Noname manuscript No.
(will be inserted by the editor)

Nonsmooth Nonconvex-Nonconcave Minimax Optimization: Primal-Dual Balancing and Iteration Complexity Analysis

Jiajin Li · Linglingzhi Zhu · Anthony Man-Cho So

Received: date / Accepted: date

Abstract Nonconvex-nonconcave minimax optimization has gained widespread interest over the last decade. However, most existing works focus on variants of gradient descent-ascent (GDA) algorithms, which are only applicable to smooth nonconvex-concave settings. To address this limitation, we propose a novel algorithm named smoothed proximal linear descent-ascent (smoothed PLDA), which can effectively handle a broad range of structured nonsmooth nonconvex-nonconcave minimax problems. Specifically, we consider the setting where the primal function has a nonsmooth composite structure and the dual problem possesses the Kurdyka-Łojasiewicz (KŁ) property with exponent $\theta \in [0, 1)$. We introduce a novel convergence analysis framework for smoothed PLDA, the key components of which are our newly developed nonsmooth primal error bound and dual error bound. Using this framework, we show that smoothed PLDA can find both ϵ-game-stationary points and ϵ-optimization-stationary points of the problems of interest in $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$ iterations. Furthermore, when $\theta \in [0, \frac{1}{2}]$, smoothed PLDA achieves the optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$. To further demonstrate the effectiveness and wide applicability of our

J. Li is supported by the University of British Columbia Start-Up Grant GR032687. A. M.-C. So is supported in part by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) project CUHK 14204823.

Jiajin Li
Sauder School of Business
University of British Columbia, Vancouver, BC, Canada
E-mail: jiajin.li@sauder.ubc.ca

Linglingzhi Zhu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, Georgia, USA
E-mail: llzzhu@gatech.edu

Anthony Man-Cho So
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
E-mail: manchoso@se.cuhk.edu.hk

analysis framework, we show that certain max-structured problem possesses the KŁ property with exponent $\theta = 0$ under mild assumptions. As a by-product, we establish algorithm-independent quantitative relationships among various stationarity concepts, which may be of independent interest.

Keywords Nonconvex-Nonconcave Minimax Optimization, Proximal Linear Scheme, Nonsmooth Composite Structure, Perturbation Analysis, Kurdyka-Łojasiewicz Property

Mathematics Subject Classification (2020) 90C47 · 90C26

1 Introduction

In this paper, we aim to solve the following minimax problem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y). \tag{P}$$

Here, $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ can be both nonconvex with respect to x and nonconcave with respect to y, $\mathcal{X} \subseteq \mathbb{R}^n$ is a nonempty closed convex set, and $\mathcal{Y} \subseteq \mathbb{R}^d$ is a nonempty compact convex set. This problem has a wide range of applications in machine learning and operations research. Examples include learning with non-decomposable loss [58, 67], training generative adversarial networks [1, 27], adversarial training [44, 59], and (distributionally) robust optimization (DRO) [7, 53].

When F is a smooth function, a natural and intuitive approach for solving (P) is gradient descent-ascent (GDA), which involves performing gradient descent on the primal variable x and gradient ascent on the dual variable y in each iteration. For nonconvex-strongly concave problems, GDA can find an ϵ-stationary point with an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ [40]. This matches the lower bound for solving (P) using a first-order oracle, as shown in [14, 37, 68]. However, when $F(x, \cdot)$ is not strongly concave for some $x \in \mathcal{X}$, GDA may encounter oscillations, even for bilinear problems. To overcome this challenge, various techniques utilizing diminishing step sizes have been proposed to guarantee convergence. Nevertheless, these techniques can only achieve a suboptimal complexity of $\mathcal{O}(\epsilon^{-6})$ at best [33, 40, 42]. To achieve a lower iteration complexity, the works [62, 66] introduce a smoothing technique to stabilize the iterates. As a result, the complexity is improved to $\mathcal{O}(\epsilon^{-4})$ for general nonconvex-concave problems. Moreover, the works [63, 66] achieve the "optimal" iteration complexity of $\mathcal{O}(\epsilon^{-2})$ for dual functions that either take a pointwise-maximum form or possess the Polyak-Łojasiewicz (PŁ) property [51].[1] On another front, multi-loop-type algorithms have been developed to achieve a lower iteration complexity for general nonconvex-concave problems [48, 61, 64, 49].

[1] As far as we know, the lowest complexity required for finding approximate stationary points of nonconvex-PŁ minimax/pointwise maximum problems remains an open question. Nevertheless, as we mentioned, for smooth nonconvex-strongly concave problems, the $\mathcal{O}(\epsilon^{-2})$ complexity is already optimal.

Among these, the two triple-loop algorithms in [39, 49] have the lowest iteration complexity of $\mathcal{O}(\epsilon^{-2.5})$.

If we take one step into the nonsmooth world with (P) possessing a separable nonsmooth structure, i.e., its objective function consists of a smooth term and a separable nonsmooth term whose proximal mapping can be readily computed, then the analytic and algorithmic framework from the purely smooth case can be easily adapted. Specifically, a class of (accelerated) proximal-GDA type algorithms have been proposed, where the gradient step is substituted by the proximal gradient step [5, 12, 16, 32]. By utilizing the gradient Lipschitz continuity condition of the smooth term, these algorithms can be shown to achieve the same iteration complexity as those for the smooth case.

As we can see, most existing algorithms can only handle the *almost* smooth case, which refers to scenarios where at least some gradient information is obtainable, such as in purely smooth or separable nonsmooth problems. By contrast, the *general* nonsmooth problem has received relatively little attention in the minimax literature. Only recently have two algorithms, namely the proximally guided stochastic subgradient method [52] and two-timescale GDA [41], been proposed to tackle general nonsmooth weakly convex-concave problems. Unfortunately, these methods suffer from the high iteration complexities of $\mathcal{O}(\epsilon^{-6})$ and $\mathcal{O}(\epsilon^{-8})$, respectively, since they only use subgradient information and neglect problem-specific structures. Thus, a natural question arises: **Q1: Are there structured nonsmooth nonconvex-concave problems whose ϵ-stationary points can be found with an iteration complexity of $\mathcal{O}(\epsilon^{-4})$, just like smooth nonconvex-concave problems?** Taking a step further, we would like to pose a more challenging and intriguing question: **Q2: Can we identify general regularity conditions to achieve the "optimal" rate of $\mathcal{O}(\epsilon^{-2})$ for nonsmooth nonconvex-nonconcave problems?**

1.1 Main Contributions

This paper provides affirmative answers to both questions **Q1** and **Q2** for a particular class of nonsmooth nonconvex-nonconcave problems. We focus on a primal function that has the composite structure $F(\cdot, y) := h_y \circ c_y$ for each $y \in \mathcal{Y}$, where h_y is convex, Lipschitz continuous, and possibly nonsmooth; c_y is continuously differentiable with a Lipschitz continuous Jacobian mapping; and the dual function is continuously differentiable and gradient Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$. Additionally, we assume that either the dual function is concave or the dual problem possesses the KŁ property with exponent $\theta \in [0, 1)$. As concavity alone cannot guarantee the KŁ property [10], we deal with the concave case separately.

To start, we introduce a new algorithm called smoothed proximal linear descent-ascent (smoothed PLDA), which builds on the smoothed GDA algorithm [66]. Unlike its predecessor, smoothed PLDA is extended to handle nonsmooth composite nonconvex-nonconcave problems. To achieve this, we leverage the proximal linear scheme to effectively manage the nonsmooth

Table 1: Comparison of the iteration complexities of smoothed PLDA and other related methods under different settings for solving $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F(x,y)$.

	Primal Func.	**Dual Func./Prob.**	**Iter. Compl.**	**Add. Asm.**
Two-timescale GDA [40]	L-smooth	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X}=\mathbb{R}^n$
Smoothed GDA [66]	L-smooth	concave	$\mathcal{O}(\epsilon^{-4})$	—
PG-SMD [52]	weakly convex	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X}$ bounded
Smoothed PLDA	nonsmooth composite	concave	$\mathcal{O}(\epsilon^{-4})$	—
GDA [40]	L-smooth	strongly concave	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{X}=\mathbb{R}^n$
Smoothed GDA [63]	L-smooth	PŁ	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{Y}=\mathbb{R}^d$
Smoothed PLDA	nonsmooth composite	KŁ with exponent θ	$\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$	$\mathcal{X}$ bounded

composite structure of the primal function. This scheme has been extensively studied in recent literature [47, 15, 30, 35, 23, 24, 29], though its application to minimax optimization has not been explored.

Although the algorithmic extension may seem intuitive and straightforward, it introduces notable challenges in the convergence analysis. These challenges arise primarily from the lack of gradient Lipschitz continuity of the primal function and the nonconcavity of the dual function. To address these challenges, we first establish a tight Lipschitz-type primal error bound property of the proximal linear scheme, which holds even without the gradient Lipschitz continuity condition (see Proposition 2). This error bound is new and plays a crucial role in demonstrating the sufficient decrease property of the designed Lyapunov function. The sufficient decrease property serves as a starting point to ensure the global convergence of smoothed PLDA.

Next, we observe that the nonconcavity of the dual function poses a fundamental challenge in achieving a favorable tradeoff between the decrease in the primal and the increase in the dual. This challenge arises from the absence of inherent dominance between the primal and dual functions, making it difficult to find an optimal balance between the primal and dual updates. As it turns out, the primal-dual tradeoff directly impacts the convergence rate. To quantify this tradeoff, we introduce a new dual error bound property based on the KŁ exponent of the dual problem (see Proposition 3). This is a notable departure from the usual approach of utilizing the KŁ exponent in pure primal nonconvex optimization [36, 3].

The aforementioned primal and dual error bounds lie at the core of our convergence analysis framework. Our main result is that, when the dual problem possesses the KŁ property with exponent $\theta \in [0,1)$, smoothed PLDA can find both ϵ-game-stationary points and ϵ-optimization-stationary points of the nonsmooth composite nonconvex-nonconcave problems introduced earlier in $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$ iterations. For concave functions, smoothed PLDA attains the same iteration complexity of $\mathcal{O}(\epsilon^{-4})$ as the smoothed GDA in [66]. Therefore, we address **Q1** by extending smoothed GDA to a nonsmooth nonconvex-nonconcave setting while preserving at least the same iteration complexity. Table 1 presents a summary of the iteration complexities of smoothed PLDA and other related methods in various scenarios. Using our analysis framework, we further show that when the KŁ exponent of the dual problem lies between 0 and $\frac{1}{2}$, smoothed PLDA achieves the "optimal" iteration complexity of $\mathcal{O}(\epsilon^{-2})$, thus addressing **Q2**.

Interestingly, our analysis framework also reveals a phase transition phenomenon: The iteration complexity depends on the slower of the primal and dual variable updates, which can be characterized by the dual error bound. Specifically, when $\theta \in [0, \frac{1}{2}]$, the dual update is faster than the primal update, which causes the primal update to dominate the optimization process. However, since the primal update involves solving a strongly convex problem, we cannot achieve an iteration complexity better than $\mathcal{O}(\epsilon^{-2})$, which is optimal for nonconvex-strongly concave problems. By contrast, when $\theta \in (\frac{1}{2}, 1)$, the dual update dominates, and the iteration complexity explicitly depends on the KŁ exponent θ, i.e., $\mathcal{O}(\epsilon^{-4\theta})$.

To further demonstrate the wide applicability and effectiveness of our analysis framework, we apply it to problems with a linear dual function and polytopal constraints on y. An important example is the max-structured problem, which takes the form

$$\min_{x \in \mathcal{X}} \max_{i \in [d]} G_i(x). \tag{1.1}$$

Here, $G_i : \mathbb{R}^n \to \mathbb{R}$, where $i \in [d]$, is a nonconvex function with the composite structure mentioned earlier. Such a problem arises frequently in machine learning applications, including distributionally robust optimization (DRO) [8, 57, 26], adversarial training [44], fairness training [45, 48], and distribution-agnostic meta-learning [17]. Under some regularity conditions, we show that problem (1.1) possesses the KŁ property with exponent $\theta = 0$.

Finally, the new dual error bound enables us to establish algorithm-independent quantitative relationships among different stationarity concepts (see Theorem 4). The relationships are obtained as a by-product of our analysis and extend the scope of previous results in [33, 54, 20] to a wider range of settings. Furthermore, they hold great promise for demystifying various notions of stationarity in the context of minimax optimization.

Structure of the paper The paper is organized as follows. In Section 2, we provide several representative applications to demonstrate not only the prevalence of problem (P) when the primal function F has the composite structure mentioned earlier but also the versatility of our proposed smoothed PLDA. In Section 3, we introduce the problem setup and key concepts used throughout the paper. We then present our proposed algorithm and a new convergence analysis framework for studying it in Sections 4 and 5, respectively. In Section 6, we verify that the max-structured problem (1.1) possesses the KŁ property with exponent $\theta = 0$. In Section 7, we clarify the relationships among various stationarity concepts, both conceptually and quantitatively. Finally, we end with some closing remarks in Section 8.

2 Motivating Applications

One important class of problems that is mostly beyond the reach of existing algorithmic approaches but can be tackled by our proposed smoothed PLDA is "two-layer" nonsmooth composite optimization, which takes the form

$$\min_{x \in \mathcal{X}} f(h(c(x))). \tag{2.1}$$

Here, $f : \mathbb{R}^d \to \mathbb{R}$ is convex, $h : \mathbb{R}^m \to \mathbb{R}^d$ is convex and Lipschitz continuous, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable with a Lipschitz continuous Jacobian mapping. Note that both f and h can be nonsmooth. Now, observe that we can isolate the first-layer nonsmoothness by reformulating (2.1) as the minimax problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^d} y^\top h(c(x)) - f^\star(y), \tag{2.2}$$

where $f^\star : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is the conjugate function of f. Consequently, we can apply our proposed smoothed PLDA to solve (2.2).

Various applications admit formulations that feature the nonsmooth composite structure in (2.1). Let us discuss two representative ones here.

(i) **Variation-Regularized Wasserstein DRO [26]**

The DRO methodology aims to find optimal decisions under the most adverse distribution in an ambiguity set, which consists of all probability distributions that fit the observed data with high confidence. In recent years, several works have interpreted regularization from a DRO perspective; see, e.g., [46, 9, 57, 18, 53, 26]. These works provide a probabilistic justification for existing regularization techniques and offer an alternative approach to tackle risk minimization problems. For instance, the work [26] establishes the asymptotic equivalence between Wasserstein distance-induced DRO and the following variation-regularized problem:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i) + \rho \max_{i \in [N]} \|\nabla_x f_\theta(x_i)\|_q. \tag{2.3}$$

Here, $\{(x_i, y_i)\}_{i \in [N]}$ are the feature-label pairs, $\ell : \mathbb{R} \to \mathbb{R}$ is the loss function, f_θ is the feature mapping (e.g., a neural network parameterized by θ), $\rho > 0$ is the constraint radius, and $q \in [1, +\infty]$ is a parameter. It can be shown that problem (2.3) takes the form in (2.1), where the assumptions on f, h, c are satisfied when, e.g., ℓ is convex and Lipschitz, and $f_\theta(x_i)$ has a Lipschitz continuous gradient and $\nabla_x f_\theta(x_i)$ has a Lipschitz continuous Jacobian with respect to θ for all $i \in [N]$. The variation regularization can be regarded as an empirical alternative to Lipschitz regularization, where the goal is to promote smoothness of f_θ over the entire training dataset. Since problem (2.3) is closely related to the task of controlling the Lipschitz constant of deep neural networks, it has attracted significant interest. In particular, problem (2.3) with $\ell(\cdot) = |\cdot|$ being the absolute loss and f_θ being a linear mapping is thoroughly investigated in [8].

(ii) ℓ_1**-Regression with Heavy-Tailed Distributions** [67]

In linear regression, when the input and output follow heavy-tailed distributions, empirical risk minimization may no longer be a suitable approach. This observation has led to recent research on learning with heavy-tailed distributions [4]. For instance, the work [67] proposes the following truncated minimization problem for ℓ_1-regression with heavy-tailed distributions:

$$\min_{\theta\in\Theta} \frac{1}{\alpha N}\sum_{i=1}^{N} \psi\left(\alpha|y_i-\theta^\top x_i|\right). \tag{2.4}$$

Here, $\alpha>0$ is a parameter and $\psi:\mathbb{R}\to\mathbb{R}$ is the nondecreasing truncation function defined by

$$\psi(t)=\begin{cases} \log\left(1+t+\frac{t^2}{2}\right), & t\geq 0\\ -\log\left(1-t+\frac{t^2}{2}\right), & t\leq 0. \end{cases}$$

Note that ψ is a $\frac{1}{4}$-weakly convex function. Then, problem (2.4) can be reformulated as

$$\min_{\theta\in\Theta}\max_{q\in\mathbb{R}^N} \frac{1}{\alpha N}\sum_{i=1}^{N}\left[\alpha q_i|y_i-\theta^\top x_i| - \psi_\mu^\star(q_i) - \frac{\mu\alpha^2}{2}\|y_i-\theta^\top x_i\|^2\right],$$

where $\psi_\mu(\cdot)=\psi(\cdot)+\frac{\mu}{2}\|\cdot\|^2$ and $\mu>\frac{1}{4}$ so as to enforce the strong convexity of ψ_μ. To the best of our knowledge, there is no provably efficient algorithm for solving (2.4).

Besides the nonsmooth composite optimization problem (2.1), a variety of minimax problems can also be tackled by smoothed PLDA. Let us illustrate this by considering the general ϕ-divergence DRO problem studied in [34].

(iii) **General** ϕ**-Divergence DRO** [34]

Building on the problem setup in (i), we consider a slight generalization of ϕ-divergence DRO, which is given by

$$\min_{\theta\in\Theta}\max_{q\in\Delta_N:\sum_{i=1}^N\phi(Nq_i)\leq n\rho} \frac{1}{N}\sum_{i=1}^{N} q_i\ell(f_\theta(x_i,y_i)) - \psi(Nq_i). \tag{2.5}$$

Here, the functions $\phi,\psi:\mathbb{R}\to\mathbb{R}$ are convex and satisfy $\phi(1)=\psi(1)=0$, $\rho\geq 0$ is the constraint radius, and Δ_N is the N-dimensional standard simplex. When $\phi=0$ and $\psi=\iota_{[0,1/\alpha)}$[2] with $\alpha\in[0,1)$, problem (2.5) reduces to Conditional Value-at-Risk minimization [55]. When $\psi=0$ and $\phi(t)=t\log t-t+1$, problem (2.5) reduces to Kullback-Leibler divergence DRO [31]. While the algorithmic approach in [34] only applies to the case where ℓ and f_θ are both smooth, our proposed smoothed PLDA is able to handle a wide range of loss functions, including nonsmooth losses such as the absolute loss.

