1–33

Doubly Smoothed GDA: Global Convergent Algorithm for Constrained Nonconvex-Nonconcave Minimax Problems

Taoli Zheng TLZHENG@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

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

José Blanchet JOSE.BLANCHET@STANFORD.EDU
Stanford University

Jiajin Li JIAJINLI@STANFORD.EDU
Stanford University

Abstract

Nonconvex-nonconcave minimax optimization has been the focus of intense research over the last decade due to its broad applications in machine learning and operation research. Unfortunately, most existing algorithms cannot guarantee convergence and always suffer from *limit cycles*. Their global convergence relies on certain uncheckable conditions, including but not limited to the global Polyak-Łojasiewicz condition, the existence of a solution satisfying the weak Minty variational inequality and α-interaction dominant condition. In this paper, we develop the first provably convergent algorithm (i.e., *doubly smoothed gradient descent ascent method*) that gets rid of the limit cycle **without** requiring any additional conditions. We further show that the algorithm has an iteration complexity of $\mathcal{O}(\epsilon^{-4})$ to game stationary points, which matches the best iteration complexity of single-loop algorithms under nonconcave-concave settings. In sum, the algorithm presented here opens up a new path for designing provable algorithms for nonconvex-nonconcave minimax optimization problems.

Keywords: nonconvex-nonconcave minimax optimization; limit cycle; global convergence

1. Introduction

In this paper, we are interested in studying nonconvex-nonconcave minimax problems of the form

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

where $f:\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}$ is nonconvex in x and nonconcave in y, and $\mathcal{X}\in\mathbb{R}^n$, $\mathcal{Y}\in\mathbb{R}^d$ are convex compact sets. Such problems have found significant applications in machine learning and operation research, including generative adversarial networks training (Goodfellow et al., 2020; Arjovsky et al., 2017), adversarial training (Madry et al., 2017; Sinha et al., 2017), multi-agent reinforcement learning (Dai et al., 2018; Omidshafiei et al., 2017), and (distributionally) robust optimization (Ben-Tal et al., 2009; Delage and Ye, 2010; Levy et al., 2020; Gao et al., 2022; Bertsimas et al., 2011), to name a few.

In practice, with computational tractability in mind, we often use first-order (gradient-based) algorithms. Unfortunately, all existing first-order algorithms cannot be guaranteed to converge to game stationary points (i.e., see Definition 1 later) and, they can even suffer from the *limit cycle* issue. That is, the generated trajectories of all these algorithms will converge to cycling orbits that do not contain any game stationary point of f, see Figure 1(a), 1(d) and 1(e) in Section 2 as illustrative examples. Such *spurious* convergence phenomena arise from the minimax structure of (P) and have no counterpart in pure minimization problems. Conceptually, nonconvex-nonconcave minimax optimization problems can be understood as a seesaw game, which means no player inherently dominates the other. More explicitly, the key difficulty lies in adjusting the primal and dual updates to achieve a good balance. Nevertheless, all existing works try to add additional regularity conditions to restrict the problem class so that the developed algorithms can converge. Those regularity conditions are usually uncheckable in practice and can be grouped into three different categories. In the first category, a host of works add the global Polyak-Łojasiewicz (PŁ) condition on the dual function $f(x, \cdot)$ (Yang et al., 2022; Nouiehed et al., 2019; Doan, 2022; Yang et al., 2020) to ensure the convergence. However, this condition naturally avoids the main difficulty since the resulting the max function $\max_{y\in\mathcal{Y}} f(\cdot, y)$ is L-smooth. We can thus easily adopt the algorithms developed for nonconvex-concave minimax problems (Zhang et al., 2020; Yang et al., 2020, 2022; Lin et al., 2020a,b). In other words, the dual update has already been automatically controlled by the primal update. On another front, variational inequality provides a unified framework for the study of equilibrium/minimax problems (Nemirovski, 2004; Korpelevich, 1976; Gidel et al., 2018; Yoon and Ryu, 2021; Mertikopoulos et al., 2018). Most of the efforts in this line of work establish the convergence results under the weak minty monotone variational inequality condition (weak MVI) or its variants (Diakonikolas et al., 2021; Gorbunov et al., 2022; Mertikopoulos et al., 2018; Liu et al., 2021, 2019; Dang and Lan, 2015; Song et al., 2020; Böhm, 2022; Dou and Li, 2021). Nonetheless, weak MVI is hard to check in practice and is inapplicable to many functions (See two examples in Figure 1) because it is identical to check the existence of solutions for a certain variational inequality. Finally, a more closely related effort to our focus is the α-dominance condition developed in Grimmer et al. (2020); Hajizadeh et al. (2022). Intuitively, this condition is to characterize how the interaction part of f affects its saddle envelope (Attouch and Wets, 1983) and thus identifies the dominant variable as our prior information. Therefore, it is easy to design an algorithm to escape from the recurrence behavior. In short, all of these regularity conditions restrict the problem class.

In this paper, instead of further enlarging the problem class by adding some weaker regularity conditions, we try to give a pure algorithmic solution. That is, we develop a provably convergent algorithm for nonconvex-nonconcave minimax optimization problems without any regularity condition. The key insight here is to allow the algorithm automatically balance the primal and dual updates. To do so, we add two extrapolations (averaging) sequences for both primal and dual variables on the conventional gradient descent ascent (GDA). All hyperparameters, including the stepsize for gradient descent and ascent steps, and extrapolation parameters, are carefully and explicitly controlled by ensuring the sufficient decrease property of a novel Lyapunov function that we introduced in our paper. Furthermore, we show that the proposed doubly smoothed gradient descent ascent method (doubly smoothed GDA) converges to the game-stationary point at the iteration complexity $\mathcal{O}(\epsilon^{-4})$, which matches the best iteration complexity of single-loop algorithms under nonconcave-concave settings. Notably, our theoretical findings do not contradict the negative results in (Daskalakis et al., 2021) but, rather, give a positive answer from a complimentary perspective. As pointed out by (Daskalakis et al., 2021; Jin et al., 2020), finding a local minimax point of smooth nonconvex-nonconcave optimization problems is a PPAD-complete problem, and any first-order method requires exponentially many queries to function values and gradients. However, the stationary concept we achieved here is simply a game stationary point, which is just a necessary condition for local minimax equilibrium.

In sum, then, our paper provides a new path and theoretical framework for designing provable algorithms for nonconvex-nonconcave minimax optimization problems.

2. Motivating Examples

In this section, we demonstrate the effectiveness of the proposed doubly smoothed GDA on two illustrative examples that do not satisfy all regularity conditions (i.e., PŁ condition, weak MVI, and α-dominant condition). We refer the readers to Appendix H for details on how to check these conditions failed for these two examples. Thus, all existing algorithms cannot be guaranteed to converge theoretically.

"Forsaken" example The first one is "Forsaken" example considered in (Hsieh et al., 2021, Example 5.2), i.e.,

$$\min_{x} \max_{y} x(y - 0.45) + \phi(x) - \phi(y), \tag{1}$$

where $\phi(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$ and $\mathcal{X} = \mathcal{Y} = \{z : -1.5 \le z \le 1.5\}$. (Hsieh et al., 2021) claims that (1) contains two spurious limit cycles on the whole domain. The one that is closer to the optimal solution $[x^\star; y^\star] \simeq [0.08; 0.4]$ is unstable, which will potentially push the trajectories to fall into the recurrent orbit.

