Nonsmooth Composite Nonconvex-Concave Minimax Optimization

Jiajin Li JIAJINLI@STANFORD.EDU
Stanford University

Linglingzhi Zhu LLZZHU@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

Abstract

Nonconvex-concave minimax optimization has received intense interest in machine learning, including learning with robustness to data distribution, learning with non-decomposable loss, adversarial learning, to name a few. Nevertheless, most existing works focus on the gradient-descent-ascent (GDA) variants that can only be applied in smooth settings. In this paper, we consider a family of minimax problems whose objective function enjoys the nonsmooth composite structure in the variable of minimization and is concave in the variables of maximization. By fully exploiting the composite structure, we propose a smoothed proximal linear descent ascent (*smoothed* PLDA) algorithm and further establish its $\mathcal{O}(\epsilon^{-4})$ iteration complexity, which matches that of smoothed GDA [37] under smooth settings. Moreover, under the mild assumption that the objective function satisfies the one-sided Kurdyka-Łojasiewicz condition with exponent $\theta \in (0,1)$, we can further improve the iteration complexity to $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$. To the best of our knowledge, this is the first provably efficient algorithm for nonsmooth nonconvex-concave problems that can achieve the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$ if $\theta \in (0, 1/2]$.

1. Introduction

Recently, the class of nonconvex-(non)concave minimax optimization problems has attracted intense attention across both optimization and machine learning communities [27], mainly as it appears in applications such as (distributionally) robust optimization [5, 9, 12, 15, 22, 30], learning with non-decomposable loss [29, 38], adversarial learning [1, 13], to just name a few. In this paper, we are interested in studying nonconvex concave minimax problems of the form

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x,y), \tag{1.1}$$

where $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is nonconvex in x but concave in y, $\mathcal{X} \subseteq \mathbb{R}^n$ is closed convex and $\mathcal{Y} \subseteq \mathbb{R}^d$ is convex compact.

When $F(x,y)$ is smooth, a simple yet natural method to solve (1.1) is gradient descent ascent (GDA). At each iteration, this algorithm performs gradient descent over the variable x and gradient ascent over the variable y. On the positive side, GDA can generate an ϵ-stationary solution for a nonconvex strongly concave problem with iteration complexity $\mathcal{O}(\epsilon^{-2})$ [18], which already matches the best-known lower bound for solving (1.1) via first-order algorithms [7, 17, 39]. Nevertheless, without the strongly concave condition, GDA will suffer from oscillation. To address this issue, various diminishing step size techniques have been proposed to guarantee the convergence but can

only achieve the suboptimal complexity of at most $\mathcal{O}(\epsilon^{-4})$ [18, 20, 33]. By further exploiting problem-specific structures, [37] invokes the Nesterov smoothing trick to the vanilla GDA, and the resulting smoothed GDA can achieve the best iteration complexity of $\mathcal{O}(\epsilon^{-2})$ for minimizing the point-wise maximum of a finite collection of nonconvex smooth functions under some assumptions. Complement to [37], the work [35] obtains the same complexity under the Polyak-Łojasiewicz (PŁ) condition [25]. On another front, multi-loop type algorithms with acceleration in the subproblems [23, 24, 32, 34] have advantages over GDA variants in terms of iteration complexity for general nonconvex concave problems. The best iteration complexity among them is $\mathcal{O}(\epsilon^{-2.5})$, which is achieved by two triple-loop algorithms [19, 24]. Furthermore, if (1.1) admits a separable nonsmooth structure — the objective function consists of a smooth term plus a separable nonsmooth term with an easily computed proximal mapping, the analysis and algorithmic framework from the pure smoothness case can be adopted. The authors of [4, 6, 8, 14] introduce a class of (accelerated) proximal-GDA type algorithms, in which the gradient step is simply replaced by the proximal gradient step. Armed with the gradient Lipschitz continuity condition of the smoothness part, similar iteration complexity results have also been obtained.

Although nonconvex minimax optimization has already been extensively investigated in the literature, most existing works focus on the *almost* smooth case. In fact, it is only recently that researchers [26] have proposed the proximally guided stochastic subgradient method for general nonsmooth weakly convex-concave problems. However, it suffers from the slow iteration complexity of $\mathcal{O}(\epsilon^{-6})$. The main reason is that the method does not take any problem-specific structure into account but just utilizes the subgradient information. As a result, there is a huge theoretical gap between the lower and upper bounds. Thus, one natural question to ask here is **Can we design a provably efficient algorithm to address nonsmooth nonconvex-concave problems, which matches the lower bound $\mathcal{O}(\epsilon^{-2})$?** In this paper, we will answer the above question in the affirmative for a specific class of nonsmooth problems, that is, $F(\cdot, y) := h_y \circ c_y$ enjoys the nonsmooth composite structure. Here, h_y is convex and Lipschitz continuous (possibly nonsmooth) and c_y is continuously differentiable with Lipschitz continuous Jacobian map for all $y \in \mathcal{Y}$.

Due to the nonsmooth composite structure $h_y \circ c_y$, there is no available gradient information to rely on. As such, we are motivated to leverage the proximal linear scheme [11] for the primal update. This leads us to a new algorithm, which we call smoothed proximal linear descent ascent (smoothed PLDA) and can be regarded as a natural extension of smoothed GDA [37] to solve nonsmooth composite nonconvex-concave problems. Unfortunately, the analysis framework in [37] cannot be adopted. The key difficulty lies in the lack of the gradient Lipschitz smoothness condition for the primal function. To circumvent this difficulty, we prove that a tight Lipschitz-type primal error bound condition (i.e., Proposition 3.1) holds for the proximal linear scheme even without the gradient Lipschitz condition. Such an error bound is new and acts as a crucial step for establishing the sufficient decrease property of a potential function for the problem.

Next, to provide a comprehensive study of the iteration complexity and convergence behavior of smoothed PLDA for nonconvex-concave problems, we prove that a dual error bound condition (i.e., Proposition 3.3) holds for the dual ascent scheme. Again, such an error bound is new. The key idea for establishing this result is to employ the Kurdyka-Łojasiewicz (KŁ) property with an explicit exponent $\theta \in (0, 1)$ [16] in the dual variable y. This is a notable departure from the usual approach of utilizing the KŁ exponent in pure primal nonconvex optimization [2, 3, 16]. Specifically, the KŁ exponent θ here is used to explicitly control the trade-off between the decrease in the primal and the increase in the dual. As a result, we are able to achieve $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$ iteration complexity if the

Table 1: Comparison of the iteration complexities of smoothed PLDA proposed in this paper 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.**	**Iter. Compl.**[1]	**Add. Asm.**
GDA [18]	L-smooth	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X}=\mathbb{R}^n$
Smoothed GDA [37]	L-smooth	concave	$\mathcal{O}(\epsilon^{-4})$	—
PG-SMD [26]	weakly-convex	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X}$ bounded
This paper	nonsmooth composite	concave	$\mathcal{O}(\epsilon^{-4})$	—
GDA [18]	L-smooth	strongly-concave	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{X}=\mathbb{R}^n$
Smoothed GDA [35]	L-smooth	PŁ condition	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{Y}=\mathbb{R}^d$
This paper	nonsmooth composite	KŁ exponent $\theta=\frac{1}{2}$	$\mathcal{O}(\epsilon^{-2})$	—

dual function satisfies the KŁ property with exponent $\theta\in(0,1)$. In particular, when $\theta\in(0,1/2]$, the proposed smoothed PLDA achieves the optimal iteration complexity $\mathcal{O}(\epsilon^{-2})$. To the best of our knowledge, this is the first provably efficient algorithm for nonsmooth nonconvex-concave problems that can achieve the optimal order in convergence rate. For general concave problems, the iteration complexity of $\mathcal{O}(\epsilon^{-4})$ has also been shown for the proposed smoothed PLDA, which achieves the same complexity as the smoothed GDA studied in [37]. Table 1 summarizes the comparison of the iteration complexities of smoothed PLDA and other related methods under various setups.