[2] For any set $\mathcal{S}\subseteq\mathbb{R}^\ell$, we use $\iota_{\mathcal{S}}:\mathbb{R}^\ell\to\{0,+\infty\}$ to denote the indicator function associated with $\mathcal{S}$.

3 Preliminaries

Let us introduce the basic problem setup and key concepts that will serve as the basis for our subsequent analysis.

Assumption 1 (Problem setup) *The following assumptions on the objective function F of problem (P) hold throughout the paper.*

(a) (**Primal function**) *For all $y \in \mathcal{Y}$, the function $F(\cdot, y) : \mathbb{R}^n \to \mathbb{R}$ takes the form $F(\cdot, y) := h_y \circ c_y(\cdot)$, where $c_y : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable with L_c-Lipschitz continuous Jacobian mapping, i.e.,*

$$\|\nabla c_y(x) - \nabla c_y(x')\| \le L_c\|x - x'\| \quad \text{for all } x, x' \in \mathcal{X};$$

and $h_y : \mathbb{R}^m \to \mathbb{R}$ is convex and L_h-Lipschitz continuous, i.e.,

$$|h_y(z) - h_y(z')| \le L_h\|z - z'\| \quad \text{for all } z, z' \in \mathbb{R}^m.$$

(b) (**Dual function**) *For all $x \in \mathcal{X}$, the function $F(x, \cdot) : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable with $\nabla_y F(\cdot, \cdot)$ being L-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$, i.e.,*

$$\|\nabla_y F(x, y) - \nabla_y F(x', y')\| \le L\|(x, y) - (x', y')\| \text{ for all } (x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}.$$

Without loss of generality, we may take $L = L_h L_c$.

Assumption 2 (KŁ property with exponent θ for dual problem) *For all $x \in \mathcal{X}$, the problem $\max_{y\in\mathcal{Y}} F(x, y)$ has a nonempty solution set and a finite optimal value. Moreover, there exist $\mu > 0$ and $\theta \in [0, 1)$ such that*

$$\operatorname{dist}(0, -\nabla_y F(x, y) + \partial\iota_{\mathcal{Y}}(y)) \ge \mu \left(\max_{y'\in\mathcal{Y}} F(x, y') - F(x, y) \right)^{\theta}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x)$, where $\mathcal{Y}^\star(x) := \operatorname{argmax}_{y'\in\mathcal{Y}} F(x, y')$.

Remark 1 *The PŁ property has been widely used as a standard assumption in nonconvex-nonconcave minimax optimization [48, 63]. However, its usage is restricted to smooth unconstrained settings. To overcome this limitation, we employ the KŁ property, which has been demonstrated in [2] to be a nonsmooth extension of the PŁ property. The KŁ exponent, a crucial quantity in the convergence rate analysis of first-order methods for nonconvex optimization [25, 36], plays a vital role in establishing the explicit convergence rate of smoothed PLDA.*

Let us now examine the stationarity measures discussed in this paper. We define the value function $f : \mathbb{R}^n \to \mathbb{R}$, the potential function $F_r : \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$, and the dual potential value function $d_r : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ by

$$f(x) := \max_{y\in\mathcal{Y}} F(x, y),$$

$$F_r(x, y, z) := F(x, y) + \frac{r}{2}\|x - z\|^2,$$

$$d_r(y, z) := \min_{x\in\mathcal{X}} F_r(x, y, z),$$

respectively, where we assume $r > L$ in the rest of this paper.

As the function $F(\cdot, y)$ is weakly convex for each $y \in \mathcal{Y}$ [24, Lemma 4.2], the value function f is also weakly convex. Thus, motivated by the development in [21], we may use the measure in Definition 1(a), which we call *optimization stationarity*, as a (primal) stationarity measure for (P). On the other hand, it is shown in [24, Lemma 4.3] that $\nabla_z d_r(y, x) = r(x - \text{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(x))$, where prox denotes the proximal mapping (see Definition 4 in Appendix A). Thus, we may use the measure in Definition 1(b), which we call *game stationarity*, as a (primal-dual) stationarity measure for (P).

Definition 1 (Stationarity measures) *Let $\epsilon \geq 0$ be given.*

(a) *The point $x \in \mathcal{X}$ is an ϵ-optimization-stationary point* (ϵ-OS) *of problem* (P) *if*

$$\| \text{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x) - x\| \leq \epsilon.$$

(b) *The pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-game-stationary point* (ϵ-GS) *of problem* (P) *if*

$$\|\nabla_z d_r(y, x)\| \leq \epsilon \quad \textit{and} \quad \text{dist}(0, -\nabla_y F(x, y) + \partial\iota_{\mathcal{Y}}(y)) \leq \epsilon.$$ [3]

Since weakly convex functions and smooth functions are subdifferentially regular, we can utilize the Fréchet subdifferential in the above definitions. A brief overview of various subdifferential constructions in nonsmooth nonconvex optimization can be found in [38]. In Section 7, we will explore the quantitative relationship between the two stationarity measures in Definition 1.

4 Proposed Algorithm — Smoothed PLDA

GDA is a commonly used method for solving smooth nonconvex-concave problems. However, its vanilla implementation can lead to oscillations, and a conventional approach to mitigate this issue is to use diminishing step sizes. Unfortunately, one cannot achieve the optimal iteration complexity with this approach. In view of this, the work [66] proposes a Nesterov-type smoothing technique that stabilizes the primal sequence and achieves a better balance between primal and dual updates. However, the technique crucially relies on the smoothness of the objective function.

The nonsmoothness in our problem setting poses a key challenge to algorithm design. To overcome this challenge, we fully leverage the composite structure of the primal function and adopt the proximal linear scheme for the primal update [24]. Specifically, given $r > L$ and $\lambda > 0$, consider the update

$$x^{k+1} \in \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ F_{x^k,\lambda}(x, y^k) + \frac{r}{2}\|x - z^k\|^2 \right\}, \tag{4.1}$$

[3] For any $x \in \mathbb{R}^\ell$ and $\mathcal{S} \subseteq \mathbb{R}^\ell$, we use $\text{dist}(x, \mathcal{S}) := \inf_{z\in\mathcal{S}} \|x - z\|$ to denote the distance from x to $\mathcal{S}$.

where

$$F_{x^k,\lambda}(x, y^k) := h_{y^k}\left(c_{y^k}(x^k) + \nabla c_{y^k}(x^k)^\top (x - x^k)\right) + \frac{1}{2\lambda}\|x - x^k\|^2.$$

Here, $\{z^k\}$ is an auxiliary sequence. In particular, when $h(t) = t$, the primal update (4.1) reduces to the standard gradient descent step as described in [66]. The dual update and other smoothing steps are the same as those in the smoothed GDA developed in [66].

Our smoothed PLDA algorithm is formally presented in Algorithm 1.

Algorithm 1: Smoothed PLDA

Input: Initial point (x^0, y^0, z^0) and parameters $r > L$, $\lambda > 0$, $\alpha > 0$, $\beta \in (0, 1)$

for $k = 0, 1, 2, \dots$ **do**

$\quad x^{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F_{x^k,\lambda}(x, y^k) + \frac{r}{2}\|x - z^k\|^2 \right\}$

$\quad y^{k+1} := \mathrm{proj}_{\mathcal{Y}}\left(y^k + \alpha \nabla_y F(x^{k+1}, y^k)\right)$

$\quad z^{k+1} := z^k + \beta(x^{k+1} - z^k)$

end

Remark 2 (Primal update) *Finding a closed-form solution for the primal update (4.1) is challenging due to the nonsmooth composite structure. However, problem (4.1) is strongly convex. If it in addition possesses certain error bound property, then an ϵ-optimal solution can be found by suitable first-order methods in at most $\mathcal{O}(\log(\epsilon^{-1}))$ iterations; see [69] and [19, Chapter 8] for details and further references. In general, the lower bound on the iteration complexity of first-order methods for solving nonsmooth strongly convex problems is $\Omega(\epsilon^{-1})$ [13, Theorem 3.13]. We emphasize that the optimal approach to solving (4.1) depends on the problem's specific structure. For simplicity and to focus on our main contributions, we assume the exactness of the primal update and concentrate on the iteration complexity of the outer loop in subsequent analyses.*

5 Convergence Analysis of Smoothed PLDA

In this section, we present the main theoretical contributions of this paper. Our goal is to investigate the convergence rate of smoothed PLDA (see Theorem 1) under different settings, including nonconvex-KŁ and nonconvex-concave.

Before proceeding, let us fix the notation as in Table 2.

5.1 Analysis Framework

We start by presenting the key ideas in the analysis.

Table 2: Notation

Notation	Definition	Notes
$F_r(x,y,z)$	$F(x,y)+\frac{r}{2}\\|x-z\\|^2$	potential function
$d_r(y,z)$	$\min_{x\in\mathcal{X}} F_r(x,y,z)$	dual potential value function
$p_r(z)$	$\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} F_r(x,y,z)$	proximal function
$x_r(y,z)$	$\operatorname{argmin}_{x\in\mathcal{X}} F_r(x,y,z)$	—
$x_r^\star(z)$	$\operatorname{argmin}_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F_r(x,y,z)$	—
$Y(z)$	$\operatorname{argmax}_{y\in\mathcal{Y}} d_r(y,z)$	$y(z)\in Y(z)$
$y_+(z)$	$\operatorname{proj}_{\mathcal{Y}}(y+\alpha\nabla_y F(x_r(y,z),y))$	one-step projected gradient ascent on the dual function

Step 1: Construct a *Lyapunov function* that possesses the sufficient decrease property.

Building on [66], we define the Lyapunov function $\Phi_r : \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ as follows:

$$\Phi_r(x,y,z) := \underbrace{F_r(x,y,z) - d_r(y,z)}_{\text{Primal Descent}} + \underbrace{p_r(z) - d_r(y,z)}_{\text{Dual Ascent}} + \underbrace{p_r(z)}_{\text{Proximal Descent}}.$$

The sole distinction between the above Lyapunov function and the one in [66] is that we swap the order of min and max in the definition of p_r, transitioning from $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F_r(x,y,\cdot)$ to $\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} F_r(x,y,\cdot)$. This alteration enables a more comprehensive analysis of cases where the dual function is nonconcave. Each term in the potential function Φ_r is tightly connected to the algorithmic updates. Specifically, the updates for the primal, dual, and auxiliary variables can be viewed as a descent step on the primal function F_r, an approximate ascent step on the dual potential value function d_r, and an approximate descent step on the proximal function p_r, respectively.

Next, our task is to quantify the change in each term of the Lyapunov function after one round of updates. To do so, the key step is the following:

Step 2: Establish a primal error bound (i.e., Proposition 2) to quantify the primal descent.

Intuitively, the primal error bound provides an estimate of the distance between the current point x^{k+1} and the optimal solution

$$x_r(y^k,z^k) = \operatorname*{argmin}_{x\in\mathcal{X}} F_r(x,y^k,z^k)$$

in terms of the iterates gap $\|x^{k+1}-x^k\|$ that results from the proximal linear scheme.

Equipped with the primal error bound, we can obtain the following basic descent estimate of the Lyapunov function:

Proposition 1 (Basic descent estimate of Φ_r) *Let*

$$r \geq 3L,\ \lambda \leq \frac{1}{L},\ \alpha \leq \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\},\ \beta \leq \min\left\{\frac{1}{28}, \frac{(r-L)^2}{14\alpha r(2r-L)^2}\right\},$$

where $\zeta > 0$ is the constant defined in Proposition 2. Then, for any $k \geq 0$, we have

$$\begin{aligned}\Phi_r^k - \Phi_r^{k+1} \geq \frac{3}{8\lambda}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^{k+1})\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 \\ - 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|^2,\end{aligned}$$

where $\Phi_r^k := \Phi_r(x^k, y^k, z^k)$.

The proof of Proposition 1, which is based on the primal descent, dual ascent, and proximal descent properties, is given in Appendix B.

Now, we proceed to bound the negative term

$$\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|$$

in the basic descent estimate using some positive terms, so as to establish the sufficient decrease property. To achieve this goal, we explicitly quantify the primal-dual relationship by a dual error bound, which will allow us to bound the primal update $\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|$ by the dual update $\|y^k - y_+^k(z^{k+1})\|$.

Step 3: Establish a dual error bound (i.e., Proposition 3) to show that the primal and dual updates are balanced in the sense that

$$\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\| \leq \omega\|y^k - y_+^k(z^{k+1})\|^{\upsilon} \quad (5.1)$$

for some constants $\omega > 0$ and $\upsilon > 0$.

We will demonstrate later that the growth power υ in (5.1) is directly determined by the KŁ exponent of the dual problem.

The dual error bound allows us to establish the global convergence rate of smoothed PLDA. To illustrate the main idea, we consider two distinct regimes:

(a) When $\upsilon \in [1, \infty)$, we can show that

$$\begin{aligned}\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\| &\leq \omega\|y^k - y_+^k(z^{k+1})\|^{\upsilon} \\ &\leq \omega \cdot \mathrm{diam}(\mathcal{Y})^{\upsilon-1} \cdot \|y^k - y_+^k(z^{k+1})\|.\end{aligned}$$

Owing to the above **homogeneous** error bound, there exist constants $a, b, c > 0$ such that

$$\Phi_r^k - \Phi_r^{k+1} \geq\ a\|x^k - x^{k+1}\|^2 + b\|y^k - y_+^k(z^{k+1})\|^2 + c\|z^k - z^{k+1}\|^2.$$

Subsequently, it becomes straightforward to achieve an $\mathcal{O}(\epsilon^{-2})$ iteration complexity.

(b) When $\upsilon \in (0,1)$, we encounter an **inhomogeneous** error bound instead. If the negative term $\|x_r(y(z^{k+1}),z^{k+1})-x_r(y_+^k(z^{k+1}),z^{k+1})\|$ is large, then we can still establish the sufficient decrease property by choosing a sufficiently small β. The final iteration complexity will explicitly depend on υ.

The details of Step 3 can be found in Subsection 5.4.

5.2 Primal Error Bound

In this subsection, we establish the primal error bound property of smoothed PLDA. Recall that for the smoothed GDA in [66], its primal error bound property essentially follows from the standard Luo-Tseng error bound for structured strongly convex problems [43, 69, 65]. Specifically, it is shown in [66, Lemma B.2] that

$$\|x^{k+1} - x_r(y^k, z^k)\| \le \zeta \|x^k - \underbrace{\mathrm{proj}_{\mathcal{X}}(x^k - \lambda \nabla_x F_r(x^k, y^k, z^k))}_{x^{k+1}}\|,$$

where $\lambda > 0$ is the step size for primal descent and $\zeta > 0$ is a constant. However, in our problem setting, the primal function does not satisfy the gradient Lipschitz continuity condition. One of our primary contributions is to demonstrate that even so, a Lipschitz-type primal error bound still holds. The primal error bound is crucial to establishing the sufficient decrease property of the Lyapunov function we mentioned earlier.

Proposition 2 (Lipschitz-type primal error bound) *For any $k \ge 0$, we have*

$$\|x^{k+1} - x_r(y^k, z^k)\| \le \zeta \|x^k - x^{k+1}\|, \tag{5.2}$$

where $\zeta := \frac{2(r-L)^{-1}+(\lambda^{-1}+L)^{-1}}{(\lambda^{-1}+L)^{-1}}\left(\sqrt{\frac{2L}{\lambda^{-1}+L}}+1\right)$.

Proof For ease of notation, we define $\hat{F}_{y,z}(\cdot) := F_r(\cdot, y, z) + \iota_{\mathcal{X}}(\cdot)$ for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$. As $x_r(y,z)$ is the optimal solution of $\min_{x\in\mathbb{R}^n} \hat{F}_{y,z}(x)$ and $F_r(\cdot, y, z)$ is $(r-L)$-strongly convex (see the remark after Fact 2 in Appendix A), we have

$$\hat{F}_{y^k,z^k}(x) - \hat{F}_{y^k,z^k}(x_r(y^k,z^k)) \ge \frac{r-L}{2}\|x - x_r(y^k,z^k)\|^2 \quad \text{for all } x \in \mathcal{X}. \tag{5.3}$$

In addition, by the convexity of $\hat{F}_{y^k,z^k}$, we obtain

$$\hat{F}_{y^k,z^k}(x) - \hat{F}_{y^k,z^k}(x_r(y^k,z^k)) \le \mathrm{dist}(0, \partial \hat{F}_{y^k,z^k}(x)) \cdot \|x - x_r(y^k,z^k)\| \tag{5.4}$$

for all $x \in \mathcal{X}$. Combining (5.3) and (5.4) yields

$$\|x - x_r(y^k,z^k)\| \le \frac{2}{r-L}\,\mathrm{dist}(0, \partial \hat{F}_{y^k,z^k}(x)) \quad \text{for all } x \in \mathcal{X}.$$

By utilizing the equivalence between proximal and subdifferential error bounds [23, Theorem 3.4], we conclude that

$$\|x - x_r(y^k, z^k)\| \leq \frac{2(r-L)^{-1} + t}{t} \left\| x - \operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x) \right\| \text{ for all } t > 0,\, x \in \mathcal{X}. \tag{5.5}$$

Now, let us examine the relationship between $\|x^{k+1} - x^k\|$ and $\|x^{k+1} - \operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x^{k+1})\|$. Let $\varphi_k : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the function defined by

$$\varphi_k(x) := \hat{F}_{y^k,z^k}(x) + \frac{\lambda^{-1} + L}{2}\|x - x^k\|^2 - \frac{\lambda^{-1} + L}{2}\|x - x^{k+1}\|^2.$$

It is clear that for any $x \in \mathcal{X}$,

$$\begin{aligned}
&\hat{F}_{y^k,z^k}(x) = F(x, y^k) + \frac{r}{2}\|x - z^k\|^2 \\
\geq\ & F_{x^k,\lambda}(x, y^k) - \frac{\lambda^{-1} + L}{2}\|x - x^k\|^2 + \frac{r}{2}\|x - z^k\|^2 \\
\geq\ & F_{x^k,\lambda}(x^{k+1}, y^k) + \frac{\lambda^{-1} + r}{2}\|x - x^{k+1}\|^2 - \frac{\lambda^{-1} + L}{2}\|x - x^k\|^2 + \frac{r}{2}\|x^{k+1} - z^k\|^2 \\
\geq\ & F(x^{k+1}, y^k) + \frac{\lambda^{-1} - L}{2}\|x^k - x^{k+1}\|^2 + \frac{\lambda^{-1} + r}{2}\|x - x^{k+1}\|^2 \\
&- \frac{\lambda^{-1} + L}{2}\|x - x^k\|^2 + \frac{r}{2}\|x^{k+1} - z^k\|^2 \\
=\ & \hat{F}_{y^k,z^k}(x^{k+1}) + \frac{\lambda^{-1} - L}{2}\|x^{k+1} - x^k\|^2 + \frac{\lambda^{-1} + r}{2}\|x - x^{k+1}\|^2 \\
&- \frac{\lambda^{-1} + L}{2}\|x - x^k\|^2,
\end{aligned}$$

where the first and third inequalities follow from Fact 2, and the second one is from the $(\lambda^{-1} + r)$-strong convexity of $F_{x^k,\lambda}(\cdot, y^k) + \frac{r}{2}\|\cdot - z^k\|^2$ on $\mathcal{X}$ and the definition of x^{k+1}. Since $r > L$, we see that for any $x \in \mathcal{X}$,

$$\begin{aligned}
\varphi_k(x) &\geq \hat{F}_{y^k,z^k}(x^{k+1}) + \frac{\lambda^{-1} - L}{2}\|x^{k+1} - x^k\|^2 + \frac{r - L}{2}\|x - x^{k+1}\|^2 \\
&\geq \hat{F}_{y^k,z^k}(x^{k+1}) + \frac{\lambda^{-1} - L}{2}\|x^{k+1} - x^k\|^2.
\end{aligned}$$

