(z_t), z_t\right) - x^*\left(z_t\right)\right\|^2 \leq \frac{3}{\mu(p-l)\tau_2^2}(1+\tau_2^2 l^2 + \gamma_2^2\tau_2^2 l^2)\left\|y_t^+(z_t) - y_t\right\|^2. \tag{26}$$

Proof. Since $\hat{f}_r(\cdot, y; z)$ is $(p-l)$ strongly convex for any fixed $y \in Y$ and $z \in \mathcal{X}$, and $\Phi(\cdot, z)$ is the pointwise supremum over $y \in \mathcal{Y}$, we can conclude that $\Phi(\cdot, z)$ is also $(p-l)$-strongly convex for any fixed $z \in \mathcal{X}$. Thus, for any $g_t \in \partial h(y^+(z_t))$, we get

$$\begin{aligned}
\frac{p-l}{2}\|x^*(z_t) - x^*(y^+(z_t), z_t)\|^2 \leq& \Phi(x^*(y^+(z_t), z_t); z_t) - \Phi(x^*(z_t); z_t) \\
=& \Phi(x^*(y^+(z_t), z_t); z_t) - \hat{f}_r(x^*(y^+(z_t), z_t), y^+(z_t); z_t) \\
&+ \hat{f}_r(x^*(y^+(z_t), z_t), y^+(z_t); z_t) - \Phi(x^*(z_t); z_t) \\
\leq& \frac{1}{2\mu}\|\nabla_y \tilde{f}(x^*(y^+(z_t), z_t), y^+(z_t)) - g_t\|^2,
\end{aligned} \tag{27}$$

where in the last inequality we used the fact that of $-\hat{f}_r(x, \cdot; z)$ is μ-PL for any $x \in X$, $z \in \mathcal{X}$ and that $\hat{f}_r(x^*(y^+(z_t), z_t), y^+(z_t), z_t) = \Psi_r(y^+(z_t); z_t) \leq \Psi_r(y^*(z_t); z_t) = \Phi(x^*(z_t), z_t)$, where $y^*(z_t) \in Y^*(z_t)$ and the equality follows from Lemma 7. On the other hand, $y^+(z_t) = \operatorname{prox}_{\tau_2 h}(y_t + \tau_2 \nabla_y \tilde{f}(x^*(y_t, z_t), y_t))$ is well defined for $\tau_2 \in (0, \frac{1}{\zeta})$, and we have $0 \in \partial h(y^+(z_t)) + (y^+(z_t) - y_t)/\tau_2 - \nabla_y \tilde{f}(x^*(y_t, z_t), y_t)$; hence, for $g_t \in \partial h(y^+(z_t))$ such that $g_t = \nabla_y \tilde{f}(x^*(y_t, z_t), y_t) - (y^+(z_t) - y_t)/\tau_2$, we get

$$\begin{aligned}
&\|\nabla_y \tilde{f}(x^*(y^+(z_t), z_t), y^+(z_t)) - g_t\|^2 \\
\leq& 3\|\nabla_y \tilde{f}(x^*(y_t, z_t), y_t) - g_t\|^2 + 3\|\nabla_y \tilde{f}(x^*(y_t, z_t), y_t) - \nabla_y \tilde{f}(x^*(y_t, z_t), y^+(z_t))\|^2 \\
&+ 3\|\nabla_y \tilde{f}(x^*(y_t, z_t), y^+(z_t)) - \nabla_y \tilde{f}(x^*(y^+(z_t), z_t), y^+(z_t))\|^2 \\
\leq& 3\Big(\frac{1}{\tau_2^2} + l^2 + \gamma_2^2 l^2\Big)\|y_t - y^+(z_t)\|^2,
\end{aligned} \tag{28}$$

where we use the identity $\|a+b+c\|^2 \leq 3(\|a\|_2 + \|b\|_2 + \|c\|_2)$ in the first inequality, and the second inequality is from the definition of $\operatorname{prox}_{\tau_2 h}$, Lipschitz continuity of $\nabla_y f$, and Lemma 8. Combining (27) and (28) gives the desired inequality. □

Next we state a classic result on projected gradient updates, which will be useful later in our analysis. Due to space limitations, we omit its proof.

Lemma 10. *Let f be a convex function and X be a convex set. Given $\bar{x} \in X$ and some arbitrary $t > 0$, let $x^+ = \mathcal{P}_X(\bar{x} - t\nabla f(\bar{x}))$. Then $\langle g_t(\bar{x}) - \nabla f(\bar{x}),\ x^+ - x\rangle \geq 0$ for all $x \in X$, where $g_t(\bar{x}) = (\bar{x} - x^+)/t$. In particular, $\langle g_t(\bar{x}) - \nabla f(\bar{x}),\ x^+ - \bar{x}\rangle \geq 0$.*

Now, we proceed with the proof of Theorem 1, which adopts the same potential function framework in [51, 52]. Following the analysis in [51], we separate our proof into several parts: we first present three descent results, then we show the descent property for the potential function, later we discuss the relation between our stationarity measure and the potential function, and lastly we put all things together.

Proof of Theorem 1. In the first part of the proof, we begin with showing primal descent by providing a lower bound on $\hat{f}\left(x_t, y_t; z_t\right) - \hat{f}\left(x_{t+1}, y_{t+1}; z_{t+1}\right)$; next, we show dual ascent by bounding $\Psi\left(y_{t+1}; z_{t+1}\right) - \Psi\left(y_t; z_t\right)$ from below, and finally lower bound the change in the Moreau envelope $P(z_t) - P(z_{t+1})$.

Primal descent: By the $(p+l)$-smoothness of $\hat{f}\left(\cdot, y_t; z_t\right)$, we have

$$\begin{aligned}
&\hat{f}\left(x_{t+1}, y_t; z_t\right) - \hat{f}\left(x_t, y_t; z_t\right) \\
&\leq \left\langle \nabla_x \hat{f}\left(x_t, y_t; z_t\right), x_{t+1} - x_t \right\rangle + \frac{p+l}{2}\left\|x_{t+1} - x_t\right\|^2 \\
&= \left\langle \nabla_x \hat{f}\left(x_t, y_t; z_t\right), \mathcal{P}_X\Big(x_t - \tau_1 \nabla_x \hat{f}\left(x_t, y_t; z_t\right)\Big) - x_t \right\rangle + \frac{p+l}{2}\left\|x_{t+1} - x_t\right\|^2.
\end{aligned} \tag{29}$$

Note that it follows from Lemma 10 that

$$\left\langle \nabla_x \hat{f}(x_t, y_t; z_t), \mathcal{P}_X(x_t - \tau_1 \nabla_x \hat{f}\left(x_t, y_t; z_t\right)) - x_t \right\rangle \leq -\frac{1}{\tau_1}\|x_{t+1} - x_t\|^2. \tag{30}$$

Thus, for the choice of $\tau_1 \leq \frac{1}{p+l}$, we get

$$\hat{f}\left(x_t, y_t; z_t\right) - \hat{f}\left(x_{t+1}, y_t; z_t\right) \geq \frac{1}{2\tau_1}\left\|x_{t+1} - x_t\right\|^2. \tag{31}$$

Moreover, since $\hat{f}\left(x_{t+1}, \cdot; z_t\right)$ is l-smooth,

$$\hat{f}\left(x_{t+1}, y_t; z_t\right) - \hat{f}\left(x_{t+1}, y_{t+1}; z_t\right) \geq \left\langle \nabla_y \hat{f}\left(x_{t+1}, y_t; z_t\right), y_t - y_{t+1} \right\rangle - \frac{l}{2}\left\|y_t - y_{t+1}\right\|^2. \tag{32}$$

Furthermore, from the definition of $\hat{f}$ and $z_{t+1} - z_t = \beta(x_{t+1} - z_t)$, and using $0 < \beta \leq 1$, we get

$$\begin{aligned}
&\hat{f}\left(x_{t+1}, y_{t+1}; z_t\right) - \hat{f}\left(x_{t+1}, y_{t+1}; z_{t+1}\right) \\
&\quad = \frac{p}{2}\left[\|x_{t+1} - z_t\|^2 - \|x_{t+1} - z_{t+1}\|^2\right] = \frac{p}{2}\left\langle z_t - z_{t+1}, z_{t+1} + z_t - 2x_{t+1}\right\rangle \\
&\quad = \frac{p}{2}\left\langle z_t - z_{t+1}, z_{t+1} - z_t + \frac{2}{\beta}(z_t - z_{t+1})\right\rangle = \frac{p}{2}\Big(\frac{2}{\beta} - 1\Big)\|z_t - z_{t+1}\|^2 \\
&\quad \geq \frac{p}{2\beta}\|z_t - z_{t+1}\|^2.
\end{aligned} \tag{33}$$

Combining (31), (32) and (33), we get

$$\begin{aligned}
&\hat{f}\left(x_t, y_t; z_t\right) - \hat{f}\left(x_{t+1}, y_{t+1}; z_{t+1}\right) \\
&\quad \geq \frac{1}{2\tau_1}\|x_{t+1} - x_t\|^2 + \frac{p}{2\beta}\left\|z_{t+1} - z_t\right\|^2 + \left\langle \nabla_y \tilde{f}(x_{t+1}, y_t), y_t - y_{t+1}\right\rangle - \frac{l}{2}\left\|y_t - y_{t+1}\right\|^2.
\end{aligned}$$

Dual Descent: Recall that the dual function of the auxiliary minimax problem, $\Psi_r(\cdot; z)$, has a composite form, i.e., $\Psi_r(\cdot; z) = \Psi(\cdot; z) - h(\cdot)$. For any $z \in \mathcal{X}$ and $y \in Y$, Lemma 6 implies that $\nabla\Psi(y; z) = \nabla_y \tilde{f}(x^*(y, z), y)$; therefore, for any $y_1, y_2 \in Y$, we have $\|\nabla\Psi(y_1; z) - \nabla\Psi(y_2; z)\| \leq \|\nabla_y \tilde{f}(x^*(y_1, z), y_1) - \nabla_y \tilde{f}(x^*(y_2, z), y_2)\| \leq l(\|x^*(y_1, z) - x^*(y_2, z)\| + \|y_1 - y_2\|) \leq l(1 + \gamma_2)\|y_1 - y_2\|$ which follows from (25c).

Thus, $\Psi(\cdot; z)$ is L_Ψ smooth for all $z \in \mathcal{X}$ with $L_\Psi := l(1 + \gamma_2)$, where $\gamma_2 := \frac{p+l}{p-l}$. Using the fact that $\nabla_y \Psi(y_t; z_t) = \nabla_y \tilde{f}(x^*(y_t, z_t), y_t)$, we get

$$\Psi\left(y_{t+1}; z_t\right) - \Psi\left(y_t; z_t\right) \geq \left\langle \nabla_y \tilde{f}(x^*(y_t, z_t), y_t), y_{t+1} - y_t \right\rangle - \frac{L_\Psi}{2}\left\|y_{t+1} - y_t\right\|^2. \tag{34}$$

Moreover, from the definition of Ψ, we also have

$$\begin{aligned}
&\Psi\left(y_{t+1}; z_{t+1}\right) - \Psi\left(y_{t+1}; z_t\right) \\
&\quad = \hat{f}\left(x^*\left(y_{t+1}, z_{t+1}\right), y_{t+1}; z_{t+1}\right) - \hat{f}\left(x^*\left(y_{t+1}, z_t\right), y_{t+1}; z_t\right) \\
&\quad \geq \hat{f}\left(x^*\left(y_{t+1}, z_{t+1}\right), y_{t+1}; z_{t+1}\right) - \hat{f}\left(x^*\left(y_{t+1}, z_{t+1}\right), y_{t+1}; z_t\right) \\
&\quad = \frac{p}{2}\left[\left\|x^*\left(y_{t+1}, z_{t+1}\right) - z_{t+1}\right\|^2 - \left\|x^*\left(y_{t+1}, z_{t+1}\right) - z_t\right\|^2\right] \\
&\quad = \frac{p}{2}\left\langle z_{t+1} - z_t,\ z_{t+1} + z_t - 2x^*\left(y_{t+1}, z_{t+1}\right)\right\rangle,
\end{aligned} \tag{35}$$

where the inequality uses the optimality of $x^*(y_{t+1}, z_t)$, i.e., $\hat{f}(x^*(y_{t+1}, z_t), y_{t+1}; z_t) \leq \hat{f}(x^*(y_{t+1}, z_{t+1}), y_{t+1}; z_t)$. Combining (35) with (34), we have

$$\begin{aligned}
&\Psi\left(y_{t+1}; z_{t+1}\right) - \Psi\left(y_t; z_t\right) \\
&\quad \geq \left\langle \nabla_y \tilde{f}\left(x^*\left(y_t, z_t\right), y_t\right), y_{t+1} - y_t \right\rangle - \frac{L_\Psi}{2}\left\|y_{t+1} - y_t\right\|^2 \\
&\qquad + \frac{p}{2}\left\langle z_{t+1} - z_t,\ z_{t+1} + z_t - 2x^*\left(y_{t+1}, z_{t+1}\right)\right\rangle.
\end{aligned}$$