2. Preliminaries

Let us introduce the basic problem setup and some essential concepts for later analysis.

Assumption 2.1 (Problem setup) *The following assumptions hold throughout the paper.*

(a) ***(Primal Function)*** *$F(\cdot,y):=h_y\circ c_y$, where $c_y:\mathbb{R}^n\to\mathbb{R}^m$ is continuously differentiable with L_c-Lipschitz continuous Jacobian map for all $y\in\mathcal{Y}$ on $\mathcal{X}$: $\|\nabla c_y(x)-\nabla c_y(x')\|\le L_c\|x-x'\|$ for all $x,x'\in\mathcal{X}$, and $h_y:\mathbb{R}^m\to\mathbb{R}$ for any $y\in\mathcal{Y}$ is a convex and L_h-Lipschitz continuous function satisfying $|h_y(z)-h_y(z')|\le L_h\|z-z'\|$, for all $z,z'\in\mathbb{R}^m$.*

(b) ***(Dual Function)*** *$F(x,\cdot)$ is concave and continuously differentiable on $\mathcal{Y}$ with $\nabla_y F(\cdot,\cdot)$ being L-Lipschitz continuous on $\mathcal{X}\times\mathcal{Y}$, i.e., $\|\nabla_y F(x,y)-\nabla_y F(x',y')\|\le L\|(x,y)-(x',y')\|$ for all $(x,y),(x',y')\in\mathcal{X}\times\mathcal{Y}$. Without loss of generality, we assume $L=L_hL_c$.*

Assumption 2.2 (Kurdyka-Łojasiewicz (KŁ) property with exponent θ for the dual function [16]) *For any fixed $x\in\mathcal{X}$, the problem $\max_{y\in\mathcal{Y}}F(x,y)$ has a nonempty solution set and a finite optimal value. There exist $\mu>0$ and $\theta\in(0,1)$ such that*

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

Now, we introduce the stationarity measures considered in this paper. Denote $f:=\max_{y\in\mathcal{Y}}F(\cdot,y)$, the potential function $F_r:\mathbb{R}^n\times\mathbb{R}^d\times\mathbb{R}^n\to\mathbb{R}$ and the dual function $d_r:\mathbb{R}^d\times\mathbb{R}^n\to\mathbb{R}$ as

$$F_r(x,y,z):=F(x,y)+\frac{r}{2}\|x-z\|^2\quad\text{and}\quad d_r(y,z):=\min_{x\in\mathcal{X}}F_r(x,y,z),$$

respectively, where we always assume $r > L$ in the remaining parts of this paper. Observing that $\nabla_x d_r(y,x) = r(x - \mathrm{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(x))$[2] by Danskin's theorem. Thus, we may use $\|\nabla_x d_r(y,x)\|$ and $\mathrm{dist}(0, -\nabla_y F(x,y) + \partial \iota_{\mathcal{Y}}(y))$ as a primal-dual stationarity measure for (1.1). This leads to the game stationarity measure in Definition 2.3.

Definition 2.3 (Stationarity Measures) *The pair $(x,y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-game stationary point (ϵ-GS) if*

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

3. Convergence Analysis of Smoothed PLDA

In this section, we first present the proposed smoothed proximal linear descent ascent algorithm and then establish its iteration complexity with/without Assumption 2.2. For smooth nonconvex concave optimization problems, a natural and intuitive algorithm is Gradient Descent Ascent (GDA), which however may suffer from oscillations. To address this issue, [37] proposes a Nesterov-type smoothing technique to handle the primal updates. It is tempting to adapt this smoothing technique to the structured nonsmooth setting. However, the nonsmoothness will cause fundamental difficulties.

On the algorithmic side, due to the composite structure $h_y \circ c_y$, there is no available gradient information to rely on. Nevertheless, by fully exploiting the composite structure, we can leverage the proximal linear scheme [11] to handle the primal update, i.e.,

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ F_{x^k,\lambda}(x, y^k) + \frac{r}{2}\|x - z^k\|^2 \right\}. \tag{3.1}$$

Here, $F_{x^k,\lambda}(x,y^k) = h_{y^k}\left(c_{y^k}(x^k) + \nabla c_{y^k}(x^k)^\top (x - x^k)\right) + \frac{\lambda}{2}\|x - x^k\|^2$ and $\{z^k\}$ is the auxiliary sequence. We use the same auxiliary sequence, smoothing and dual update as the smoothed GDA [37]. The "Smoothed PLDA" algorithm is formally presented in Algorithm 1.

Algorithm 1: Smoothed Proximal Linear Descent Ascent (**Smoothed PLDA**)

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

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

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

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

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

end

On the theoretical side, it is far from obvious how to tackle the nonsmoothness. Before we introduce the main obstacle and our main theoretical results, let us introduce some basic definitions and concepts for the later analysis. We define the proximal function $p_r : \mathbb{R}^n \to \mathbb{R}$ as

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

and let

$$x_r(y,z) := \arg\min_{x \in \mathcal{X}} F_r(x,y,z) = \mathrm{prox}_{\frac{1}{r}F(\cdot,y)+\iota_{\mathcal{X}}}(z),$$

2. $\iota_{\mathcal{Y}} : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the indicator function of $\mathcal{Y}$, and $\mathrm{dist}(x, \mathcal{S}) := \inf_{z \in \mathcal{S}} \|x - z\|$ is the distance from $x \in \mathbb{R}^n$ to a set $\mathcal{S} \subseteq \mathbb{R}^n$.

$$x_r^*(z) := \operatorname*{argmin}_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_r(x,y,z) = \operatorname{prox}_{\frac{1}{r}f+\iota_{\mathcal{X}}}(z),$$
$$y_+(z) := \operatorname{proj}_{\mathcal{Y}}\left(y + \alpha\nabla_y F_r(x_r(y,z),y,z)\right).$$

To analyze the convergence of *smoothed PLDA*, we follow [37] to define a potential function as

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

In fact, all terms in the potential function Φ_r are closely related with the algorithmic updates. Namely, the update for the primal, dual and auxiliary variables can be understood as a primal descent for the function F_r, approximate dual ascent for the dual function d_r and approximate proximal descent for the proximal function p_r.

To start with, we study the sufficient decrease property of this potential function. One of the main obstacles that prevents us from adopting the analysis framework in [37] is the lack of the gradient Lipschitz smoothness condition for the primal function. In the smooth case, the key to guarantee the basic descent estimate (i.e., Proposition 4.1 in [37]) is the primal error bound condition (3.2), which trivially holds because it is equivalent to the Luo-Tseng error bound condition for structured strongly convex functions [21, 36]. That is, for some $\zeta > 0$ $\|x^{k+1} - x_r(y^k,z^k)\| \le \zeta\|x^k - \operatorname{proj}_{\mathcal{X}}(x^k - c\nabla_x F_r(x^k,y^k,z^k))\|$, where $c > 0$ is the step size for the primal descent. However, to the best of our knowledge, there is no analog for the nonsmooth convex composite problem with the proximal linear scheme. One of our main theoretical contributions is to show that a similar Lipschitz-type primal error bound condition still holds without L-smooth Lipschitz condition.