It then follows from the definition of φ_k that

$$\varphi_k(x^{k+1}) - \inf_{x \in \mathbb{R}^n} \varphi_k(x) \leq L\|x^{k+1} - x^k\|^2.$$

This, together with the definition of the proximal mapping, implies that for any $\rho > 0$,

$$\begin{aligned}
&\varphi_k\left(\operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1})\right) + \frac{\rho}{2}\|\operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1}) - x^{k+1}\|^2 \leq \varphi_k(x^{k+1}) \\
\leq\ & \inf_{x \in \mathbb{R}^n} \varphi_k(x) + L\|x^{k+1} - x^k\|^2 \leq \varphi_k\left(\operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1})\right) + L\|x^{k+1} - x^k\|^2,
\end{aligned}$$

which in turn implies that

$$\| \operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1}) - x^{k+1}\| \leq \sqrt{\frac{2L}{\rho}}\|x^{k+1} - x^k\|. \tag{5.6}$$

Recall that our goal is to find a relationship between $\|x^{k+1} - x^k\|$ and $\|x^{k+1} - \operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x^{k+1})\|$. To do this, it suffices to relate $\operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x^{k+1})$ to $\operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1})$. By definition, we have

$$\hat{F}_{y^k,z^k}(x) + \frac{\lambda^{-1}+L}{2}\|x - x^k\|^2 = \varphi_k(x) + \frac{\lambda^{-1}+L}{2}\|x - x^{k+1}\|^2,$$

which implies that $\operatorname{prox}_{\frac{1}{\lambda^{-1}+L}\varphi_k}(x^{k+1}) = \operatorname{prox}_{\frac{1}{\lambda^{-1}+L}\hat{F}_{y^k,z^k}}(x^k)$. By setting $t = (\lambda^{-1}+L)^{-1}$ in (5.5) and $\rho = \lambda^{-1} + L$ in (5.6), we conclude that

$$\begin{aligned}
&\|x^{k+1} - x_r(y^k, z^k)\| \\
\leq\ & \frac{2(r-L)^{-1} + (\lambda^{-1}+L)^{-1}}{(\lambda^{-1}+L)^{-1}} \left\| x^{k+1} - \operatorname{prox}_{\frac{1}{\lambda^{-1}+L}\hat{F}_{y^k,z^k}}(x^{k+1}) \right\| \\
\leq\ & \frac{2(r-L)^{-1} + (\lambda^{-1}+L)^{-1}}{(\lambda^{-1}+L)^{-1}} \left(\left\| x^{k+1} - \operatorname{prox}_{\frac{1}{\lambda^{-1}+L}\hat{F}_{y^k,z^k}}(x^k) \right\| + \|x^{k+1} - x^k\| \right) \\
=\ & \frac{2(r-L)^{-1} + (\lambda^{-1}+L)^{-1}}{(\lambda^{-1}+L)^{-1}} \left(\left\| x^{k+1} - \operatorname{prox}_{\frac{1}{\lambda^{-1}+L}\varphi_k}(x^{k+1}) \right\| + \|x^{k+1} - x^k\| \right) \\
\leq\ & \frac{2(r-L)^{-1} + (\lambda^{-1}+L)^{-1}}{(\lambda^{-1}+L)^{-1}} \left(\sqrt{\frac{2L}{\lambda^{-1}+L}} + 1 \right) \|x^{k+1} - x^k\|,
\end{aligned}$$

where the second inequality is due to the nonexpansiveness of the proximal mapping. The proof is complete. □

Remark 3 *We can also establish the primal error bound* (5.2) *using [23, Theorem 5.3 and Theorem 5.10]. However, since these two theorems are proved using Ekeland's variational principle, the resulting constant* ζ *will be larger. Our approach to establishing* (5.2) *is more elementary and yields a smaller constant* ζ*. It is worth noting that the constant* ζ *obtained in Proposition 2 plays a crucial role in controlling the step sizes for both the primal and dual updates.*

5.3 Dual Error Bound

After establishing the primal error bound and verifying the basic descent estimate in Proposition 1 (which completes Step 2), we proceed to use the KŁ exponent of the dual problem to derive a new dual error bound. Such an error bound provides a crucial relationship between the primal and dual updates and reveals an interesting phase transition phenomenon. Moreover, it furnishes an effective and theoretically justified approach for balancing the primal and dual updates and allows us to establish the global convergence rate of smoothed PLDA. We should point out that our use of the KŁ exponent represents a marked departure from typical approaches, which only focus on primal nonconvex optimization [3, 36].

Proposition 3 (Dual error bound with KŁ exponent) *Suppose that Assumption 2 holds. Then, for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$, we have*

(a) (KŁ exponent $\theta = 0$): $\|x_r(y(z), z) - x_r(y_+(z), z)\| \le \omega_1 \|y - y_+(z)\|$,
(b) (KŁ exponent $\theta \in (0,1)$): $\|x_r(y(z), z) - x_r(y_+(z), z)\| \le \omega_2 \|y - y_+(z)\|^{\frac{1}{2\theta}}$,

where $\omega_1 := \frac{\sqrt{2L\cdot \mathrm{diam}(\mathcal{Y})\cdot(1+\alpha L)}}{\alpha\mu\sqrt{r-L}}$ *and* $\omega_2 := \frac{\sqrt{2}}{\sqrt{r-L}}\left(\frac{1+\alpha L(1+\sigma_2)}{\alpha\mu}\right)^{\frac{1}{2\theta}}$ *(recall that* $y_+(z) = \mathrm{proj}_{\mathcal{Y}}(y + \alpha\nabla_y F(x_r(y,z), y))$*).*

Proof Let $\psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$\psi(x, z) = F_r(x, y(z), z).$$

Consider arbitrary $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathbb{R}^n$. Note that the function $\psi(\cdot, z)$ is $(r - L)$-strongly convex. Since

$$\operatorname*{argmin}_{x' \in \mathcal{X}} \psi(x', z) = \operatorname*{argmin}_{x' \in \mathcal{X}} F_r(x', y(z), z) = x_r(y(z), z),$$

we see that

$$\psi(x, z) - \psi(x_r(y(z), z), z) \ge \frac{r-L}{2}\|x - x_r(y(z), z)\|^2. \tag{5.7}$$

In addition, we have

$$\begin{aligned}
&\psi(x, z) - \psi(x_r(y(z), z), z) \\
&\le \psi(x, z) - F_r(x_r(y_+(z), z), y_+(z), z) \\
&\le \max_{y' \in \mathcal{Y}} F(x, y') + \frac{r}{2}\|x - z\|^2 - F_r(x_r(y_+(z), z), y_+(z), z) \\
&= \max_{y' \in \mathcal{Y}} F(x, y') - F(x_r(y_+(z), z), y_+(z)) + \frac{r}{2}\|x - z\|^2 - \frac{r}{2}\|x_r(y_+(z), z) - z\|^2,
\end{aligned} \tag{5.8}$$

where the first inequality follows from

$$\begin{aligned}
F_r(x_r(y_+(z), z), y_+(z), z) &= \min_{x' \in \mathcal{X}} \left\{ F(x', y_+(z)) + \frac{r}{2}\|x' - z\|^2 \right\} \\
&\le \max_{y' \in \mathcal{Y}} \min_{x' \in \mathcal{X}} \left\{ F(x', y') + \frac{r}{2}\|x' - z\|^2 \right\} \\
&= \min_{x' \in \mathcal{X}} \left\{ F(x', y(z)) + \frac{r}{2}\|x' - z\|^2 \right\} \\
&= \min_{x' \in \mathcal{X}} \psi(x', z) = \psi(x_r(y(z), z), z).
\end{aligned}$$

As (5.7) and (5.8) hold for any $x \in \mathcal{X}$, we obtain the intermediate relation

$$\begin{aligned} &\frac{r-L}{2}\|x_r(y(z),z) - x_r(y_+(z),z)\|^2 \\ &\leq \max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z),y') - F(x_r(y_+(z),z),y_+(z)) \end{aligned} \tag{5.9}$$

by taking $x = x_r(y_+(z),z)$.

Now, we utilize the KŁ exponent θ of the dual problem to bound the right-hand side of (5.9) in terms of $\|y - y_+(z)\|$. Consider the following two cases:

(a) Suppose that $\theta = 0$. If $F(x_r(y_+(z),z),y_+(z)) = \max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z),y')$, then the desired inequality follows trivially from (5.9). Otherwise, we have $y_+(z) \in \mathcal{Y} \setminus \mathcal{Y}^\star(x_r(y_+(z),z))$. By Assumption 2, we have

$$\mathrm{dist}(0, -\nabla_y F(x_r(y_+(z),z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) \geq \mu.$$

Since $F(x,\cdot)$ is continuously differentiable on the compact set $\mathcal{Y}$, $\|\nabla_y F(x,\cdot)\|$ is bounded on $\mathcal{Y}$. Without loss of generality, we can assume that $\|\nabla_y F(x,\cdot)\| \leq L$ on $\mathcal{Y}$. According to the mean value theorem, we have

$$\max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z),y') - F(x_r(y_+(z),z),y_+(z)) \leq L \cdot \mathrm{diam}(\mathcal{Y}).$$

Using the fact that $y_+(z) = \mathrm{proj}_{\mathcal{Y}}(y + \alpha\nabla_y F(x_r(y,z),y))$, we bound

$$\begin{aligned} &\max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z),y') - F(x_r(y_+(z),z),y_+(z)) \\ &\leq \frac{L \cdot \mathrm{diam}(\mathcal{Y})}{\mu^2} \cdot \mathrm{dist}^2(0, -\nabla_y F(x_r(y_+(z),z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) \\ &\leq \frac{Lb^2 \cdot \mathrm{diam}(\mathcal{Y})}{\mu^2}\|y_+(z) - y\|^2, \end{aligned}$$

where $b := \frac{1}{\alpha} + L > 0$ and the last inequality is due to the fact that the projected gradient ascent method satisfies the so-called *relative error condition* [3, Section 5]. It follows that

$$\|x_r(y(z),z) - x_r(y_+(z),z)\| \leq \frac{\sqrt{2L \cdot \mathrm{diam}(\mathcal{Y})} \cdot (1+\alpha L)}{\alpha\mu\sqrt{r-L}}\|y - y_+(z)\|.$$

(b) If $\theta \in (0,1)$, then we have

$$\begin{aligned} &\mu\left(\max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z),y') - F(x_r(y_+(z),z),y_+(z))\right)^\theta \\ &\leq \mathrm{dist}(0, -\nabla_y F(x_r(y_+(z),z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) \\ &\leq \mathrm{dist}(0, -\nabla_y F(x_r(y,z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) \\ &\quad + \|\nabla_y F(x_r(y,z),y_+(z)) - \nabla_y F(x_r(y_+(z),z),y_+(z))\| \\ &\leq \mathrm{dist}(0, -\nabla_y F(x_r(y,z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) + L\sigma_2\|y - y_+(z)\| \end{aligned}$$

$$\leq \left(\frac{1}{\alpha} + L + L\sigma_2\right) \|y_+(z) - y\|,$$

where the third inequality follows from the L-Lipschitz continuity of $\nabla_y F(\cdot, \cdot)$ and (A.3); the last inequality follows from the relative error condition of the projected gradient ascent method. This, together with (5.9), yields

$$\|x_r(y(z), z) - x_r(y_+(z), z)\| \leq \frac{\sqrt{2}}{\sqrt{r - L}} \left(\frac{1 + \alpha L(1 + \sigma_2)}{\alpha\mu}\right)^{\frac{1}{2\theta}} \|y - y_+(z)\|^{\frac{1}{2\theta}}.$$

The proof is complete. □

The dual error bound in Proposition 3 involves the primal-dual quantity $x_r(\cdot, z)$. As the following corollary shows, one can also derive an alternative dual error bound that involves the pure primal quantity $x_r^\star(z)$. Since $x_r^\star(z)$ is the proximal mapping of $\frac{1}{r}f + \iota_{\mathcal{X}}$ at $z \in \mathbb{R}^n$, the alternative dual error bound is closely related to the optimization-stationarity measure. As we shall see, such an error bound plays a crucial role in establishing not only the iteration complexity of smoothed PLDA for finding an OS of problem (P) (see Theorem 1) but also a quantitative relationship between the optimization-stationarity and game-stationarity measures (see Section 7).

Corollary 1 (Alternative dual error bound with KŁ exponent) *Suppose that Assumption 2 holds. Then, for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$, we have*

(a) (KŁ exponent $\theta = 0$): $\|x_r^\star(z) - x_r(y_+(z), z)\| \leq \omega_1 \|y - y_+(z)\|$,
(b) (KŁ exponent $\theta \in (0, 1)$): $\|x_r^\star(z) - x_r(y_+(z), z)\| \leq \omega_2 \|y - y_+(z)\|^{\frac{1}{2\theta}}$.

The ideas of the proof of Corollary 1 are similar to those of Proposition 3. The main difference lies in the definition of ψ, which needs to be modified as follows:

$$\psi(x, z) = \max_{y \in \mathcal{Y}} F_r(x, y, z).$$

We refer the reader to Appendix C for details.

Remark 4 (i) *Our results extend those in [63] to a more general setting, where the primal function is nonsmooth and the dual problem possesses the KŁ property with exponent taking any value in $[0, 1)$. In particular, this covers the case where there are constraints on the dual variable. More importantly, all the existing techniques cannot handle the nonsmooth structure of the primal function. Our generalization provides a deeper understanding of the relationship between the primal and dual updates and serves as a useful tool for studying various stationarity measures in Section 7.* (ii) *In the concave case where Assumption 2 is not satisfied, one can establish a similar dual error bound with $\theta = 1$ and ω being related to* diam$(\mathcal{Y})$*; see Lemma 8 in Appendix D.* (iii) *The dual error bound developed in this paper can potentially be applied to study the convergence rates of other algorithms for minimax optimization, making it a valuable contribution in its own right.*

5.4 Iteration Complexity of Smoothed PLDA

To establish the main theorem, we first develop the following lemma, which establishes a connection between various iterates gaps and the game-stationarity measure.

Lemma 1 *Let $\epsilon \geq 0$ be given. Suppose that*

$$\max\left\{\|x^{k+1}-x^k\|, \|y_+^k(z^{k+1})-y^k\|, \|x^{k+1}-z^k\|\right\} \leq \epsilon.$$

Then, $\left(x^{k+1}, y^{k+1}\right)$ is a $\rho\epsilon$-GS of problem (P), where

$$\rho = \max\left\{(\eta+1+\sigma_1\alpha\beta L)\left(\frac{1}{\alpha}+L\right), r(\zeta+\sigma_2(\eta+1+\sigma_1\alpha\beta L)+\sigma_1)\right\}$$

with $\sigma_1 := \frac{r}{r-L}$ and $\sigma_2 := \frac{2r-L}{r-L}$.

Proof By Definition 1, we only have to quantify $\|\nabla_z d_r(y^{k+1}, x^{k+1})\|$ and $\mathrm{dist}(0, -\nabla_y F(x^{k+1}, y^{k+1}) + \partial\iota_{\mathcal{Y}}(y^{k+1}))$. To begin, observe that

$$\begin{aligned}
&\|\nabla_z d_r(y^{k+1}, x^{k+1})\| \\
&= r\|x^{k+1} - x_r(y^{k+1}, x^{k+1})\| \\
&\leq r(\|x^{k+1} - x_r(y^k, z^k)\| + \|x_r(y^k, z^k) - x_r(y^{k+1}, z^k)\|) \\
&\quad + r\|x_r(y^{k+1}, z^k) - x_r(y^{k+1}, x^{k+1})\| \\
&\leq r\left(\zeta\|x^{k+1} - x^k\| + \sigma_2\|y^k - y^{k+1}\| + \sigma_1\|x^{k+1} - z^k\|\right) \\
&\leq r\left(\zeta\|x^{k+1} - x^k\| + \sigma_1\|x^{k+1} - z^k\|\right) \\
&\quad + r\sigma_2(\eta\|x^{k+1} - x^k\| + \sigma_1\alpha L\|z^{k+1} - z^k\| + \|y_+^k(z^{k+1}) - y^k\|) \\
&\leq r(\zeta + \sigma_2(\eta + 1 + \sigma_1\alpha\beta L) + \sigma_1)\epsilon,
\end{aligned}$$

where the second inequality is due to Proposition 2 and Lemma 2, the third is due to Lemma 4, and the fourth follows from the update $z^{k+1} = z^k + \beta(x^{k+1} - x^k)$ and the assumption of the lemma. Next, recall that

$$y^{k+1} = \mathrm{proj}_{\mathcal{Y}}\left(y^k + \alpha\nabla_y F(x^{k+1}, y^k)\right) = \operatorname*{argmin}_{y\in\mathcal{Y}}\left\{\|y - y^k - \alpha\nabla_y F(x^{k+1}, y^k)\|^2\right\}.$$

The necessary optimality condition yields

$$\begin{aligned}
0 &\in y^{k+1} - y^k - \alpha\nabla_y F(x^{k+1}, y^k) + \partial\iota_{\mathcal{Y}}(y^{k+1}) \\
&= y^{k+1} - y^k - \alpha\nabla_y F(x^{k+1}, y^{k+1}) + \partial\iota_{\mathcal{Y}}(y^{k+1}) \\
&\quad + \alpha(\nabla_y F(x^{k+1}, y^{k+1}) - \nabla_y F(x^{k+1}, y^k)).
\end{aligned}$$

Since $\partial\iota_{\mathcal{Y}}(y^{k+1})$ is the normal cone to $\mathcal{Y}$ at y^{k+1}, we have

$$\begin{aligned}
v &:= -\frac{1}{\alpha}(y^{k+1} - y^k) + \nabla_y F(x^{k+1}, y^k) - \nabla_y F(x^{k+1}, y^{k+1}) \\
&\in -\nabla_y F(x^{k+1}, y^{k+1}) + \partial\iota_{\mathcal{Y}}(y^{k+1}).
\end{aligned}$$

It follows that

$$\begin{aligned}\|v\| &\le \left(\frac{1}{\alpha}+L\right)\|y^{k+1}-y^k\| \\ &\le \left(\frac{1}{\alpha}+L\right)\left(\eta\|x^k-x^{k+1}\|+\sigma_1\alpha L\|z^{k+1}-z^k\|+\|y_+^k(z^{k+1})-y^k\|\right) \\ &\le (\eta+1+\sigma_1\alpha\beta L)\left(\frac{1}{\alpha}+L\right)\epsilon,\end{aligned}$$

where the first inequality is due to the L-Lipschitz continuity of $\nabla_y F(\cdot,\cdot)$ and the second is due to Lemma 4. Putting everything together yields the desired result. □

With the help of Propositions 1, 2, 3, and Lemma 1, we are now ready to develop the main theorem, which gives the iteration complexity of smoothed PLDA under various settings.