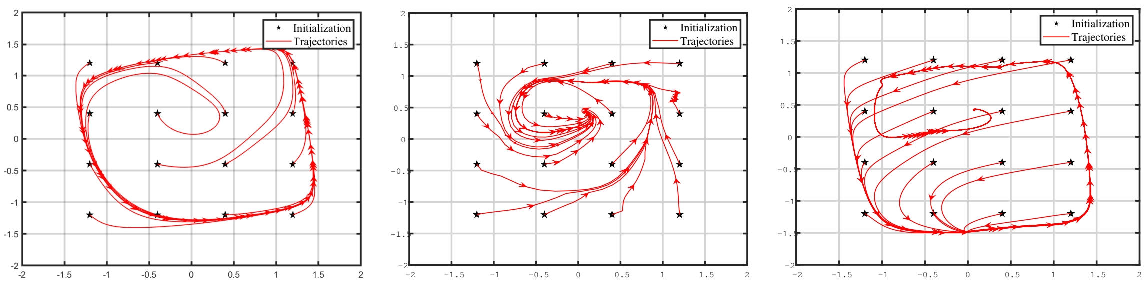
"Bilinearly-Coupled Minimax" example The other one is the "Bilinearly-Coupled Minimax" example (2) discussed in (Grimmer et al., 2020). That is,

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x) + 10xy - f(y), \tag{2}$$

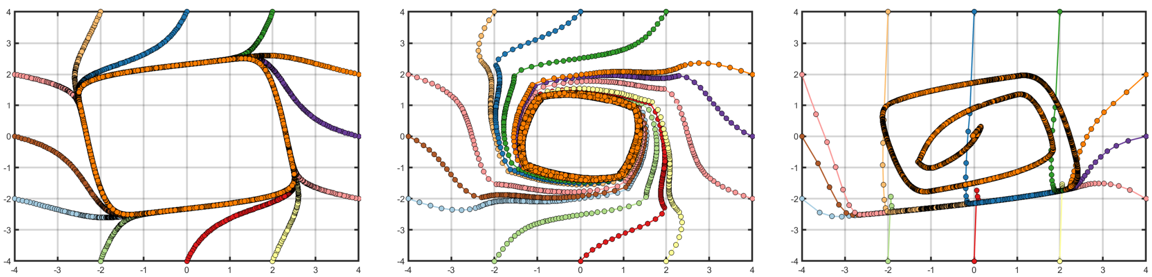
where $f(z) = (z + 1)(z - 1)(z + 3)(z - 3)$ and $\mathcal{X} = \mathcal{Y} = \{z : -4 \le z \le 4\}$. Notably, it is a well-representative example to showcase the limit cycle phenomenon as it breaks the α-dominant condition. When the bilinear intersection term between primal and dual variables x and y is moderate, we have no idea on either primal or dual variable dominates the other.

To showcase the convergence behaviors and effectiveness of the proposed doubly smoothed GDA, we compare it with other two state-of-the-art methods, that is, the damped extragradient method (Damped EGM) (Hajizadeh et al., 2022) and generalized curvature extragradient method (CurvatureEG+) (Pethick et al., 2022). Damped EGM is guaranteed to converge under one-sided α-dominant condition and CurvatureEG+ could converge under weak MVI condition, which is the weakest varitional inequality based condition as far as we know in the literature. All of these methods started from the same initialization for a fair comparison.

From Figure 1(f) and 1(c), we can easily observe that the proposed doubly smoothed GDA is able to successfully get rid of the limit cycle in these two examples. As we shall see later, this experiment result can fully corroborate our theoretical finding. Not surprisingly, Damped EGM suffered from the spurious cycling convergence phenomenon, see Figure 1(d) and 1(a) for details. Moreover, the same failure result was observed for the CurvatureEG+ on the "Bilinearly-Coupled Minimax" example in Figure 1(e). Although CurvatureEG+ can converge to a desired stationary point for "Forsaken" example (Pethick et al., 2022) (See Figure 1(b)), the global convergence of CurvatureEG+



(*a*) Damped EG (*b*) CurvatureEG+ (*c*) Doubly Smoothed GDA



(*d*) Damped EG (*e*) CurvatureEG+ (*f*) Doubly Smoothed GDA

Figure 1: Trajectories of different methods with various initialization for "Forsaken" example (see (a)-(c)) and "Bilinearly-Coupled Minimax" example (see (d)-(f)).

is still under some algorithmic dependent conditions, and unfortunately this condition is somehow uncheckable and inapplicable in practice. Moreover, CurvatureEG+ is computationally expensive since a backtracking line search procedure is performed in each step. Comparably, our algorithm is much more efficient since only the gradient step is executed at each iteration. More importantly, our algorithm can get rid of the limit cycle and enjoy global convergence without any conditions, which has been justified in these two examples.

3. Doubly Smoothed GDA

In this section, we propose our algorithm (i.e., doubly smoothed GDA) for solving (P). To start with, we introduce the blanket assumption which is needed throughout the paper.

Assumption 1 (Lipschitz gradient) *The function f is continuously differentiable and there exist positive constant $L_x, L_y > 0$ such that for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$*

$$\|\nabla_x f(x,y) - \nabla_x f(x',y')\| \le L_x(\|x - x'\| + \|y - y'\|),$$
$$\|\nabla_y f(x,y) - \nabla_y f(x',y')\| \le L_y(\|x - x'\| + \|y - y'\|).$$

For simplicity, we assume $L_y = tL_x = tL$ with $t > 0$.

For general smooth nonconvex-concave problems, a simple and natural algorithm is GDA, which suffers from oscillation even for the bilinear problem $\min_{x\in[-1,1]} \max_{y\in[-1,1]} xy$. (Zhang et al., 2020) proposed a smoothed GDA using Moreau-Yosida smoothing techniques to address the oscillation issue. Specifically, they introduce an auxiliary variable z and define a regularized function as follows:

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

The additional quadratic term smooths the primal update and consequently the algorithm can achieve a better trade-off between primal and dual updates. We adapt the smoothing technique to the nonconvex-nonconcave setting where the balance of primal and dual updates is not a trivial task. To tackle this problem, we also smooth the dual update by subtracting a quadratic term of dual variable and propose a new regularized function $F : \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ as

$$F(x, y, z, v) := f(x, y) + \frac{r_1}{2}\|x - z\|^2 - \frac{r_2}{2}\|y - v\|^2$$

with different smoothed parameters $r_1 > L_x$, $r_2 > L_y$ for x and y, respectively. Then, our doubly smoothed GDA is formally presented in Algorithm 1.

Algorithm 1 Doubly Smoothed GDA

Data: Initial x^0, y^0, z^0, v^0, stepsizes $\alpha, c > 0$, and extrapolation parameters β, μ
for $t = 0, \cdots, k$ **do**
    $x^{t+1} = \mathrm{proj}_{\mathcal{X}}(x^t - c\nabla_x F(x^t, y^t, z^t, v^t))$;
    $y^{t+1} = \mathrm{proj}_{\mathcal{Y}}(y^t + \alpha\nabla_y F(x^{t+1}, y^t, z^t, v^t))$;
    $z^{t+1} = z^t + \beta(x^{t+1} - z^t)$;
    $v^{t+1} = v^t + \mu(y^{t+1} - v^t)$;
end

The choice of r_1 and r_2 is crucial for the convergence of the algorithms in both theoretical and practical senses. In particular, when $r_1 = r_2$, it reduced to the *proximal-point mapping* proposed in (Liu et al., 2021) and inexact proximal point method (PPM) is only known convergent under certain VI conditions. Even with the exact computation of proximal mapping, PPM will diverge in the absence of regularity conditions (Grimmer et al., 2020). In contrast, with an unbalanced r_1 and r_2, our algorithm could always converge. The key insight here is to carefully adjust r_1 and r_2 to balance the primal-dual updates via ensuring the sufficient decrease property of a novel Lyapunov function introduced in our paper. In fact, as we will show later in Section 4, r_1 and r_2 are typically not equal theoretically and practically. The introduced two auxiliary variables z and v are also indispensable parts of convergence, which are updated by averaging steps. Intuitively, the exponential averaging applied to proximal variables z and v ensures they do not deviate too much from x and y, contributing to sequence stability.