Proximal Descent: Let $y^*(z_{t+1}) \in Y^*(z_{t+1})$ and $y^*(z_t) \in Y^*(z_t)$. Then, we have

$$\begin{aligned}
&P(z_t) - P(z_{t+1}) \\
&= \min_{x\in X}\max_{y\in\mathcal{Y}} \hat{f}_r(x,y;z_t) - \min_{x\in X}\max_{y\in\mathcal{Y}} \hat{f}_r(x,y;z_{t+1}) \\
&= \max_{y\in\mathcal{Y}}\min_{x\in X} \hat{f}_r(x,y;z_t) - \max_{y\in\mathcal{Y}}\min_{x\in X} \hat{f}_r(x,y;z_{t+1}) \\
&= \Psi_r(y^*(z_t);z_t) - \Psi_r(y^*(z_{t+1});z_{t+1}) \\
&\geq \Psi_r(y^*(z_{t+1});z_t) - \Psi_r(y^*(z_{t+1});z_{t+1}) \\
&= \hat{f}(x^*(y^*(z_{t+1}),z_t),y^*(z_{t+1});z_t) - \hat{f}(x^*(y^*(z_{t+1}),z_{t+1}),y^*(z_{t+1});z_{t+1}) \\
&\geq \hat{f}(x^*(y^*(z_{t+1}),z_t),y^*(z_{t+1});z_t) - \hat{f}(x^*(y^*(z_{t+1}),z_t),y^*(z_{t+1});z_{t+1}) \\
&= \frac{p}{2}\Big[\|x^*(y^*(z_{t+1});z_t) - z_t\|^2 - \|x^*(y^*(z_{t+1});z_t) - z_{t+1}\|^2\Big] \\
&= -\frac{p}{2}\langle z_{t+1} - z_t,\ z_{t+1} + z_t - 2x^*(y^*(z_{t+1}),z_t)\rangle,
\end{aligned}$$

where the second equality follows from Lemma 5 since $\hat{f}_r(x,y;z)$ is strongly convex in x and PL in y for every $z \in \mathcal{X}$, the first inequality uses the optimality of $y^*(z_t)$ for $\max_y \Psi_r(\cdot, z_t)$ and the second inequality is from the optimality of $x^*(y^*(z_{t+1}), z_{t+1})$ for $\min_{x\in X} \hat{f}(x, y^*(z_{t+1}, z_{t+1})$.

Potential Function: We adopt the potential function as in [51, 52],

$$V_t := V(x_t,y_t,z_t) = \hat{f}_r(x_t,y_t;z_t) - 2\Psi_r(y_t;z_t) + 2P(z_t), \qquad \forall\, t \geq 0. \tag{36}$$

Note that we have $P(z_t) \geq \Phi^* = \min_{x\in X}\Phi(x)$, and from weak duality, we also have $\hat{f}_r(x_t,y_t;z_t) \geq P(z_t) \geq \Psi_r(y_t;z_t)$; thus, we can conclude that $V_t \geq P(z_t) \geq \Phi^*$ for all $t \in \mathbb{Z}_+$.

Recall that $y_{t+1} = \arg\min_{y\in\mathcal{Y}} q_t(y)$ where $q_t(y) := \tau_2 h(y) + \frac{1}{2}\|y - \Big(y_t + \tau_2\nabla_y\tilde{f}(x_{t+1},y_t;z_t)\Big)\|^2$. Since $h(\cdot)$ is ζ-weakly convex, it follows that $q_t(\cdot)$ is strongly convex with modulus $(1-\tau_2\zeta)$. Therefore, we get

$$\begin{aligned}
&\tau_2 h(y_t) + \frac{1}{2}\|\tau_2\nabla_y\hat{f}(x_{t+1},y_t,z_t)\|^2 \\
&\geq \tau_2 h(y_{t+1}) + \frac{1}{2}\|y_{t+1} - (y_t + \tau_2\nabla_y\hat{f}(x_{t+1},y_t,z_t))\|^2 + \frac{1-\tau_2\zeta}{2}\|y_{t+1} - y_t\|^2,
\end{aligned}$$

which implies that

$$\Big\langle \nabla_y\tilde{f}(x_{t+1},y_t), y_{t+1} - y_t\Big\rangle + h(y_t) - h(y_{t+1}) \geq \frac{2-\tau_2\zeta}{2\tau_2}\|y_{t+1} - y_t\|^2. \tag{37}$$

Since $V_t = \hat{f}(x_t,y_t;z_t) - 2\Psi(y_t;z_t) + 2P(z_t) + h(y_t)$, combining (37) with the above descent results, we get

$$\begin{aligned}
&V_t - V_{t+1} \\
&\geq \frac{1}{2\tau_1}\|x_{t+1} - x_t\|^2 - \frac{l}{2}\|y_{t+1} - y_t\|^2 + \frac{p}{2\beta}\|z_t - z_{t+1}\|^2 \\
&\quad + \langle \nabla_y\tilde{f}(x_{t+1},y_t),\ y_{t+1} - y_t\rangle + h(y_t) - h(y_{t+1}) \\
&\quad + 2\Big\langle \nabla_y\tilde{f}(x^*(y_t,z_t),y_t) - \nabla_y\tilde{f}(x_{t+1},y_t),\ y_{t+1} - y_t\Big\rangle - L_\Psi\|y_{t+1} - y_t\|^2 \\
&\quad + p\langle z_{t+1} - z_t, z_{t+1} + z_t - 2x^*(y_{t+1},z_{t+1})\rangle \\
&\quad - p\langle z_{t+1} - z_t, z_{t+1} + z_t - 2x^*(y^*(z_{t+1}),z_t)\rangle \\
&\geq \frac{1}{2\tau_1}\|x_{t+1} - x_t\|^2 + \Big(\frac{1}{\tau_2} - \frac{l+\zeta}{2} - L_\Psi\Big)\|y_{t+1} - y_t\|^2 + \frac{p}{2\beta}\|z_t - z_{t+1}\|^2 \\
&\quad + 2\Big\langle \nabla_y\tilde{f}(x^*(y_t,z_t),y_t) - \nabla_y\tilde{f}(x_{t+1},y_t),\ y_{t+1} - y_t\Big\rangle \\
&\quad + 2p\langle z_{t+1} - z_t, x^*(y^*(z_{t+1}),z_t) - x^*(y_{t+1},z_{t+1})\rangle \\
&\geq \frac{1}{2\tau_1}\|x_{t+1} - x_t\|^2 + \frac{1}{2\tau_2}\|y_{t+1} - y_t\|^2 + \frac{p}{2\beta}\|z_t - z_{t+1}\|^2 \\
&\quad + 2\Big\langle \nabla_y\tilde{f}(x^*(y_t,z_t),y_t) - \nabla_y\tilde{f}(x_{t+1},y_t),\ y_{t+1} - y_t\Big\rangle \\
&\quad + 2p\langle z_{t+1} - z_t, x^*(y^*(z_{t+1}),z_t) - x^*(y_{t+1},z_{t+1})\rangle,
\end{aligned} \tag{38}$$

where the last inequality uses $\frac{1}{\tau_2} \geq l + \zeta + 2L_\Psi = 9l + \zeta$ which follows from $p = 2l$, implying $L_\Psi = 4l$, and from our choices of $\tau_2 = \frac{1}{16}(\frac{3}{\tau_1} + \zeta)^{-1}$ for $\tau_1 \leq \frac{1}{p+l}$.

Consider $A := 2\left\langle \nabla_y \tilde{f}\left(x^*\left(y_t, z_t\right), y_t\right) - \nabla_y \tilde{f}\left(x_{t+1}, y_t\right),\ y_{t+1} - y_t \right\rangle$. Note that

$$\begin{aligned} A &\geq -2\|\nabla_y \tilde{f}\left(x^*\left(y_t, z_t\right), y_t\right) - \nabla_y \tilde{f}\left(x_{t+1}, y_t\right)\|\|y_{t+1} - y_t\| \\ &\geq -2l\left\|x_{t+1} - x^*\left(y_t, z_t\right)\right\|\left\|y_{t+1} - y_t\right\| \\ &\geq -l\nu\left\|y_{t+1} - y_t\right\|^2 - \frac{l}{\nu}\left\|x_{t+1} - x^*\left(y_t, z_t\right)\right\|^2, \end{aligned}$$

where the second inequality uses the l-Lipschitz continuity of $\nabla_y \tilde{f}(\cdot, y)$ for any $y \in Y$, and the last inequality is due to the Young's inequality with $\nu > 0$ specified later. Moreover, using Lemma 8 to bound the term $\left\|x_{t+1} - x^*\left(y_t, z_t\right)\right\|^2$ yields

$$A \geq -l\nu\left\|y_{t+1} - y_t\right\|^2 - \frac{l}{\nu}\gamma_3^2\left\|x_{t+1} - x_t\right\|^2. \tag{39}$$

Next, consider $B := 2p\left\langle z_{t+1} - z_t, x^*\left(y^*\left(z_{t+1}\right), z_t\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\rangle$. Note that

$$\begin{aligned} B =& 2p\left\langle z_{t+1} - z_t, x^*\left(y^*\left(z_{t+1}\right), z_t\right) - x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right)\right\rangle \\ &+ 2p\left\langle z_{t+1} - z_t, x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\rangle \\ \geq& -2p\gamma_1\left\|z_{t+1} - z_t\right\|^2 + 2p\left\langle z_{t+1} - z_t, x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\rangle \\ \geq& -\left(2p\gamma_1 + \frac{p}{6\beta}\right)\left\|z_{t+1} - z_t\right\|^2 - 6p\beta\left\|x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\|^2, \end{aligned} \tag{40}$$

where in the first inequality we use Lemma 8 and the Cauchy-Schwarz inequality, and for the second inequality we used Young's inequality for $\beta > 0$ specified later, i.e., $2\langle a, b\rangle \leq \frac{1}{\theta}\|a\|^2 + \theta\|b\|^2$ for any $\theta > 0$.

Plugging (39) and (40) into (38) leads to

$$\begin{aligned} V_t - V_{t+1} \geq& \left(\frac{1}{2\tau_1} - \frac{l}{\nu}\gamma_3^2\right)\left\|x_{t+1} - x_t\right\|^2 + \left(\frac{1}{2\tau_2} - l\nu\right)\left\|y_{t+1} - y_t\right\|^2 \\ &+ \left(\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta}\right)\left\|z_t - z_{t+1}\right\|^2 - 6p\beta\left\|x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\|^2. \end{aligned} \tag{41}$$

Next, we bound $\|y_{t+1} - y_t\|$ by bounding $\|y^+(z_t) - y_{t+1}\|$. First, recall that y_{t+1} and $y^+(z_t)$ have the form: $y_{t+1} = \operatorname{prox}_{\tau_2 h}\left(y_t + \tau_2 \nabla_y \tilde{f}(x_{t+1}, y_t)\right)$ and $y^+(z_t) = \operatorname{prox}_{\tau_2 h}\left(y_t + \tau_2 \nabla_y \tilde{f}(x^*(y_t, z_t), y_t)\right)$; hence,

$$\begin{aligned} \left\|y^+(z_t) - y_{t+1}\right\| &\leq \frac{\tau_2}{1 - \tau_2\zeta}\|\nabla_y \tilde{f}(x_{t+1}, y_t) - \nabla_y \tilde{f}(x^*(y_t, z_t), y_t)\| \\ &\leq \frac{\tau_2 l}{1 - \tau_2\zeta}\|x_{t+1} - x^*(y_t, z_t)\| \leq \frac{\tau_2 l\gamma_3}{1 - \tau_2\zeta}\|x_{t+1} - x_t\|, \end{aligned} \tag{42}$$

where in the first inequality we used the fact that $\operatorname{prox}_{\tau_2 h}(\cdot)$ is $\frac{1}{1-\tau_2\zeta}$-Lipschitz continuous for $0 < \tau_2 < \frac{1}{\zeta}$, the second inequality and the last one follow from the Lipschitz continuity of $\nabla_y \tilde{f}(\cdot, y_t)$ and Lemma 8, respectively. Thus,

$$\begin{aligned} \left\|y_{t+1} - y_t\right\|^2 \geq& \frac{1}{2}\left\|y^+(z_t) - y_t\right\|^2 - \left\|y^+(z_t) - y_{t+1}\right\|^2 \\ \geq& \frac{1}{2}\left\|y^+(z_t) - y_t\right\|^2 - \frac{\tau_2^2 l^2}{(1 - \tau_2\zeta)^2}\gamma_3^2\left\|x_{t+1} - x_t\right\|^2, \end{aligned} \tag{43}$$

where in the first inequality we use $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and the second one follows from (42).