Proposition 3.1 (Lipschitz-type primal error bound) *For any $k \ge 0$, it holds that*

$$\|x^{k+1} - x_r(y^k,z^k)\| \le \zeta\|x^k - x^{k+1}\|, \tag{3.2}$$

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

It is worth noting that a similar error bound condition has also been investigated in [10]. However, the approach therein did not provide any explicit constant estimation. The explicit constant we derive in Proposition 3.1 plays an important role in controlling the step size for both primal and dual updates.

Armed with Proposition 3.1, we can establish the sufficient decrease property of the potential function.

Proposition 3.2 (Sufficient decrease property) *Let $r \ge 3L$, $\lambda \ge L$, $\alpha \le \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\}$, and $\beta \le \min\left\{\frac{1}{28}, \frac{(r-L)^2}{32\alpha r(r+L)^2}\right\}$. Then for any $k \ge 0$,*

$$\Phi_r^k - \Phi_r^{k+1} \ge \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 - 28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k),z^k)\|^2,$$

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

A basic principle to conduct the convergence analysis is to make the potential function decrease sufficiently at each iteration. The key obstacle here is to bound the negative term $\|x_r^*(z^k) -$

$x_r(y_+^k(z^k), z^k)\|$. Conceptually, this term is related to $\|y^k - y_+^k(z^k)\|$. That is, if $\|y^k - y_+^k(z^k)\| = 0$, then y^k is the optimal solution of $\max_{y\in\mathcal{Y}} d_r(y, z^k)$ and thus $x_r^*(z^k) = x_r(y_+^k(z^k), z^k) = x_r(y^k, z^k)$. Consequently, it is a natural idea to bound $\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|$ by $\|y^k - y_+^k(z^k)\|$. The remaining question is how to determine the explicit growth rate. At the heart of our analysis is to exploit the KŁ exponent to characterize the primal-dual perturbation quantity explicitly, which leads to the following dual error bound condition. As we shall see later, it plays a vital role in analyzing the explicit convergence rate or iteration complexity of the proposed smoothed PLDA algorithm.

Proposition 3.3 (Dual error bound condition with KŁ exponent) *Suppose Assumption 2.2 holds. Then*

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

where $\omega := \frac{\sqrt{2}}{\sqrt{r-L}} \left(\frac{1+\alpha L(1+\sigma_2)}{\alpha\mu}\right)^{\frac{1}{2\theta}}$.

Remark 3.4 (i) *This result generalizes the one in [35], which corresponds to the special case of* $\theta = \frac{1}{2}$ *and* $\mathcal{Y} = \mathbb{R}^d$*, to the full range of* θ *with/without constraints.* (ii) *For the general concave case without Assumption 2.2, there is a similar dual error bound as Proposition 3.3 with* $\theta = 1$ *and* ω *being related to the diameter of the compact set* $\mathcal{Y}$*, see Lemma 7 for details.* (iii) *The analysis framework of Proposition 3.3 cannot include the sharpness case (i.e.,* $\theta = 0$*). [37] provides a certain dual error bound condition for max-structure problems under several strong assumptions, which satisfies the KŁ condition with* $\theta = 0$*. We leave this case as our future work.*

Armed with Proposition 3.3, we present the main theorem concerning the iteration complexity of smoothed PLDA under various settings.

Theorem 3.5 *Suppose that* $r \ge 3L$, $\lambda \ge L$, $\alpha \le \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\}$, $\beta \le \min\left\{\frac{1}{28}, \frac{(r-L)^2}{32\alpha r(r+L)^2}\right\}$. *Then for any integer* $K > 0$*, there exists a* $k \in \{1, 2, \cdots, K\}$ *such that* $\left(x^{k+1}, y^{k+1}\right)$ *is an* $\mathcal{O}(K^{-\frac{1}{4}})$*-game stationary if* $\beta \le K^{-\frac{1}{2}}$*. If we further suppose Assumption 2.2 holds, then*

(a) (KŁ exponent $\theta \in (\frac{1}{2}, 1)$): $\mathcal{O}(K^{-\frac{1}{4\theta}})$-game stationary if $\beta \le K^{-\frac{2\theta-1}{2\theta}}$;

(b) (KŁ exponent $\theta \in (0, \frac{1}{2}]$): $\mathcal{O}(K^{-\frac{1}{2}})$-game stationary if $\beta \le \frac{\mathrm{diam}(\mathcal{Y})^{\frac{2\theta-1}{\theta}}}{448\alpha r\omega^2}$.

Remark 3.6 *We claim that our algorithm can achieve the optimal iteration complexity of* $\mathcal{O}(\epsilon^{-2})$ *when the dual function satisfies the KŁ condition with exponent* $\theta \in (0, \frac{1}{2}]$*, which certainly includes the nonconvex-strongly concave problem as a special case. To the best of our knowledge, this is the first provably efficient algorithm for nonsmooth nonconvex-concave problems, which can achieve the same results as the smooth case.*

4. Conclusion and Future Directions

In this paper, we propose a smoothed proximal linear descent ascent (*smoothed* PLDA) algorithm to solve a class of nonsmooth composite nonconvex concave problems, which can achieve $\mathcal{O}(\epsilon^{-4})$ iteration complexity for general concave problems. To further arm with the one-sided Kurdyka-Łojasiewicz condition with the exponent $\theta \in (0, 1)$ in the dual variable, we can establish the iteration complexity as $\mathcal{O}(\epsilon^{-2\max\{2\theta,1\}})$, which matches the optimal order as $\mathcal{O}(\epsilon^{-2})$. It would be super

interesting to further investigate whether the analysis framework introduced in this paper can be extended to include the sharpness case (i.e., $\theta = 0$) and thus address the max structure problem studied in [37]. Another natural future direction is to extend our algorithm into the stochastic setting so that the modern machine learning tasks can be benefited from our methods.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[2] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

[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] Babak Barazandeh and Meisam Razaviyayn. Solving non-convex non-differentiable min-max games using proximal gradient method. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3162–3166. IEEE, 2020.

[5] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust optimization. In *Robust optimization*. Princeton university press, 2009.

[6] Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

[7] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

[8] Ziyi Chen, Yi Zhou, Tengyu Xu, and Yingbin Liang. Proximal gradient descent-ascent: variable convergence under KŁ geometry. In *International Conference on Learning Representations*, 2021.

[9] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[10] 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.

[11] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

[12] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*, 2017.

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

[14] Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] 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.

[17] Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

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

[20] 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.

[21] 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.

[22] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

[23] 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. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] 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.

[25] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[26] 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*, pages 1–35, 2021.

[27] 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.

[28] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[29] Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: how and why. In *International Conference on Machine Learning*, pages 793–801. PMLR, 2016.

[30] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[31] Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[32] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] 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. *arXiv preprint arXiv:2006.02032*, 2020.

[34] Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[36] 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.

[37] Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] Lijun Zhang and Zhi-Hua Zhou. ℓ_1-regression with heavy-tailed distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

[39] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

Appendix A. Organization of the Appendix

We organize the appendix as follows:

- All notations used are summarized in A.1 and useful technical lemmas in A.2.
- The proof of Lipschitz type perturbation bound (Proposition 3.1) is included in Section B.
- Sufficient decrease property of potential function Φ_r (Proposition 3.2) is given in Section C.
- Dual error bound condition under two different settings is established in Section D.
- The proof details of the main theorem 3.5 are given in Section E.

A.1. Notations

Before we present all proof details of theorems and lemmas, we adopt the following notations:

- $F_r(x,y,z) := F(x,y) + \frac{r}{2}\|x-z\|^2$: smoothed potential function;
- $d_r(y,z) := \min_{x\in\mathcal{X}} F_r(x,y,z)$: dual function of $F_r(x,y,z)$
- $p_r(z) := \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_r(x,y,z)$: proximal function of $F_r(x,y,z)$;
- $x_r(y,z) := \operatorname{argmin}_{x\in\mathcal{X}} F_r(x,y,z)$;
- $x_r^*(z) := \operatorname{argmin}_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_r(x,y,z)$;
- $Y(z) := \operatorname{argmax}_{y\in\mathcal{Y}} d_r(y,z)$;
- $y_+(z) := \operatorname{proj}_{\mathcal{Y}}(y + \alpha\nabla_y F_r(x_r(y,z),y,z))$: one-step projected gradient ascent of dual function.