We would like to highlight that the way we use the Moreau-Yosida smoothing techniques is a notable departure from usual. The smoothing techniques are commonly invoked in solving nonconvex-concave problems to achieve a better iteration complexity (Zhang et al., 2020; Li et al., 2022; Yang et al., 2022). However, we target at smoothing the primal and dual variables with different magnitudes to ensure global convergence.

4. Convergence Results

The convergence result of the proposed doubly smoothed GDA (i.e., Algorithm 1) will be discussed in this section. To illustrate the main result, we first list some notations in the following Table 1 and the stationary measure is provided in Definition 1.

Optimization problems	Function values	Optimal solutions
$\min_{x\in\mathcal{X}} F(x,y,z,v)$	$d(y,z,v)$	$x(y,z,v)$
$\max_{y\in\mathcal{Y}} F(x,y,z,v)$	$h(x,z,v)$	$y(x,z,v)$
$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x,y,z,v)$	$p(z,v)$	
$\min_{x\in\mathcal{X}} h(x,z,v)$	$p(z,v)$	$x(z,v)=x(y(z,v),z,v)$
$\max_{y\in\mathcal{Y}} d(y,z,v)$	$p(z,v)$	$y(z,v)=y(x(z,v),z,v)$
$\min_{z\in\mathbb{R}^n} p(z,v)$	$g(v)$	$z(v)$
$\max_{y\in\mathcal{Y}} \min_{(x,z)\in\mathcal{X}\times\mathbb{R}^n} F(x,y,z,v)$	$g(v)$	$y(v)=y(z(v),v)$
$\max_{v\in\mathbb{R}^d} g(v)$	$\bar{F}$	

Table 1: Notations

Definition 1 (Stationary measure) *The point $(x,y)\in\mathcal{X}\times\mathcal{Y}$ is said to be a ϵ-game stationary point* (ϵ-GS) *if*

$$\operatorname{dist}(\mathbf{0},\nabla_x f(x,y)+\partial\mathbf{1}_{\mathcal{X}}(x))\le\epsilon \quad \textit{and} \quad \operatorname{dist}(\mathbf{0},-\nabla_y f(x,y)+\partial\mathbf{1}_{\mathcal{Y}}(y))\le\epsilon.$$

Remark 1 *The definition of the game stationary point is a natural extension of the first-order stationary point in minimization problems. It is a necessary condition for local minimax point (Jin et al., 2020) and has been widely used in nonconvex-nonconcave optimization (Diakonikolas et al., 2021; Lee and Kim, 2021). There is another notion of ϵ-stationary point proposed in (Liu et al., 2021), which is called "nearly ϵ-stationary point". It is also derived from the pure minimization problem, which is an extension of the one proposed in (Davis and Drusvyatskiy, 2019) designed for weakly-convex functions. We will discuss the relationship among different stationary concepts in Appendix G and actually show these notions are equivalent.*

Inspired by Zhang et al. (2020); Li et al. (2022), we consider a novel Lyapunov function $\Phi:\mathbb{R}^n\times\mathbb{R}^d\times\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}$ defined as follows:

$$\Phi(x,y,z,v)=\underbrace{F(x,y,z,v)-d(y,z,v)}_{\text{Primal descent}}+\underbrace{p(z,v)-d(y,z,v)}_{\text{Dual ascent}}+\underbrace{p(z,v)-g(v)}_{\text{Proximal descent}}+\underbrace{\bar{F}-g(v)}_{\text{Proximal ascent}}+\bar{F}.$$

Actually, the Lyapunov function is highly related to the iterate updates. The primal update corresponds to the "primal descent" and gradient ascent in dual variable induces the "dual ascent" part. The averaging updates of proximal variables could be understood as an approximate gradient descent of $p(z,v)$ and an approximate gradient ascent of $g(v)$, resulting in the "proximal descent" and "proximal ascent" terms in the Lyapunov function. Compared with that in Zhang et al. (2020); Li et al. (2022), we have an additional "proximal ascent" term. It is introduced by the regularized term for dual variable in F and the update of proximal variable v. Essentially, the "nonconcavity" of $f(x,\cdot)$ brings the additional term. With this constructed Lyapunov function, we can establish the following sufficient decrease theorem as our first important result.

Theorem 1 (Sufficient decrease property) *Suppose that* $3(1+t)L \le r_1 \le 4(1+t)L$, $(\frac{t^2}{2+3t} + 4t+4)L \le r_2 \le (6t+4)L$ *with the parameters*

- (Stepsize) $c \in \left[\frac{1}{4(1+t)L}, \frac{1}{\max\{L+r_1, 4tL\}}\right]$, $\alpha \in \left[\frac{r_1-L}{3t^2L^2+(7tL+3r_2)(r_1-L)}, \frac{r_1-L}{3t^2L^2+(6tL+3r_2)(r_1-L)}\right]$;
- (Extrapolation) $\beta \in (0, \frac{1}{500(t+1)}]$, $\mu \in (0, \frac{1}{180(3t+2)}]$.

Then for any $t \ge 0$,

$$\begin{aligned}
&\Phi(x^t, y^t, z^t, v^t) - \Phi(x^{t+1}, y^{t+1}, z^{t+1}, v^{t+1}) \\
\ge\ & \frac{6r_1}{125}\|x^{t+1} - x^t\|^2 + \frac{2r_2}{25}\|y^t - y_+^t(z^t, v^t)\|^2 + \frac{r_2}{2\mu}\|v^t - v^{t+1}\|^2 + \frac{49r_1}{100\beta}\|z_+^t(v^t) - z^t\|^2 - \\
& 4r_2(t+2)\mu\|y(v^t) - y(z_+^t(v^t), v^t)\|^2,
\end{aligned}$$

where $y_+(z,v) := \mathrm{proj}_{\mathcal{Y}}(y + \alpha\nabla_y F(x(y,z,v), y, z, v))$ *and* $z_+(v) := z + \beta(x(y(z,v), z, v) - z)$.

We know that Φ is lower bounded by $\bar{F}$ by its construction, so the crux of establishing the subsequence convergence is to prove the decreasing property of the Lyapunov function. Although Theorem 1 quantifies the variation of the Lyapunov function values between two consecutive iterates, there is a negative error term $\|y(v^t) - y(z_+^t(v^t), v^t)\|$ that makes the decreasing property of Φ unclear. Therefore, we first characterize the negative error term by other positive terms and then prove the sufficient decrease property by bounding the coefficients. Conceptually, the error term is related to $\|z_+(v) - z(v)\|$ by the Lipschitz property of $y(z,v)$. However, $\|z_+(v) - z(v)\|$ may not be a suitable surrogate since it deviates from the existing positive terms. To remedy this, we note that $\|z - z_+(v)\| = 0$ implies z is the optimal solution to $\min_{z\in\mathbb{R}^n} p(z,v)$, i.e., $z = z(v)$, which provides the possibility of bounding the error term by $\|z - z_+(v)\|$. We provide the explicit form in the following proposition.

Proposition 1 (Proximal error bound) *Under the assumption is Theorem 1, for any* $z \in \mathbb{R}^n, v \in \mathbb{R}^d$, *it follows that*

$$\|y(z_+(v), v) - y(z(v), v)\|^2 \le \omega\|z - z_+(v)\|,$$

where $\omega := \frac{2tLr_1^2(tL^2(1-t) - L(tr_1+r_2) + r_1r_2)\,\mathrm{diam}(\mathcal{X})}{(r_2-tL)(r_1-L)^2(tL^2(2-t) - L(2tr_1+r_2) + r_1r_2)} > 0$[1].