On the other hand, since $x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) = x^*\left(z_{t+1}\right)$, we have

$$\begin{aligned} &\left\|x^*\left(y^*\left(z_{t+1}\right), z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\|^2 \\ &= \left\|x^*\left(z_{t+1}\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\|^2 \\ &\leq 4\left\|x^*\left(z_{t+1}\right) - x^*\left(z_t\right)\right\|^2 + 4\left\|x^*\left(z_t\right) - x^*\left(y^+\left(z_t\right), z_t\right)\right\|^2 \\ &\quad + 4\left\|x^*\left(y^+\left(z_t\right), z_t\right) - x^*\left(y_{t+1}, z_t\right)\right\|^2 + 4\left\|x^*\left(y_{t+1}, z_t\right) - x^*\left(y_{t+1}, z_{t+1}\right)\right\|^2 \\ &\leq 4\left\|x^*\left(z_t\right) - x^*\left(y_t^+\left(z_t\right), z_t\right)\right\|^2 + 4\gamma_2^2\left\|y_t^+\left(z_t\right) - y_{t+1}\right\|^2 + 8\gamma_1^2\left\|z_t - z_{t+1}\right\|^2 \\ &\leq 4\left\|x^*\left(z_t\right) - x^*\left(y_t^+\left(z_t\right), z_t\right)\right\|^2 + \frac{4\gamma_2^2\tau_2^2 l^2\gamma_3^2}{(1 - \tau_2\zeta)^2}\left\|x_{t+1} - x_t\right\|^2 + 8\gamma_1^2\left\|z_t - z_{t+1}\right\|^2, \end{aligned} \tag{44}$$

which follows from Lemma 8 and (42). Finally, plugging (43) and (44) into (41), we get

$$\begin{aligned} V_t - V_{t+1} \geq & \left[\frac{1}{2\tau_1} - \frac{l}{\nu}\gamma_3^2 - \left(\frac{1}{2\tau_2} - l\nu\right)\frac{\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2 - \frac{24p\beta\gamma_2^2\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2\right]\|x_{t+1}-x_t\|^2 \\ & - 24p\beta\left\|x^*(z_t) - x^*\left(y_t^+(z_t), z_t\right)\right\|^2 + \frac{1}{2}\left(\frac{1}{2\tau_2} - l\nu\right)\left\|y^+(z_t) - y_t\right\|^2 \\ & + \left[\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta} - 48p\beta\gamma_1^2\right]\|z_t - z_{t+1}\|^2 \\ \geq & \frac{3}{10\tau_1}\|x_{t+1}-x_t\|^2 + \frac{1}{8\tau_2}\left\|y^+(z_t) - y_t\right\|^2 + \frac{p}{4\beta}\|z_t - z_{t+1}\|^2 \\ & - 24p\beta\left\|x^*(z_t) - x^*\left(y_t^+(z_t), z_t\right)\right\|^2, \end{aligned} \tag{45}$$

where the last inequality holds due to our choice of `sm-MGDA` parameters, i.e., $\tau_1 \in (0, \frac{1}{p+l}]$, $\tau_2 = \frac{1}{16}(\frac{3}{\tau_1} + \zeta)^{-1}$, $\beta = \alpha\min\{\mu, l\}\tau_2$ for any $\alpha \in (0, \frac{1}{406}]$ and $p = 2l$, whenever $0 < \nu \leq \frac{1}{4l\tau_2}$. Indeed, our choice implies that $\gamma_1 = 2$, $\gamma_2 = 3$ and $\gamma_3^2 \leq \frac{2}{\tau_1^2 l^2} + 2$; moreover, choosing $\nu = \frac{1}{4l\tau_2} \geq \frac{12}{l\tau_1}$, we have $\frac{1}{2\tau_2} - l\nu = \frac{1}{4\tau_2}$ and this particular selection of parameters implies that

$$\begin{aligned} & \frac{l}{\nu}\gamma_3^2 + \left(\frac{1}{2\tau_2} - l\nu\right)\frac{\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2 + 24p\beta\gamma_2^2\frac{\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2 \\ & = \left[\nu^{-1}\left(l\tau_1\gamma_3^2\right) + \frac{\tau_2}{4\tau_1}\frac{1}{(1-\tau_2\zeta)^2}\left(l^2\tau_1^2\gamma_3^2\right) + 48\frac{\tau_2}{\tau_1}\cdot 9l\beta\frac{\tau_2}{(1-\tau_2\zeta)^2}\left(l^2\tau_1^2\gamma_3^2\right)\right]\frac{1}{\tau_1} \\ & \leq \left[\frac{2\nu^{-1}}{\tau_1 l} + \frac{1}{96}\frac{1}{(1-\tau_2\zeta)^2} + 18l^2\cdot\frac{\min\{\mu,l\}}{l}\frac{\tau_2^2}{(1-\tau_2\zeta)^2}\cdot\alpha\right]\frac{1+\tau_1^2 l^2}{\tau_1} \\ & \leq \left[\frac{2\nu^{-1}}{\tau_1 l} + \frac{1}{96}\frac{1}{(1-\tau_2\zeta)^2} + \frac{\alpha}{12}\cdot\frac{1}{96}\frac{1}{(1-\tau_2\zeta)^2}\right]\frac{1+\tau_1^2 l^2}{\tau_1} \\ & \leq \left[\frac{1}{6} + \frac{1}{84} + \frac{\alpha}{12}\cdot\frac{1}{84}\right]\frac{10}{9}\cdot\frac{1}{\tau_1} \leq \frac{1}{5\tau_1} \end{aligned}$$

holds for any $\alpha \in (0, \frac{1}{2}]$, where in the first inequality we used $\gamma_3^2 \leq \frac{2}{\tau_1^2 l^2} + 2$ and $\frac{\tau_2}{\tau_1} \leq \frac{1}{48}$, the second inequality follows from $\tau_2^2 \leq \frac{\tau_1^2}{48^2}$ and $\tau_1 \leq \frac{1}{3l}$, finally, for the last inequality we use the bounds $\nu^{-1} \leq \frac{l\tau_1}{12}$ and $1 + \tau_1^2 l^2 \leq \frac{10}{9}$ together with $\frac{1}{(1-\tau_2\zeta)^2} \leq (\frac{16}{15})^2$; hence, $\frac{1}{96}\cdot\frac{1}{(1-\tau_2\zeta)^2} \leq \frac{1}{84}$, which follows from $\tau_2 \leq \frac{1}{16\zeta}$. On the other hand, we also have

$$\frac{p}{2\beta} - 2p\gamma_1 - \frac{p}{6\beta} - 48p\beta\gamma_1^2 = \left[\frac{1}{3} - 4\beta - 192\beta^2\right]\frac{p}{\beta} \geq \frac{p}{4\beta}$$

holds for all $\beta \in (0, \frac{1}{78}]$, and since $\beta = \alpha\min\{\mu, l\}\tau_2 \leq \alpha\min\{\mu, l\}\frac{\tau_1}{48} \leq \frac{\alpha}{144}\frac{\min\{\mu,l\}}{l} \leq \frac{\alpha}{144}$, we conclude that the inequality holds for all $\alpha \in (0, 1]$.

Next, we bound the last term on the right hand side of (45) using Lemma 9 as follows:

$$\begin{aligned} & 24p\beta\left\|x^*(z_t) - x^*\left(y^+(z_t), z_t\right)\right\|^2 \\ & \leq \frac{72p\beta}{\mu(p-l)\tau_2^2}(1 + \tau_2^2 l^2 + \gamma_2^2\tau_2^2 l^2)\|y^+(z_t) - y_t\|^2 \leq \frac{1}{16\tau_2}\|y^+(z_t) - y_t\|^2, \end{aligned} \tag{46}$$

where in the second inequality we use $p = 2l$, $\gamma_2 = 3$, $\tau_2 l \leq 1/144$, $\beta = \alpha\min\{\mu, l\}\tau_2$ for $\alpha \in (0, \frac{1}{2306}]$ implying that

$$\frac{72p\beta}{\mu(p-l)\tau_2^2}(1 + \tau_2^2 l^2 + \gamma_2^2\tau_2^2 l^2) \leq \frac{144\beta}{\mu\tau_2^2}\Big(1 + \frac{10}{144^2}\Big) \leq \frac{\alpha}{\tau_2}\Big(144 + \frac{10}{144}\Big) \leq \frac{1}{16\tau_2}.$$

Plugging the bound in (46) into (45) and using the fact $\|z_{t+1} - z_t\| = \beta\|x_{t+1} - z_t\|$, we get

$$V_t - V_{t+1} \geq \frac{3}{10\tau_1}\|x_{t+1} - x_t\|^2 + \frac{1}{16\tau_2}\left\|y^+(z_t) - y_t\right\|^2 + \frac{p\beta}{4}\|x_{t+1} - z_t\|^2. \tag{47}$$

Stationarity Measure: From the first-order optimality conditions, for all $t \geq 1$, we have $G_t^x \in \nabla_x\tilde{f}(x_t, y_t) + \partial\delta_X(x_t)$ and $G_t^y \in \nabla_y\tilde{f}(x_t, y_t) - \partial h(y_t)$, where G_t^x and G_t^y are defined in (11). Note that given $a, b, c \in \mathcal{X}$ and a closed convex

set $X \subset \mathcal{X}$, it holds that $\|\mathcal{P}_X(a+b)-c\| \leq \|\mathcal{P}_X(a)-c\|+\|b\|$; hence,

$$\begin{aligned}\tau_1\|G^x_{t+1}\| \leq& \|x_{t+1}-x_t\|+\tau_1\|\nabla_x\tilde{f}(x_{t+1},y_{t+1})-\nabla_x\tilde{f}(x_t,y_t)\|+\tau_1 p\|z_t-x_t\| \\ \leq& \|x_{t+1}-x_t\|+\tau_1 l(\|x_{t+1}-x_t\|+\|y_{t+1}-y_t\|)+\tau_1 p\|x_t-z_t\| \\ \leq& (1+\tau_1 p+\tau_1 l)\|x_{t+1}-x_t\|+\tau_1 p\|x_{t+1}-z_t\|+\tau_1 l\|y_{t+1}-y_t\|.\end{aligned}$$

Then, using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we get

$$\|G^x_{t+1}\|^2 \leq 3\frac{(1+\tau_1 p+\tau_1 l)^2}{\tau_1^2}\|x_{t+1}-x_t\|^2+3p^2\|x_{t+1}-z_t\|^2+3l^2\|y_{t+1}-y_t\|^2. \tag{48}$$

In addition, for $0 < \tau_2 < \frac{1}{\zeta}$, $\mathrm{prox}_{\tau_2 h}(\cdot)$ is $\frac{1}{1-\tau_2\zeta}$-Lipschitz continuous, and for any given $a, b, c \in \mathcal{Y}$, it also holds that $\|\mathrm{prox}_{\tau_2 h}(a+b)-c\| \leq \|\mathrm{prox}_{\tau_2 h}(a)-c\|+\frac{1}{1-\tau_2\zeta}\|b\|$; hence, together with the definition of $y^+(z_t) = \mathrm{prox}_{\tau_2 h}\Big(y_t+\tau_2\nabla_y\tilde{f}(x^*(y_t,z_t),y_t)\Big)$ and using Lemma 8, it also holds that

$$\begin{aligned}\|y_{t+1}-y_t\| =& \Big\|\mathrm{prox}_{\tau_2 h}(y_t+\tau_2\nabla_y\tilde{f}(x_{t+1},y_t))-y_t\Big\| \\ \leq& \|y^+(z_t)-y_t\|+\frac{\tau_2}{1-\tau_2\zeta}\Big\|\nabla_y\tilde{f}\left(x_{t+1},y_t\right)-\nabla_y\tilde{f}\left(x^*(y_t,z_t),y_t\right)\Big\| \\ \leq& \|y^+(z_t)-y_t\|+\frac{\tau_2 l}{1-\tau_2\zeta}\left\|x_{t+1}-x^*\left(y_t,z_t\right)\right\|, \\ \leq& \|y^+(z_t)-y_t\|+\frac{\tau_2 l}{1-\tau_2\zeta}\gamma_3\left\|x_{t+1}-x_t\right\|.\end{aligned} \tag{49}$$

Note that $\tau_2\left\|G^y_{t+1}\right\| \leq \|y_{t+1}-y_t\|+\tau_2 l\|y_{t+1}-y_t\|$; therefore, we get

$$\left\|G^y_{t+1}\right\|^2 \leq \frac{2(1+\tau_2 l)^2}{\tau_2^2}\left(\|y^+(z_t)-y_t\|^2+\frac{\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2\left\|x_{t+1}-x_t\right\|^2\right). \tag{50}$$