A.2. Useful Technical Lemmas

To begin with, we introduce the weakly convex function which plays an important role in our following analysis.

Definition A.1 *The function $\ell : \mathbb{R}^n \to \mathbb{R}$ is ρ-weakly convex on $\mathcal{X} \subseteq \mathbb{R}^n$ if for any $x, y \in \mathcal{X}$ and $\tau \in [0,1]$,*

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

When ℓ is locally Lipschitz, it is equivalent to $\ell + \frac{\rho}{2}\|\cdot\|^2$ is convex on $\mathcal{X}$.

By assumption of the problem (1.1) (recalling $L = L_h L_c$ and $r > L$) with [11, Lemma 3.2, Lemma 4.2], we directly have the following useful results.

Fact A.2 *The functions $F(\cdot, y)$ for any $y \in \mathbb{R}^d$ and f are L-weakly convex on $\mathcal{X}$.*

Fact A.3 *Let $y \in \mathbb{R}^d$. For all $x, \bar{x} \in \mathcal{X}$ it follows that*

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

The definition of the weakly convex function together with Fact A.2 implies that $F_r(\cdot, y, z)$ is $(r-L)$-strongly convex for any $(y, z) \in \mathcal{Y} \times \mathbb{R}^n$.

Lemma 1 *Suppose that f_i is ρ_i-weakly convex functions for all $i \in [N]$ and $\mathcal{A}$ is a bounded set , i.e., $\operatorname{diam}(\mathcal{A}) \leq B$, we have $\sup_{y\in\mathcal{A}} \sum_{i=1}^N y_i f_i(x)$ is $\left(B \max_{i\in[N]} \rho_i\right)$-weakly convex.*

The following Lemma 2 and 3 are also needed in our analysis, and the proof of which is similar to that of Lemma B.2 and B.3 in [37], respectively. For the sake of completeness, we present the proof here.

Lemma 2 *For any $y, y' \in \mathcal{Y}$ and $z, z' \in \mathbb{R}^n$, the following inequalies hold:*

$$\|x_r(y, z) - x_r(y, z')\| \leq \sigma_1 \|z - z'\|, \tag{A.1}$$
$$\|x_r^*(z) - x_r^*(z')\| \leq \sigma_1 \|z - z'\|, \tag{A.2}$$
$$\|x_r(y, z) - x_r(y', z)\| \leq \sigma_2 \|y - y'\|, \tag{A.3}$$

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

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

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

Since from Fact A.2 one has that $F(\cdot, y)$ is L-weakly convex, and consequently we know for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathbb{R}^n$ that

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

Thus, combining (A.4) and (A.5) one has that

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

On the other hand, again by (A.5) we have

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

and this together with (A.6) implies that

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

which by the Cauchy-Schwarz inequality further implies

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

Hence, (A.1) holds with Lipschitz modulus $\sigma_1 = \frac{r}{r-L}$. The inequality (A.2) holds with the similar argument as above since $f = \max_{y\in\mathcal{Y}} F(\cdot, y)$ is also L-weakly convex by Fact A.2.

Now, we start to prove (A.3). Using (A.5) we have that

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

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

Moreover, by the concavity of $F_r(x,\cdot,z)$ we have

$$F_r(x_r(y,z),y',z) - F_r(x_r(y,z),y,z) \leq \langle \nabla_y F_r(x_r(y,z),y,z), y' - y\rangle, \tag{A.9}$$

and from the Lipschitz continuity of $\nabla_y F_r(x,\cdot,z)$, we have

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

Combining (A.7)-(A.10) it follows that

$$(r-L)\|x_r(y,z) - x_r(y',z)\|^2 \leq \langle \nabla_y F_r(x_r(y,z),y,z) - \nabla_y F_r(x_r(y',z),y',z), y' - y\rangle + \frac{L}{2}\|y - y'\|^2.$$

This together with the consequence of the Lipschitz continuity of $\nabla_y F_r(\cdot,\cdot,z)$ that

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

implies that

$$(r-L)\|x_r(y,z) - x_r(y',z)\|^2 \leq L\|x_r(y',z) - x_r(y,z)\|\|y - y'\| + \frac{3L}{2}\|y - y'\|^2. \tag{A.11}$$

Let $\zeta := \frac{\|x_r(y,z) - x_r(y',z)\|}{\|y - y'\|}$. Then we know from (A.11) that

$$\zeta^2 \leq \frac{r+L}{r-L}\zeta + \frac{3(r+L)}{2(r-L)} \leq \frac{1}{2}\zeta^2 + \frac{1}{2}\left(\frac{r+L}{r-L}\right)^2 + \frac{3}{2}\left(\frac{r+L}{r-L}\right)^2 = \frac{1}{2}\zeta^2 + 2\left(\frac{r+L}{r-L}\right)^2.$$

where the second inequality is due to the basic inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ for $a, b \in \mathbb{R}$ and $\frac{r+L}{r-L} > 1$. Thus,

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

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

Lemma 3 *The dual function $d_r(\cdot,\cdot)$ is differentiable on $\mathcal{Y}\times\mathbb{R}^n$, and for each $y\in\mathcal{Y}$, $z\in\mathbb{R}^n$*

$$\nabla_y d_r(y,z)=\nabla_y F(x_r(y,z),y),\quad \nabla_z d_r(y,z)=\nabla_z F_r(x_r(y,z),y,z)=r(z-x_r(y,z)).$$

Moerover, $\nabla_y d_r(\cdot,\cdot)$ is Lipschitz continuous, i.e.,

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

with $L_{d_r}:=\max\{\sigma_1+r,(\sigma_2+1)L\}$.

Proof Since $F_r(\cdot,y,z)$ is strongly convex, $F_r(x,\cdot,z)$ is concave on $\mathcal{Y}$ and $F_r(x,y,\cdot)$ is (strongly) convex, by [28, Theorem 10.31] we know that

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

is differentiable on $\mathcal{Y}\times\mathbb{R}^n$. Then for each $y\in\mathcal{Y}$, $z\in\mathbb{R}^n$, one has

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

and from [11, Lemma 4.3] we know that

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

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

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

where the third inequality is due to (A.3). Also,

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

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

Appendix B. Proof of Proposition 3.1 — Lipschitz type Primal Error Bound Condition

Recall that

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

For simplicity, we denote $\hat{F}_{y,z} := F_r(\cdot, y, z) + \iota_{\mathcal{X}}$ for any $y \in \mathbb{R}^d$ and $z \in \mathbb{R}^n$. Note that

$$\hat{F}_{y^k,z^k}(x) - \hat{F}_{y^k,z^k}(x_r(y^k,z^k)) \geq \frac{r-L}{2}\|x - x_r(y^k,z^k)\|^2, \quad \text{for each } x \in \mathcal{X}. \tag{B.1}$$