Theorem 1 (Main theorem) *Under the setting of Proposition 1, for any integer $K>0$, there exists an index $k\in\{1,2,\ldots,K\}$ such that the following hold:*

(a) (Concave): Suppose that $F(x,\cdot)$ is concave for all $x\in\mathcal{X}$. If $\max_{y\in\mathcal{Y}}F(\cdot,y)$ is bounded below on $\mathcal{X}$ and $\beta\le\mathcal{O}(K^{-\frac{1}{2}})$, then $\left(x^{k+1},y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{4}})$-GS and z^{k+1} is an $\mathcal{O}(K^{-\frac{1}{4}})$-OS of problem (P).
(b) (KŁ exponent $\theta\in(\frac{1}{2},1)$): Suppose that Assumption 2 holds with $\theta\in(\frac{1}{2},1)$. If $\mathcal{X}$ is compact and $\beta\le\mathcal{O}(K^{-\frac{2\theta-1}{2\theta}})$, then $\left(x^{k+1},y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{4\theta}})$-GS and z^{k+1} is an $\mathcal{O}(K^{-\frac{1}{4\theta}})$-OS of problem (P).
(c) (KŁ exponent $\theta\in[0,\frac{1}{2}]$): Suppose that Assumption 2 holds with $\theta\in[0,\frac{1}{2}]$. If $\mathcal{X}$ is compact and $\beta\le\frac{\mathrm{diam}(\mathcal{Y})^{\frac{2\theta-1}{\theta}}}{224\alpha r\omega_2^2}$ when $\theta\in(0,\frac{1}{2}]$ (resp., $\beta\le\frac{1}{224\alpha r\omega_1^2}$ when $\theta=0$), then $\left(x^{k+1},y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{2}})$-GS and z^{k+1} is an $\mathcal{O}(K^{-\frac{1}{2}})$-OS of problem (P).

Remark 5 (i) *Note that concavity alone is not sufficient to guarantee the KŁ property. An example can be found in [10, Theorem 36]. Therefore, in Theorem 1, we need to treat the concave case separately.* (ii) *When the dual function is concave, the results in cases (b) and (c) of Theorem 1 remain valid under the weaker assumption that the KŁ property holds locally around any GS of problem (P), cf. Assumption 2. The proof follows that in [66] and is omitted here.* (iii) *The compactness assumption on $\mathcal{X}$ implies that p_r is bounded below. For the case where the dual function is concave, we can relax this assumption to that of the lower boundedness of $\max_{y\in\mathcal{Y}}F(\cdot,y)$ on $\mathcal{X}$. The latter allows for the possibility that $\mathcal{X}$ is unbounded and is in line with the assumptions made in the literature on smooth nonconvex-concave minimax optimization; see, e.g., [66] and [63].*

Remark 6 *For smooth nonconvex-concave problems, the work [40] introduces a two-timescale GDA algorithm that computes an ϵ-OS with an iteration complexity of $\mathcal{O}(\epsilon^{-6})$. Such an ϵ-OS can then be converted into an ϵ-GS with an* *additional cost of $\mathcal{O}(\epsilon^{-2})$ (depending on the specific algorithm employed). By contrast, Theorem 1 shows that smoothed PLDA can find both an ϵ-GS and an ϵ-OS of a structured nonsmooth nonconvex-concave problem with the same iteration complexity of $\mathcal{O}(\epsilon^{-4})$. Furthermore, when the dual problem satisfies the KŁ property with exponent $\theta \in [0, \frac{1}{2}]$, smoothed PLDA has the optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$ for finding both an ϵ-GS and an ϵ-OS; see Footnote 1.*

Proof As mentioned in Section 5.1, the global convergence rate of smoothed PLDA depends on the specific regime of the growth power v in (5.1). Let us begin by considering cases (a) and (b), i.e., either the dual function is concave or the dual problem possesses the KŁ property with exponent $\theta \in (\frac{1}{2}, 1)$.

In view of the basic descent estimate in Proposition 1, we consider the following two scenarios separately:

(1) There exists a $k \in \{0, 1, \dots, K-1\}$ such that

$$\begin{aligned}&\frac{1}{2}\max\left\{\frac{3}{8\lambda}\|x^k - x^{k+1}\|^2, \frac{1}{8\alpha}\|y^k - y^k_+(z^{k+1})\|^2, \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2\right\}\\ &\leq 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y^k_+(z^{k+1}), z^{k+1})\|^2.\end{aligned}$$

(2) For any $k \in \{0, 1, \dots, K-1\}$, we have

$$\begin{aligned}&\frac{1}{2}\max\left\{\frac{3}{8\lambda}\|x^k - x^{k+1}\|^2, \frac{1}{8\alpha}\|y^k - y^k_+(z^{k+1})\|^2, \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2\right\}\\ &\geq 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y^k_+(z^{k+1}), z^{k+1})\|^2.\end{aligned}$$

Scenario (1). For the case where the KŁ exponent $\theta \in (\frac{1}{2}, 1)$, we deduce from Proposition 3(b) that

$$\begin{aligned}\|y^k - y^k_+(z^{k+1})\|^2 &\leq 224r\alpha\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y^k_+(z^{k+1}), z^{k+1})\|^2\\ &\leq 224r\alpha\beta\omega_2^2\|y^k - y^k_+(z^{k+1})\|^{\frac{1}{\theta}},\end{aligned}$$

which gives $\|y^k - y^k_+(z^{k+1})\| \leq \rho_1\beta^{\frac{\theta}{2\theta-1}}$ with $\rho_1 := (224r\alpha\omega_2^2)^{\frac{\theta}{2\theta-1}}$. Additionally, using the update $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$ and Proposition 3(b), we have

$$\begin{aligned}\|x^{k+1} - z^k\|^2 &= \frac{1}{\beta^2}\|z^{k+1} - z^k\|^2\\ &\leq 49\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y^k_+(z^{k+1}), z^{k+1})\|^2\\ &\leq 49\omega_2^2\|y^k - y^k_+(z^{k+1})\|^{\frac{1}{\theta}} \leq \rho_2^2\beta^{\frac{1}{2\theta-1}},\end{aligned}$$

where $\rho_2 := \sqrt{49\omega_2^2\rho_1^{\frac{1}{\theta}}}$. Lastly, we have

$$\begin{aligned}\|x^{k+1} - x^k\|^2 &\leq \frac{224}{3}r\beta\lambda\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y^k_+(z^{k+1}), z^{k+1})\|^2\\ &\leq \frac{224}{3}r\omega_2^2\beta\lambda\|y^k - y^k_+(z^{k+1})\|^{\frac{1}{\theta}} \leq \rho_3^2\beta^{\frac{2\theta}{2\theta-1}},\end{aligned}$$

where $\rho_3 := \sqrt{\frac{224}{3} r\omega_2^2 \lambda \rho_1^{\frac{1}{\theta}}}$. Combining the above inequalities, we have

$$\begin{aligned}&\max\left\{\|x^k - x^{k+1}\|, \|y^k - y_+^k(z^{k+1})\|, \|z^k - x^{k+1}\|\right\}\\ \le\ &\max\left\{\rho_1 \beta^{\frac{\theta}{2\theta-1}}, \rho_2 \beta^{\frac{1}{4\theta-2}}, \rho_3 \beta^{\frac{\theta}{2\theta-1}}\right\}.\end{aligned}$$

It then follows from Lemma 1 that $\left(x^{k+1}, y^{k+1}\right)$ is a $\rho \cdot \max\{\rho_1 \beta^{\theta/(2\theta-1)}, \rho_2 \beta^{1/(4\theta-2)}, \rho_3 \beta^{\theta/(2\theta-1)}\}$-GS of problem (P).

For the concave case, we can replace the primal-dual quantity $x_r(y(z^{k+1}), z^{k+1})$ by the primal quantity $x_r^\star(z^{k+1})$ on the right-hand side of the inequality for Scenario (1) and apply Lemma 8 in Appendix D (so that θ and ω_2 are replaced by 1 and $\kappa = \frac{1+\alpha L\sigma_2+\alpha L}{\alpha(r-L)} \cdot \mathrm{diam}(\mathcal{Y})$, respectively) to conclude that $\left(x^{k+1}, y^{k+1}\right)$ is a $\max\{\rho_1\beta, \rho_2\beta^{1/2}, \rho_3\beta\}$-GS of problem (P).

Furthermore, observe that for the case where $\theta \in (\frac{1}{2}, 1)$, the optimization-stationarity measure of z^{k+1} satisfies

$$\begin{aligned}&\|z^{k+1} - x_r^\star(z^{k+1})\|\\ \le\ &\|z^{k+1} - x^{k+1}\| + \|x^{k+1} - x_r(y^k, z^k)\| + \|x_r(y^k, z^k) - x_r(y^k, z^{k+1})\|\\ &+ \|x_r(y^k, z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\| + \|x_r(y_+^k(z^{k+1}), z^{k+1}) - x_r^\star(z^{k+1})\|\\ \le\ &(1+\sigma_1)\|z^k - x^{k+1}\| + \zeta\|x^k - x^{k+1}\| + \sigma_2\|y^k - y_+^k(z^{k+1})\|\\ &+ \omega_2\|y^k - y_+^k(z^{k+1})\|^{\frac{1}{2\theta}}\\ \le\ &\max\left\{\left(\rho_2(1+\sigma_1) + \omega_2\rho_1^{\frac{1}{2\theta}}\right)\beta^{\frac{1}{4\theta-2}}, (\zeta\rho_3 + \sigma_2\rho_1)\beta^{\frac{\theta}{2\theta-1}}\right\},\end{aligned} \tag{5.10}$$

where the second inequality follows from the update $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$, Proposition 2, Corollary 1, Lemma 2, and the fact that $0 < \beta \le \frac{1}{28}$. It is evident that the dependence on β is identical to that of the game-stationarity measure of (x^{k+1}, y^{k+1}). For the concave case, the optimization-stationarity measure of z^{k+1} can be bounded similarly.

Scenario (2). Proposition 1 implies that for any $k \in \{0, 1, \ldots, K-1\}$, we have

$$\Phi_r^k - \Phi_r^{k+1} \ge \frac{3}{16\lambda}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^{k+1})\|^2 + \frac{2r}{7\beta}\|z^k - z^{k+1}\|^2. \tag{5.11}$$

The assumptions in case (a) imply that

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F_r(x, y, z) = \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} F_r(x, y, z)$$

for all $z \in \mathbb{R}^n$ (cf. the proof of [66, Lemma B.7]) and lead to the lower boundedness of p_r, while the assumptions in case (b) directly imply that p_r is lower bounded. In both cases, there exists a $\underline{\Phi} > -\infty$ such that for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathbb{R}^n$,

$$\Phi_r(x, y, z)$$

$$= p_r(z) + (F_r(x,y,z) - d_r(y,z)) + (p_r(z) - d_r(y,z)) \geq p_r(z) \geq \underline{\Phi} > -\infty.$$

It follows that

$$\Phi_r^0 - \underline{\Phi} \geq \sum_{k=0}^{K-1} \Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1})$$
$$\geq \sum_{k=0}^{K-1} \min\left\{\frac{3}{16\lambda}, \frac{1}{16\alpha}, \frac{2\beta r}{7}\right\} \left(\|x^k - x^{k+1}\|^2 + \|y^k - y_+^k(z^{k+1})\|^2 + \|z^k - x^{k+1}\|^2\right),$$

where the last inequality is due to (5.11) and the update $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$. In particular, there exists a $k \in \{0, 1, \dots, K-1\}$ such that

$$\max\left\{\|x^k - x^{k+1}\|^2, \|y^k - y_+^k(z^{k+1})\|^2, \|x^{k+1} - z^k\|^2\right\} \leq \frac{\Phi_r^0 - \underline{\Phi}}{\min\left\{\frac{3}{16\lambda}, \frac{1}{16\alpha}, \frac{2\beta r}{7}\right\} K}.$$

Based on Lemma 1 and the fact that $0 < \beta \leq \frac{1}{28}$, we conclude that $\left(x^{k+1}, y^{k+1}\right)$ is an $\mathcal{O}(\sqrt{1/K\beta})$-GS of problem (P). Moreover, by using the same argument as in (5.10), we can show that z^{k+1} is an $\mathcal{O}(\sqrt{1/K\beta})$-OS of problem (P). Upon setting $\beta = CK^{-\frac{1}{2}}$ for the concave case and $\beta = CK^{-\frac{2\theta-1}{2\theta}}$ for the case where the KŁ exponent $\theta \in (\frac{1}{2}, 1)$, with C being a suitably chosen constant, we see that the bounds obtained for the optimization-stationarity and game-stationarity measures are of the same order for both Scenarios (1) and (2). This establishes cases (a) and (b).

Next, let us consider case (c). First, suppose that the KŁ exponent $\theta \in (0, \frac{1}{2}]$. Once again, we can use Proposition 1 to obtain

$$\begin{aligned}\Phi_r^k - \Phi_r^{k+1} \geq \frac{3}{8\lambda}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^{k+1})\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 \\ - 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|^2.\end{aligned}$$

Using Proposition 3(b) and the boundedness of $\mathcal{Y}$, we have

$$\begin{aligned}&\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\| \\ &\leq \omega_2\|y^k - y_+^k(z^{k+1})\|^{\frac{1}{2\theta}} \\ &= \omega_2\|y^k - y_+^k(z^{k+1})\|^{\frac{1}{2\theta}-1}\|y^k - y_+^k(z^{k+1})\| \\ &\leq \omega_2 \cdot \mathrm{diam}(\mathcal{Y})^{\frac{1}{2\theta}-1} \cdot \|y^k - y_+^k(z^{k+1})\|. \end{aligned} \tag{5.12}$$

It follows that

$$\begin{aligned}\Phi_r^k - \Phi_r^{k+1} \geq \frac{3}{8\lambda}\|x^k - x^{k+1}\|^2 + \left(\frac{1}{8\alpha} - 14r\beta\omega_2^2 \cdot \mathrm{diam}(\mathcal{Y})^{\frac{1-2\theta}{\theta}}\right)\|y^k - y_+^k(z^{k+1})\|^2 \\ + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2.\end{aligned}$$

Since $\beta \le \frac{\operatorname{diam}(\mathcal{Y})^{\frac{2\theta-1}{\theta}}}{224\alpha r\omega_2^2}$, we have

$$\Phi_r^k - \Phi_r^{k+1} \ge \frac{3}{8\lambda}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^{k+1})\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2.$$

The desired result can then be obtained by following a similar argument as that in the paragraph proceeding (5.11).

When the KŁ exponent $\theta = 0$, we can apply Proposition 3(a) to derive the inequality (5.12) with θ and ω_2 being replaced by $\frac{1}{2}$ and ω_1, respectively. The desired result then follows by adapting the argument in the preceding paragraph. □

5.5 Phase Transition Phenomenon

Theorem 1 shows that smoothed PLDA achieves the optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$ when the KŁ exponent of the dual problem lies in $[0, \frac{1}{2}]$. What is more, it reveals a surprising phase transition phenomenon at the boundary $\theta = \frac{1}{2}$, where the iteration complexity of the method changes. Such a phenomenon can be explained using the dual error bound we developed in Proposition 3, which characterizes the inherent tradeoff between the primal and dual updates. Indeed, such an error bound aims to control the negative term in the basic descent estimate in Proposition 1. The "degree" of this control depends on the KŁ exponent of the dual problem, which decides the final iteration complexity of our method. Let us consider the two regimes separately.

1. When $\theta \in (0, \frac{1}{2}]$, the primal update dominates the optimization process as the dual problem is already nice enough. Conceptually, since the primal update still requires the solution of a strongly convex problem, we cannot surpass the optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$. Quantitatively, the dual update yields a faster decrease in the Lyapunov function value than the primal one, as we have

   $$\|x_r(y(z), z) - x_r(y_+(z), z)\| \le \omega_2 \|y - y_+(z)\|^{\frac{1}{2\theta}}.$$

   Thus, the main bottleneck in the basic descent estimate is the three positive quantities, and we are only able to achieve an iteration complexity of $\mathcal{O}(\epsilon^{-2})$ by using the following sufficient decrease inequality:

   $$\Phi_r^k - \Phi_r^{k+1} = \Omega\left(\|x^k - x^{k+1}\|^2 + \|y^k - y_+^k(z^{k+1})\|^2 + \|z^k - z^{k+1}\|^2\right).$$

2. When $\theta \in (\frac{1}{2}, 1)$, the dual update dominates the optimization process. This is because the growth condition of the dual problem is worse compared to that of the strongly convex function in the primal update. Quantitatively, the dual update yields a slower decrease in the Lyapunov function value than the primal one. When we incorporate the dual error bound into the basic descent estimate, it becomes evident that the main challenge lies in managing the quantity $\|y^k - y_+^k(z^{k+1})\|^{\frac{1}{\theta}}$. In order to effectively balance the three positive quantities in the basic descent estimate, it is necessary to optimally set $\beta = \mathcal{O}(\epsilon^{4\theta})$ in our proof of Theorem 1.

The underlying insight of the aforementioned phase transition phenomenon is that the overall iteration complexity of smoothed PLDA is determined by the slower of the primal and dual updates. As we have previously discussed, the dual error bound in Proposition 3 offers an effective and theoretically justified approach for balancing the primal and dual updates. It is thus natural to ask whether the primal-dual relationship within our dual error bound is optimal or not.