Then, let $\bar{\kappa} := \max\{\kappa, 1\}$, using (48), (49) and (50), we have

$$\begin{aligned}&\|G^x_{t+1}\|^2+\bar{\kappa}\left\|G^y_{t+1}\right\|^2-3p^2\|x_{t+1}-z_t\|^2 \\ \leq& \frac{3}{\tau_1^2}\left((1+\tau_1 p+\tau_1 l)^2+\Big(l^2+\frac{\bar{\kappa}(1+\tau_2 l)^2}{3\tau_2^2}\Big)\frac{2\tau_1^2\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2\right)\left\|x_{t+1}-x_t\right\|^2 \\ &+\frac{2}{\tau_2^2}\Big(\bar{\kappa}(1+\tau_2 l)^2+3\tau_2^2 l^2\Big)\left\|y^+(z_t)-y_t\right\|^2, \\ \leq& \frac{172}{10}\frac{\bar{\kappa}}{\tau_1^2}\|x_{t+1}-x_t\|^2+\frac{33}{16}\frac{\bar{\kappa}}{\tau_2^2}\|y^+(z_t)-y_t\|^2,\end{aligned} \tag{51}$$

which follows from the following bounds: $2\bar{\kappa}(1+\tau_2 l)^2+6\tau_2^2 l^2 \leq 2\bar{\kappa}[(1+\frac{1}{144})^2+\frac{3}{144^2}] \leq \frac{33}{16}\bar{\kappa}$, $3(1+\tau_1 p+\tau_1 l)^2 = (1+3l\tau_1)^2 \leq 12$ and

$$\begin{aligned}3\Big(l^2+\frac{\bar{\kappa}(1+\tau_2 l)^2}{3\tau_2^2}\Big)\frac{2\tau_1^2\tau_2^2 l^2}{(1-\tau_2\zeta)^2}\gamma_3^2 \leq& \Big(\tau_2^2 l^2+\frac{\bar{\kappa}}{3}(1+\tau_2 l)^2\Big)\cdot\frac{12}{(1-\tau_2\zeta)^2}\cdot(1+\tau_1^2 l^2) \\ \leq& 12\Big(\frac{1}{144^2}+\frac{\bar{\kappa}}{3}\Big(1+\frac{1}{144}\Big)^2\Big)\Big(\frac{16}{15}\Big)^2\Big(1+\frac{1}{9}\Big) \leq \frac{52}{10}\bar{\kappa},\end{aligned}$$

where we used $\tau_1 l \leq \frac{1}{3}$ and $\tau_2 l \leq \frac{1}{144}$ for $p = 2l$, $\gamma_3^2 \leq 2(1+\frac{1}{\tau_1^2 l^2})$ and $\frac{1}{(1-\tau_2\zeta)^2} \leq (\frac{16}{15})^2$ together with $\bar{\kappa} \geq 1$.

Putting pieces together: Combining (47) and (51), we get

$$\begin{aligned}&\left\|G^x_{t+1}\right\|^2+\bar{\kappa}\left\|G^y_{t+1}\right\|^2 \\ \leq& \frac{172}{10}\frac{\bar{\kappa}}{\tau_1^2}\left\|x_{t+1}-x_t\right\|^2+\frac{33}{16}\frac{\bar{\kappa}}{\tau_2^2}\left\|y^+(z_t)-y_t\right\|^2+3p^2\left\|x_{t+1}-z_t\right\|^2 \\ \leq& \max\left\{\frac{58\bar{\kappa}}{\tau_1}, \frac{33\bar{\kappa}}{\tau_2}, \frac{12p}{\beta}\right\}(V_t-V_{t+1}) \\ \leq& \frac{O(1)\bar{\kappa}}{\tau_2}\left(V_t-V_{t+1}\right),\end{aligned} \tag{52}$$

where in the third inequality we use $\frac{1}{\tau_1} \leq \frac{1}{48\tau_2} = \mathcal{O}(\frac{1}{\tau_2})$ and $\frac{p}{\beta} = \frac{2}{\alpha}\frac{l}{\min\{\mu,l\}} \cdot \frac{1}{\tau_2} = \mathcal{O}\Big(\frac{\bar{\kappa}}{\tau_2}\Big)$ since $\frac{2}{\alpha} \leq 4612$ and $\frac{l}{\min\{\mu,l\}} = \bar{\kappa}$. Thus, (52) directly implies that

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\big\|G_{t+1}^x\big\|^2 + \bar{\kappa}\big\|G_{t+1}^y\big\|^2\Big) \leq \frac{O(1)\bar{\kappa}}{T}\Big(\frac{1}{\tau_1}+\zeta\Big)\Big[V_0 - \min_{x\in X, y\in Y, z\in\mathcal{X}} V(x,y;z)\Big].$$

For any $x \in X$, $y \in Y$ and $z \in \mathcal{X}$, let

$$\Delta(x,y;z) := \hat{f}_r(x,y;z) - \Psi_r(y;z) + P(z) - \Psi_r(y;z). \tag{53}$$

Fix an arbitrary $z \in \mathcal{X}$, weak duality implies that $\Delta(x,y;z) \geq 0$ for all $x \in X$ and $y \in Y$; moreover, according to Lemma 7, we can find $x \in X$ and $y \in Y$ such that $\Delta(x,y;z) = 0$. Therefore, for any $z \in \mathcal{X}$, we have $\min_{x\in X, y\in Y}\Delta(x,y;z) = 0$, which implies that $\min_{x\in X,y\in\mathcal{Y},z\in\mathcal{X}} V(x,y;z) = \min_{z\in\mathcal{X}}\{P(z) + \min_{x\in X,y\in Y}\Delta(x,y;z)\} = \min_{z\in\mathcal{X}} P(z)$. This observation leads to the following identity:

$$\begin{aligned}
&V_0 - \min_{x\in X,y\in\mathcal{Y},z\in\mathcal{X}} V(x,y;z)\\
=&P(z_0) + \Delta(x_0,y_0;z_0) - \min_{x\in X,y\in\mathcal{Y},z\in\mathcal{X}}\Big\{P(z) + \Delta(x,y;z)\Big\}\\
=&P(z_0) - \min_{z\in\mathcal{X}} P(z) + \Delta(x_0,y_0;z_0).
\end{aligned}$$

Furthermore, for any $z \in X$, we have

$$P(z) = \min_{x\in X}\max_{y\in Y}\tilde{f}_r(x,y) + \frac{p}{2}\|x-z\|^2 = \min_{x\in X}\Phi(x) + \frac{p}{2}\|x-z\|^2 \leq \Phi(z), \tag{54}$$

and we also have $P(z) \geq \min_{x\in X}\Phi(x)$ for all $z \in X$; therefore, $\min_{z\in\mathcal{X}} P(z) = \min_{x\in X}\Phi(x)$. Hence, for any $x_0, z_0 \in X$ and $y_0 \in Y$, we have

$$V_0 - \min_{x\in X,y\in Y,z\in\mathcal{X}} V(x,y,z) = P(z_0) - \min_{z\in X}\Phi(z) + \Delta_0,$$

where $\Delta_0 := \Delta(x_0,y_0;z_0)$. Thus, when we initialize $z_0 \in \mathcal{M}$, for the final complexity bound we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\big\|G_{t+1}^x\big\|^2 + \bar{\kappa}\big\|G_{t+1}^y\big\|^2\Big) \leq \frac{O(1)\bar{\kappa}}{T}\Big(\frac{1}{\tau_1}+\zeta\Big)\Big(P(z_0) - \bar{F} + \Delta_0\Big),$$

where we used the fact that $\min_{z\in X}\Phi(z) \geq \min_{z\in X} F(A(z)) = \bar{F}$. Moreover, since we have $G_t^x \in \nabla_x\tilde{f}(x_t,y_t) + \partial\delta_X(x_t)$ and $G_t^y \in \nabla_y\tilde{f}(x_t,y_t) - \partial h(y_t)$ for all $t \geq 1$, it follows that

$$\operatorname{dist}\Big(0, \nabla_x\tilde{f}(x_t,y_t) + \partial\delta_X(x_t)\Big) \leq \|G_t^x\|, \quad \operatorname{dist}\Big(0, -\nabla_y\tilde{f}(x_t,y_t) + \partial h(y_t)\Big) \leq \|G_t^y\|.$$

Therefore, for any $T \in \mathbb{Z}_+$, we can conclude that $\min_{1\leq t\leq T} D_t = \mathcal{O}\Big(\frac{\bar{\kappa}}{T}\Big)$, where

$$D_t := \operatorname{dist}^2\Big(0, \nabla_x\tilde{f}(x_t,y_t) + \partial\delta_X(x_t)\Big) + \bar{\kappa}\operatorname{dist}^2\Big(0, -\nabla_y\tilde{f}(x_t,y_t) + \partial h(y_t)\Big), \quad \forall\, t \in \mathbb{Z}_+.$$

□

8 Proof of Theorem 2

Proof. Note that $z_0 \in X$ and $\{x_t\} \subset X$ by the construction; hence, through induction one can argue that $\{z_t\} \subset X$. Moreover, from (47) in the proof of Theorem 1, we know that $\{V_t\}_{t\geq 0}$ is non-increasing sequence, where $V_t \triangleq \hat{f}_r(x_t,y_t;z_t) + 2P(z_t) - 2\Psi_r(y_t;z_t)$. We clearly have $P(z_t) - \Psi_r(y_t;z_t) \geq 0$ from weak duality; therefore, it holds for all $t \geq 0$, $\hat{f}_r(x_t,y_t;z_t) \leq V_t \leq V_0$.

For any $x \in X$, define $Y \supset R^*(x) \triangleq \arg\max_{y\in\mathcal{Y}}\tilde{f}_r(x,y) = \arg\max_{y\in\mathcal{Y}} f(A(x),y) - h(y)$. Define $\bar{l}_y := \max_{x\in X,y\in Y}\|\nabla_y f(A(x),y)\| < \infty$, and let $r^*(x_t) \in R^*(x_t)$ be an arbitrary maximizer of $\tilde{f}_r(x_t,\cdot)$. Since $\|\nabla_y\hat{f}(x_t,y;z_t)\| \leq \bar{l}_y$ for all $y \in Y$, mean-value theorem implies that

$$\hat{f}(x_t,r^*(x_t);z_t) - \hat{f}(x_t,y_t;z_t) \leq \bar{l}_y\|y_t - r^*(x_t)\|;$$

moreover, h being Lipschitz on Y implies that $h(y_t) - h(r^*(x_t)) \leq l_h\|y_t - r^*(x_t)\|$. Thus, summing the two inequalities, we get

$$\Phi(x_t) + \frac{p}{2}\|x_t - z_t\|^2 - (\bar{l}_y + l_h)\|y_t - r^*(x_t)\| \leq \hat{f}_r(x_t, y_t; z_t), \tag{55}$$

where we used $\Phi(x_t) + \frac{p}{2}\|x_t - z_t\|^2 = \Phi(x_t; z_t) = \max_{y\in\mathcal{Y}} \hat{f}_r(x_t, y; z_t) = \hat{f}_r(x_t, r^*(x_t); z_t) = \hat{f}(x_t, r^*(x_t); z_t) - h(r^*(x_t))$. Therefore, for all $t \geq 0$, $\Phi(x_t) \leq V_0 + (\bar{l}_y + l_h)D_Y$ with $D_Y := \max_{y_1,y_2\in Y}\|y_1 - y_2\|$. Thus, by definition of Φ, we get $F(A(x_t)) + \frac{\rho}{4}\|c(x_t)\|^2 = \Phi(x_t) \leq V_0 + (\bar{l}_y + l_h)D_Y$ for $t \geq 0$. Note that $F(A(x_t)) \geq \min_{x\in X} F(A(x)) = \bar{F}$; therefore, we get

$$\|c(x_t)\|^2 \leq \frac{4}{\rho}\left(V_0 - \bar{F} + (\bar{l}_y + l_h)D_Y\right), \quad \forall\, t \geq 0. \tag{56}$$

Recall the definition of the gap function $\Delta(\cdot,\cdot;\cdot)$ given in (53). If we initialize $x_0 = x^*(z_0)$ and $y_0 \in Y^*(z_0)$ for some arbitrary $z_0 \in \mathcal{M}$, we have $\Delta(x_0, y_0; z_0) = 0$; hence, together with (54) and $z_0 \in \mathcal{M}$, it implies that $V_0 = P(z_0) \leq \max_{y\in\mathcal{Y}} \tilde{f}_r(z_0, y) = \max_{y\in\mathcal{Y}} f_r(z_0, y) = F(z_0)$, which is independent of ρ –the first equality follows from the fact that $z_0 \in \mathcal{M}$ implies $A(z_0) = z_0$ and $c(z_0) = 0$. On the other hand, if we initialize $x_0 = z_0$ and $y_0 \in Y^*(z_0)$ for some arbitrary $z_0 \in \mathcal{M}$, then $\Delta_0 = \hat{f}_r(z_0, y_0; z_0) - P(z_0)$ since $P(z_0) = \Psi_r(y_0; z_0)$ whenever $y_0 \in Y^*(z_0)$. Moreover, since $V_0 = P(z_0) + \Delta_0$, we have $V_0 = \hat{f}_r(z_0, y_0; z_0) = \tilde{f}_r(z_0, y_0) = f_r(z_0, y_0) \leq F(z_0)$, which is independent of ρ as well. Therefore, for $\rho \geq 16\Big(F(z_0) - \bar{F} + (\bar{l}_y + l_h)D_Y\Big)$, we can conclude that $\|c(x_t)\| \leq \frac{1}{2}$ for all $t \geq 0$.