By the convexity of $\hat{F}_{y^k,z^k}$, we obtain for each $x \in \mathcal{X}$ that

$$\begin{aligned}\hat{F}_{y^k,z^k}(x) - \hat{F}_{y^k,z^k}(x_r(y^k,z^k)) &\leq g_x^\top (x - x_r(y^k,z^k)) \\ &= \operatorname{dist}(0, \partial \hat{F}_{y^k,z^k}(x)) \cdot \|x - x_r(y^k,z^k)\|,\end{aligned} \tag{B.2}$$

where $g_x \in \partial \hat{F}_{y^k,z^k}(x)$ satisfying $\|g_x\| = \min_{g \in \partial \hat{F}_{y^k,z^k}(x)} \|g\|$. Together (B.1) and (B.2), we have

$$\|x - x_r(y^k,z^k)\| \leq 2(r-L)^{-1} \operatorname{dist}(0, \partial \hat{F}_{y^k,z^k}(x)), \quad \text{for each } x \in \mathcal{X}.$$

By invoking [10, Theorem 3.4], it follows that

$$\|x - x_r(y^k,z^k)\| \leq \frac{2(r-L)^{-1} + t}{t} \left\| x - \operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x) \right\|. \tag{B.3}$$

Next, we exploit the precise quantitative relationship between the norm of the residual term $\|x^{k+1} - x^k\|$ and $\|x^{k+1} - \operatorname{prox}_{t\hat{F}_{y^k,z^k}}(x^{k+1})\|$. We define the function $\varphi_k : \mathbb{R}^n \to \mathbb{R}$ for any $k \in \mathbb{N}$ as

$$\varphi_k(x) := \hat{F}_{y^k,z^k}(x) + \frac{\lambda+L}{2}\|x - x^k\|^2 - \frac{\lambda+L}{2}\|x - x^{k+1}\|^2 \quad \text{for each } x \in \mathbb{R}^n.$$

From Fact A.3 and the strong convexity of $F_{x^k,\lambda}(\cdot, y^k)$ we know that for any $x \in \mathcal{X}$,

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

Then, from $r - L > 0$ we have for any $x \in \mathcal{X}$

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

which combines with the definition of φ_k implies

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

Consequently we can observe that for any $\rho > 0$

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

which implies that

$$\|\operatorname{prox}_{\frac{1}{\rho}\varphi_k}(x^{k+1}) - x^{k+1}\| \le \sqrt{\frac{2L}{\rho}}\|x^{k+1} - x^k\|. \tag{B.4}$$

Furthermore, it is interesting to observe that

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

and consequently $\operatorname{prox}_{\frac{1}{\lambda+L}\varphi_k}(x^{k+1}) = \operatorname{prox}_{\frac{1}{\lambda+L}\hat{F}_{y^k,z^k}}(x^k)$. This together with (B.4) (letting $\rho = \lambda + L$) and (B.3) (letting $t = (\lambda+L)^{-1}$) implies that

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

which completes the proof for (3.2). Furthermore, by the nonexpansiveness of the projection operator, we have

$$\begin{aligned}
&\|y^{k+1} - y_+^k(z^k)\| \\
=\ & \|\operatorname{proj}_{\mathcal{Y}}(y^k + \alpha\nabla_y F_r(x^{k+1}, y^k, z^k)) - \operatorname{proj}_{\mathcal{Y}}(y^k + \alpha\nabla_y F_r(x_r(y^k, z^k), y^k, z^k))\| \\
\le\ & \|y^k + \alpha\nabla_y F_r(x^{k+1}, y^k, z^k) - (y^k + \alpha\nabla_y F_r(x_r(y^k, z^k), y^k, z^k))\| \\
\le\ & \alpha L\|x^{k+1} - x_r(y^k, z^k)\| \\
\le\ & \alpha L\zeta\|x^k - x^{k+1}\|,
\end{aligned} \tag{B.5}$$

where the first inequality is due to the nonexpansiveness of the projection operator. The proof is complete.

Appendix C. Sufficient Decrease Property of Potential Function

Lemma 4 (Primal Descent) *For any $k \geq 0$, it follows that*

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

Proof From the definition we know that $F_{x^k,\lambda}(\cdot, y^k)$ is λ-strongly convex, and then one has that

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

Due to Fact A.3, we have

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

Hence, it follows that

$$\begin{aligned} F_r(x^k, y^k, z^k) &\geq F(x^{k+1}, y^k) + \frac{r}{2}\|x^{k+1} - z^k\|^2 + \frac{2\lambda + r - L}{2}\|x^k - x^{k+1}\|^2 \\ &= F_r(x^{k+1}, y^k, z^k) + \frac{2\lambda + r - L}{2}\|x^k - x^{k+1}\|^2. \end{aligned} \tag{C.3}$$

Next, as $F_r(x, \cdot, z)$ is concave, we have

$$F_r(x^{k+1}, y^k, z^k) - F_r(x^{k+1}, y^{k+1}, z^k) \geq \langle \nabla_y F_r(x^{k+1}, y^k, z^k), y^k - y^{k+1}\rangle. \tag{C.4}$$

At last, on top of the update of variable z^{k+1}, i.e. $z^{k+1} = z^k + \beta(x^{k+1} - z^k)$, we can verify

$$F_r(x^{k+1}, y^{k+1}, z^k) - F_r(x^{k+1}, y^{k+1}, z^{k+1}) = \frac{(2-\beta)r}{2\beta}\|z^k - z^{k+1}\|^2. \tag{C.5}$$

Summing up (C.3), (C.4) and (C.5), the desired result is obtained. ■

Lemma 5 (Dual Ascent) *For any $k \geq 0$, it follows that*

$$\begin{aligned} d_r(y^{k+1}, z^{k+1}) - d_r(y^k, z^k) \geq &\langle \nabla_y F_r(x_r(y^k, z^k), z^k; y^k), y^{k+1} - y^k\rangle - \frac{L_{d_r}}{2}\|y^k - y^{k+1}\|^2 + \\ &\frac{r}{2}\left\langle z^{k+1} - z^k, z^{k+1} + z^k - 2x_r(y^{k+1}, z^{k+1})\right\rangle \end{aligned} \tag{C.6}$$

Proof From Lemma 3 we know $\nabla_y d_r(\cdot, z)$ is Lipschitz continuous with L_{d_r}. Then we know that

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

On the other hand, one has that

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

Finally, combining above inequalities we know (C.6) holds. ■

Lemma 6 (Proximal Descent (Smoothness)) *For any $k \geq 0$, it follows that*

$$p_r(z^k) - p_r(z^{k+1}) \geq \frac{r}{2}\left\langle z^{k+1} - z^k, 2x_r(y(z^{k+1}), z^k) - z^k - z^{k+1}\right\rangle, \tag{C.7}$$

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

Proof From Sion's minimax theorem [31], we know that

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

which implies

$$p_r(z) = \max_{y\in\mathcal{Y}} d_r(y, z).$$

Consequently, it follows from the definition of $y(z^{k+1})$ that

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

where the second inequality is from that $F_r(x', y, z) \geq \min_{x\in\mathcal{X}} F_r(x, y, z) = d_r(y, z)$ holds for any $x' \in \mathcal{X}$. The proof is complete. ■

Proof Proposition 3.2 From Lemma 4, 5, 6, we know that

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

$$\underbrace{2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^k)-x_r(y^{k+1},z^{k+1})\right\rangle}_{②}. \tag{C.8}$$

Subsequently, we simplify the terms ① and ②. First, for ① we know that

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

For the left term, one has that

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

where the last inequality essentially follows from the property of the projection operator and the update of dual variable y^k that $y^{k+1} = \mathrm{proj}_{\mathcal{Y}}(y^k + \alpha\nabla_y F_r(x^{k+1},y^k,z^k))$. On the other hand, for the right term we have

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

where the third inequality follows from $2|x||y| \leq \tau x^2 + \frac{1}{\tau}y^2$ and the last inequality follows from Proposition 3.1. Together them, we obtain,

$$① \geq \left(\frac{1}{\alpha} - L\zeta^2\right)\|y^k - y^{k+1}\|^2 - L\|x^{k+1}-x^k\|^2. \tag{C.9}$$