While a complete answer to this question remains elusive, we now show that an approach different from the one we used in Section 5.3 will lead to a dual error bound that gives suboptimal iteration complexity results. By Lemma 2 in Appendix A, we know that

$$\|x_r(y(z), z) - x_r(y_+(z), z)\| \leq \sigma_2 \|y(z) - y_+(z)\| \tag{5.13}$$

for any $z \in \mathbb{R}^n$ (recall that $y(z) \in \mathrm{argmax}_{y \in \mathcal{Y}} d_r(y, z)$ and $y_+(z) = \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x_r(y, z), y)))$. This suggests that one may obtain a dual error bound by relating $\|y(z) - y_+(z)\|$ to $\|y - y_+(z)\|$ for any $y \in \mathcal{Y}$. To implement this approach, we develop the following new KŁ calculus rule for the max operator, which is motivated by the definition of $y(z)$ and could be of independent interest.

Proposition 4 (KŁ exponent of max operator) *Suppose that Assumption 2 holds. Then, for any $z \in \mathbb{R}^n$, the function $-d_r(\cdot, z) + \iota_{\mathcal{Y}}(\cdot)$ possesses the KŁ property with exponent θ. Specifically, for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$, we have*

$$\mathrm{dist}(0, -\nabla_y d_r(y, z) + \partial \iota_{\mathcal{Y}}(y)) \geq \mu \left(\max_{y' \in \mathcal{Y}} d_r(y', z) - d_r(y, z) \right)^{\theta}.$$

Proof Let $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$ be arbitrary. We bound

$$\begin{aligned}
& \mathrm{dist}(0, -\nabla_y d_r(y, z) + \partial \iota_{\mathcal{Y}}(y)) \\
= \; & \mathrm{dist}(0, -\nabla_y F_r(x_r(y, z), y, z) + \partial \iota_{\mathcal{Y}}(y)) \\
\geq \; & \mu \left(\max_{y' \in \mathcal{Y}} F(x_r(y, z), y') - F(x_r(y, z), y) \right)^{\theta} \\
= \; & \mu \left(\max_{y' \in \mathcal{Y}} F(x_r(y, z), y') + \frac{r}{2}\|x_r(y, z) - z\|^2 - d_r(y, z) \right)^{\theta} \\
\geq \; & \mu \left(\max_{y' \in \mathcal{Y}} \min_{x \in \mathcal{X}} \left\{ F(x, y') + \frac{r}{2}\|x - z\|^2 \right\} - d_r(y, z) \right)^{\theta} \\
= \; & \mu \left(\max_{y' \in \mathcal{Y}} d_r(y', z) - d_r(y, z) \right)^{\theta},
\end{aligned}$$

where the first equality holds due to the strong concavity of $-F_r(\cdot, y, z)$ and [56, Theorem 10.31], the first inequality follows directly from Assumption 2 by taking $x = x_r(y, z)$, and the second equality follows from the fact that

$$d_r(y, z) = \min_{x \in \mathcal{X}} \left\{ F(x, y) + \frac{r}{2}\|x - z\|^2 \right\} = F(x_r(y, z), y) + \frac{r}{2}\|x_r(y, z) - z\|^2.$$

The proof is then complete. □

To proceed, we make use of [22, Theorem 3.7], which connects the KŁ property and the slope error bound. Specifically, let $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$ be arbitrary. By Lemma 3 in Appendix A, the function $d_r(\cdot, z)$ is differentiable. Thus, the Fréchet and limiting subdifferentials of the function $-d_r(\cdot, z) + \iota_{\mathcal{Y}}(\cdot)$ coincide (see, e.g., [38]), and the slope of this function equals $\text{dist}(0, -\nabla_y d_r(\cdot, z) + \partial \iota_{\mathcal{Y}}(\cdot))$ (see, e.g., [22]). Consequently, we can apply Proposition 4, [22, Theorem 3.7], and the relative error condition of the projected gradient ascent method to get

$$\begin{aligned} \text{dist}(y_+(z), Y(z)) &= \mathcal{O}\left(\text{dist}^{\frac{1-\theta}{\theta}}(0, -\nabla_y d_r(y_+(z), z) + \partial \iota_{\mathcal{Y}}(y_+(z))\right) \\ &= \mathcal{O}\left(\|y - y_+(z)\|^{\frac{1-\theta}{\theta}}\right). \end{aligned} \tag{5.14}$$

Combining (5.14) with (5.13) yields the bound

$$\|x_r(y(z), z) - x_r(y_+(z), z)\| = \mathcal{O}\left(\|y - y_+(z)\|^{\frac{1-\theta}{\theta}}\right),$$

which we shall refer to as the pure dual error bound.

Let us now compare the exponents of the dual error bound in Proposition 3 and the pure dual error bound; see Figure 1. When $\theta \in (\frac{1}{2}, 1)$, the exponent of the former is smaller; when $\theta \in (0, \frac{1}{2}]$, the opposite is true. However, as we have noted earlier, the exponent of the dual error bound does not affect the final iteration complexity when $\theta \in (0, \frac{1}{2}]$, as the primal update dominates the optimization process. Thus, we see that for the purpose of studying the iteration complexity of smoothed PLDA, the pure dual error bound is inferior to the dual error bound in Proposition 3.

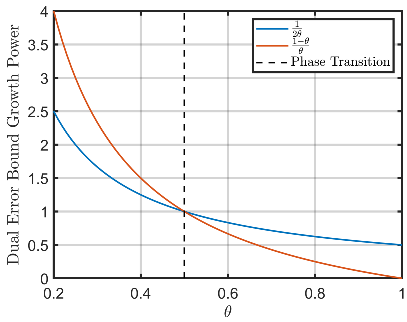

Fig. 1: Comparison between the dual error bound (Proposition 3) and the pure dual error bound.

6 Verification of KŁ Property

As we have seen from the last section, the KŁ property (Assumption 2) is crucial to establishing the iteration complexity of our proposed smoothed PLDA. In this section, we show that Assumption 2 holds when the dual function is linear, i.e., problem (P) takes the form

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{F(x,y)=y^\top G(x)\right\}, \tag{6.1}$$

where $G:\mathbb{R}^n\to\mathbb{R}^d$ is an arbitrary mapping and $\mathcal{Y}$ (assumed to be nonempty) is a polytope defined by

$$\mathcal{Y}=\{y\in\mathbb{R}^d : a_i^\top y\le b_i,\ i\in[l]\},$$

with $l\in\mathbb{N}$ and $a_i\in\mathbb{R}^d$, $b_i\in\mathbb{R}$ for $i\in[l]$.

Theorem 2 (KŁ exponent of linear dual problem) *Consider problem (6.1). Suppose there exists a $\delta>0$ such that for any $x\in\mathcal{X}$ and $y^\star(x)\in\operatorname{argmax}_{y\in\mathcal{Y}} y^\top G(x)$, we have $\lambda_i(x)\ge\delta$ for all $i\in\mathcal{I}(x)$, where $\mathcal{I}(x)=\{i : a_i^\top y^\star(x)=b_i\}\subseteq[l]$ is the active index set and $\lambda_i(x)$ is the dual variable associated with the constraint $a_i^\top y-b_i\le 0$ for $i\in[l]$. Then, Assumption 2 holds with $\theta=0$.*

Proof Our goal is to show that there exists a $\mu>0$ satisfying

$$\operatorname{dist}(0,-G(x)+\partial\iota_{\mathcal{Y}}(y))\ge\mu\quad\text{for all } x\in\mathcal{X},\ y\in\mathcal{Y}\setminus\mathcal{Y}^\star(x),$$

where $\mathcal{Y}^\star(x)=\operatorname{argmax}_{y'\in\mathcal{Y}}F(x,y')=\{y\in\mathcal{Y}:\max_{y'\in\mathcal{Y}}F(x,y')\le F(x,y)\}$. In view of the equivalence between the weak sharp minimum property and the KŁ property with exponent $\theta=0$ for convex functions [11, Theorem 5], it suffices to show that

$$\mu\cdot\operatorname{dist}\left(y,\mathcal{Y}^\star(x)\right)\le\max_{y'\in\mathcal{Y}}F(x,y')-F(x,y)\quad\text{for all } x\in\mathcal{X},\ y\in\mathcal{Y}.$$

Towards that end, let $x\in\mathcal{X}$ be arbitrary and consider the following linear programming problem:

$$\begin{aligned}&\max_{y'\in\mathbb{R}^d}\ F(x,y')\\ &\ \text{s.t.}\ \ a_i^\top y'-b_i\le 0,\quad i\in[l].\end{aligned} \tag{6.2}$$

For any $\bar{y}\in\mathcal{Y}^\star(x)$, using the KKT condition of (6.2), we can find $\lambda_i(x)\ge 0$ for $i\in[l]$ such that

$$\begin{cases}-G(x)+\sum_{i=1}^{l}\lambda_i(x)a_i=0,\\ \lambda_i(x)\cdot(a_i^\top\bar{y}-b_i)=0,\quad i\in[l].\end{cases} \tag{6.3}$$

Moreover, if we let $\mathcal{L}(x):=\left\{y'\in\mathbb{R}^d:(\lambda_i(x)a_i)^\top y'=\lambda_i(x)b_i,\ i\in[l]\right\}$, then $\mathcal{Y}\cap\mathcal{L}(x)=\mathcal{Y}^\star(x)$.

Now, observe that as x varies over $\mathcal{X}$, there are at most 2^l different sets of active constraints in problem (6.2). This, together with the complementarity condition in (6.3), implies that there are at most 2^l different polyhedra in the set $\{\mathcal{L}(x) : x \in \mathcal{X}\}$. Thus, by the linear regularity property of polyhedral sets [6, Corollary 5.26], there exists an absolute constant $\gamma_0 > 0$ such that for all $y \in \mathcal{Y}$,

$$\operatorname{dist}(y, \mathcal{Y}^\star(x)) = \operatorname{dist}(y, \mathcal{Y} \cap \mathcal{L}(x)) \leq \gamma_0 \cdot \operatorname{dist}(y, \mathcal{L}(x)). \tag{6.4}$$

Similarly, by Hoffman's error bound [28], there exists an absolute constant $\gamma_1 > 0$ such that for all $y \in \mathcal{Y}$,

$$\begin{aligned} \operatorname{dist}(y, \mathcal{L}(x)) &\leq \gamma_1 \cdot \sum_{\lambda_i(x)>0} \left|a_i^\top y - b_i\right| \\ &= \gamma_1 \cdot \sum_{\lambda_i(x)>0} \frac{1}{\lambda_i(x)} \left|\lambda_i(x) \cdot (a_i^\top y - b_i)\right|. \end{aligned} \tag{6.5}$$

Since $\lambda_i(x) \geq \delta$ if $\lambda_i(x) \neq 0$ by assumption, we get from (6.4) and (6.5) that

$$\operatorname{dist}\left(y, \mathcal{Y}^\star(x)\right) \leq \frac{\gamma}{\delta} \cdot \sum_{i=1}^{l} |\lambda_i(x) \cdot (a_i^\top y - b_i)|, \tag{6.6}$$

where $\gamma := \gamma_0\gamma_1$. In addition, we have

$$\begin{aligned} &\sum_{i=1}^{l} |\lambda_i(x) \cdot (a_i^\top y - b_i)| = -\sum_{i=1}^{l} \lambda_i(x) \cdot (a_i^\top y - b_i) \\ = \; &-y^\top G(x) + \sum_{i=1}^{l} \lambda_i(x) \cdot a_i^\top \bar{y} = -F(x,y) + \max_{y' \in \mathcal{Y}} F(x, y'), \end{aligned} \tag{6.7}$$

where the second and last equalities follow from (6.3). Putting (6.6) and (6.7) together yields

$$\mu \cdot \operatorname{dist}\left(y, \mathcal{Y}^\star(x)\right) \leq \max_{y' \in \mathcal{Y}} F(x, y') - F(x, y)$$

for all $y \in \mathcal{Y}$, where $\mu = \frac{\delta}{\gamma}$. The proof is then complete. □

Corollary 3 (Max-structured problem) *Consider problem (6.1), where $\mathcal{Y} = \{y \in \mathbb{R}^d : \sum_{i=1}^d y_i = 1,\ y \geq 0\}$ is the standard simplex. Suppose there exists a $\delta > 0$ such that for all $x \in \mathcal{X}$ and $j \in \mathcal{I}(x)$,*

$$\max_{i \in [d]} G_i(x) \geq G_j(x) + \delta. \tag{6.8}$$

Then, Assumption 2 holds with $\theta = 0$.

Proof Let $x \in \mathcal{X}$ and $y^\star(x) \in \operatorname{argmax}_{y\in\mathcal{Y}} y^\top G(x)$ be arbitrary. Since $\mathcal{Y}$ is the standard simplex, the KKT condition (6.3) implies that

$$G(x) = (u(x) - \lambda_1(x), \ldots, u(x) - \lambda_d(x)),$$

where $\lambda_i(x) \geq 0$ (resp., $u(x) \in \mathbb{R}$) is the dual variable associated with the constraint $y_i \geq 0$ for $i \in [d]$ (resp., $\sum_{i=1}^d y_i = 1$). If $y_i^\star(x) > 0$ for some $i \in [d]$, then $\lambda_i(x) = 0$, which implies that $\max_{i\in[d]} G_i(x) = u(x)$. On the other hand, condition (6.8) implies that for any $j \in [d]$ with $y_j^\star(x) = 0$, we have

$$\lambda_j(x) = u(x) - (u(x) - \lambda_j(x)) = \max_{i\in[d]} G_i(x) - G_j(x) \geq \delta.$$

Therefore, by applying Theorem 2, we conclude that problem (6.1) satisfies Assumption 2 with $\theta = 0$. □

Remark 7 (i) *The results in Theorem 2 and Corollary 3, which require minimal assumption on the mapping G, concern the KL exponent of problem* (6.1) *and do not depend on the algorithm used to solve the problem. However, if we are interested in solving problem* (6.1) *using our smoothed PLDA algorithm, then the mapping G needs to possess additional structure, so that the objective function F satisfies Assumption 1. In particular, if $\mathcal{Y}$ is an arbitrary polytope, then the mapping G should be continuously differentiable and have a Lipschitz continuous Jacobian. If $\mathcal{Y}$ is the standard simplex (which is the case in the max-structured problem), then the mapping G can take the more general form $G(x) = (h_1(c_1(x)), \ldots, h_d(c_d(x)))$, where $h_i : \mathbb{R}^m \to \mathbb{R}$ is convex Lipschitz and $c_i : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable with a Lipschitz continuous Jacobian for $i \in [d]$.* (ii) *If the assumptions in Theorem 2 and Corollary 3 hold only locally around every GS of problem* (6.1)*, then it is still possible to show that Assumption 2 holds with $\theta = 0$ locally around every GS of problem* (6.1)*. Such a local property is already sufficient to yield a convergence rate guarantee similar to that in Theorem 1 for smoothed PLDA when applied to problem* (6.1)*; cf. Remark 5*(ii). (iii) *The work [66] establishes a dual error bound for the max-structured problem under the assumptions that the primal function satisfies the gradient Lipschitz continuity condition and certain strict complementarity condition (see [66, Assumption 3.5]) holds. It can be shown that the latter implies condition* (6.8) *in Corollary 3. However, it should be noted that the proof technique used in [66] cannot be extended to the nonsmooth, possibly even discontinuous, setting considered here.*

7 Quantitative Relationships among Different Stationarity Concepts

A fundamental question in the study of minimax optimization is how to define the concept of stationarity. One approach is to consider various natural optimality conditions of the minimax problem and extract from them the corresponding stationarity concepts. In the nonconvex-nonconcave setting, the different stationarity concepts obtained via this approach may not coincide, and their relationships are still not well understood. In this section, we aim to elucidate the relationships among several well-known stationarity concepts for problem (P). Interestingly, the dual error bound we developed in Corollary 1 plays an important role in obtaining our results.

To begin, let us consider the following exact stationarity concepts for problem (P):

Definition 2 (see, e.g., [33, 54]) *The point $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ is called a*

(a) *minimax point* (MM) *of problem* (P) *if*

$$F(x^\star, y) \le F(x^\star, y^\star) \le \max_{y' \in \mathcal{Y}} F(x, y') \quad \textit{for any } (x, y) \in \mathcal{X} \times \mathcal{Y};$$

(b) *game-stationary point* (GS) *of problem* (P) *if*

$$0 \in \partial_x F(x^\star, y^\star) + \partial \iota_{\mathcal{X}}(x^\star) \quad \textit{and} \quad 0 \in -\nabla_y F(x^\star, y^\star) + \partial \iota_{\mathcal{Y}}(y^\star).$$

Furthermore, the point $x^\star \in \mathcal{X}$ is called an

(c) *optimization-stationary point* (OS) *of problem* (P) *if*

$$0 \in \partial (f + \iota_{\mathcal{X}})(x^\star) = \partial \left(\max_{y \in \mathcal{Y}} F(\cdot, y) + \iota_{\mathcal{X}} \right)(x^\star).$$

The extreme value theorem guarantees the existence of an MM when $\mathcal{X}$ is compact, even if $F(\cdot,\cdot)$ is nonconvex-nonconcave. In addition, the weak convexity of $F(\cdot, y)$ and $-F(x, \cdot)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ ensures the existence of a GS [50, Proposition 4.2]. However, it is worth noting that finding an MM of a nonconvex-nonconcave optimization problem is hard, as it includes the task of finding a global maximum of a nonconcave function as special case. Also, since most applications in machine learning, such as GANs and adversarial training, involve sequential games, the optimization-stationarity concept may be more attractive.

While the above definition of OS is standard, it only characterizes the primal variable x. As such, the concept of OS is not directly comparable to that of MM or GS. To circumvent this difficulty, we introduce the concept of an extended OS (eOS). Specifically, the point $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ is called an eOS of problem (P) if $x^\star$ is an OS and $F(x^\star, y) \le F(x^\star, y^\star)$ for any $y \in \mathcal{Y}$. From Definition 2, we can immediately conclude that every MM is an eOS. To gain some intuition on the possible relationships among the three types of stationarity points MM, GS, and eOS, it is instructive to consider the following example.

Example 1 *The following instances of problem* (P) *give rise to the relationships among MM, GS, and eOS shown in Figure 2.*

(a) *If $F(x, y) = x^3 - 2xy - y^2$ and $\mathcal{X} \times \mathcal{Y} = [-1, 1] \times [-1, 1]$, then*

$$\mathrm{MM} = \{(-1, 1), (0, 0)\}, \ \ \mathrm{GS} = \mathrm{eOS} = \left\{ (-1, 1), \left(-\frac{2}{3}, \frac{2}{3} \right), (0, 0) \right\}.$$

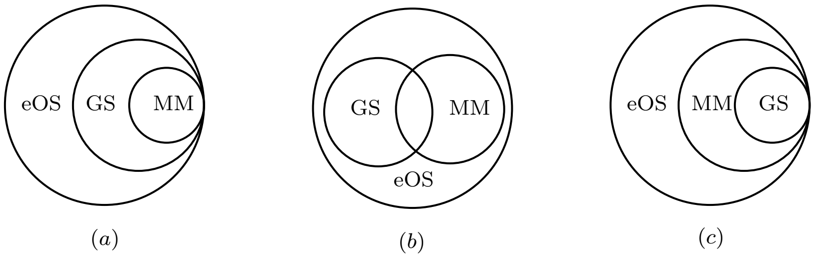

Fig. 2: Possible relationships among the three types of stationarity points MM, GS, and eOS: (a) $F(x,y) = x^3 - 2xy - y^2$ with $\mathcal{X} \times \mathcal{Y} = [-1,1] \times [-1,1]$; (b) $F(x,y) = \sin(x)y$ with $\mathcal{X} \times \mathcal{Y} = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-1,1]$; (c) $F(x,y) = xy$ with $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times [-1,1]$.