Armed with Theorem 1 and Proposition 1, we can establish the main theorem concerning the iteration complexity of doubly smoothed GDA with respect to the following standard stationary measure for nonconvex-nonconcave minimax optimization problems.

Theorem 2 (Convergence theorem) *Under the assumption of Theorem 1, for any* $T > 0$ *there exists a* $t \in \{1, 2, \cdots, T\}$ *such that* (x^{t+1}, y^{t+1}) *is a* $\mathcal{O}(T^{-\frac{1}{4}})$-GS.

5. Related Works

There are three representative types of regularity conditions in the literature to restrict the problem class such that we can develop algorithms to get rid of the limit cycle.

1. The $\mathrm{diam}(\mathcal{X})$ denotes the diameter of the set $\mathcal{X}$.

Polyak-Łojasiewica (PŁ) condition The PŁ condition (3) was originally proposed by (Polyak, 1964) and is a crucial tool in unveiling linear convergence of first-order algorithms for pure minimization problems (Karimi et al., 2016). That is, the problem $\max_{x\in\mathbb{R}^d} h(x)$ has a nonempty solution set and a finite optimal value. There exist a constant $\mu > 0$ such that for any $x \in \mathbb{R}^d$,

$$\frac{1}{2}\|\nabla h(x)\|^2 \geq \mu(h(x) - \min_{x\in\mathbb{R}^d} h(x)). \tag{3}$$

There are a host of works trying to invoke the PŁ condition on the dual function $f(x, \cdot)$ (Yang et al., 2022; Nouiehed et al., 2019; Doan, 2022; Yang et al., 2020). Unfortunately, we would like to point out that this condition is too restrictive and inherently avoid the main difficulty in addressing general nonconvex-nonconcave minimax problems. With PŁ condition imposed on the dual function, the inner maximization value function $\phi(\cdot) = \max_{y\in\mathcal{Y}} f(\cdot, y)$ is L-smooth (Nouiehed et al., 2019, Lemma A.5). Thus, the dual update can naturally be controlled by the primal since we can regard minimax problems as pure smooth (weakly convex) minimization problems over x. However, for general cases, the inner value function ϕ will be even not Lipschitz. Recently, the authors of (Nouiehed et al., 2019) propose a so-called multi-step GDA method with the iteration complexity as $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-2})$. (Doan, 2022) further develops the single-loop two-timescale GDA method to better take the computational tractability into account and the complexity is improved to $\mathcal{O}(\epsilon^{-2})$. Following the smoothing (extrapolation) technique developed in (Zhang et al., 2020), (Yang et al., 2022) extends the proposed smoothed GDA to the stochastic setting and gets the iteration complexity as $\mathcal{O}(\epsilon^{-4})$.

Varitional Inequality (VI) Variational inequalities can be regarded as generalizations of minimax optimization problems (Dem'yanov and Pevnyi, 1972). In convex-concave minimax optimization, finding a saddle point is equivalent to solving the Stampacchia Variational Inequality (SVI):

$$\langle G(u^\star), u - u^\star\rangle \geq 0, \quad \forall u \in \mathcal{U}. \tag{4}$$

Here $u := [x; y]$, $u^\star$ is the optimal solution, and the operator G is a gradient operator: $G(u) := [\nabla_x f(x, y); -\nabla_y f(x, y)]$ with $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$. The solution to (4) is referred to as a strong solution to the VI corresponding to G and $\mathcal{U}$ (Hartman and Stampacchia, 1966). For the nonconvex-nonconcave minimax problem, without the monotonicity of G, the solution of SVI may not even exist. One alternative condition in the literature is to assume the existence of solutions $u^\star$ for Minty Variational Inequality (MVI):

$$\langle G(u), u - u^\star\rangle \geq 0, \quad \forall u \in \mathcal{U}. \tag{5}$$

The solution of (5) is called a weak solution of the VI (Facchinei and Kanzow, 2007). In the setting where G is continuous and monotone, the solution sets of (4) and (5) are equivalent. However, these two solution sets are different in general and a weak solution may not exist when a strong solution exists. A large number of literature have established the convergence results under the MVI condition or its variants (Diakonikolas et al., 2021; Gorbunov et al., 2022; Mertikopoulos et al., 2018; Liu et al., 2021, 2019; Dang and Lan, 2015; Song et al., 2020; Böhm, 2022; Dou and Li, 2021). Although MVI leads to convergence, it is hard to check in practice and is inapplicable to many functions (See two examples in Figure 1). A natural question is *Can we further relax the MVI condition to ensure convergence*? One possible way is to relax the nonnegative lower bound to be a negative one (Iusem et al., 2017; Lee and Kim, 2021; Cai et al., 2022; Cai and Zheng, 2022;

Diakonikolas et al., 2021), so-called the weak MVI condition:

$$\langle G(u), u - u^\star \rangle \geq -\frac{\rho}{2}\|G(u)\|^2, \quad \forall u \in \mathcal{U}. \tag{6}$$

Here, we restricted $\rho \in [0, \frac{1}{4L})$ and (Diakonikolas et al., 2021) proposed a Generalized extragradient method (Generalized EGM) with $\mathcal{O}(\epsilon^{-2})$ iteration complexity. To include a wider function class, (Pethick et al., 2022) enlarged the range of ρ to $[0, \frac{1}{L})$ and ρ can be larger if more curvature information of $f(x, y)$ is involved. However, for general smooth nonconvex-nonconcave problems, variant VI conditions are hard to check and ρ would easily violate the constraints. In this case, the proposed CurvatureEG+ still suffers from the limit cycle issue (See Figure 1(e)).

α-interaction dominant condition Another line of work is to impose the α-interaction dominant conditions (7a), (7b) on f. That is

$$\nabla^2_{xx} f(x,y) + \nabla^2_{xy} f(x,y)(\eta \boldsymbol{I} - \nabla^2_{yy} f(x,y))^{-1} \nabla^2_{yx} f(x,y) \succeq \alpha \boldsymbol{I}, \tag{7a}$$

$$-\nabla^2_{yy} f(x,y) + \nabla^2_{yx} f(x,y)(\eta \boldsymbol{I} + \nabla^2_{xx} f(x,y))^{-1} \nabla^2_{xy} f(x,y) \succeq \alpha \boldsymbol{I}. \tag{7b}$$

Intuitively, this condition is to characterize how the interaction part of $f(x, y)$ affects the landscape of saddle envelope $f_\eta(x, y) = \min_{z\in\mathcal{X}} \max_{v\in\mathcal{Y}} f(z, v) + \frac{\eta}{2}\|x - z\|^2 - \frac{\eta}{2}\|y - v\|^2$ (Attouch and Wets, 1983). We say α is in interaction dominant regimes if α in (7a), (7b) is a sufficiently large positive number and in interaction weak regimes when α is a small but nonzero positive number. The convergence results can only be guaranteed for these two regimes (Grimmer et al., 2020). Otherwise, the proposed Damped proximal point method (Damped PPM) may fall into the limit cycle or even diverge (See Figure 1(d)). Unfortunately, such conditions only hold with $\alpha = -\rho < 0$ for general smooth nonconvex-nonconcave function, which will dramatically restrict the problem class. Moreover, second-order information of f is required. For instance, if we choose Exponential Linear Units (ELU) with $a = 1$ (Clevert et al., 2015) as an active function in neural networks, f is L-smooth but not second-order differentiable. (Grimmer et al., 2020) studied the convergence for Damped PPM and showed that in the interaction dominate regimes, this algorithm converges with only one-sided dominance. In the interaction weak regime, their method also guarantees the local convergence with $\mathcal{O}(\log(\epsilon^{-1}))$. Taking computational efficiency into consideration, (Hajizadeh et al., 2022) developed the Damped EGM method which converges with $\mathcal{O}(\log(\epsilon^{-1}))$ iteration complexity under two-sided dominance conditions.