Then, we continue to bound ②,

$$\begin{aligned}
② &= 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^k) - x_r(y^{k+1},z^{k+1})\right\rangle \\
&= 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^k) - x_r(y(z^{k+1}),z^{k+1})\right\rangle + \\
&\quad 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^{k+1}) - x_r(y^{k+1},z^{k+1})\right\rangle \\
&\geq -2r\sigma_1\|z^{k+1}-z^k\|^2 + 2r\left\langle z^{k+1}-z^k, x_r(y(z^{k+1}),z^{k+1}) - x_r(y^{k+1},z^{k+1})\right\rangle \\
&\geq -2r\sigma_1\|z^{k+1}-z^k\|^2 - \frac{r}{7\beta}\|z^{k+1}-z^k\|^2 - 7r\beta\|x_r(y(z^{k+1}),z^{k+1}) - x_r(y^{k+1},z^{k+1})\|^2,
\end{aligned} \tag{C.10}$$

where the inequality is from (A.1) with the Cauchy-Schwarz inequality and the AM-GM inequality. Thus, the inequalities (C.8)-(C.9) above imply that

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \frac{2\lambda + r - 3L}{2}\|x^k - x^{k+1}\|^2 + \left(\frac{1}{\alpha} - L_{d_r} - L\zeta^2\right)\|y^k - y^{k+1}\|^2 + \\
& \left(\frac{(2-\beta)r}{2\beta} - 2r\sigma_1 - \frac{r}{7\beta}\right)\|z^k - z^{k+1}\|^2 - 7r\beta\|x_r^*(z^{k+1}) - x_r(y^{k+1}, z^{k+1})\|^2. \qquad \text{(C.11)}
\end{aligned}$$

On top of (B.5), we have

$$\begin{aligned}
\|y^{k+1} - y^k\|^2 &= \|y^{k+1} - y_+^k(z^k) + y_+^k(z^k) - y^k\|^2 \\
&\geq \frac{1}{2}\|y^k - y_+^k(z^k)\|^2 - \|y^{k+1} - y_+^k(z^k)\|^2 \\
&\geq \frac{1}{2}\|y^k - y_+^k(z^k)\|^2 - \eta^2\|x^k - x^{k+1}\|^2. \qquad \text{(C.12)}
\end{aligned}$$

Similarly, by Lemma 2 and (B.5), we have

$$\begin{aligned}
&\|x_r^*(z^{k+1}) - x_r(y^{k+1}, z^{k+1})\|^2 \\
\leq\ & 4\|x_r^*(z^{k+1}) - x_r^*(z^k)\|^2 + 4\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2 + \\
& 4\|x_r(y_+^k(z^k), z^k) - x_r(y^{k+1}, z^k)\|^2 + 4\|x_r(y^{k+1}, z^k) - x_r(y^{k+1}, z^{k+1})\|^2 \\
\leq\ & 8\sigma_1^2\|z^k - z^{k+1}\|^2 + 4\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2 + 4\sigma_2^2\eta^2\|x^k - x^{k+1}\|^2. \qquad \text{(C.13)}
\end{aligned}$$

Substituting (C.12)-(C.13) to (C.11) yields

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \left(\frac{2\lambda + r - 3L}{2} - 28r\beta\sigma_2^2\eta^2\right)\|x^k - x^{k+1}\|^2 + \\
& \left(\frac{1}{\alpha} - L_{d_r} - L\zeta^2\right)\left(\frac{1}{2}\|y^k - y_+^k(z^k)\|^2 - \eta^2\|x^k - x^{k+1}\|^2\right) + \\
& \left(\frac{(2-\beta)r}{2\beta} - 2r\sigma_1 - \frac{r}{7\beta} - 56r\beta\sigma_1^2\right)\|z^k - z^{k+1}\|^2 - 28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2.
\end{aligned}$$

Suppose that $r \geq 3L$ (which implies $L_{d_r} \leq 5L$), we have

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

$$\begin{aligned}
\frac{(2-\beta)r}{2\beta} - 2r\sigma_1 - \frac{r}{7\beta} - 56r\beta\sigma_1^2 &\geq \frac{6r}{7\beta} - \frac{7r}{2} - 126r\beta \\
&= \frac{r}{\beta}\left(\frac{6}{7} - \frac{7}{2}\beta - 126\beta^2\right) \geq \frac{4r}{7\beta}.
\end{aligned}$$

- As $\lambda \geq L$, we have

$$\alpha \leq \frac{1}{4L\zeta^2} = \frac{\lambda}{4L^2\zeta^2}\frac{L}{\lambda} \leq \frac{\lambda}{4L^2\zeta^2} \quad \text{and} \quad \frac{\eta^2}{4\alpha} = \frac{\alpha L^2\zeta^2}{4} \leq \frac{\lambda}{16}.$$

Moreover, due to $\beta \leq \min\left\{\frac{1}{28}, \frac{1}{8\alpha r\sigma_2^2}\right\}$, we can obtain

$$28r\beta\sigma_2^2\eta^2 \leq \frac{14\eta^2}{4\alpha} \leq \frac{14\lambda}{16} \quad \text{and} \quad \frac{2\lambda + r - 3L}{2} - 28r\beta\sigma_2^2\eta^2 - \frac{\eta^2}{4\alpha} \geq \frac{\lambda}{16}.$$

Together all pieces, we get

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ &\frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 - \\
&28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2.
\end{aligned}$$

The proof is complete.

Appendix D. Dual Error Bound Condition

We mainly discuss two cases — the general concave and the KŁ functions with the exponent θ, i.e., Assumption 2.2.

Lemma 7 *For any $z \in \mathbb{R}^n$, it follows that*

$$\|x_r^*(z) - x_r(y_+(z), z)\|^2 \leq \kappa\|y - y_+(z)\|, \tag{D.1}$$

where $y_+(z) := \mathrm{proj}_{\mathcal{Y}}\left(y + \alpha\nabla_y F_r(x_r(y,z), y, z)\right)$ and $\kappa := \frac{1+\alpha L\sigma_2+\alpha L}{\alpha(r-L)} \cdot \mathrm{diam}(\mathcal{Y})$.

Proof Recall that $y(z)$ is an arbitrary vector in $Y(z)$. By the strong convexity of $F_r(\cdot, y, z)$ we have

$$\begin{aligned}
F_r(x_r^*(z), y_+(z), z) - F_r(x_r(y_+(z), z), y_+(z), z) &\geq \frac{r-L}{2}\|x_r(y_+(z), z) - x_r^*(z)\|^2, \\
F_r(x_r(y_+(z), z), y(z), z) - F_r(x_r^*(z), y(z), z) &\geq \frac{r-L}{2}\|x_r(y_+(z), z) - x_r^*(z)\|^2,
\end{aligned}$$

which together with $F_r(x_r^*(z), y(z), z) \geq F_r(x_r^*(z), y_+(z), z)$ implies that

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

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

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

For simplicity, we denote the function

$$\xi(\cdot) := \alpha F_r(x_r(y_+(z), z), \cdot, z) - \left\langle \alpha\nabla_y F_r(x_r(y_+(z), z), y_+(z), z), \cdot \right\rangle -$$

$$\left\langle y_+(z) - y - \alpha\nabla_y F_r(x_r(y,z),y,z), \cdot \right\rangle.$$