(b) *If* $F(x,y) = \sin(x)y$ *and* $\mathcal{X} \times \mathcal{Y} = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-1,1]$*, then*

$$\begin{aligned} &\mathrm{MM} = \{(0,[-1,1])\}\,, \ \mathrm{GS} = \left\{\left(-\frac{\pi}{2}, -1\right), (0,0), \left(\frac{\pi}{2}, 1\right)\right\}, \\ &\mathrm{eOS} = \left\{\left(-\frac{\pi}{2}, -1\right), (0,[-1,1]), \left(\frac{\pi}{2}, 1\right)\right\}. \end{aligned}$$

(c) *If* $F(x,y) = xy$ *and* $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times [-1,1]$*, then*

$$\mathrm{GS} = \{(0,0)\}\,, \ \mathrm{MM} = \mathrm{eOS} = \{(0,[-1,1])\}\,.$$

As we can see from Figure 2, the set of GS can fail to capture that of MM. Now, in view of the hardness of computing an MM, let us focus on expounding the relationship between the concepts of game stationarity and optimization stationarity. In fact, we shall address the more general problem of relating the approximate versions of these concepts as introduced in Definition 1.

To begin, let $x \in \mathcal{X}$ be arbitrary. Observe that since $f + \iota_{\mathcal{X}}$ is L-weakly convex on $\mathcal{X}$ (see Fact 1 in Appendix A) and $r > L$ by assumption, we have

$$\mathrm{dist}\left(0, \partial(f + \iota_{\mathcal{X}})\left(\mathrm{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x)\right)\right) \le r \left\|\mathrm{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x) - x\right\|;$$

see, e.g., [21, Section 2.2]. It follows from Definition 1 that for problem (P), a 0-OS is an OS. Similarly, since $F(\cdot,y) + \iota_{\mathcal{X}}(\cdot)$ is L-weakly convex on $\mathcal{X}$ for any $y \in \mathcal{Y}$, we have

$$\begin{aligned} \mathrm{dist}\left(0, \partial(F(\cdot,y) + \iota_{\mathcal{X}})\left(\mathrm{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(x)\right)\right) &\le r\left\|\mathrm{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(x) - x\right\| \\ &= \|\nabla_z d_r(y,x)\|\,, \end{aligned}$$

where the last equality is due to Lemma 3 in Appendix A. It follows from Definition 1 that for problem (P), a 0-GS is a GS. Based on the above, we may regard ϵ-OS and ϵ-GS as smoothed surrogates of OS and GS, respectively.

Remark 8 *When $F(x,\cdot)$ is concave for any $x \in \mathcal{X}$, each OS $x^\star$ has a corresponding GS whose x-coordinate is $x^\star$. To be more precise, if $x^\star$ is an OS, then we have $x_r^\star(x^\star) = \operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x^\star) = x^\star$. Consequently, for any $y \in \mathcal{Y}$, we can bound the game-stationarity measure of $(x^\star, y)$ as*

$$\begin{aligned}
&\frac{1}{r}\|\nabla_z d_r(y, x^\star)\| = \|x_r(y, x^\star) - x^\star\| \\
\le\ & \|x_r(y, x^\star) - x_r(y_+(x^\star), x^\star)\| + \|x_r(y_+(x^\star), x^\star) - x_r^\star(x^\star)\| \\
&+ \|\operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x^\star) - x^\star\| \\
\le\ & \sigma_2\|y - y_+(x^\star)\| + \kappa^{\frac{1}{2}}\|y - y_+(x^\star)\|^{\frac{1}{2}} \\
=\ & \mathcal{O}\left(\operatorname{dist}(0, -\nabla_y d_r(y, x^\star) + \partial\iota_{\mathcal{Y}}(y)) + \operatorname{dist}^{\frac{1}{2}}(0, -\nabla_y d_r(y, x^\star) + \partial\iota_{\mathcal{Y}}(y))\right),
\end{aligned}$$

where the second inequality follows from Lemma 2, Lemma 8, and the fact that $\operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x^\star) = x^\star$, and the last equality is due to [36, Lemma 4.1] and Lemma 3 in Appendix A. Now, observe that

$$\begin{aligned}
\max_{y\in\mathcal{Y}} d_r(y, x^\star) &= \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\left\{F(x,y) + \frac{r}{2}\|x - x^\star\|^2\right\} \\
&= \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{F(x,y) + \frac{r}{2}\|x - x^\star\|^2\right\} = \max_{y\in\mathcal{Y}} F(x^\star, y),
\end{aligned}$$

where the second equality is due to the convexity of $F_r(\cdot, y, x^\star)$ for any $y \in \mathcal{Y}$ (recall that $r > L$ by assumption) and concavity of $F(x,\cdot)$ for any $x \in \mathcal{X}$, and the last equality is from $\operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x^\star) = x^\star$. Since an optimal solution $y^\star(x^\star)$ of the above problem satisfies

$$0 \in -\nabla_y d_r(y^\star(x^\star), x^\star) + \partial\iota_{\mathcal{Y}}(y^\star(x^\star)) = -\nabla_y F(x^\star, y^\star(x^\star)) + \partial\iota_{\mathcal{Y}}(y^\star(x^\star))$$

and thus also $\|\nabla_z d_r(y^\star(x^\star), x^\star)\| = 0$, we conclude that $(x^\star, y^\star(x^\star))$ is a GS.

Now, let us state the main result in this section, which establishes a quantitative, algorithm-independent relationship between the ϵ-game-stationarity and ϵ-optimization-stationarity concepts for any $\epsilon \in [0,1)$. The key tool used is the alternative dual error bound (Corollary 1) presented in Section 5.3.

Theorem 4 *Suppose that $(x,y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-GS of problem (P) for some $\epsilon \in [0,1)$. Then, the point x is*

(a) (KŁ exponent $\theta \in [0,1)$): an $\mathcal{O}(\epsilon^{\min\{1,\frac{1}{2\theta}\}})$-OS of problem (P);
(b) (Concave): an $\mathcal{O}(\epsilon^{\frac{1}{2}})$-OS of problem (P).

Proof For case (a), we first compute

$$\begin{aligned}
&\|\operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(x) - x\| \\
=\ & \|x_r^\star(x) - x\| \\
\le\ & \|x_r^\star(x) - x_r(y_+(x), x)\| + \|x_r(y_+(x), x) - x_r(y, x)\| + \|x_r(y, x) - x\| \\
\le\ & \omega\|y - y_+(x)\|^\gamma + \sigma_2\|y - y_+(x)\| + \frac{1}{r}\|\nabla_z d_r(y, x)\|,
\end{aligned} \tag{7.1}$$

where the last inequality follows from Corollary 1, Lemma 2, and Lemma 3 with ω being equal to ω_2 (resp., ω_1) and γ being equal to $\frac{1}{2\theta}$ (resp., 1) when $\theta \in (0,1)$ (resp., $\theta = 0$).

Next, we estimate $\|y - y_+(x)\|$ in terms of the game-stationarity measure of (x, y). Let $y_\#(x) := \mathrm{proj}_\mathcal{Y}(y + \alpha\nabla_y F(x, y))$. By [36, Lemma 4.1], we have

$$\|y - y_\#(x)\| \le \mathrm{dist}(0, -\nabla_y F(x, y) + \partial\iota_\mathcal{Y}(y)).$$

Moreover, as $\mathcal{Y}$ is a convex set, we have

$$\begin{aligned}
&\|y_\#(x) - y_+(x)\| \\
&= \|\,\mathrm{proj}_\mathcal{Y}\,(y + \alpha\nabla_y F(x, y)) - \mathrm{proj}_\mathcal{Y}\,(y + \alpha\nabla_y F(x_r(y, x), y))\,\| \\
&\le \alpha\|\nabla_y F(x, y) - \nabla_y F(x_r(y, x), y)\| \\
&\le \alpha L\|x_r(y, x) - x\|,
\end{aligned}$$

where the first inequality holds due to the nonexpansiveness of $\mathrm{proj}_\mathcal{Y}$ and the second inequality is due to the L-Lipschitz continuity of $\nabla_y F(\cdot, y)$ on $\mathcal{X}$. Putting the above together yields

$$\begin{aligned}
\|y - y_+(x)\| &\le \|y - y_\#(x)\| + \|y_\#(x) - y_+(x)\| \\
&\le \mathrm{dist}(0, -\nabla_y F(x, y) + \partial\iota_\mathcal{Y}(y)) + \alpha L\|x_r(y, x) - x\|. \qquad (7.2)
\end{aligned}$$

Since $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-GS, we have

$$\|\nabla_z d_r(y, x)\| \le \epsilon \quad \text{and} \quad \mathrm{dist}(0, -\nabla_y F(x, y) + \partial\iota_\mathcal{Y}(y)) \le \epsilon.$$

It follows that the optimization-stationarity measure of x can be bounded as

$$\begin{aligned}
&\|\,\mathrm{prox}_{\frac{1}{r}f+\iota_\mathcal{X}}(x) - x\| \\
&\le \omega\|y - y_+(x)\|^\gamma + \sigma_2\|y - y_+(x)\| + \frac{1}{r}\|\nabla_z d_r(y, x)\| \\
&\le \omega\left(\mathrm{dist}(0, -\nabla_y F(x, y) + \partial\iota_\mathcal{Y}(y)) + \frac{\alpha L}{r}\|\nabla_z d_r(y, x)\|\right)^\gamma \\
&\quad + \sigma_2\left(\mathrm{dist}(0, -\nabla_y F(x, y) + \partial\iota_\mathcal{Y}(y)) + \frac{\alpha L}{r}\|\nabla_z d_r(y, x)\|\right) + \frac{1}{r}\|\nabla_z d_r(y, x)\| \\
&\le \omega\left(1 + \frac{\alpha L}{r}\right)^\gamma \epsilon^\gamma + \sigma_2\left(1 + \frac{\alpha L}{r}\right)\epsilon + \frac{\epsilon}{r} = \mathcal{O}(\epsilon^{\min\{1, \frac{1}{2\theta}\}}),
\end{aligned}$$

where the first inequality follows from (7.1) and the second from (7.2). This implies that x is an $\mathcal{O}(\epsilon^{\min\{1, \frac{1}{2\theta}\}})$-OS of problem (P).

Now, for case (b), we can apply Lemma 8 to derive the inequality (7.1) with ω and γ being replaced by $\sqrt{\kappa}$ and $\frac{1}{2}$, respectively. The remainder of the proof follows the same argument as that for case (a). □

Remark 9 *Theorem 4 expands on the findings of Propositions 4.11 and 4.12 in [40] by covering a broader range of scenarios. Specifically, it applies to settings where* (i) *$\mathcal{X}$ is not necessarily the entire space $\mathbb{R}^n$;* (ii) *the primal function is nonsmooth or lacks gradient Lipschitz continuity;* (iii) *the dual problem may not be concave and only satisfies the KŁ property with exponent θ. Furthermore, our results apply to the problems studied in [66, 63].*

8 Closing Remarks

In this paper, we proposed smoothed PLDA for solving a class of nonsmooth composite nonconvex-nonconcave problems. When the dual function is concave, we showed that our algorithm can find both an ϵ-GS and an ϵ-OS of the problem in $\mathcal{O}(\epsilon^{-4})$ iterations, which is a first step towards matching the complexity results for smooth minimax problems. Moreover, when the dual problem possesses the KŁ property with exponent $\theta \in [0,1)$, we showed that our algorithm has an iteration complexity of $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$. As it turns out, this complexity is determined by the slower of the primal and dual updates in smoothed PLDA and reveals an interesting phase transition phenomenon: When $\theta \in [0, \frac{1}{2}]$, the primal update, which involves solving a strongly convex problem, dominates. As such, we cannot break the optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$. On the other hand, when $\theta \in (\frac{1}{2}, 1)$, the dual update dominates and explicitly depends on θ, resulting in an iteration complexity of $\mathcal{O}(\epsilon^{-4\theta})$. The insights gained from our analysis suggest a new algorithm design principle—namely, primal-dual balancing—which holds promise for the design of more efficient algorithms in the context of minimax optimization.

Our work suggests several directions for further study. We mention some of them here.

(a) Extend our algorithm and its analysis to the stochastic setting to benefit modern machine learning tasks.
(b) Investigate the lower complexity bounds of first-order methods for nonconvex-KŁ problems and their dependence on the KŁ exponent θ. We conjecture that our proposed smoothed PLDA already has the optimal complexity, at least in terms of the dependence on θ.
(c) Identify additional structured nonconvex-nonconcave problems that satisfy the KŁ property and characterize their KŁ exponents.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 214–223. PMLR, 2017.
2. Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

3. Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.
4. Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.
5. Babak Barazandeh and Meisam Razaviyayn. Solving non-convex non-differentiable min-max games using proximal gradient method. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3162–3166. IEEE, 2020.
6. Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.
7. Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.
8. Jose Blanchet, Peter W. Glynn, Jun Yan, and Zhengqing Zhou. Multivariate distributionally robust convex regression under absolute error loss. In *Advances in Neural Information Processing Systems 32*, 2019.
9. Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.
10. Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities and applications. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
11. Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.
12. Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM Journal on Optimization*, 33(3):1884–1913, 2023.
13. Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
14. Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.
15. Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.
16. Ziyi Chen, Yi Zhou, Tengyu Xu, and Yingbin Liang. Proximal gradient descent-ascent: Variable convergence under KŁ geometry. In *Proceedings of the 9th of International Conference on Learning Representations (ICLR 2021)*, 2021.
17. Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. In *Advances in Neural Information Processing Systems 33*, 2020.

18. Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. Generalised Lipschitz regularisation equals distributional robustness. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pages 2178–2188. PMLR, 2021.
19. Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics and the Mathematical Optimization Society, Philadelphia, 2022.
20. Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.
21. Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.
22. Dmitriy Drusvyatskiy, Alexander D. Ioffe, and Adrian S. Lewis. Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria. *Mathematical Programming*, 185(1):357–383, 2021.
23. Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
24. Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.
25. Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.
26. Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.
27. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
28. Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, pages 174–176, 2003.
29. Yaohua Hu, Chong Li, Jinhua Wang, Xiaoqi Yang, and Linglingzhi Zhu. Linearized proximal algorithms with adaptive stepsizes for convex composite optimization with applications. *Applied Mathematics & Optimization*, 87(3):52, 2023.
30. Yaohua Hu, Chong Li, and Xiaoqi Yang. On convergence rates of linearized proximal algorithms for convex composite optimization with applications. *SIAM Journal on Optimization*, 26(2):1207–1235, 2016.
31. Zhaolin Hu and L. Jeff Hong. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*,

1(2):9, 2013.
32. Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. In *Advances in Neural Information Processing Systems 34*, 2021.
33. Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the 37th of International Conference on Machine Learning (ICML 2020)*, pages 4880–4889. PMLR, 2020.
34. Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems 33*, 2020.
35. Adrian S. Lewis and Stephen J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158:501–546, 2016.
36. Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018.
37. Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *Advances in Neural Information Processing Systems 34*, 2021.
38. Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020.
39. Tianyi Lin, Chi Jin, and Michael Jordan. Near-optimal algorithms for minimax optimization. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT 2020)*, pages 2738–2779. PMLR, 2020.
40. Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 6083–6093. PMLR, 2020.
41. Tianyi Lin, Chi Jin, and Michael Jordan. Two-timescale gradient descent ascent algorithms for nonconvex minimax optimization. *arXiv preprint arXiv:2408.11974*, 2024.
42. 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.
43. Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993.
44. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
45. Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on*

Machine Learning (ICML 2019), pages 4615–4625. PMLR, 2019.
46. Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.
47. Yu Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimization Methods and Software*, 22(3):469–483, 2007.
48. Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, 2019.
49. Dmitrii M. Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.
50. Jong-Shi Pang and Meisam Razaviyayn. A unified distributed algorithm for non-cooperative games. In Shuguang Cui, Alfred O. Hero, III, Zhi-Quan Luo, and José M.F. Moura, editors, *Big Data over Networks*, chapter 4. Cambridge University Press, 2016.
51. B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.
52. Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.
53. Hamed Rahimian and Sanjay Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:article no. 4, 2022.
54. Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.
55. R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.
56. R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media, second edition, 2009.
57. Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.
58. Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: How and why. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, pages 793–801. PMLR, 2016.
59. Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, 2018.

60. Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.
61. Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32*, 2019.
62. 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*, 201:635–706, 2023.
63. Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022)*, pages 5485–5517. PMLR, 2022.
64. Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems 33*, 2020.
65. Jiawei Zhang and Zhi-Quan Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020.
66. Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave minmax problems. In *Advances in Neural Information Processing Systems 33*, 2020.
67. Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems 31*, 2018.
68. Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021)*, pages 482–492. PMLR, 2021.
69. Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

A Useful Technical Lemmas

To begin, we introduce the concept of weakly convex functions, which plays an important role in our subsequent analysis.

Definition 3 (Weak convexity) *A function $\ell : \mathbb{R}^n \to \mathbb{R}$ is said to be ρ-weakly convex on a set $\mathcal{X} \subseteq \mathbb{R}^n$ for some constant $\rho \geq 0$ if for any $x, y \in \mathcal{X}$ and $\tau \in [0,1]$, we have*

$$\ell(\tau x + (1-\tau)y) \leq \tau\ell(x) + (1-\tau)\ell(y) + \frac{\rho\tau(1-\tau)}{2}\|x-y\|^2.$$

The above definition is equivalent to the convexity of the function $\ell(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ on $\mathcal{X}$.

Definition 4 (Proximal mapping) *Let $\ell : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function. The proximal mapping of ℓ with parameter $\mu > 0$ at the point $x \in \mathbb{R}^n$ is defined by*

$$\operatorname{prox}_{\mu\ell}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \ell(y) + \frac{1}{2\mu}\|y-x\|^2 \right\}.$$

By our assumptions on problem (P) (recall that $L = L_h L_c$ and $r > L$) and [24, Lemma 3.2 and Lemma 4.2], we have the following useful results:

Fact 1 *The functions $F(\cdot, y)$ for any $y \in \mathcal{Y}$ and f are L-weakly convex on $\mathcal{X}$.*

Fact 2 *Let $y \in \mathbb{R}^d$ and $r > 0$ be given. For all $x, \bar{x} \in \mathcal{X}$, we have*

$$-\frac{r^{-1}+L}{2}\|x-\bar{x}\|^2 \leq F(x,y) - F_{\bar{x},r}(x,y) \leq \frac{-r^{-1}+L}{2}\|x-\bar{x}\|^2.$$

Using Definition 3 and Fact 1, we can deduce that $F_r(\cdot, y, z)$ is $(r-L)$-strongly convex for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$.