6. Conclusion

In this paper, we propose a doubly smoothed gradient descent ascent method (i.e., doubly smoothed GDA), which is the first provable single-loop algorithm for solving nonconvex-nonconcave constrained problems without any regularity condition. This algorithm is easy to implement and has an iteration complexity of $\mathcal{O}(\epsilon^{-4})$, which matches the best complexity result in general nonconvex-concave setting. Our algorithm opens a new line of studying convergence behaviors of nonconvex-nonconcave problems. One possible direction is to analyze the last-iterate convergence of our algorithm, which attracts a lot of attention in the area of minimax optimization. Another natural direction is to extend our algorithm to the stochastic setting so that we can tackle large-scale tasks in modern machine learning.

Acknowledgement

Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF grants 1915967 and 2118199. Anthony Man-Cho So is supported by in part by the Hong Kong RGC GRF project CUHK 14203920.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Hédy Attouch and Roger J-B Wets. A convergence theory for saddle functions. *Transactions of the American Mathematical Society*, 280(1):1–41, 1983.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Axel Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *arXiv preprint arXiv:2201.12247*, 2022.

Yang Cai and Weiqiang Zheng. Accelerated single-call methods for constrained min-max optimization. *arXiv preprint arXiv:2210.03096*, 2022.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2206.05248*, 2022.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

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

Cong D Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310, 2015.

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.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

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

Vladimir Fedorovich Dem'yanov and Aleksandr Borisovich Pevnyi. Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52, 1972.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

Thinh Doan. Convergence rates of two-time-scale gradient descent-ascent dynamics for solving nonconvex min-max problems. In *Learning for Dynamics and Control Conference*, pages 192–206. PMLR, 2022.

Zehao Dou and Yuanzhi Li. On the one-sided convergence of adam-type algorithms in non-convex non-concave min-max optimization. *arXiv preprint arXiv:2109.14213*, 2021.

Francisco Facchinei and Christian Kanzow. Generalized nash equilibrium problems. *4or*, 5(3): 173–210, 2007.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

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.

Eduard Gorbunov, Adrien Taylor, Samuel Horváth, and Gauthier Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. *arXiv preprint arXiv:2210.13831*, 2022.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020.

Saeed Hajizadeh, Haihao Lu, and Benjamin Grimmer. On the linear convergence of extra-gradient methods for nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2201.06167*, 2022.

Philip Hartman and Guido Stampacchia. On some non-linear elliptic differential-functional equations. *Acta mathematica*, 115:271–310, 1966.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR, 2021.

Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR, 2020.

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

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34:22588–22600, 2021.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

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.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth composite nonconvex-concave minimax optimization. *arXiv preprint arXiv:2209.10825*, 2022.

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, 2020a.

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

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *J. Mach. Learn. Res.*, 22:169–1, 2021.

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

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Arkadi Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

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.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

Jong-Shi Pang. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.

Thomas Pethick, Panagiotis Patrinos, Olivier Fercoq, Volkan Cevherå, et al. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

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.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $o(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR, 2021.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.

Appendix A. Organization of the Appendix

We organize the appendix as follows:

- Some useful Lipschitz error bounds are provided in Section B.
- The characterization of changes in the Lyapunov function between successive iterations is established in Section C.
- The proof of Theorem 1 is given in Section D.
- The proof of Proposition 1 is given in Section E.
- The proof of Theorem 2 is given in Section F.
- The quantitative relationship between different notions of stationary point is provided in Section G.
- The "weak MVI" and "α-interaction dominant" conditions of two examples in Figure 1 are checked in Section H.

Appendix B. Useful Lemmas

In this section, some technical lemmas are presented. We always assume that $r_1 > L_x$ and $r_2 > L_y$.

Lemma 1 *For any $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, $z \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$, it follows that*

$$\frac{r_1 - L_x}{2}\|x - x'\|^2 \le F(x', y, z, v) - F(x, y, z, v) - \langle \nabla_x F(x, y, z, v), x' - x\rangle \le \frac{L_x + r_1}{2}\|x - x'\|^2,$$
$$-\frac{L_y + r_2}{2}\|y - y'\|^2 \le F(x, y', z, v) - F(x, y, z, v) - \langle \nabla_y F(x, y, z, v), y' - y\rangle \le \frac{L_y - r_2}{2}\|y - y'\|^2.$$

Proof Since f is L-smooth (from the Assumption 1), we have

$$\begin{aligned} &-\frac{L_x}{2}\|x - x'\|^2 \le f(x', y) - f(x, y) - \langle \nabla_x f(x, y), x' - x\rangle \le \frac{L_x}{2}\|x - x'\|^2, \\ &-\frac{L_y}{2}\|y - y'\|^2 \le f(x, y') - f(x, y) - \langle \nabla_y f(x, y), y' - y\rangle \le \frac{L_y}{2}\|y - y'\|^2. \end{aligned} \tag{8}$$

On the other hand, we know that

$$\begin{aligned} &F(x', y, z, v) - F(x, y, z, v) - \langle \nabla_x F(x, y, z, v), x' - x\rangle \\ =& f(x', y) - f(x, y) - \langle \nabla_x f(x, y) + r_1(x - z), x' - x\rangle + \frac{r_1}{2}\|x' - z\|^2 - \frac{r_1}{2}\|x - z\|^2 \\ =& f(x', y) - f(x, y) - \langle \nabla_x f(x, y), x' - x\rangle + \frac{r_1}{2}\|x' - x\|^2 \end{aligned} \tag{9}$$

and similarly

$$\begin{aligned} &F(x, y', z, v) - F(x, y, z, v) - \langle \nabla_y F(x, y, z, v), y' - y\rangle \\ =& f(x, y') - f(x, y) - \langle \nabla_y f(x, y) - r_2(y - v), y' - y\rangle - \frac{r_2}{2}\|y' - v\|^2 + \frac{r_2}{2}\|y - v\|^2 \\ =& f(x, y') - f(x, y) - \langle \nabla_y f(x, y), y' - y\rangle - \frac{r_2}{2}\|y' - y\|^2. \end{aligned} \tag{10}$$

Combing (8), (9) and (10), we directly obtain the desired results. ■

Lemma 2 (Lipschitz type error bound conditions) *Suppose that $r_2 > (\frac{L_y}{r_1-L_x}+2)L_y$, then for any $x,x' \in \mathcal{X}$, $y,y' \in \mathcal{Y}$, $z,z' \in \mathbb{R}^n$ and $v,v' \in \mathbb{R}^d$. Then the following inequalities hold:*

(i) $\|x(y',z,v)-x(y,z,v)\| \le \sigma_1\|y'-y\|$,

(ii) $\|x(y,z',v)-x(y,z,v)\| \le \sigma_2\|z-z'\|$,

(iii) $\|x(z',v)-x(z,v)\| \le \sigma_2\|z-z'\|$,

(iv) $\|y(z,v)-y(z',v)\| \le \sigma_3\|z-z'\|$,

(v) $\|y(x,z,v)-y(x',z,v)\| \le \sigma_4\|x-x'\|$,

(vi) $\|y(x,z,v)-y(x,z,v')\| \le \sigma_5\|v-v'\|$,

(vii) $\|y(z,v)-y(z,v')\| \le \sigma_5\|v-v'\|$,

(viii) $\|y(v)-y(v')\| \le \sigma_5\|v-v'\|$,

where $\sigma_1 = \frac{L_y+r_1-L_x}{r_1-L_x}$, $\sigma_2 = \frac{r_1}{r_1-L_x}$, $\sigma_3 = \frac{\sigma_2(L_x+r_2-L_y\sigma_1-L_y)}{r_2-L_y\sigma_1-L_y}$, $\sigma_4 = \frac{L_x+r_2-L_y}{r_2-L_y}$, and $\sigma_5 = \frac{r_2}{r_2-L_y}$.