Then we know that

$$\begin{aligned}\max_{y'\in\mathcal{Y}} \xi(y') &\le \alpha F_r(x_r(y_+(z),z),y_+(z),z) - \left\langle \alpha\nabla_y F_r(x_r(y_+(z),z),y_+(z),z), y_+(z)\right\rangle + \\ &\quad \max_{y'\in\mathcal{Y}} \left\langle y + \alpha\nabla_y F_r(x_r(y,z),y,z) - y_+(z), y'\right\rangle \\ &\le \xi(y_+(z)), \end{aligned} \tag{D.3}$$

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

$$\begin{aligned}&F_r(x_r(y_+(z),z),y(z),z) - F_r(x_r(y_+(z),z),y_+(z),z) \\ &\le \frac{1}{\alpha}\Big\langle y(z) - y_+(z), \alpha\nabla_y F_r(x_r(y_+(z),z),y_+(z),z) + y_+(z) - y - \alpha\nabla_y F_r(x_r(y,z),y,z)\Big\rangle \\ &\le \frac{1}{\alpha}\|y_+(z) - y(z)\|\|y - y_+(z)\| + L\|y_+(z) - y(z)\|(\|x_r(y_+(z),z) - x_r(y,z)\| + \|y_+(z) - y\|) \\ &\le \left(\frac{1}{\alpha} + L\sigma_2 + L\right)\|y_+(z) - y(z)\|\|y - y_+(z)\|, \end{aligned} \tag{D.4}$$

where the last inequality is from (A.3). Hence, we know from (D.2) and (D.4) that

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

which is the desired result. ■

Proof of Proposition 3.3 **Proof** At first, we define the function $\psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

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

where $\psi(\cdot,z)$ is $(r-L)$ strongly convex. As such, one has that

$$\frac{r-L}{2}\|x_r^*(z) - x_r(y_+(z),z)\|^2 \le \psi(x_r(y_+(z),z),z) - \psi(x_r^*(z),z). \tag{D.5}$$

Furthermore, note that,

$$\begin{aligned}\psi(x_r(y_+(z),z),z) - \psi(x_r^*(z),z) &\le \psi(x_r(y_+(z),z),z) - F_r(x_r(y_+(z),z),y_+(z),z) \\ &= \max_{y\in\mathcal{Y}} F(x_r(y_+(z),z),y) - F(x_r(y_+(z),z),y_+(z)), \end{aligned} \tag{D.6}$$

where the inequality holds as we have

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

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

Incorporating (D.5) and (D.6), we obtain the following intermediate relationship as a starting point:

$$\frac{r-L}{2}\|x_r^*(z) - x_r(y_+(z),z)\|^2 \le \max_{y\in\mathcal{Y}} F(x_r(y_+(z),z),y) - F(x_r(y_+(z),z),y_+(z)). \tag{D.7}$$

Next, by fully exploiting the KŁ exponent of the dual function, we target at quantifying the right-hand side term of (D.7) by $\|y - y_+(z)\|$.

By fully exploiting the KŁ condition, we have

$$\begin{aligned}
&\mu\left(\max_{y\in\mathcal{Y}} F(x_r(y_+(z),z),y) - F(x_r(y_+(z),z),y_+(z))\right)^\theta \\
\le\ & \operatorname{dist}(0, -\nabla_y F(x_r(y_+(z),z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) \\
\le\ & \operatorname{dist}(0, -\nabla_y F(x_r(y,z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) + \\
& \|\nabla_y F(x_r(y,z),y_+(z)) - \nabla_y F(x_r(y_+(z),z),y_+(z))\| \\
\overset{①}{\le}\ & \operatorname{dist}(0, -\nabla_y F(x_r(y,z),y_+(z)) + \partial\iota_{\mathcal{Y}}(y_+(z))) + L\sigma_2\|y - y_+(z)\| \\
\overset{②}{\le}\ & \left(\frac{1}{\alpha} + L + L\sigma_2\right)\|y_+(z) - y\|,
\end{aligned}$$

where ① follows from the gradient Lipschitz continuity property of $\nabla_y F(x,y)$ and (A.3); ② follows from the fact that projected gradient ascent method satisfied the so-called *relative error condition* [3] and recall that $y_+(z) = \operatorname{proj}_{\mathcal{Y}}(y + \alpha\nabla_y F(x_r(y,z),y))$.

Incorporating with (D.7), we have

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

■

Appendix E. Proof of Theorem 3.5

To prove the main theorem, we rely on the following lemma, which connects the decrease quantities and the game stationary concept.

Lemma 8 *Let $\epsilon \ge 0$. Suppose that*

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

then $\left(x^{k+1}, y^{k+1}\right)$ is a $\rho\epsilon$-game stationary solution for some $\rho > 0$.

Proof Based on Definition 2.3, we have to quantify the two terms $\|\nabla_x d_r(y^{k+1}, x^{k+1})\|$ and $\mathrm{dist}(0, \nabla_y F(x^{k+1}, y^{k+1}) + \partial \iota_{\mathcal{Y}}(y^{k+1}))$. From Lemma 3, we know

$$\begin{aligned}
&\|\nabla_x d_r(y^{k+1}, x^{k+1})\| \\
&= r\|x^{k+1} - x_r(y^{k+1}, x^{k+1})\| \\
&= r\|x^{k+1} - x_r(y^k, z^k) + x_r(y^k, z^k) - x_r(y^{k+1}, z^k) + x_r(y^{k+1}, z^k) - x_r(y^{k+1}, x^{k+1})\| \\
&\le r\left(\|x^{k+1} - x_r(y^k, z^k)\| + \|x_r(y^k, z^k) - x_r(y^{k+1}, z^k)\| + \|x_r(y^{k+1}, z^k) - x_r(y^{k+1}, x^{k+1})\|\right) \\
&\le r\left(\zeta\|x^{k+1} - x^k\| + \sigma_2\|y^k - y^{k+1}\| + \sigma_1\|x^{k+1} - z^k\|\right) \\
&\le r\left(\zeta\|x^{k+1} - x^k\| + \sigma_2(\eta\|x^{k+1} - x^k\| + \|y_+^k(z^k) - y^k\|) + \sigma_1\|x^{k+1} - z^k\|\right) \\
&\le r(\zeta + \sigma_2(\eta + 1) + \sigma_1)\epsilon.
\end{aligned}$$

On the other hand, we have

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

and the necessary optimality condition yields

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

Let $v := -\frac{1}{\alpha}(y^{k+1} - y^k) \in -\nabla_y F(x^{k+1}, y^{k+1}) + \partial\iota_{\mathcal{Y}}(y^{k+1})$. Then from (B.5) we have

$$\|v\| = \frac{1}{\alpha}\|y^{k+1} - y^k\| \le \frac{\eta}{\alpha}\|x^k - x^{k+1}\| + \frac{1}{\alpha}\|y_+^k(z^k) - y^k\| \le \frac{1}{\alpha}(1+\eta)\epsilon.$$

Hence, we finish the proof with $\rho = \max\left\{\frac{1}{\alpha}(1+\eta), r(\zeta + \sigma_2(\eta+1) + \sigma_1)\right\}$. ■

Proof of Theorem 3.5 First, for both general concave case and the case KŁ exponent $\theta \in (\frac{1}{2}, 1)$, we consider two phrases separately:
(1) there exists $k \in \{0, 1, \cdots, K-1\}$ such that

$$\begin{aligned}
&\frac{1}{2}\max\left\{\frac{\lambda}{16}\|x^k - x^{k+1}\|^2, \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2, \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2\right\} \\
&\le 28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2;
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{2}\max\left\{\frac{\lambda}{16}\|x^k - x^{k+1}\|^2, \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2, \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2\right\} \\
&\ge 28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2.
\end{aligned}$$