Let $x_t = u_t s_t v_t^\top$ be the compact singular value decomposition of x_t. Then, $\|c(x_t)\| = \|x_t^\top x_t - I_r\| = \|v_t(s_t^2 - I_r)v_t^\top\| = \|s_t^2 - I_r\|$. Note that whenever $\operatorname{dist}(x_t, \mathcal{M}) \leq \frac{1}{2}$, projection on to $\mathcal{M}$ is well-defined, i.e., single-valued and Lipschitz continuous; moreover, $\operatorname{dist}(x_t, \mathcal{M}) = \|x_t - \mathcal{P}_{\mathcal{M}}(x_t)\| = \|u_t s_t v_t^\top - u_t v_t^\top\| = \|s_t - I_r\|$ and $\|s_t^2 - I_r\| \geq \|s_t - I_r\|$, i.e., $\|c(x_t)\| \geq \operatorname{dist}(x_t, \mathcal{M})$. Hence, we can conclude that for any $\rho \geq 16\Big(F(z_0) - \bar{F} + (\bar{l}_y + l_h)D_Y\Big)$, we have $\operatorname{dist}(x_t, \mathcal{M}) \leq \frac{1}{2}$ for all $t \geq 0$. On the other hand, since $\mathcal{M}$ is compact, for any $C > 0$ such that $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$, it must hold that $\|x_t\| < C$ for all $t \geq 0$. Thus, Theorem 1 implies that $\sum_{t=1}^{\infty}\|\nabla_x \tilde{f}(x_t, y_t)\|^2 + \bar{\kappa}\operatorname{dist}(0, -\nabla_y \tilde{f}(x_t, y_t) + \partial h(y_t))^2 = \mathcal{O}(\bar{\kappa})$, and we have $\min\{\|\nabla_x \tilde{f}(x_t, y_t)\|^2 + \bar{\kappa}\operatorname{dist}(0, -\nabla_y \tilde{f}(x_t, y_t) + \partial h(y_t))^2 : t = 1, \ldots, T\} = \mathcal{O}(\bar{\kappa}/T)$ for all $T \geq 1$. Therefore, for any $\epsilon > 0$ given, the algorithm can generate $(x_\epsilon, y_\epsilon) \in X \times Y$ such that $\|\nabla_x \tilde{f}(x_\epsilon, y_\epsilon)\| \leq \epsilon$, $\bar{\kappa}\operatorname{dist}(0, -\nabla_y \tilde{f}(x_\epsilon, y_\epsilon) + \partial h(y_\epsilon)) \leq \epsilon$, and $\operatorname{dist}(x_\epsilon, \mathcal{M}) \leq \frac{1}{2}$ within $\mathcal{O}(\frac{\bar{\kappa}}{\epsilon^2})$ iterations of `sm-MGDA`. Hence, setting $\rho \geq \max\{16\Big(F(z_0) - \bar{F} + (\bar{l}_y + l_h)D_Y\Big), 36\bar{L}_x\}$ where $\bar{L}_x := \sup_{y\in Y} L_x(y) < \infty$ with $L_x(y) := \max\{\|\nabla_x f(x, y)\| : \|x\|_2 \leq 1\}$ for $y \in Y$, and invoking Lemma 2 implies that $\|x_\epsilon - \mathcal{P}_{\mathcal{M}}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon$, $\|\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon)\| = \mathcal{O}(\epsilon)$, and $\operatorname{dist}(0, -\nabla_y f(\mathcal{P}_{\mathcal{M}}(x_\epsilon), y_\epsilon) + \partial h(y_\epsilon)) = \mathcal{O}(\epsilon)$.

Next, we argue for the asymptotic stationarity. According to the proof of Theorem 1, for all $t \geq 1$, we have $G_t^x \in \nabla_x \tilde{f}(x_t, y_t) + \partial\delta_X(x_t)$ and $G_t^y \in -\nabla_y \tilde{f}(x_t, y_t) + \partial h(y_t)$. Since $\|x_t\| < C$ for all $t \geq 1$, we also get $G_t^x = \nabla_x \tilde{f}(x_t, y_t)$ for all $t \geq 1$. Furthermore, Theorem 1 guarantees that $\sum_{t=1}^{+\infty}\|G_t^x\|^2 + \bar{\kappa}\|G_t^y\|^2 = \mathcal{O}(\bar{\kappa})$, which implies that $\nabla_x \tilde{f}(x_t, y_t) \to 0$ and $G_t^y \to 0$ as $t \to \infty$. Moreover, for all $t \geq 1$, since $\operatorname{dist}(\bar{x}_t, \mathcal{M}) \leq \frac{1}{2}$, Lemma 2 implies that $\|x_t - \mathcal{P}_{\mathcal{M}}(x_t)\| \to 0$ and $\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_t), y_t) \to 0$. Since `sm-MGDA` sequence $\{(x_t, y_t)\}_{t\geq 0} \subset X \times Y$ is bounded, it has at least one limit point. Let (x^*, y^*) be an arbitrary limit point of $\{(x_t, y_t)\}_{t\geq 1}$ and let $\{(x_{t_k}, y_{t_k})\}_{k\geq 1}$ be a subsequence such that $(x_{t_k}, y_{t_k}) \to (x^*, y^*)$ as $k \to \infty$. One can conclude that $x^* = \mathcal{P}_{\mathcal{M}}(x^*) \in \mathcal{M}$ and $\operatorname{grad}_x f(x^*, y^*) = 0$ since $\|x_{t_k} - \mathcal{P}_{\mathcal{M}}(x_{t_k})\| \to 0$ and $\operatorname{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_{t_k}), y_{t_k}) \to 0$ as $k \to \infty$; furthermore, since $G_{t_k}^y \in -\nabla_y \tilde{f}(x_{t_k}, y_{t_k}) + \partial h(y_{t_k})$ for all $k \geq 0$, $\nabla_y \tilde{f}(x_{t_k}, y_{t_k}) \to \nabla_y f(x^*, y^*)$ (this is due to $A(x^*) = x^*$ as $x^* \in \mathcal{M}$) and $G_{t_k}^y \to 0$, it follows from [41, Theorem 24.4] that $0 \in -\nabla_y f(x^*, y^*) + \partial h(y^*)$.

□

9 Proof of Theorem 3

9.1 Proof of Lemma 3

Since we select $z_0 \in X$ and $\{x_t\} \subset X$ by the construction, through induction one can argue that $\{z_t\} \subset X$. Moreover, from (47) in the proof of Theorem 1, we know that $\{V_t\}_{t\geq 0}$ is non-increasing sequence where $V_t \triangleq \hat{f}_r(x_t, y_t; z_t) - 2\Psi_r(y_t; z_t) + 2P(z_t)$. We clearly have $\hat{f}_r(x_t, y_t; z_t) - \Psi_r(y_t; z_t) \geq 0$ and $P(z_t) - \Psi_r(y_t; z_t) \geq 0$

from weak duality, and we also have $P(z_t) - \Phi^* \geq 0$ since $\Phi^* = \min_{z\in\mathcal{X}} P(z)$ as $P(\cdot)$ is a Moreau envelope of $\Phi(\cdot)$; hence, $V_0 - \Phi^* \geq V_t - \Phi^* \geq 0$ for $t \geq 0$. Therefore, we can conclude that

$$0 \leq P(z_t) - \Psi_r(y_t; z_t) \leq V_t - \Phi^* \leq V_0 - \Phi^*, \quad \forall t \geq 0. \tag{57}$$

For any $z \in \mathcal{X}$, recall that $\Psi_r(y; z) \triangleq \min_{x\in X} \tilde{f}_r(x, y) + \frac{p}{2}\|x - z\|^2$, and that $\tilde{f}_r(x, \cdot)$ is μ-concave over $\mathcal{Y}$ for all $x \in X$; thus, it follows that $\Psi_r(\cdot; z)$ is μ-concave. Indeed, for any $z \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\Psi_r(y; z) + \frac{\mu}{2}\|y\|^2 = \min_{x\in X} H(x, y; z), \quad H(x, y; z) := \tilde{f}_r(x, y) + \frac{\mu}{2}\|y\|^2 + \frac{p}{2}\|x - z\|^2,$$

and $H(x, \cdot; z)$ is concave for all $x \in X$ and $z \in \mathcal{X}$. Since the pointwise infimum of concave functions is concave, $\Psi_r(y; z) + \frac{\mu}{2}\|y\|^2$ is concave in y; hence, $\Psi_r(\cdot; z)$ is μ-concave for all $z \in \mathcal{X}$. Note that for any $t \geq 0$, we have

$$P(z_t) = \max_{y\in\mathcal{Y}} \Psi_r(y; z_t) = \Psi_r(y^*(z_t); z_t) \;\Rightarrow\; \frac{\mu}{2}\|y_t - y^*(z_t)\|^2 \leq P(z_t) - \Psi_r(y_t; z_t) \leq V_0 - \Phi^*. \tag{58}$$

Moreover, it follows from $x^*(z) \triangleq \arg\min_{x\in X} \Phi(x; z) = \arg\min_{x\in X} \max_{y\in\mathcal{Y}} \hat{f}_r(x, y; z)$ and the definition of $r^*(\cdot)$ that $x^*(z) = \arg\min_{x\in X} \hat{f}_r(x, r^*(x); z)$, which leads to $P(z) = \Phi(x^*(z); z) = \hat{f}_r(x^*(z), r^*(x^*(z)); z)$. Finally, since $(x^*(z), y^*(z))$ is the unique saddle point of $\hat{f}_r(x, y; z)$, we can conclude that $y^*(z) = r^*(x^*(z))$ for all $z \in \mathcal{X}$.

It is known that $r^*(\cdot)$ is κ_{yx}-Lipschitz with $\kappa_{yx} := \frac{l_{yx}}{\mu}$, e.g., see [54, Lemma A.4]. Consequently, we have

$$\|y^*(z_t) - y^*(z_0)\| = \|r^*(x^*(z_t)) - r^*(x^*(z_0))\| \leq \kappa_{yx}\|x^*(z_t) - x^*(z_0)\|.$$

Furthermore, using Lemma 8, we get $\|x^*(z) - x^*(z')\| \leq \frac{p}{p-l}\|z - z'\|$, which implies that $\|y^*(z_t) - y^*(z_0)\| \leq \kappa_{yx}\frac{p}{p-l}\|z_t - z_0\|$. Combining this bound with (58) and using triangular inequality leads to $\|y_t - y^*(z_0)\| \leq \sqrt{\frac{2}{\mu}(V_0 - \Phi^*)} + 2\kappa_{yx}\|z_t - z_0\|$ for all $t \geq 0$. Since `sm-MGDA` is initialized from (x_0, y_0, z_0) as in Theorem 2, according to the proof of Theorem 2, one has $V_0 \leq F(z_0)$, which completes the proof of Lemma 3.

9.2 Proof of Lemma 4

The first result directly follows from [54, Lemma E.4]. Moreover, Theorem 1 implies that $\sum_{t=0}^{\infty}\|x_{t+1} - x_t\|^2 = \mathcal{O}(\kappa\tau_1^2)$, and $\tau_2 < \frac{1}{l} \leq \frac{1}{\mu}$ implies that $1 - \tau_2\mu/2 \in (0, 1)$, which implies that $\sum_{t=0}^{\infty}\delta_t < \infty$; hence, $\delta_t \to 0$.

Now we are ready to prove Theorem 3.

Proof of Theorem 3. Lemma 3 implies that $\{x_t, y_t, z_t\}$ stays in a bounded set; therefore, it has at least one limit point (x^*, y^*, z^*). Consider the potential function values at (x_t, y_t, z_t), i.e., $V_t = \hat{f}_r(x_t, y_t; z_t) - 2\Psi_r(y_t; z_t) + 2P(z_t)$ for $t \geq 0$. It follows from (57) that for all $t \geq 0$, we have $\hat{f}_r(x_t, y_t; z_t) \leq V_t \leq V_0$ where we used $P(z_t) \geq \Psi_r(y_t; z_t)$.