Proof (i) From Lemma 1, we know that

$$
\begin{aligned}
&F(x(y,z,v),y',z,v) - F(x(y',z,v),y',z,v) \ge \frac{r_1-L_x}{2}\|x(y,z,v)-x(y',z,v)\|^2,\\
&F(x(y,z,v),y',z,v) - F(x(y,z,v),y,z,v) \le \langle \nabla_y F(x(y,z,v),y,z,v), y'-y\rangle + \frac{L_y-r_2}{2}\|y-y'\|^2,\\
&F(x(y',z,v),y,z,v) - F(x(y',z,v),y',z,v) \le \langle \nabla_y F(x(y',z,v),y,z,v), y-y'\rangle + \frac{L_y+r_2}{2}\|y-y'\|^2,\\
&F(x(y,z,v),y,z,v) - F(x(y',z,v),y,z,v) \le \frac{L_x-r_1}{2}\|x(y,z,v)-x(y',z,v)\|^2.
\end{aligned}
$$

Combining above inequalities one has that

$$
\begin{aligned}
&(r_1-L_x)\|x(y,z,v)-x(y',z,v)\|^2\\
&\le \langle \nabla_y F(x(y,z,v),y,z,v) - \nabla_y F(x(y',z,v),y,z,v), y'-y\rangle + L_y\|y-y'\|^2\\
&\le L_y\|x(y',z,v)-x(y,z,v)\|\|y'-y\| + L_y\|y-y'\|^2,
\end{aligned}
$$

where the second inequality is from Cauchy-Schwarz inequality and L-smooth property. Let $\xi := \|x(y',z,v)-x(y,z,v)\|/\|y-y'\|$. Then it follows that

$$
\xi^2 \le \frac{L_y}{r_1-L_x} + \frac{L_y}{r_1-L_x}\xi.
$$

Consequently, utilizing AM-GM inequality we derive $\xi \le \frac{\sqrt{L_y^2+2r_1L_y-2L_xL_y}}{r_1-L_x} \le \frac{L_y+r_1-L_x}{r_1-L_x} = \sigma_1$.

(ii-iii) Again from Lemma 1, we know that

$$
\begin{aligned}
F(x(y,z,v),y,z',v) - F(x(y,z',v),y,z',v) &\geq \frac{r_1 - L_x}{2}\|x(y,z,v) - x(y,z',v)\|^2, \\
F(x(y,z,v),y,z',v) - F(x(y,z,v),y,z,v) &= \frac{r_1}{2}\langle z' + z - 2x(y,z,v), z' - z\rangle, \\
F(x(y,z,v),y,z,v) - F(x(y,z',v),y,z,v) &\leq \frac{L_x - r_1}{2}\|x(y,z,v) - x(y,z',v)\|^2, \\
F(x(y,z',v),y,z,v) - F(x(y,z',v),y,z',v) &= \frac{r_1}{2}\langle z + z' - 2x(y,z',v), z - z'\rangle.
\end{aligned}
$$

and consequently we have

$$
\begin{aligned}
(r_1 - L_x)\|x(y,z,v) - x(y,z',v)\|^2 &\leq r_1\langle x(y,z',v) - x(y,z,v), z' - z\rangle \\
&\leq r_1\|x(y,z',v) - x(y,z,v)\|\|z' - z\|,
\end{aligned}
$$

which completes the proof of (ii). Moreover, since $\max_{y\in\mathcal{Y}} F(\cdot,y,\cdot,\cdot)$ is $(r_1 - L)$-weakly convex in x, the similar argument leads to (iii).

(iv-v) From Lemma 1 and (ii) we know that

$$
\begin{aligned}
&F(x(y(z,v),z,v),y(z,v),z,v) - F(x(y(z,v),z,v),y(z',v),z,v) \geq \frac{r_2 - L_y}{2}\|y(z,v) - y(z',v)\|^2, \\
&F(x(y(z,v),z,v),y(z,v),z,v) - F(x(y(z,v),z',v),y(z,v),z,v) \\
\leq\ &\langle \nabla_x F(x(y(z,v),z,v),y(z,v),z,v), x(y(z,v),z,v) - x(y(z,v),z',v)\rangle + \frac{L_x - r_1}{2}\sigma_2^2\|z - z'\|^2, \\
&F(x(y(z,v),z',v),y(z,v),z,v) - F(x(y(z,v),z',v),y(z',v),z,v) \\
\leq\ &\langle \nabla_y F(x(y(z,v),z',v),y(z',v),z,v), y(z,v) - y(z',v)\rangle + \frac{L_y - r_2}{2}\|y(z,v) - y(z',v)\|^2, \\
&F(x(y(z,v),z',v),y(z',v),z,v) - F(x(y(z,v),z,v),y(z',v),z,v) \\
\leq\ &\langle \nabla_x F(x(y(z,v),z,v),y(z',v),z,v), x(y(z,v),z',v) - x(y(z,v),z,v)\rangle + \frac{r_1 + L_x}{2}\sigma_2^2\|z - z'\|^2.
\end{aligned}
$$

Armed with these inequalities, we conclude that

$$
\begin{aligned}
&(r_2 - L_y)\|y(z,v) - y(z',v)\|^2 \\
\leq\ &L_x\sigma_2^2\|z - z'\|^2 + \langle \nabla_y F(x(y(z,v),z',v),y(z',v),z,v), y(z,v) - y(z',v)\rangle \\
&\langle \nabla_x F(x(y(z,v),z,v),y(z',v),z,v), x(y(z,v),z',v) - x(y(z,v),z,v)\rangle - \\
&\langle \nabla_x F(x(y(z,v),z,v),y(z,v),z,v), x(y(z,v),z',v) - x(y(z,v),z,v)\rangle \\
\leq\ &L_x\sigma_2^2\|z - z'\|^2 + L_y\sigma_1\|y(z,v) - y(z',v)\|^2 + L_x\sigma_2\|z - z'\|\|y(z,v) - y(z',v)\|,
\end{aligned}
$$

where the last inequality is from Cauchy-Schwarz inequality with (i) and (ii). Thus,

$$
\|y(z,v) - y(z',v)\|^2 \leq \frac{L_x\sigma_2}{r_2 - L_y\sigma_1 - L_y}\|y(z',v) - y(z,v)\|\|z - z'\| + \frac{L_x\sigma_2^2}{r_2 - L_y\sigma_1 - L_y}\|z - z'\|^2,
$$

which implies (iii) that

$$\begin{aligned}\frac{\|y(z,v)-y(z',v)\|}{\|z-z'\|} &\le \frac{\sigma_2\sqrt{L_x^2+2r_2L_x-2L_xL_y-2L_xL_y\sigma_1}}{r_2-L_y\sigma_1-L_y}\\ &\le \frac{\sigma_2(L_x+r_2-L_y\sigma_1-L_y)}{r_2-L_y\sigma_1-L_y}=\sigma_3.\end{aligned}$$