Lemma 2 and Lemma 3 are also essential to our analysis. The proofs of these lemmas are similar to those of Lemma B.2 and Lemma B.3 in [66], respectively. For completeness, we present the proofs here.

Lemma 2 *For any $y, y' \in \mathcal{Y}$ and $z, z' \in \mathbb{R}^n$, we have*

$$\|x_r(y,z) - x_r(y,z')\| \leq \sigma_1\|z-z'\|, \tag{A.1}$$

$$\|x_r^\star(z) - x_r^\star(z')\| \leq \sigma_1\|z-z'\|, \tag{A.2}$$

$$\|x_r(y,z) - x_r(y',z)\| \leq \sigma_2\|y-y'\|, \tag{A.3}$$

where $\sigma_1 := \frac{r}{r-L}$ and $\sigma_2 := \frac{2r-L}{r-L}$.

Proof From the definition of F_r, we know that for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z, z' \in \mathbb{R}^n$,

$$F_r(x,y,z')-F_r(x,y,z) = \frac{r}{2}(\|x-z'\|^2-\|x-z\|^2) = \frac{r}{2}(\|z'\|^2-2(z'-z)^\top x-\|z\|^2). \tag{A.4}$$

Fact 1 implies that $F(\cdot, y)$ is L-weakly convex for any $y \in \mathcal{Y}$. Therefore, for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathbb{R}^n$, we have

$$F_r(x,y,z) - F_r(x_r(y,z),y,z) \geq \frac{r-L}{2}\|x - x_r(y,z)\|^2. \tag{A.5}$$

Combining (A.4) and (A.5) yields

$$\begin{aligned}
&F_r(x_r(y,z),y,z') - F_r(x_r(y,z'),y,z') \\
=& F_r(x_r(y,z),y,z') - F_r(x_r(y,z),y,z) + F_r(x_r(y,z),y,z) - F_r(x_r(y,z'),y,z) \\
&- (F_r(x_r(y,z'),y,z') - F_r(x_r(y,z'),y,z)) \\
\leq& \frac{r}{2}(\|z'\|^2 - 2(z'-z)^\top x_r(y,z) - \|z\|^2) - \frac{r-L}{2}\|x_r(y,z) - x_r(y,z')\|^2 \\
&- \frac{r}{2}(\|z'\|^2 - 2(z'-z)^\top x_r(y,z') - \|z\|^2) \\
\leq& r(z'-z)^\top (x_r(y,z') - x_r(y,z)) - \frac{r-L}{2}\|x_r(y,z) - x_r(y,z')\|^2.
\end{aligned} \tag{A.6}$$

On the other hand, (A.5) implies that

$$F_r(x_r(y,z),y,z') - F_r(x_r(y,z'),y,z') \geq \frac{r-L}{2}\|x_r(y,z) - x_r(y,z')\|^2.$$

Combining this inequality with (A.6), we get

$$(r-L)\|x_r(y,z) - x_r(y,z')\|^2 \leq r(z'-z)^\top (x_r(y,z') - x_r(y,z)).$$

Using the Cauchy-Schwarz inequality, we obtain

$$\|x_r(y,z) - x_r(y,z')\| \leq \frac{r}{r-L}\|z - z'\|.$$

Therefore, we conclude that (A.1) holds with $\sigma_1 = \frac{r}{r-L}$. Since $f = \max_{y \in \mathcal{Y}} F(\cdot, y)$ is also L-weakly convex by Fact 1, we can prove that (A.2) holds by using a similar argument as above.

We now proceed to prove (A.3). Using (A.5), we have

$$F_r(x_r(y',z),y,z) - F_r(x_r(y,z),y,z) \geq \frac{r-L}{2}\|x_r(y,z) - x_r(y',z)\|^2, \tag{A.7}$$

$$F_r(x_r(y,z),y',z) - F_r(x_r(y',z),y',z) \geq \frac{r-L}{2}\|x_r(y,z) - x_r(y',z)\|^2. \tag{A.8}$$

Moreover, by the L-Lipschitz continuity of $\nabla_y F_r(x,\cdot,z)$ for any $x \in \mathcal{X}$ and $z \in \mathbb{R}^n$, we have

$$\begin{aligned}
F_r(x_r(y,z),y',z) - F_r(x_r(y,z),y,z) \leq& \langle \nabla_y F_r(x_r(y,z),y,z), y'-y\rangle \\
&+ \frac{L}{2}\|y - y'\|^2
\end{aligned} \tag{A.9}$$

and

$$F_r(x_r(y',z),y,z)-F_r(x_r(y',z),y',z)\le\langle\nabla_y F_r(x_r(y',z),y',z),y-y'\rangle + \frac{L}{2}\|y-y'\|^2. \tag{A.10}$$

Incorporating (A.7)–(A.10), we obtain

$$\begin{aligned}&(r-L)\|x_r(y,z)-x_r(y',z)\|^2\\ \le\ &\langle\nabla_y F_r(x_r(y,z),y,z)-\nabla_y F_r(x_r(y',z),y',z),y'-y\rangle+L\|y-y'\|^2.\end{aligned}$$

Since $\nabla_y F_r(\cdot,\cdot,z)=\nabla_y F(\cdot,\cdot)$ is L-Lipschitz for any $z\in\mathbb{R}^n$, we have

$$\begin{aligned}&\|\nabla_y F_r(x_r(y,z),y,z)-\nabla_y F_r(x_r(y',z),y',z)\|\\ \le\ &\|\nabla_y F_r(x_r(y,z),y,z)-\nabla_y F_r(x_r(y,z),y',z)\|\\ &+\|\nabla_y F_r(x_r(y,z),y',z)-\nabla_y F_r(x_r(y',z),y',z)\|\\ \le\ &L(\|y-y'\|+\|x_r(y',z)-x_r(y,z)\|).\end{aligned}$$

It follows that

$$(r-L)\|x_r(y,z)-x_r(y',z)\|^2\le L\|x_r(y',z)-x_r(y,z)\|\cdot\|y-y'\|+2L\|y-y'\|^2.$$

Let $\zeta:=\frac{\|x_r(y,z)-x_r(y',z)\|}{\|y-y'\|}$. Then, the above inequality gives

$$\begin{aligned}\zeta^2\ &\le\frac{L}{r-L}\zeta+\frac{2L}{(r-L)}\le\frac{1}{2}\zeta^2+\frac{L^2}{2(r-L)^2}+\frac{2L}{(r-L)}\\ &\le\frac{1}{2}\zeta^2+\frac{L^2+4L(r-L)}{2(r-L)^2}\le\frac{1}{2}\zeta^2+\frac{(L+2(r-L))^2}{2(r-L)^2},\end{aligned}$$

where the second inequality holds because $ab\le\frac{1}{2}(a^2+b^2)$ for any $a,b\in\mathbb{R}$. Thus, we get

$$\|x_r(y,z)-x_r(y',z)\|\le\frac{2r-L}{r-L}\|y-y'\|,$$

which shows that (A.3) holds with $\sigma_2=\frac{2r-L}{r-L}$. The proof is complete. □

Lemma 3 *The dual potential value function d_r is differentiable on $\mathcal{Y}\times\mathbb{R}^n$ with*

$$\begin{aligned}&\nabla_y d_r(y,z)=\nabla_y F(x_r(y,z),y),\\ &\nabla_z d_r(y,z)=\nabla_z F_r(x_r(y,z),y,z)=r(z-x_r(y,z))\end{aligned}$$

for any $y\in\mathcal{Y}$ and $z\in\mathbb{R}^n$. Moreover, for any $y\in\mathcal{Y}$ and $z\in\mathbb{R}^n$, the gradients $\nabla_y d_r(\cdot,z)$ and $\nabla_z d_r(y,\cdot)$ are Lipschitz continuous, i.e.,

$$\begin{aligned}&\|\nabla_y d_r(y',z)-\nabla_y d_r(y'',z)\|\le L_{d_r}\|y'-y''\|\quad\text{for all } y',y''\in\mathcal{Y},\\ &\|\nabla_z d_r(y,z')-\nabla_z d_r(y,z'')\|\le L'_{d_r}\|z'-z''\|\quad\text{for all } z',z''\in\mathbb{R}^n,\end{aligned}$$

where $L_{d_r}:=(\sigma_2+1)L$ and $L'_{d_r}:=(\sigma_1+1)r$.

Proof Since $F_r(\cdot, y, z)$ is strongly convex, $F_r(x, \cdot, z)$ is weakly concave, and $F_r(x, y, \cdot)$ is strongly convex for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathbb{R}^n$, the function

$$d_r(\cdot,\cdot) = \min_{x\in\mathcal{X}} F_r(x,\cdot,\cdot)$$

is differentiable on $\mathcal{Y} \times \mathbb{R}^n$ due to [56, Theorem 10.31]. In particular, for any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$, one has

$$\nabla_y d_r(y, z) = \nabla_y F_r(x_r(y, z), y, z) = \nabla_y F(x_r(y, z), y).$$

Using [24, Lemma 4.3], we have

$$\nabla_z d_r(y, z) = \nabla_z F_r(x_r(y, z), y, z) = r(z - \mathrm{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(z)) = r(z - x_r(y, z)).$$

It follows that for any $y', y'' \in \mathcal{Y}$,

$$\begin{aligned}
&\|\nabla_y d_r(y', z) - \nabla_y d_r(y'', z)\| \\
&= \|\nabla_y F_r(x_r(y', z), y', z) - \nabla_y F_r(x_r(y'', z), y'', z)\| \\
&\le \|\nabla_y F_r(x_r(y', z), y', z) - \nabla_y F_r(x_r(y', z), y'', z)\| \\
&\quad + \|\nabla_y F_r(x_r(y', z), y'', z) - \nabla_y F_r(x_r(y'', z), y'', z)\| \\
&\le L\|y' - y''\| + L\|x_r(y', z) - x_r(y'', z)\| \\
&\le L\|y' - y''\| + L\sigma_2\|y' - y''\| \le L_{d_r}\|y' - y''\|,
\end{aligned}$$

where the third inequality is due to (A.3). Moreover, we have

$$\begin{aligned}
\|\nabla_z d_r(y, z') - \nabla_z d_r(y, z'')\| &= r\|z' - x_r(y, z') - (z'' - x_r(y, z''))\| \\
&\le r\|z' - z''\| + r\|x_r(y, z') - x_r(y, z'')\| \\
&\le r\|z' - z''\| + r\sigma_1\|z' - z''\| \le L'_{d_r}\|z' - z''\|,
\end{aligned}$$

where the second inequality is due to (A.1). The proof is complete. □

Lemma 4 *For any $k \ge 0$, we have*

$$\|y^{k+1} - y_+^k(z^{k+1})\| \le \eta\|x^k - x^{k+1}\| + \sigma_1\alpha L\|z^{k+1} - z^k\|,$$

where $\eta := \alpha L\zeta$.

Proof We compute

$$\begin{aligned}
&\|y^{k+1} - y_+^k(z^{k+1})\| \\
&= \|\,\mathrm{proj}_{\mathcal{Y}}(y^k + \alpha\nabla_y F(x^{k+1}, y^k)) - \mathrm{proj}_{\mathcal{Y}}(y^k + \alpha\nabla_y F(x_r(y^k, z^{k+1}), y^k))\| \\
&\le \|y^k + \alpha\nabla_y F(x^{k+1}, y^k) - (y^k + \alpha\nabla_y F(x_r(y^k, z^{k+1}), y^k))\| \\
&\le \alpha L(\|x^{k+1} - x_r(y^k, z^k)\| + \|x_r(y^k, z^k) - x_r(y^k, z^{k+1})\|) \\
&\le \eta\|x^k - x^{k+1}\| + \sigma_1\alpha L\|z^{k+1} - z^k\|,
\end{aligned}$$

where the first inequality is due to the nonexpansiveness of the projection operator, the second is due to the L-Lipschitz continuity of $\nabla_y F(\cdot,\cdot)$, and the third follows from Proposition 2 and (A.1). The proof is complete. □

B Sufficient Decrease Property of Lyapunov Function

Lemma 5 (Primal descent) *For any $k \geq 0$, we have*

$$
\begin{aligned}
&F_r(x^k, y^k, z^k) - F_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \frac{2\lambda^{-1} + r - L}{2}\|x^k - x^{k+1}\|^2 + \langle \nabla_y F_r(x^{k+1}, y^k, z^k), y^k - y^{k+1}\rangle - \frac{L}{2}\|y^k - y^{k+1}\|^2 \\
&+ \frac{(2-\beta)r}{2\beta}\|z^k - z^{k+1}\|^2.
\end{aligned}
$$

Proof One can infer from the definition that $F_{x^k,\lambda}(\cdot, y^k)$ is λ^{-1}-strongly convex. Therefore, we have

$$
\begin{aligned}
F_r(x^k, y^k, z^k) &= F(x^k, y^k) + \frac{r}{2}\|x^k - z^k\|^2 = F_{x^k,\lambda}(x^k, y^k) + \frac{r}{2}\|x^k - z^k\|^2 \\
&\geq F_{x^k,\lambda}(x^{k+1}, y^k) + \frac{r}{2}\|x^{k+1} - z^k\|^2 + \frac{\lambda^{-1} + r}{2}\|x^k - x^{k+1}\|^2.
\end{aligned}
$$

Moreover, Fact 2 implies that

$$
F_{x^k,\lambda}(x^{k+1}, y^k) \geq F(x^{k+1}, y^k) + \frac{\lambda^{-1} - L}{2}\|x^{k+1} - x^k\|^2.
$$

It follows that

$$
\begin{aligned}
F_r(x^k, y^k, z^k) &\geq F(x^{k+1}, y^k) + \frac{r}{2}\|x^{k+1} - z^k\|^2 + \frac{2\lambda^{-1} + r - L}{2}\|x^k - x^{k+1}\|^2 \\
&= F_r(x^{k+1}, y^k, z^k) + \frac{2\lambda^{-1} + r - L}{2}\|x^k - x^{k+1}\|^2. \qquad \text{(B.1)}
\end{aligned}
$$

Next, as $\nabla_y F_r(x, \cdot, z)$ is L-Lipschitz continuous for any $x \in \mathcal{X}$ and $z \in \mathbb{R}^n$, we have

$$
\begin{aligned}
F_r(x^{k+1}, y^k, z^k) - F_r(x^{k+1}, y^{k+1}, z^k) \geq &\langle \nabla_y F_r(x^{k+1}, y^k, z^k), y^k - y^{k+1}\rangle \\
&- \frac{L}{2}\|y^k - y^{k+1}\|^2. \qquad \text{(B.2)}
\end{aligned}
$$

At last, based on the update $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$, we get

$$
F_r(x^{k+1}, y^{k+1}, z^k) - F_r(x^{k+1}, y^{k+1}, z^{k+1}) = \frac{(2-\beta)r}{2\beta}\|z^k - z^{k+1}\|^2. \qquad \text{(B.3)}
$$

By summing up (B.1), (B.2), and (B.3), we obtain the desired result. □

Lemma 6 (Dual ascent) *For any $k \geq 0$, we have*

$$
\begin{aligned}
&d_r(y^{k+1}, z^{k+1}) - d_r(y^k, z^k) \\
\geq\ & \langle \nabla_y F_r(x_r(y^k, z^k), y^k, z^k), y^{k+1} - y^k\rangle - \frac{L_{d_r}}{2}\|y^k - y^{k+1}\|^2 \\
&+ \frac{r}{2}\left\langle z^{k+1} - z^k, z^{k+1} + z^k - 2x_r(y^{k+1}, z^{k+1})\right\rangle.
\end{aligned}
$$

Proof Based on Lemma 3, $\nabla_y d_r(\cdot, z)$ is L_{d_r}-Lipschitz continuous for any $z \in \mathbb{R}^n$. Thus, we have

$$d_r(y^{k+1}, z^k) - d_r(y^k, z^k) \geq \langle \nabla_y d_r(y^k, z^k), y^{k+1} - y^k \rangle - \frac{L_{d_r}}{2} \|y^k - y^{k+1}\|^2$$
$$= \langle \nabla_y F_r(x_r(y^k, z^k), y^k, z^k), y^{k+1} - y^k \rangle - \frac{L_{d_r}}{2} \|y^k - y^{k+1}\|^2.$$

In addition, one has

$$\begin{aligned} & d_r(y^{k+1}, z^{k+1}) - d_r(y^{k+1}, z^k) \\ = & F_r(x_r(y^{k+1}, z^{k+1}), y^{k+1}, z^{k+1}) - F_r(x_r(y^{k+1}, z^k), y^{k+1}, z^k) \\ \geq & F_r(x_r(y^{k+1}, z^{k+1}), y^{k+1}, z^{k+1}) - F_r(x_r(y^{k+1}, z^{k+1}), y^{k+1}, z^k) \\ = & \frac{r}{2}\|x_r(y^{k+1}, z^{k+1}) - z^{k+1}\|^2 - \frac{r}{2}\|x_r(y^{k+1}, z^{k+1}) - z^k\|^2 \\ = & \frac{r}{2} \left\langle z^{k+1} - z^k, z^{k+1} + z^k - 2x_r(y^{k+1}, z^{k+1}) \right\rangle. \end{aligned}$$

Finally, by combining the above inequalities, the proof is complete. □

Lemma 7 (Proximal descent (smoothness)) *For any $k \geq 0$, we have*

$$p_r(z^k) - p_r(z^{k+1}) \geq \frac{r}{2} \left\langle z^{k+1} - z^k, 2x_r(y(z^{k+1}), z^k) - z^k - z^{k+1} \right\rangle,$$

where $y(z^{k+1}) \in Y(z^{k+1})$.