Let $\bar{Y} \subset \mathcal{Y}$ denote a compact set for which $\{y_t\}_{t\geq 0} \subset \bar{Y}$. Define $\bar{l}_y := \max_{x\in X, y\in\bar{Y}}\|\nabla_y f(A(x), y)\|$ and let l_h be the Lipschitz constant of h over $\operatorname{dom} h \cap \bar{Y}$ –Assumption 1.**(i)** implies that such l_h exists. Then, from the same arguments we used for (55), we get

$$\Phi(x_t) + \frac{p}{2}\|x_t - z_t\|^2 - (\bar{l}_y + l_h)\|y_t - r^*(x_t)\| \leq \hat{f}_r(x_t, y_t; z_t). \tag{59}$$

Therefore, for all $t \geq 0$, it holds that $\Phi(x_t) \leq V_0 + (\bar{l}_y + l_h)\sqrt{\delta_t}$. Thus, from definition 7, we get $F(A(x_t)) + \frac{\rho}{4}\|c(x_t)\|^2 \leq V_0 + (\bar{l}_y + l_h)\sqrt{\delta_t}$ for $t \geq 0$. Note that $F(A(x_t)) \geq \bar{F}$ and we get

$$\|c(x_t)\|^2 \leq \frac{4}{\rho}\left(V_0 - \bar{F} + (\bar{l}_y + l_h)\sqrt{\delta_t}\right), \quad \forall\, t \geq 0. \tag{60}$$

Since `sm-MGDA` is initialized from (x_0, y_0, z_0) as in Theorem 2, we have $V_0 = P(z_0) \leq F(z_0)$, which is independent of the parameter $\rho > 0$. Suppose we fix $\rho > 32(F(z_0) - \bar{F})$. Recall that $\delta_t \to 0$ as $t \to \infty$; therefore, there exists $\bar{T}_\rho \in \mathbb{Z}_+$ such that $\delta_t \leq \frac{\rho^2}{1024(\bar{l}_y + l_h)^2}$ for all $t \geq \bar{T}_\rho$. Thus, for all $t \geq \bar{T}_\rho$, we can conclude that $\|c(x_t)\| \leq \frac{1}{2}$.

For any fixed $t \geq \bar{T}_\rho$, let $x_t = u_t s_t v_t^\top$ be the compact singular value decomposition of x_t. Then, $\|c(x_t)\| = \|x_t^\top x_t - I_r\| = \|v_t(s_t^2 - I_r)v_t^\top\| = \|s_t^2 - I_r\|$. Note that whenever $\operatorname{dist}(x_t, \mathcal{M}) \leq \frac{1}{2}$, projection on to $\mathcal{M}$ is well-defined, i.e., single-valued and Lipschitz continuous, and $\operatorname{dist}(x_t, \mathcal{M}) = \|x_t - \mathcal{P}_{\mathcal{M}}(x_t)\| = \|u_t s_t v_t^\top - u_t v_t^\top\| = \|s_t - I_r\|$ and $\|s_t^2 - I_r\| \geq \|s_t - I_r\|$, i.e., $\|c(x_t)\| \geq \operatorname{dist}(x_t, \mathcal{M})$. Hence, we can conclude that

for any $\rho \geq 32(F(z_0)-\bar{F})$, we have $\mathrm{dist}(x_t,\mathcal{M}) \leq \frac{1}{2}$ for all $t \geq \bar{T}_\rho$. On the other hand, since $\mathcal{M}$ is compact, for any $C > 0$ such that $C > \frac{1}{2} + \sup_{x\in\mathcal{M}}\|x\|$, it must hold that $\|x_t\| < C$ for all $t \geq \bar{T}_\rho$. Thus, Theorem 1 implies that $\sum_{t=\bar{T}_\rho}^{\infty}\|\nabla_x\tilde{f}(x_t,y_t)\|^2+\bar{\kappa}\mathrm{dist}(0,-\nabla_y\tilde{f}(x_t,y_t)+\partial h(y_t))^2 = \mathcal{O}(\bar{\kappa})$, and we have $\min\{\|\nabla_x\tilde{f}(x_t,y_t)\|^2+\bar{\kappa}\mathrm{dist}(0,-\nabla_y\tilde{f}(x_t,y_t)+\partial h(y_t))^2 : t = \bar{T}_\rho,\ldots,\bar{T}_\rho+T-1\} = \mathcal{O}(\bar{\kappa}/T)$ for all $T \geq 1$. Therefore, for any $\epsilon > 0$ given, the algorithm can generate (x_ϵ,y_ϵ) such that $\|\nabla_x\tilde{f}(x_\epsilon,y_\epsilon)\| \leq \epsilon$, $\bar{\kappa}\mathrm{dist}(0,-\nabla_y\tilde{f}(x_\epsilon,y_\epsilon)+\partial h(y_\epsilon)) \leq \epsilon$, and $\mathrm{dist}(x_\epsilon,\mathcal{M}) \leq \frac{1}{2}$ within $\bar{T}_\rho + \mathcal{O}(\frac{1}{\epsilon^2})$ iterations of `sm-MGDA`. Moreover, since $\bar{F} \leq \Phi^*$ and $\{z_t\} \subset X$, it follows from Lemma 3 that $y_t \in Y_c := \{y \in \mathcal{Y} : \|y-y_0\| \leq \sqrt{\frac{2}{\mu}(F(z_0)-\bar{F})} + 4\kappa_{yx}C\}$ for all $t \geq 0$ and note that the bound does not depend on ρ parameter since l_{yx} and so is $\kappa_{yx} = l_{yx}/\mu$ independent of ρ. Hence, setting $\rho \geq \max\{32(F(z_0)-\bar{F}), 36\bar{l}_x\}$ where $\bar{l}_x := \sup_{y\in Y_c} l_x(y) < \infty$, and invoking Lemma 2 implies that $\|x_\epsilon - \mathcal{P}_{\mathcal{M}}(x_\epsilon)\| \leq \frac{3}{\rho}\epsilon$, $\|\mathrm{grad}_x f(\mathcal{P}_{\mathcal{M}}(x_\epsilon),y_\epsilon)\| = \mathcal{O}(\epsilon)$, and $\mathrm{dist}(0,-\nabla_y\tilde{f}(x_\epsilon,y_\epsilon)+\partial h(y_\epsilon)) = \mathcal{O}(\epsilon)$. Moreover, for any limit point (x^*,y^*) of the `sm-MGDA` sequence (x_t,y_t), it holds that $x^* \in \mathcal{M}$, $\mathrm{grad}_x f(x^*,y^*) = 0$ and $0 \in -\nabla_y f(x^*,y^*) + \partial h(y^*)$. □

10 Proof of Theorem 4

Given arbitrary $x_0 \in \mathcal{M}$ and $y_0 \in Y$. We use a slightly modified version[6] of the potential function defined in [49, Lemma 3.4], i.e., $V_0 := f_r(x_0,y_0)$ and for all $t \geq 0$,

$$V_{t+1} := \tilde{f}_r(x_{t+1},y_{t+1}) + \frac{1}{2\tau_2}\left(\frac{16}{\tau_2\theta_{t+1}} - 15\right)\|y_{t+1}-y_t\|^2 + \left(\frac{8}{\tau_2}\left(1-\frac{\theta_t}{\theta_{t+1}}\right) - \frac{\theta_t}{2}\right)\|y_{t+1}\|^2. \tag{61}$$

For $t \geq 0$, recall that $y_{t+1} = \mathrm{prox}_{\tau_2 h}\Big(y_t + \tau_2\nabla_y\tilde{f}(x_{t+1},y_t) - \tau_2\theta_t y_t\Big)$; hence, it holds that $0 \in \frac{1}{\tau_2}(y_{t+1}-y_t) - (\nabla_y\tilde{f}(x_{t+1},y_t) - \theta_t y_t) + \partial h(y_{t+1})$, which also implies

$$\mathcal{G}^y_{t+1} := \frac{1}{\tau_2}(y_{t+1}-y_t) + \nabla_y\tilde{f}(x_{t+1},y_{t+1}) - \nabla_y\tilde{f}(x_{t+1},y_t) + \theta_t y_t \in \nabla_y\tilde{f}(x_{t+1},y_{t+1}) - \partial h(y_{t+1}).$$

Then, $D^y_{t+1} := \mathrm{dist}\Big(0,-\nabla_y\tilde{f}(x_{t+1},y_{t+1}) + \partial h(y_{t+1})\Big) \leq \|\mathcal{G}^y_{t+1}\| \leq (\frac{1}{\tau_2}+l)\|y_{t+1}-y_t\| + \theta_t\|y_t\|$ for all $t \geq 0$. Similarly, for $t \geq 0$, since $x_{t+1} = \mathcal{P}_X\Big(x_t - \tau_{1,t}\nabla_x\tilde{f}(x_t,y_t)\Big)$, it holds $0 \in \frac{1}{\tau_{1,t}}(x_{t+1}-x_t) + \nabla_x\tilde{f}(x_t,y_t) + \partial\delta_X(x_{t+1})$; therefore, we get

$$\mathcal{G}^x_{t+1} := \frac{1}{\tau_{1,t}}(x_t-x_{t+1}) + \nabla_x\tilde{f}(x_{t+1},y_{t+1}) - \nabla_x\tilde{f}(x_t,y_t) \in \nabla_x\tilde{f}(x_{t+1},y_{t+1}) + \partial\delta_X(x_{t+1}).$$

Then, $D^x_{t+1} := \mathrm{dist}\Big(0,\nabla_x\tilde{f}(x_{t+1},y_{t+1}) + \partial\delta_X(x_{t+1})\Big) \leq \|\mathcal{G}^x_{t+1}\| \leq (\frac{1}{\tau_{1,t}}+l)\|x_{t+1}-x_t\| + l\|y_{t+1}-y_t\|$ for all $t \geq 0$. Thus, combining the two inequality above, we get for all $t \geq 0$ that

$$\begin{aligned} D^2_{t+1} &:= (D^x_{t+1})^2 + (D^y_{t+1})^2 \leq \|\mathcal{G}^x_{t+1}\|^2 + \|\mathcal{G}^y_{t+1}\|^2 \\ &\leq 2\Big(\frac{1}{\tau_{1,t}}+l\Big)^2\|x_{t+1}-x_t\|^2 + 2\Big(l^2 + \Big(\frac{1}{\tau_2}+l\Big)^2\Big)\|y_{t+1}-y_t\|^2 + 2\theta_t^2\|y_t\|^2. \end{aligned} \tag{62}$$

Given an arbitrary $\tilde{b} > \max\{\frac{1}{16}\frac{19^2}{20^2}\cdot(\frac{2}{\tau_2 l}-1),\ 2\}$, and for all $t \geq 0$ let $\alpha_t := \frac{8(\tilde{b}-2)l^2}{\tau_2\theta_t^2}$ and $\beta_t := \tau_2 l^2 + \frac{16l^2}{\tau_2\theta_t^2}\tilde{b} - l$. Note that since $\theta_t = \frac{19}{20}\cdot\frac{1}{\tau_2}\cdot\frac{1}{(t+1)^{1/4}}$, the definition of $\tilde{b}$ implies that $\beta_t > l$ for $t \geq 0$. Using these definitions, we get the following identities:

$$\tau_{1,t} = (\beta_t + l/2)^{-1}, \quad (\beta_t+l)/2 = \alpha_t + \tau_2 l^2/2 + \frac{16l^2}{\tau_2\theta_t^2}, \quad \forall\, t \geq 0. \tag{63}$$

Moreover, let $d_1 := \frac{8\tilde{b}^2}{(\tilde{b}-2)^2} + \frac{1}{32}\frac{19^4}{20^4}\cdot\frac{(\tau_2-\frac{1}{2l})^2+1}{\tau_2^2(\tilde{b}-2)^2 l^2}$. Then, it follows from [49, Equation (3.52)] that

$$\frac{2(\frac{1}{\tau_{1,t}}+l)^2}{\alpha_t^2} \leq \frac{(2\beta_t+l)^2+4l^2}{\alpha_t^2} \leq 4d_1, \quad \forall\, t \geq 0. \tag{64}$$

[6] The potential function we use in this paper also involves the closed convex function $h(\cdot)$.