Next, we consider (iv). Still from Lemma 1, we have,

$$\begin{aligned}
F(x,y(x,z,v),z,v)-F(x,y(x',z,v),z,v) &\ge \frac{r_2-L_y}{2}\|y(x,z,v)-y(x',z,v)\|^2,\\
F(x',y(x,z,v),z,v)-F(x',y(x',z,v),z,v) &\le \frac{L_y-r_2}{2}\|y(x,z,v)-y(x',z,v)\|^2,\\
F(x,y(x,z,v),z,v)-F(x',y(x,z,v),z,v) &\le \langle\nabla_xF(x',y(x,z,v),z,v),x-x'\rangle+\frac{L_x+r_1}{2}\|x-x'\|^2,\\
F(x',y(x',z,v),z,v)-F(x,y(x',z,v),z,v) &\le \langle\nabla_xF(x',y(x',z,v),z,v),x'-x\rangle+\frac{L_x-r_1}{2}\|x-x'\|^2.
\end{aligned}$$

Summing them up, we derive that

$$(r_2-L_y)\|y(x,z,v)-y(x',z,v)\|^2\le L_x\|x-x'\|^2+L_x\|x-x'\|\|y(x,z,v)-y(x',z,v)\|.$$

Let $\zeta := \|y(x,z,v)-y(x',z,v)\|/\|x-x'\|$. Then $\zeta^2 \le \frac{L_x}{r_2-L_y}+\frac{L_x}{r_2-L_y}\zeta$ and consequently $\zeta\le\frac{L_x+r_2-L_y}{r_2-L_y}=\sigma_4$.

(vi-viii) We know from Lemma 1 that

$$\begin{aligned}
&F(x,y(x,z,v),z,v)-F(x,y(x,z,v'),z,v)\ge\frac{r_2-L_y}{2}\|y(x,z,v)-y(x,z,v')\|^2,\\
&F(x,y(x,z,v),z,v)-F(x,y(x,z,v),z,v')=\frac{r_2}{2}\langle v'+v-2y(x,z,v),v'-v\rangle,\\
&F(x,y(x,z,v'),z,v')-F(x,y(x,z,v'),z,v)=\frac{r_2}{2}\langle v+v'-2y(x,z,v'),v-v'\rangle,\\
&F(x,y(x,z,v),z,v')-F(x,y(x,z,v'),z,v')\\
\le&\frac{L_y-r_2}{2}\|y(x,z,v)-y(x,z,v')\|^2+\langle\nabla_yF(x,y(x,z,v'),z,v'),y(x,z,v)-y(x,z,v')\rangle.
\end{aligned}$$

Armed with these inequalities, we conclude that

$$\begin{aligned}(r_2-L_y)\|y(x,z,v)-y(x,z,v')\|^2 &\le \langle\nabla_yF(x,y(x,z,v'),z,v'),y(x,z,v)-y(x,z,v')\rangle+\\ &\quad r_2\langle y(x,z,v)'-y(x,z,v),v'-v\rangle\\ &\le r_2\|y(x,z,v')-y(x,z,v)\|\|v-v'\|,\end{aligned}$$

which indicates the inequality (vi).

Since $d(\cdot,z,v)=\min_{x\in\mathcal{X}}F(x,\cdot,z,v)$ and $\ell(\cdot,v)=\min_{x\in\mathcal{X},z\in\mathbb{R}^n}F(x,\cdot,z,v)$ are (r_2-L_y)-strongly concave, similarly we can derive the Lipschitz property of $y(z,v)$ and $y(v)$. ■

Lemma 3 (**L-smooth property of dual function**) *For any fixed $z \in \mathbb{R}^n$, $v \in \mathbb{R}^d$, the dual function $d(\cdot, z, v)$ is continuously differentiable with the gradient $\nabla_y d(y, z, v) = \nabla_y F(x(y, z, v), y, z, v)$ and*

$$\|\nabla_y d(y, z, v) - \nabla_y d(y', z, v)\| \leq L_d \|y - y'\|,$$

where $L_d := L_y \sigma_1 + L_y + r_2$.

Proof Using Danskin's theorem, we know that $d(\cdot, z, v)$ is differentiable with $\nabla_y d(y, z, v) = \nabla_y F(x(y, z, v), y, z, v)$. Also, we know from the L-smooth property of f that

$$\begin{aligned}
\|\nabla_y d(y, z, v) - \nabla_y d(y', z, v)\| &= \|\nabla_y F(x(y, z, v), y, z, v) - \nabla_y F(x(y', z, v), y', z, v)\| \\
&\leq \|\nabla_y F(x(y, z, v), y, z, v) - \nabla_y F(x(y', z, v), y, z, v)\| + \\
&\quad \|\nabla_y F(x(y', z, v), y, z, v) - \nabla_y F(x(y', z, v), y', z, v)\| \\
&\leq L_y \|x(y', z, v) - x(y, z, v)\| + (L_y + r_2)\|y - y'\| \\
&\leq (L_y \sigma_1 + L_y + r_2)\|y - y'\| = L_d \|y - y'\|,
\end{aligned}$$

where the last inequality is due to the error bound in Lemma 2 (i). ∎

Recall that $y_+(z, v) = \text{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x(y, z, v), y, z, v))$. Incorporating the iterates of doubly smoothed GDA, we have the following error bounds.

Lemma 4 *For any $t \geq 0$, the following inequalities hold:*

(i) $\|x^{t+1} - x(y^t, z^t, v^t)\| \leq \sigma_6 \|x^{t+1} - x^t\|$,

(ii) $\|y^{t+1} - y(x^t, z^t, v^t)\| \leq \sigma_7 \|y^{t+1} - y^t\|$,

(iii) $\|y(z^t, v^t) - y^t\| \leq \sigma_8 \|y^t - y_+^t(z^t, v^t)\|$,

(iv) $\|y^{t+1} - y_+^t(z^t, v^t)\| \leq L_y \alpha \sigma_6 \|x^t - x^{t+1}\|$,

where $\sigma_6 = \frac{2cr_1+1}{cr_1 - cL_x}$, $\sigma_7 = \frac{2\alpha r_2+1}{\alpha r_2 - \alpha L_y}$ and $\sigma_8 = \frac{1+\alpha L_d}{\alpha(r_2 - L_y)}$.

Proof (i-ii) First, we consider (i) which is also called "primal error bound". Adopting the proof in (Pang, 1987, Theorem 3.1), we can easily derive that

$$\|x^t - x(y^t, z^t, v^t)\| \leq \frac{cL_x + cr_1 + 1}{cr_1 - cL_x}\|x^{t+1} - x^t\|,$$

which implies that

$$\|x^{t+1} - x(y^t, z^t, v^t)\| \leq \|x^{t+1} - x^t\| + \|x^t - x(y^t, z^t, v^t)\| \leq \frac{2cr_1 + 1}{cr_1 - cL_x}\|x^{t+1} - x^t\|.$$

Similarly, we can derive the "primal error bound" for y_t in the inequality (ii).