To begin with, we consider the first phrase. For the case KŁ exponent $\theta \in (\frac{1}{2}, 1)$, from Proposition 3.3 we know that

$$\|y^k - y_+^k(z^k)\|^2 \le 224r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2 \le 224r\beta\omega^2\|y^k - y_+^k(z^k)\|^{\frac{1}{\theta}}.$$

Then we have $\|y^k - y_+^k(z^k)\| \le \rho_1 \beta^{\frac{\theta}{2\theta-1}}$ where $\rho_1 := (224r\omega^2)^{\frac{\theta}{2\theta-1}}$. Additionally,

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

where $\rho_2 := 49\omega^2\rho_1^{\frac{1}{\theta}}$. We also have

$$\|x^{k+1} - x^k\|^2 \le \frac{448r\beta}{\lambda}\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2 \le \frac{448r\omega^2\beta}{\lambda}\|y^k - y_+^k(z^k)\|^{\frac{1}{\theta}} \le \rho_3 \beta^{\frac{2\theta}{2\theta-1}},$$

where $\rho_3 := \frac{448r\omega^2}{\lambda}\rho_1^{\frac{1}{\theta}}$. Combining the above inequalities we have

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

According to Lemma 8, we know that there exists $\rho > 0$ such that $\left(x^{k+1}, y^{k+1}\right)$ is a $\rho\max\{\rho_1\beta^{\frac{\theta}{2\theta-1}}, \rho_2^{\frac{1}{2}}\beta^{\frac{1}{4\theta-2}}, \rho_3^{\frac{1}{2}}\beta^{\frac{\theta}{2\theta-1}}\}$-solution. For the general concave case, from Lemma 7 we know that above analysis holds with $\theta = 1$ and κ replacing ω, i.e., $\left(x^{k+1}, y^{k+1}\right)$ is a $\rho\max\{\rho_1\beta, \rho_2^{\frac{1}{2}}\beta^{\frac{1}{2}}, \rho_3^{\frac{1}{2}}\beta\}$-solution.

In the second phrase, from Proposition 3.2 we have for any $k \in \{0, 1, \cdots, K-1\}$ that

$$\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \ge \frac{\lambda}{32}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{2r}{7\beta}\|z^k - z^{k+1}\|^2 \tag{E.1}$$

By the definition of $d_r(\cdot)$ and $p_r(\cdot)$, we have

$$\begin{aligned}F_r(x, y, z) &\ge \min_{x\in\mathcal{X}} F_r(x, y, z) = d_r(y, z), \\ p_r(z) &= \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} F_r(x, y, z) \ge \min_{x\in\mathcal{X}} F_r(x, y, z) = d_r(y, z), \\ p_r(z) &= \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left\{F(x, y) + \frac{r}{2}\|x - z\|^2\right\} \ge \underline{\Phi}.\end{aligned}$$

Hence, we have

$$\Phi_r(x, y, z) = p_r(z) + (F_r(x, y, z) - d_r(y, z)) + (p_r(z) - d_r(y, z)) \ge p_r(z) \ge \underline{\Phi}.$$

Consequently, it follows that

$$\begin{aligned}&\Phi_r(x^0, y^0, z^0) - \underline{\Phi} \\ \ge &\sum_{k=0}^{K-1} \Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\ \ge &\sum_{k=0}^{K-1} \frac{\lambda}{32}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{2r}{7\beta}\|z^k - z^{k+1}\|^2 \\ \ge &\sum_{k=0}^{K-1} \min\left\{\frac{\lambda}{32}, \frac{1}{16\alpha}, \frac{2r}{7}\right\}\left(\|x^k - x^{k+1}\|^2 + \|y^k - y_+^k(z^k)\|^2 + \beta\|z^k - x^{k+1}\|^2\right).\end{aligned}$$

Therefore, there exists a $k \in \{0, 1, \cdots, K-1\}$ such that

$$\max\left\{\|x^k - x^{k+1}\|^2, \|y^k - y_+^k(z^k)\|^2, \beta\|x^{k+1} - z^k\|^2\right\} \leq \frac{\Phi_r(x^0, y^0, z^0) - \underline{\Phi}}{\min\left\{\frac{\lambda}{32}, \frac{1}{16\alpha}, \frac{2r}{7}\right\} K}.$$

Hence, by Lemma 8 and $0 < \beta < 1$, $\left(x^{k+1}, y^{k+1}\right)$ is a $\sqrt{\frac{\rho(\Phi_r(x^0,y^0,z^0)-\underline{\Phi})}{\min\left\{\frac{\lambda}{32}, \frac{1}{16\alpha}, \frac{2r}{7}\right\} K\beta}}$ - game stationary solution. Then, we take $\beta = CK^{-\frac{2\theta-1}{2\theta}}$ (C is a constant) for the case KŁ exponent $\theta \in (\frac{1}{2}, 1)$ (resp. $\beta = CK^{-\frac{1}{2}}$ for the general concave case), and consequently the results for two phrases coincide which yields the desirable result $\mathcal{O}(K^{-\frac{1}{4\theta}})$ (resp. $\mathcal{O}(K^{-\frac{1}{4}})$).

Now, we consider the case KŁ exponent $\theta \in (0, \frac{1}{2}]$. Again based on Proposition 3.2, we have

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 - \\
& 28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2.
\end{aligned}$$

Armed with the dual error bound (see Proposition 3.3) with $\frac{1}{2\theta} \in [1, +\infty)$ and the boundedness of $\mathcal{Y}$, then

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

Thus,

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \left(\frac{1}{8\alpha} - 28r\beta\omega^2\mathrm{diam}(\mathcal{Y})^{\frac{1-2\theta}{\theta}}\right)\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2.
\end{aligned}$$

Since $\beta \leq \frac{\mathrm{diam}(\mathcal{Y})^{\frac{2\theta-1}{\theta}}}{448\alpha r\omega^2}$, we have

$$\begin{aligned}
&\Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2.
\end{aligned}$$

Similarly, we know the potential function is bounded below and

$$\begin{aligned}
&\Phi_r(x^0, y^0, z^0) - \underline{\Phi} \\
\geq\ & \sum_{k=0}^{K-1} \Phi_r(x^k, y^k, z^k) - \Phi_r(x^{k+1}, y^{k+1}, z^{k+1}) \\
\geq\ & \sum_{k=0}^{K-1} \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{16\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 \\
\geq\ & \sum_{k=0}^{K-1} \min\left\{\frac{\lambda}{16}, \frac{1}{16\alpha}, \frac{4r}{7}\right\}\left(\|x^k - x^{k+1}\|^2 + \|y^k - y_+^k(z^k)\|^2 + \beta\|z^k - x^{k+1}\|^2\right).
\end{aligned}$$

Therefore, due to $\beta < 1$, there exists a $k \in \{0, 1, \cdots, K-1\}$ such that

$$\max\left\{\|x^k - x^{k+1}\|^2, \|y^k - y_+^k(z^k)\|^2, \|x^{k+1} - z^k\|^2\right\} \leq \frac{(\Phi_r(x^0, y^0, z^0) - \underline{\Phi})}{\min\left\{\frac{\lambda}{16}, \frac{1}{16\alpha}, \frac{4r}{7}\right\} K\beta}.$$

Hence, by Lemma 8, $\left(x^{k+1}, y^{k+1}\right)$ is a $\sqrt{\frac{\rho(\Phi_r(x^0,y^0,z^0)-\underline{\Phi})}{\min\left\{\frac{\lambda}{16},\frac{1}{16\alpha},\frac{4r}{7}\right\}K\beta}}$ - game stationary solution which implies the result $\mathcal{O}(K^{-\frac{1}{2}})$. Finally, the case $\theta = 0$ is similar to $\theta = \frac{1}{2}$ with ω' replacing ω with regard to Proposition 3.3. The proof is complete.