Therefore, (62) and (64) together imply that

$$D_{t+1}^2 \leq \|\mathcal{G}_{t+1}^x\|^2+\|\mathcal{G}_{t+1}^y\|^2 \leq 4d_1\alpha_t^2\|x_{t+1}-x_t\|^2+\Big(\frac{4}{\tau_2^2}+6l^2\Big)\|y_{t+1}-y_t\|^2+2\theta_t^2\|y_t\|^2, \quad \forall\, t\geq 0.$$

In addition, from [49, Equation (3.48)], for all $t \geq 1$, we have

$$\begin{aligned} &\alpha_t\|x_{t+1}-x_t\|^2+\frac{9}{10\tau_2}\|y_{t+1}-y_t\|^2 \leq V_t - V_{t+1} + B_t,\\ &B_t := \frac{8}{\tau_2}\Big(\frac{\theta_{t-1}}{\theta_t}-\frac{\theta_t}{\theta_{t+1}}\Big)\|y_{t+1}\|^2+\frac{\theta_{t-1}-\theta_t}{2}\|y_{t+1}\|^2. \end{aligned} \tag{65}$$

Next, we upper bound $V_1 - V_0$; the analysis in [49] does not require this bound; however, to establish asymptotic stationarity we need this bound as $V_0 = f_r(x_0, y_0)$ does not depend on ρ whenever $x_0 \in \mathcal{M}$ and this observation is essential for our analysis. Indeed, we first bound $\tilde{f}(x_1, y_1) - \tilde{f}(x_0, y_0)$ using the following two inequalities:

$$\tilde{f}(x_1,y_0)-\tilde{f}(x_0,y_0) \leq \Big(\frac{l}{2}-\frac{1}{\tau_{1,0}}\Big)\|x_2-x_1\|^2 \leq -\frac{\beta_0+l}{2}\|x_1-x_0\|^2, \tag{66}$$

$$\tilde{f}_r(x_1,y_1)-\tilde{f}_r(x_1,y_0) \leq \langle \nabla_y \tilde{f}(x_1,y_0)-g_0,\, y_1-y_0\rangle \leq \frac{1}{2\tau_2}\|y_1-y_0\|^2+\tau_2(l_y^2(y_0)+l_h^2), \tag{67}$$

where $g_0 \in \partial h(y_0)$, $\bar{l}_y(y_0) := \max_{x\in X}\|\nabla_y f(A(x), y_0)\| < \infty$, and l_h is the Lipschitz constant of h. The inequality in (66) follows from the similar arguments we used in (29) and (30); furthermore, (67) follows from concavity of $f_r(x_1, \cdot)$ and using Young's inequality. Therefore, summing (66) and (67), we get

$$\tilde{f}_r(x_1,y_1)-\tilde{f}_r(x_0,y_0) \leq -\frac{\beta_0+l}{2}\|x_1-x_0\|^2+\frac{1}{2\tau_2}\|y_1-y_0\|^2+\tau_2(l_y^2(y_0)+l_h^2);$$

hence, using the definition of V_1 given in (61) together with $V_0 := \tilde{f}_r(x_0, y_0) = f_r(x_0, y_0)$ for $x_0 \in \mathcal{M}$, it follows that

$$V_1-V_0 \leq -\frac{\beta_0+l}{2}\|x_1-x_0\|^2+\Big(\frac{1}{2\tau_2}+\frac{8}{\tau_2^2\theta_1}-\frac{15}{2\tau_2}\Big)\|y_1-y_0\|^2+\left(\frac{8}{\tau_2}\left(1-\frac{\theta_0}{\theta_1}\right)-\frac{\theta_0}{2}\right)\|y_1\|^2+\tau_2(l_y^2(y_0)+l_h^2).$$

Note that from (63), we have $\frac{\beta_0+l}{2} \geq \alpha_0$; moreover, since $\frac{8}{\tau_2^2\theta_1} \leq \frac{13}{10\tau_2}$, we also have $\frac{9}{10\tau_2}+\frac{1}{2\tau_2}+\frac{8}{\tau_2^2\theta_1}-\frac{15}{2\tau_2} \leq 0$, and finally we also observe that $1-\frac{\theta_0}{\theta_1} < 0$; therefore, after using these bounds in the above inequality and dropping the non-positive terms from the right hand side, we get

$$\alpha_0\|x_1-x_0\|^2+\frac{9}{10\tau_2}\|y_1-y_0\|^2 \leq V_0 - V_1 + B_0, \quad \text{where} \quad B_0 := \tau_2(l_y^2(y_0)+l_h^2). \tag{68}$$

For $t \geq 0$, we have $\|\mathcal{G}_{t+1}^x\|^2+\|\mathcal{G}_{t+1}^y\|^2 \leq A_t(V_t - V_{t+1} + B_t) + 2\theta_t^2\|y_t\|^2$ for $A_t := \max\{4d_1\alpha_t, \frac{10}{9}(\frac{4}{\tau_2}+6l^2\tau_2)\}$. Therefore, we can conclude that

$$\sum_{t=0}^{T-1}\frac{1}{A_t}\Big(\|\mathcal{G}_{t+1}^x\|^2+\|\mathcal{G}_{t+1}^y\|^2\Big) \leq \sum_{t=0}^{T-1}\Big(V_t - V_{t+1} + B_t + \frac{2\theta_t^2}{A_t}\|y_t\|^2\Big), \qquad \forall\, T \geq 1. \tag{69}$$

Note that from (61), we can lower bound the potential value at time $t+1$ as follows:

$$V_{t+1} \geq \tilde{f}_r(x_{t+1},y_{t+1}) - \frac{15}{2\tau_2}D_Y^2 + \frac{8}{\tau_2}(1-\frac{\theta_0}{\theta_1})R_Y^2 - \frac{\theta_0}{2}R_Y^2, \qquad \forall\, t \geq 0, \tag{70}$$

where $D_Y := \max_{y,y'\in Y}\|y-y'\|$ and $R_Y := \max_{y\in Y}\|y\|$; therefore, for $\underline{f} := \inf_{x\in X, y\in Y} f_r(A(x), y)$, we have $V_{t+1} \geq \underline{V} := \underline{f} - \frac{15}{2\tau_2}D_Y^2 + \frac{8}{\tau_2}(1-\frac{\theta_0}{\theta_1})R_Y^2 - \frac{\theta_0}{2}R_Y^2$. Thus, for any $T \geq 1$, we have

$$\sum_{t=0}^{T-1}(V_t - V_{t+1}) = V_0 - V_T \leq V_0 - \underline{V}; \tag{71}$$

moreover, using the definition of $\{B_t\}_{t\geq 0}$ given in (65) and (68), it follows that for any $T \geq 1$, we also have

$$\begin{aligned} \sum_{t=0}^{T-1} B_t &= \frac{8}{\tau_2}\Big(\frac{\theta_0}{\theta_1}-\frac{\theta_{T-1}}{\theta_T}\Big)R_Y^2 + \frac{\theta_0-\theta_{T-1}}{2}R_Y^2 + \tau_2(l_y^2(y_0)+l_h^2),\\ &\leq \Big(\frac{8}{\tau_2}\frac{\theta_0}{\theta_1}+\frac{\theta_0}{2}\Big)R_Y^2 + \tau_2(l_y^2(y_0)+l_h^2). \end{aligned} \tag{72}$$

Therefore, we can conclude that for all $T \geq 1$,

$$\begin{aligned}\sum_{t=0}^{T-1}(V_t - V_{t+1} + B_t) &\leq f_r(x_0,y_0) - \underline{f} + \frac{15}{2\tau_2}D_Y^2 + \Big(\frac{16}{\tau_2}\frac{\theta_0}{\theta_1} - \frac{8}{\tau_2} + \theta_0\Big)R_Y^2 + \tau_2(l_y^2(y_0) + l_h^2) \\ &\leq f_r(x_0,y_0) - \underline{f} + \frac{15}{2\tau_2}D_Y^2 + \left(8\cdot\frac{3}{2} + 1\right)\frac{R_Y^2}{\tau_2} + \tau_2(l_y^2(y_0) + l_h^2) := d_2,\end{aligned} \tag{73}$$

where in the last inequality we used $2\frac{\theta_0}{\theta_1} - 1 = 2\cdot 2^{1/4} - 1 \leq \frac{3}{2}$. Note that d_2 is independent of the penalty parameter ρ.

Let $\bar{A} := \max\{4d_1,\ \frac{10}{9}(\frac{4}{\tau_2} + 6l^2\tau_2)\frac{1}{\alpha_0}\}$. Then, $A_t \leq \bar{A}\cdot\alpha_t$ for all $t \geq 0$; hence, (69) implies that for all $T \geq 1$,

$$\sum_{t=0}^{T-1}\frac{\|\mathcal{G}_{t+1}^x\|^2 + \|\mathcal{G}_{t+1}^y\|^2}{\alpha_t} \leq \bar{A}d_2 + 2\bar{A}R_Y^2\cdot\sum_{t=0}^{T-1}\frac{\theta_t^2}{A_t}, \quad \text{and} \quad \sum_{t=0}^{T-1}\frac{\theta_t^2}{A_t} \leq \frac{1}{4d_1}\sum_{t=0}^{T-1}\frac{\theta_t^2}{\alpha_t} = \frac{\tau_2}{32d_1(\tilde{b}-2)l^2}\sum_{t=0}^{T-1}\theta_t^4.$$

Thus, $\sum_{t=0}^{T-1}\frac{\theta_t^2}{A_t} = \mathcal{O}(1)\sum_{t=0}^{T-1}\frac{1}{t+1} = \mathcal{O}(\log(T)+1)$. Moreover, noting that $\sum_{t=0}^{T-1}\frac{1}{\alpha_t} = \Omega(\sqrt{T})$, we have

$$\min_{t=0,\ldots,T-1}\{\|\mathcal{G}_{t+1}^x\|^2 + \|\mathcal{G}_{t+1}^y\|^2\} = \mathcal{O}\Big(\log(T)/\sqrt{T}\Big), \qquad \forall\, T \geq 1. \tag{74}$$

Moreover, for any given $\epsilon > 0$, define $T_\epsilon := \inf\{t \geq 0 : D_{t+1} \leq \epsilon\}$, for which it holds that

$$\epsilon^2\,\Omega(\sqrt{T_\epsilon}) = \sum_{t=0}^{T_\epsilon-1}\frac{\epsilon^2}{\alpha_t} \leq \sum_{t=0}^{T_\epsilon-1}\frac{1}{\alpha_t}D_{t+1}^2 \leq \sum_{t=0}^{T_\epsilon-1}\frac{\|\mathcal{G}_{t+1}^x\|^2 + \|\mathcal{G}_{t+1}^y\|^2}{\alpha_t} = \mathcal{O}(1+\log T_\epsilon); \tag{75}$$

therefore, within $T_\epsilon = \mathcal{O}(\frac{1}{\epsilon^4}\log^2(\frac{1}{\epsilon}))$ iterations, `sm-MGDA` will return a pair (x_ϵ, y_ϵ) satisfying

$$\mathrm{dist}^2\Big(0, \nabla_x\tilde{f}(x_\epsilon,y_\epsilon) + \partial\delta_X(x_\epsilon)\Big) + \mathrm{dist}^2\Big(0, -\nabla_y\tilde{f}(x_\epsilon,y_\epsilon) + \partial h(y_\epsilon)\Big) \leq \epsilon^2. \tag{76}$$

To establish the stationarity of (x_ϵ, y_ϵ) for the original problem (1), we need to argue that the norm constraint $x \in X$ is never active. Indeed, from (65), (68) and (72), it follows that $V_T \leq V_0 + \sum_{t=0}^{T-1}B_t \leq f_r(x_0,y_0) + \Big(\frac{8}{\tau_2}\frac{\theta_0}{\theta_1} + \frac{\theta_0}{2}\Big)R_Y^2 + \tau_2(l_y^2(y_0) + l_h^2)$. Thus, for $t \geq 0$, since $\{\theta_t/\theta_{t+1}\}_{t\geq 0}$ is decreasing and $\theta_t \leq \frac{1}{\tau_2}$, (61) implies that

$$\begin{aligned}\tilde{f}_r(x_{t+1},y_{t+1}) &\leq V_{t+1} + \frac{15}{2\tau_2}\|y_{t+1} - y_t\|^2 + \Big(\frac{8}{\tau_2}\frac{\theta_t}{\theta_{t+1}} + \frac{\theta_t}{2}\Big)\|y_{t+1}\|^2 \\ &\leq f_r(x_0,y_0) + \frac{15}{2\tau_2}D_Y^2 + \Big(\frac{16}{\tau_2}\frac{\theta_0}{\theta_1} + \theta_0\Big)R_Y^2 + \tau_2(l_y^2(y_0) + l_h^2) := \bar{f}.\end{aligned} \tag{77}$$

Using a similar argument that we adopted for deriving (55), for any $y_t^* \in \arg\max_{y\in\mathcal{Y}}\tilde{f}_r(x_t,y)$, it holds that

$$\Phi(x_t) := \arg\min_{y\in\mathcal{Y}}\tilde{f}_r(x_t,y) \leq \tilde{f}_r(x_t,y_t) + (\bar{l}_y + l_h)\|y_t - y_t^*\| \leq \bar{f} + (\bar{l}_y + l_h)D_Y := \bar{\Phi},$$

where $\bar{l}_y := \max_{x\in X, y\in Y}\|\nabla_y f(A(x),y)\| < \infty$; therefore, since $\Phi(x_t) = F(A(x_t)) + \frac{\rho}{4}\|c(x)\|^2$ and $\min_{x\in X}F(A(x)) = \bar{F} > -\infty$, we have

$$\|c(x_t)\|^2 \leq \frac{4}{\rho}\Big(\bar{\Phi} - \bar{F}\Big).$$

Then, for $\rho > 16(\bar{\Phi} - \bar{F}) = 16(\bar{f} + (\bar{l}_y + l_h)D_Y - \bar{F})$, we have $\|c(x_t)\| \leq \frac{1}{2}$, which further implies $\|x_t - \mathcal{P}_{\mathcal{M}}(x_t)\| \leq \frac{1}{2}$. Thus, whenever $C > 0$ is sufficiently large, $x_t \in \mathrm{int}(X)$ for all $t \geq 0$. Then, (76) reads

$$\|\nabla_x\tilde{f}(x_\epsilon,y_\epsilon)\|^2 + \mathrm{dist}^2\Big(0, -\nabla_y\tilde{f}(x_\epsilon,y_\epsilon) + \partial h(y_\epsilon)\Big) \leq \epsilon^2.$$