Proof Recall that $p_r(z) = \max_{y \in \mathcal{Y}} d_r(y, z)$. Due to the definition of $y(z^{k+1})$, we have

$$\begin{aligned} & p_r(z^{k+1}) - p_r(z^k) \\ \leq & d_r(y(z^{k+1}), z^{k+1}) - d_r(y(z^{k+1}), z^k) \\ \leq & F_r(x_r(y(z^{k+1}), z^k), y(z^{k+1}), z^{k+1}) - F_r(x_r(y(z^{k+1}), z^k), y(z^{k+1}), z^k) \\ = & \frac{r}{2} \left\langle z^{k+1} - z^k, z^{k+1} + z^k - 2x_r(y(z^{k+1}), z^k) \right\rangle, \end{aligned}$$

where the second inequality follows from the fact that

$$F_r(x', y, z) \geq \min_{x \in \mathcal{X}} F_r(x, y, z) = d_r(y, z)$$

for any $x' \in \mathcal{X}$. The proof is complete. □

Proof of Proposition 1 From Lemmas 5, 6, and 7, we have

$$\begin{aligned}
&\Phi_r(x^k,y^k,z^k)-\Phi_r(x^{k+1},y^{k+1},z^{k+1})\\
=&F_r(x^k,y^k,z^k)-F_r(x^{k+1},y^{k+1},z^{k+1})+2(d_r(y^{k+1},z^{k+1})-d_r(y^k,z^k))\\
&+2(p_r(z^k)-p_r(z^{k+1}))\\
\geq&\frac{2\lambda^{-1}+r-L}{2}\|x^k-x^{k+1}\|^2+\frac{(2-\beta)r}{2\beta}\|z^k-z^{k+1}\|^2-\left(L_{d_r}+\frac{L}{2}\right)\|y^k-y^{k+1}\|^2\\
&+\underbrace{\langle\nabla_y F_r(x^{k+1},y^k,z^k),y^k-y^{k+1}\rangle+2\langle\nabla_y F_r(x_r(y^k,z^k),y^k,z^k),y^{k+1}-y^k\rangle}_{①}\\
&+\underbrace{2r\left\langle z^{k+1}-z^k,x_r(y(z^{k+1}),z^k)-x_r(y^{k+1},z^{k+1})\right\rangle}_{②}.
\end{aligned}$$

First, we simplify the term ①. We know that

$$\begin{aligned}
①=&\langle\nabla_y F_r(x^{k+1},y^k,z^k),y^k-y^{k+1}\rangle+2\langle\nabla_y F_r(x_r(y^k,z^k),y^k,z^k),y^{k+1}-y^k\rangle\\
=&\langle\nabla_y F_r(x^{k+1},y^k,z^k),y^{k+1}-y^k\rangle\\
&+2\langle\nabla_y F_r(x_r(y^k,z^k),y^k,z^k)-\nabla_y F_r(x^{k+1},y^k,z^k),y^{k+1}-y^k\rangle.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\langle\nabla_y F_r(x^{k+1},y^k,z^k),y^{k+1}-y^k\rangle\\
=&\langle\nabla_y F_r(x^{k+1},y^k,z^k)+\frac{1}{\alpha}(y^k-y^{k+1}),y^{k+1}-y^k\rangle+\frac{1}{\alpha}\|y^k-y^{k+1}\|^2\\
=&\frac{1}{\alpha}\langle y^k+\alpha\nabla_y F_r(x^{k+1},y^k,z^k)-y^{k+1},y^{k+1}-y^k\rangle+\frac{1}{\alpha}\|y^k-y^{k+1}\|^2\\
\geq&\frac{1}{\alpha}\|y^k-y^{k+1}\|^2,
\end{aligned}$$

where the last inequality follows from the property of the projection operator and the dual update $y^{k+1}=\mathrm{proj}_{\mathcal{Y}}(y^k+\alpha\nabla_y F_r(x^{k+1},y^k,z^k))$ (recall that $\nabla_y F_r(x,y,z)=\nabla_y F(x,y)$ for any $x\in\mathcal{X}$, $y\in\mathcal{Y}$, and $z\in\mathbb{R}^n$). For the remaining terms, we have

$$\begin{aligned}
&2\langle\nabla_y F_r(x_r(y^k,z^k),y^k,z^k)-\nabla_y F_r(x^{k+1},y^k,z^k),y^{k+1}-y^k\rangle\\
\geq&-2\|\nabla_y F_r(x_r(y^k,z^k),y^k,z^k)-\nabla_y F_r(x^{k+1},y^k,z^k)\|\cdot\|y^{k+1}-y^k\|\\
\geq&-2L\|x^{k+1}-x_r(y^k,z^k)\|\cdot\|y^k-y^{k+1}\|\\
\geq&-L\zeta^2\|y^k-y^{k+1}\|^2-L\zeta^{-2}\|x^{k+1}-x_r(y^k,z^k)\|^2\\
\geq&-L\zeta^2\|y^k-y^{k+1}\|^2-L\|x^{k+1}-x^k\|^2,
\end{aligned}$$

where the third inequality holds because $2|a||b|\leq\tau a^2+\frac{1}{\tau}b^2$ for any $a,b\in\mathbb{R}$ and $\tau>0$, and the last inequality follows from Proposition 2. Putting the above together, we obtain

$$①\geq\left(\frac{1}{\alpha}-L\zeta^2\right)\|y^k-y^{k+1}\|^2-L\|x^{k+1}-x^k\|^2. \tag{B.4}$$

Next, we bound the term ② by

$$
\begin{aligned}
② &= 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^k)-x_r(y^{k+1},z^{k+1})\right\rangle\\
&= 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^k)-x_r(y(z^{k+1}),z^{k+1})\right\rangle\\
&\quad+2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\right\rangle\\
&\geq -2r\sigma_1\|z^{k+1}-z^k\|^2+2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\right\rangle\\
&\geq -2r\sigma_1\|z^{k+1}-z^k\|^2-\frac{r}{7\beta}\|z^{k+1}-z^k\|^2\\
&\quad-7r\beta\|x_r(y(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\|^2,
\end{aligned}\tag{B.5}
$$

where the first inequality is due to (A.1) and the Cauchy-Schwarz inequality, and the second inequality again follows from the fact that $2|a||b|\leq \tau a^2+\frac{1}{\tau}b^2$ for any $a,b\in\mathbb{R}$ and $\tau>0$. Now, the inequalities (B.4) and (B.5) imply that

$$
\begin{aligned}
&\Phi_r(x^k,y^k,z^k)-\Phi_r(x^{k+1},y^{k+1},z^{k+1})\\
&\geq \frac{2\lambda^{-1}+r-3L}{2}\|x^k-x^{k+1}\|^2+\left(\frac{1}{\alpha}-L_{d_r}-\frac{L}{2}-L\zeta^2\right)\|y^k-y^{k+1}\|^2\\
&\quad+\left(\frac{(2-\beta)r}{2\beta}-2r\sigma_1-\frac{r}{7\beta}\right)\|z^k-z^{k+1}\|^2\\
&\quad-7r\beta\|x_r(y(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\|^2.
\end{aligned}\tag{B.6}
$$

By Lemma 4 and the fact that $\|u+v\|^2\leq 2(\|u\|^2+\|v\|^2)$ for any $u,v\in\mathbb{R}^d$, we have

$$
\begin{aligned}
\|y^{k+1}-y^k\|^2 &= \|y^{k+1}-y_+^k(z^{k+1})+y_+^k(z^{k+1})-y^k\|^2\\
&\geq \frac{1}{2}\|y^k-y_+^k(z^{k+1})\|^2-\|y^{k+1}-y_+^k(z^{k+1})\|^2\\
&\geq \frac{1}{2}\|y^k-y_+^k(z^{k+1})\|^2-2\eta^2\|x^{k+1}-x^k\|^2-2\sigma_1^2\alpha^2L^2\|z^{k+1}-z^k\|^2.
\end{aligned}\tag{B.7}
$$

Similarly, by Lemma 2 and Lemma 4, we have

$$
\begin{aligned}
&\|x_r(y(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\|^2\\
&\leq 2\|x_r(y(z^{k+1}),z^{k+1})-x_r(y_+^k(z^{k+1}),z^{k+1})\|^2\\
&\quad+2\|x_r(y_+^k(z^{k+1}),z^{k+1})-x_r(y^{k+1},z^{k+1})\|^2\\
&\leq 2\|x_r(y(z^{k+1}),z^{k+1})-x_r(y_+^k(z^{k+1}),z^{k+1})\|^2+2\sigma_2^2\|y^{k+1}-y_+^k(z^{k+1})\|^2\\
&\leq 2\|x_r(y(z^{k+1}),z^{k+1})-x_r(y_+^k(z^{k+1}),z^{k+1})\|^2\\
&\quad+4\sigma_2^2\eta^2\|x^{k+1}-x^k\|^2+4\sigma_2^2\sigma_1^2\alpha^2L^2\|z^{k+1}-z^k\|^2.
\end{aligned}\tag{B.8}
$$

Substituting (B.7) and (B.8) into (B.6) yields

$$\Phi_r(x^k,y^k,z^k)-\Phi_r(x^{k+1},y^{k+1},z^{k+1})$$

$$\geq \left(\frac{2\lambda^{-1}+r-3L}{2} - 28r\beta\sigma_2^2\eta^2\right)\|x^k - x^{k+1}\|^2 + \left(\frac{1}{\alpha} - L_{d_r} - \frac{L}{2} - L\zeta^2\right)\cdot$$
$$\left(\frac{1}{2}\|y^k - y_+^k(z^{k+1})\|^2 - 2\eta^2\|x^{k+1}-x^k\|^2 - 2\sigma_1^2\alpha^2L^2\|z^{k+1}-z^k\|^2\right)$$
$$+ \left(\frac{(2-\beta)r}{2\beta} - 2r\sigma_1 - \frac{r}{7\beta} - 28r\beta\sigma_1^2\sigma_2^2\alpha^2L^2\right)\|z^k - z^{k+1}\|^2$$
$$- 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|^2.$$

Suppose that $r \geq 3L$, which implies that $L_{d_r} + \frac{L}{2} \leq 5L$. We observe the following:

- As $\alpha \leq \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\}$, we have $\frac{1}{\alpha} - L_{d_r} - \frac{L}{2} \geq \frac{1}{2\alpha}$ and $\frac{1}{2\alpha} - L\zeta^2 \geq \frac{1}{4\alpha}$.
- As $\beta \leq \frac{1}{28}$ and $\sigma_1 \leq \frac{3}{2}$, $\sigma_2 \leq \frac{5}{2}$, we have

$$\begin{aligned}
&\frac{(2-\beta)r}{2\beta} - 2r\sigma_1 - \frac{r}{7\beta} - 28r\beta\sigma_1^2\sigma_2^2\alpha^2L^2 - \frac{1}{2\alpha}\sigma_1^2\alpha^2L^2 \\
&\geq \frac{6r}{7\beta} - \frac{7r}{2} - 28r\beta\left(\frac{3}{2}\right)^2\left(\frac{5}{2}\right)^2\left(\frac{1}{10L}\right)^2L^2 - \frac{r}{24} \\
&\geq \frac{r}{\beta}\left(\frac{6}{7} - 4\beta - 14\beta^2\right) \geq \frac{4r}{7\beta}.
\end{aligned}$$

- As $\lambda^{-1} \geq L$ and recalling that $\eta = \alpha L\zeta$, we have

$$\alpha \leq \frac{1}{4L\zeta^2} = \frac{\lambda^{-1}}{4L^2\zeta^2}\frac{L}{\lambda^{-1}} \leq \frac{\lambda^{-1}}{4L^2\zeta^2} \quad \text{and} \quad \frac{\eta^2}{2\alpha} = \frac{\alpha L^2\zeta^2}{2} \leq \frac{1}{8\lambda}.$$

Moreover, since $\beta \leq \frac{(r-L)^2}{14\alpha r(2r-L)^2} = \frac{1}{14\alpha r\sigma_2^2}$, we have

$$28r\beta\sigma_2^2\eta^2 \leq \frac{2\eta^2}{\alpha} \leq \frac{1}{2\lambda} \quad \text{and} \quad \frac{2\lambda^{-1}+r-3L}{2} - 28r\beta\sigma_2^2\eta^2 - \frac{\eta^2}{2\alpha} \geq \frac{3}{8\lambda}.$$

Putting everything together yields

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
&\geq \frac{3}{8\lambda}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^{k+1})\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 \\
&\quad - 14r\beta\|x_r(y(z^{k+1}), z^{k+1}) - x_r(y_+^k(z^{k+1}), z^{k+1})\|^2.
\end{aligned}$$

The proof is complete.

C Proof of Corollary 1

Proof The proof follows closely that of Proposition 3. Let $\psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$\psi(x,z) = \max_{y \in \mathcal{Y}} F_r(x,y,z)$$

and consider arbitrary $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathbb{R}^n$. Again, note that $\psi(\cdot, z)$ is $(r-L)$-strongly convex. This implies that

$$\psi(x,z) - \psi(x_r^\star(z), z) \geq \frac{r-L}{2}\|x - x_r^\star(z)\|^2. \tag{C.1}$$

Moreover, we have

$$\begin{aligned}
&\psi(x,z) - \psi(x_r^\star(z), z) \\
&\leq \psi(x,z) - F_r(x_r(y_+(z),z), y_+(z), z) \\
&= \max_{y' \in \mathcal{Y}} F(x,y') + \frac{r}{2}\|x - z\|^2 - F_r(x_r(y_+(z),z), y_+(z), z) \\
&= \max_{y' \in \mathcal{Y}} F(x,y') - F(x_r(y_+(z),z), y_+(z)) + \frac{r}{2}\|x - z\|^2 - \frac{r}{2}\|x_r(y_+(z),z) - z\|^2,
\end{aligned} \tag{C.2}$$

where the inequality follows from

$$\begin{aligned}
F_r(x_r(y_+(z),z), y_+(z), z) &= \min_{x' \in \mathcal{X}} \left\{ F(x', y_+(z)) + \frac{r}{2}\|x' - z\|^2 \right\} \\
&\leq \max_{y' \in \mathcal{Y}} \min_{x' \in \mathcal{X}} \left\{ F(x', y') + \frac{r}{2}\|x' - z\|^2 \right\} \\
&\leq \min_{x' \in \mathcal{X}} \max_{y' \in \mathcal{Y}} \left\{ F(x', y') + \frac{r}{2}\|x' - z\|^2 \right\} = \psi(x_r^\star(z), z).
\end{aligned}$$

As (C.1) and (C.2) hold for any $x \in \mathcal{X}$, we obtain the intermediate relation

$$\frac{r-L}{2}\|x_r^\star(z) - x_r(y_+(z),z)\|^2 \leq \max_{y' \in \mathcal{Y}} F(x_r(y_+(z),z), y') - F(x_r(y_+(z),z), y_+(z))$$

by taking $x = x_r(y_+(z), z)$.

The remaining steps of the proof are the same as those of Proposition 3. For brevity, we omit them here. □

D Dual Error Bound for Concave Case

The following lemma provides a dual error bound for the case where $F(x, \cdot)$ is concave for any $x \in \mathbb{R}^n$. Its proof is based on Lemma B.10 in [66].

Lemma 8 *For any $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$, we have*

$$\|x_r^\star(z) - x_r(y_+(z), z)\|^2 \leq \kappa \|y - y_+(z)\|,$$

where $\kappa := \frac{1 + \alpha L \sigma_2 + \alpha L}{\alpha(r-L)} \cdot \operatorname{diam}(\mathcal{Y})$.

Proof Let $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$ be arbitrary. Recall that $y(z)$ is an arbitrary vector in $Y(z)$. By the $(r-L)$-strong convexity of $F_r(\cdot, y, z)$, we have

$$F_r(x_r^\star(z), y_+(z), z) - F_r(x_r(y_+(z), z), y_+(z), z) \geq \frac{r-L}{2}\|x_r(y_+(z), z) - x_r^\star(z)\|^2,$$
$$F_r(x_r(y_+(z), z), y(z), z) - F_r(x_r^\star(z), y(z), z) \geq \frac{r-L}{2}\|x_r(y_+(z), z) - x_r^\star(z)\|^2.$$

Moreover, since $(x_r^\star(z), y(z))$ is a saddle point of the convex-concave function $F_r(\cdot, \cdot, z)$ [60], we have $F_r(x_r^\star(z), y(z), z) \geq F_r(x_r^\star(z), y_+(z), z)$, which implies that

$$\begin{aligned} &F_r(x_r(y_+(z), z), y(z), z) - F_r(x_r(y_+(z), z), y_+(z), z) \\ \geq\ &(r-L)\|x_r(y_+(z), z) - x_r^\star(z)\|^2. \end{aligned} \tag{D.1}$$

Note that $y_+(z)$ is the maximizer of

$$\max_{y' \in \mathcal{Y}} \left\langle y + \alpha \nabla_y F_r(x_r(y, z), y, z) - y_+(z), y' \right\rangle.$$

For simplicity, we define the function $\xi : \mathbb{R}^d \to \mathbb{R}$ by

$$\begin{aligned} \xi(\cdot) := \alpha F_r(x_r(y_+(z), z), \cdot, z) &- \left\langle \alpha \nabla_y F_r(x_r(y_+(z), z), y_+(z), z), \cdot \right\rangle \\ &- \left\langle y_+(z) - y - \alpha \nabla_y F_r(x_r(y, z), y, z), \cdot \right\rangle. \end{aligned}$$

Then, we have

$$\begin{aligned} &\max_{y' \in \mathcal{Y}}\ \xi(y') \\ &\leq \alpha F_r(x_r(y_+(z), z), y_+(z), z) - \left\langle \alpha \nabla_y F_r(x_r(y_+(z), z), y_+(z), z), y_+(z) \right\rangle \\ &\quad + \max_{y' \in \mathcal{Y}} \left\langle y + \alpha \nabla_y F_r(x_r(y, z), y, z) - y_+(z), y' \right\rangle \\ &\leq \xi(y_+(z)), \end{aligned}$$

where the first inequality holds because $F_r(x_r(y_+(z), z), \cdot, z)$ is concave. Therefore, we have $\xi(y(z)) \leq \xi(y_+(z))$, which implies that

$$\begin{aligned} &F_r(x_r(y_+(z), z), y(z), z) - F_r(x_r(y_+(z), z), y_+(z), z) \\ &\leq \frac{1}{\alpha}\Big\langle y(z) - y_+(z), \alpha \nabla_y F_r(x_r(y_+(z), z), y_+(z), z) - \alpha \nabla_y F_r(x_r(y, z), y, z) \Big\rangle \\ &\quad + \frac{1}{\alpha}\Big\langle y(z) - y_+(z), y_+(z) - y \Big\rangle \\ &\leq L\|y_+(z) - y(z)\|(\|x_r(y_+(z), z) - x_r(y, z)\| + \|y_+(z) - y\|) \\ &\quad + \frac{1}{\alpha}\|y_+(z) - y(z)\| \cdot \|y - y_+(z)\| \\ &\leq \left(\frac{1}{\alpha} + L\sigma_2 + L\right)\|y_+(z) - y(z)\| \cdot \|y - y_+(z)\|. \end{aligned} \tag{D.2}$$

Here, the second inequality is due to the L-Lipschitz continuity of $\nabla_y F_r(\cdot, \cdot, z)$ and the last inequality is from (A.3). Hence, by combining (D.1) and (D.2), we obtain

$$(r - L)\|x_r^\star(z) - x_r(y_+(z), z)\|^2 \le \left(\frac{1}{\alpha} + L\sigma_2 + L\right) \mathrm{diam}(\mathcal{Y}) \cdot \|y - y_+(z)\|,$$

which proves the desired result. □