Thus, we can conclude that (x_ϵ, y_ϵ) is indeed $\mathcal{O}(\epsilon)$-stationary point of the NCMC minimax problem in (1) by invoking Lemma 2 for $\rho > 0$ sufficiently large, i.e., $\rho \geq \max\{16(\bar{\Phi} - \bar{F}), 36\max_{y\in Y}L_x(y)\}$, where $L_x(y) := \max\{\|\nabla_x f(x,y)\|:\ \|x\|_2 \leq 1\}$. Moreover, define $\{(\bar{x}_t, \bar{y}_t)\}_{t\geq 0}$ such that $(\bar{x}_t, \bar{y}_t) = (x_{T(t)}, y_{T(t)})$ where $T(t) := \arg\min\{\|\mathcal{G}_{k+1}^x\|^2 + \|\mathcal{G}_{k+1}^y\|^2:\ k = 0,\ldots,t-1\}$ defined for all $t \geq 1$. Since $\{(\bar{x}_t,\bar{y}_t)\}_{t\geq 0}$ is a bounded sequence, it has at least one limit point, and any of its limit points is a stationary point of stationary point of the NCMC minimax problem in (1).

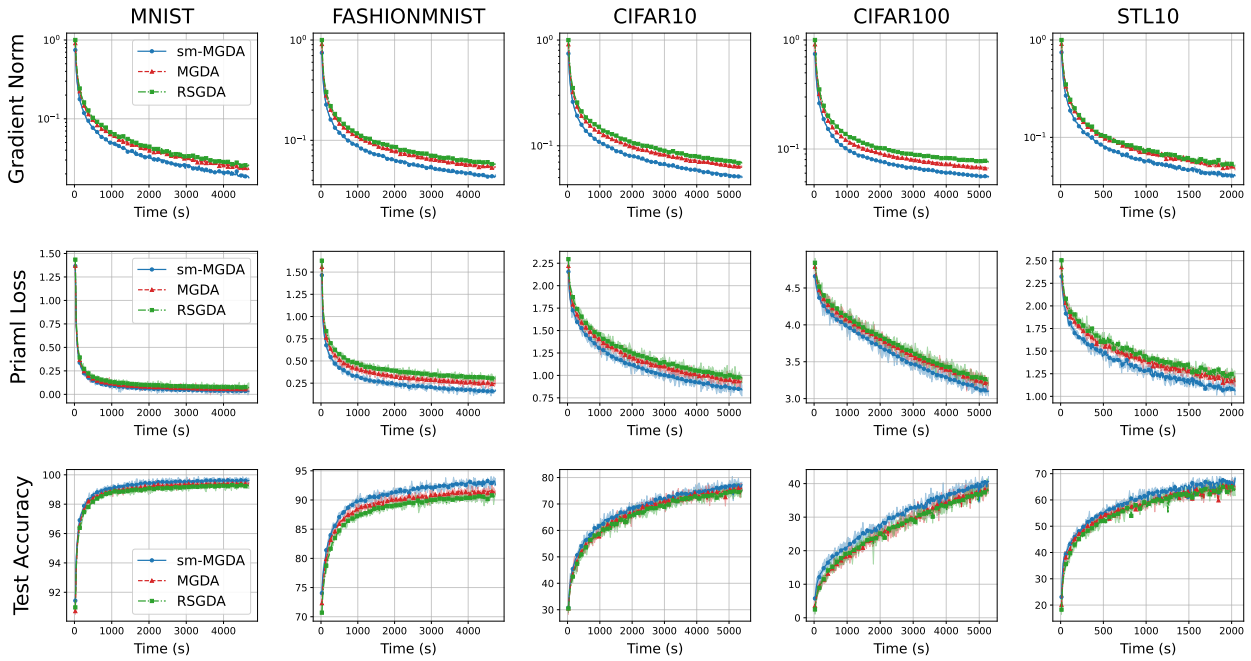

Figure 7: Primal loss, gradient norm, and test accuracy of tested algorithms over 3 runs.

11 Addition experiments

Here we provide additional experiments on superquantile-based learning. Indeed, we focus on distributionally robust optimization (DRO) over Riemannian manifold. Given a set of data samples $\{\xi_i\}_{i=1}^n$, the DRO over Riemannian manifold $\mathcal{M}$ can be written as the following minimax problem:

$$\min_{x\in\mathcal{M}} \max_{w\in\mathcal{S}} \left\{ \sum_{i=1}^{n} w_i \ell(x;\xi_i) - \alpha \|w - \frac{\mathbf{1}}{n}\|^2 \right\}, \tag{78}$$

where $\alpha > 0$ denotes the coefficient, $w = (w_1, \cdots, w_n)$, $\mathcal{S} = \{w \in \mathbb{R}^n : \sum_{i=1}^n w_i = 1,\ w_i \geq 0\}$. Here $\ell(x;\xi_i)$ denotes the loss function over the Riemannian manifold $\mathcal{M}$, which applies to many machine learning problems such as ICA [19], dictionary learning [42], neural network training [24], structured low-rank matrix learning [15], among others. For example, the task of PCA can be cast on a Grassmann manifold.

In the experiment, we use Stiefel manifold $\mathcal{M} = \mathrm{St}(r,d) = \left\{X \in \mathbb{R}^{d\times r} \ :\ X^\top X = I_r\right\}$ on parameters x of DNNs (convolution layers and linear layers), see Table 2 for details. Different algorithms are tested on CIFAR-10, CIFAR-100, STL-10, Fashion MNIST, and MNIST datasets. We set $\tau_1 = \tau_2 = 10^{-3}$, $\beta = 0.9$, $p = 1$, $\rho = 10, C = 1000$ for `sm-MGDA` with the same τ_1 and τ_2 for MGDA and RSGDA. The batch size is 512, and the model is trained for 200 epochs. The results are listed in Figure 7, where the primal loss denotes $F(x)$. It is shown that `sm-MGDA` not only converges the fastest but also has the highest test accuracy compared with other tested algorithms. Furthermore, the gradient norm and primal loss are also the lowest compared with the tested algorithms. The final test accuracy of the compared algorithms are list in Table 4. It is shown that `sm-MGDA` has the highest test accuracy compared with the other two algorithms.

Table 4: Test accuracy (%) of different algorithms on five datasets after 200 epochs.

Algorithm	**MNIST**	**FashionMNIST**	**CIFAR-10**	**CIFAR-100**	**STL-10**
MGDA	99.28	91.85	74.42	37.82	64.01
RSGDA	99.26	90.95	74.95	38.12	63.81
`sm-MGDA`	**99.42**	**94.12**	**76.95**	**41.12**	**65.81**

Since Algorithm 1 is a retration-free algorithm, the heatmaps of parameters $W^\top W$ across different layers of the model training by Algorithm 1 after 200 epochs are shown in Figure 9, which demonstrates that the parameters of the models are indeed lies in the Stiefel manifold. The figures of manifold error with epoch for superquantile-based learning and robust DNN training task are shown in Figures 8 and 6. respectively. It is shown that the manifold error decreases with the epochs, which validates our theoretical result.

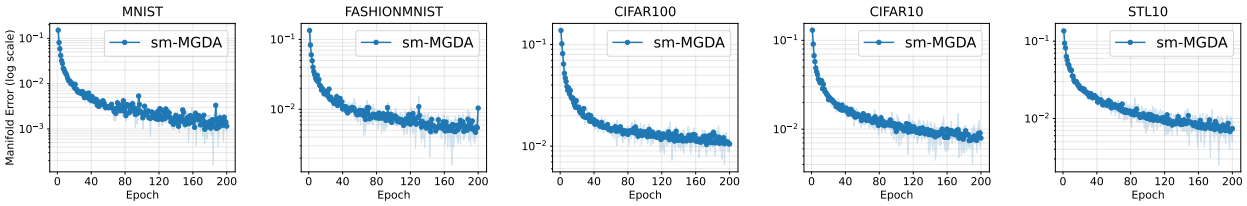

Figure 8: Manifold error of the model with epoch on superquantile-based learning problem.

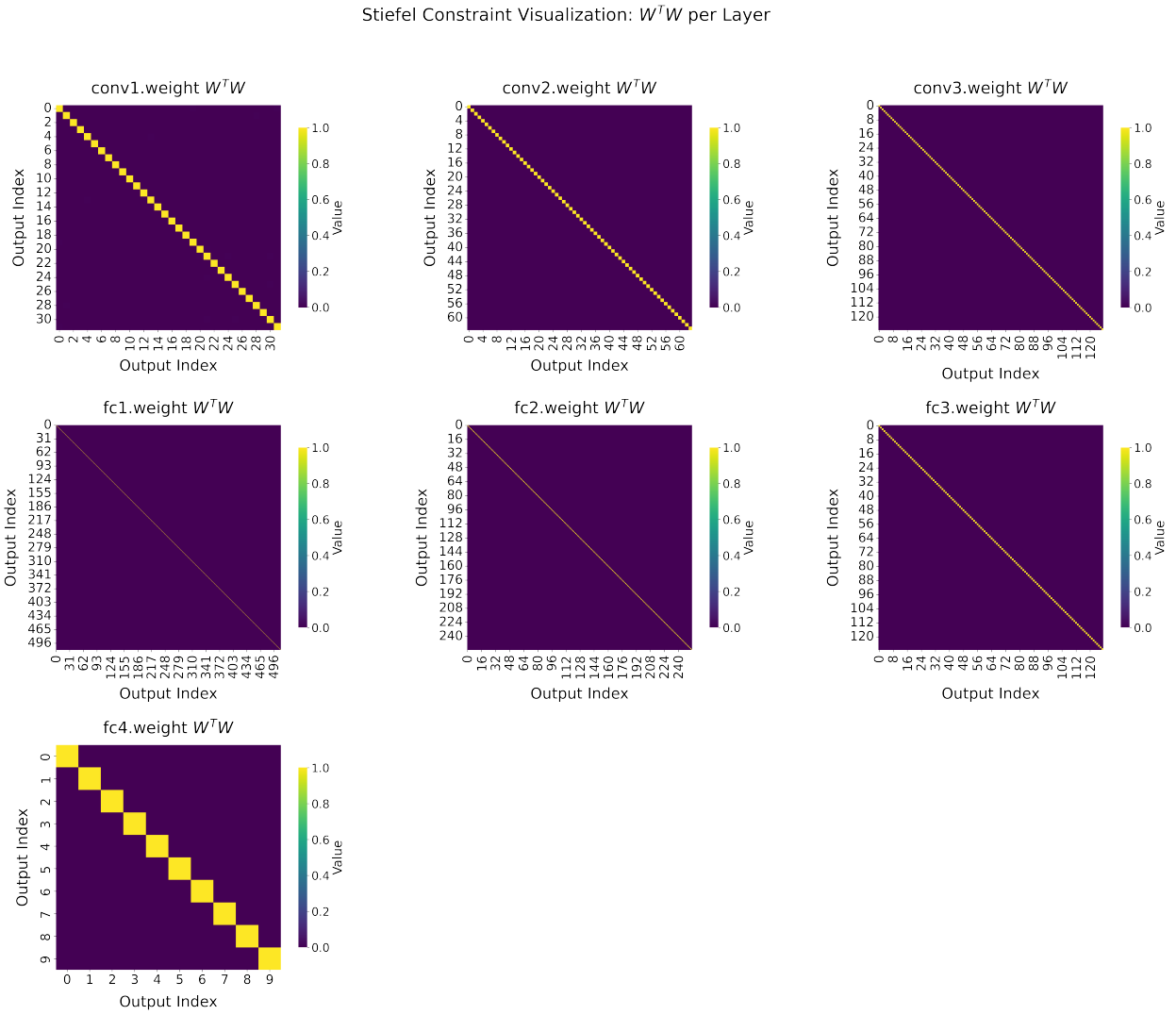

Figure 9: Heatmaps of $W^\top W$ across different layers (from left to right: Layer 1 to Layer 7) after training 200 epochs. The diagonal dominance in each block demonstrates the approximate satisfaction of Stiefel manifold constraints ($W^\top W \approx I$).