(iii) Let $u^t := y^t - y_+^t(z^t, v^t)$. Then it follows from the projection theorem to convex sets and the optimality condition that

$$\begin{aligned}
&\langle \alpha \nabla_y F(x(y^t, z^t, v^t), y^t, z^t, v^t) + u^t, y(z^t, v^t) - y^t + u^t \rangle \leq 0, \\
&\langle \alpha \nabla_y F(x(y(z^t, v^t), z^t, v^t), y(z^t, v^t), z^t, v^t), y^t - u^t - y(z^t, v^t) \rangle \leq 0.
\end{aligned}$$

Adding and rearranging above two inequalities, we have

$$\begin{aligned}
&\alpha\langle\nabla_y F(x(y(z^t,v^t),z^t,v^t),y(z^t,v^t),z^t,v^t)-\nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t),y^t-y(z^t,v^t)\rangle\\
&\le\langle u^t,y^t-y(z^t,v^t)-\alpha\nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t)+\alpha\nabla_y F(x(y(z^t,v^t),z^t,v^t),y(z^t,v^t),z^t,v^t)\rangle\\
&\le\|u^t\|\|y^t-y(z^t,v^t)\|(1+\alpha L_d),
\end{aligned}$$

where the last inequality is from Lemma 3. Also, note that $d(\cdot,z,v)=\min_{x\in\mathcal{X}}F(x,\cdot,z,v)=F(x(\cdot,z,v),\cdot,z,v)$ is (r_2-L_y)-strongly concave, then

$$\langle\nabla_y d(y(z^t,v^t),z^t,v^t)-\nabla_y d(y^t,z^t,v^t),z^t,v^t),y^t-y(z^t,v^t)\rangle\ge(r_2-L_y)\|y(z^t,v^t)-y^t\|^2.$$

Combining the upper and lower bound provided above, we get

$$\|y(z^t,v^t)-y^t\|\le\frac{1+\alpha L_d}{\alpha(r_2-L_y)}\|u^t\|.$$

(iv) Utilizing the inequality (i), we can further bound the desired term

$$\begin{aligned}
&\|y^{t+1}-y_+^t(z^t,v^t)\|\\
&=\|\operatorname{proj}_{\mathcal{Y}}(y^t+\alpha\nabla_y F(x^{t+1},y^t,z^t,v^t))-\operatorname{proj}_{\mathcal{Y}}(y^t+\alpha\nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t))\|\\
&\le\alpha\|\nabla_y F(x^{t+1},y^t,z^t,v^t)-\nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t)\|\\
&\le\alpha L_y\|x^{t+1}-x(y^t,z^t,v^t)\|\le L_y\alpha\sigma_6\|x^t-x^{t+1}\|.
\end{aligned}$$

The proof is complete. ■

Appendix C. Sufficient Decrease Lemmas

Lemma 5 (Primal descent) *For any $t\ge 0$, the following inequality holds:*

$$\begin{aligned}
F(x^t,y^t,z^t,v^t)\ge F(x^{t+1},y^{t+1},z^{t+1},v^{t+1})+\left(\frac{1}{c}-\frac{L_x+r_1}{2}\right)\|x^{t+1}-x^t\|^2+\\
\langle\nabla_y F(x^{t+1},y^t,z^t,v^t),y^t-y^{t+1}\rangle-\frac{L_y-r_2}{2}\|y^{t+1}-y^t\|^2+\\
\frac{2-\beta}{2\beta}r_1\|z^{t+1}-z^t\|^2+\frac{\mu-2}{2\mu}r_2\|v^{t+1}-v^t\|^2
\end{aligned}$$

Proof We firstly split the target into four parts as follows:

$$\begin{aligned}
&F(x^t,y^t,z^t,v^t)-F(x^{t+1},y^{t+1},z^{t+1},v^{t+1})\\
&=\underbrace{F(x^t,y^t,z^t,v^t)-F(x^{t+1},y^t,z^t,v^t)}_{①}+\underbrace{F(x^{t+1},y^t,z^t,v^t)-F(x^{t+1},y^{t+1},z^t,v^t)}_{②}+\\
&\quad\underbrace{F(x^{t+1},y^{t+1},z^t,v^t)-F(x^{t+1},y^{t+1},z^{t+1},v^t)}_{③}+\\
&\quad\underbrace{F(x^{t+1},y^{t+1},z^{t+1},v^t)-F(x^{t+1},y^{t+1},z^{t+1},v^{t+1})}_{④}.
\end{aligned}$$

As for ①, we have that

$$\begin{aligned} F(x^{t+1}, y^t, z^t, v^t) - F(x^t, y^t, z^t, v^t) &\leq \langle \nabla_x F(x^t, y^t, z^t, v^t), x^{t+1} - x^t \rangle + \frac{L_x + r_1}{2} \|x^{t+1} - x^t\|^2 \\ &\leq \left(-\frac{1}{c} + \frac{L_x + r_1}{2} \right) \|x^{t+1} - x^t\|^2, \end{aligned}$$

where the first inequality is from Lemma 1 and the second one is due to the projection update of x^{t+1}, i.e., $\langle x^t - c\nabla_x F(x^t, y^t, z^t, v^t) - x^{t+1}, x^t - x^{t+1} \rangle \leq 0$. Next, one has for the inequality ② that

$$F(x^{t+1}, y^{t+1}, z^t, v^t) - F(x^{t+1}, y^t, z^t, v^t) \leq \langle \nabla_y F(x^{t+1}, y^t, z^t, v^t), y^{t+1} - y^t \rangle + \frac{L_y - r_2}{2} \|y^{t+1} - y^t\|^2.$$

For ③, it follows that

$$\begin{aligned} F(x^{t+1}, y^{t+1}, z^t, v^t) - F(x^{t+1}, y^{t+1}, z^{t+1}, v^t) &= \frac{r_1}{2} (\|x^{t+1} - z^t\|^2 - \|x^{t+1} - z^{t+1}\|^2) \\ &= \frac{2 - \beta}{2\beta} r_1 \|z^{t+1} - z^t\|^2. \end{aligned}$$

Here, the second equality is from the update of z^{t+1}, i.e, $z^{t+1} = z^t + \beta(x^{t+1} - z^t)$. Similarly, we consider ④ as follow:

$$\begin{aligned} F(x^{t+1}, y^{t+1}, z^{t+1}, v^t) - F(x^{t+1}, y^{t+1}, z^{t+1}, v^{t+1}) &= \frac{r_2}{2} (\|y^{t+1} - v^{t+1}\|^2 - \|y^{t+1} - v^t\|^2) \\ &= \frac{\mu - 2}{2\mu} r_2 \|v^{t+1} - v^t\|^2. \end{aligned}$$

Combine all above bounds and then it leads to the conclusion. ■

Lemma 6 (Dual ascent) *For any $t \geq 0$, the following inequality holds:*

$$\begin{aligned} d(y^{t+1}, z^{t+1}, v^{t+1}) \geq d(y^t, z^t, v^t) &+ \frac{(2 - \mu) r_2}{2\mu} \|v^{t+1} - v^t\|^2 + \\ &\frac{r_1}{2} \langle z^{t+1} + z^t - 2x(y^{t+1}, z^{t+1}, v^t), z^{t+1} - z^t \rangle + \\ &\langle \nabla_y F(x(y^t, z^t, v^t), y^t, z^t, v^t), y^{t+1} - y^t \rangle - \frac{L_d}{2} \|y^{t+1} - y^t\|^2 \end{aligned}$$

Proof The difference of the update for the dual function is controlled by following three parts:

$$\begin{aligned} &d(y^{t+1}, z^{t+1}, v^{t+1}) - d(y^t, z^t, v^t) \\ = \;& \underbrace{d(y^{t+1}, z^{t+1}, v^{t+1}) - d(y^{t+1}, z^{t+1}, v^t)}_{①} + \underbrace{d(y^{t+1}, z^{t+1}, v^t) - d(y^{t+1}, z^t, v^t)}_{②} + \\ &\underbrace{d(y^{t+1}, z^t, v^t) - d(y^t, z^t, v^t)}_{③}. \end{aligned}$$

For the first part,

$$①=\frac{r_2}{2}\left(\|y^{t+1}-v^t\|^2-\|y^{t+1}-v^{t+1}\|^2\right)=\frac{(2-\mu)r_2}{2\mu}\|v^{t+1}-v^t\|^2.$$

For the second part,

$$\begin{aligned}②&=F(x(y^{t+1},z^{t+1},v^t),y^{t+1},z^{t+1},v^t)-F(x(y^{t+1},z^t,v^t),y^{t+1},z^t,v^t)\\&\geq F(x(y^{t+1},z^{t+1},v^t),y^{t+1},z^{t+1},v^t)-F(x(y^{t+1},z^{t+1},v^t),y^{t+1},z^t,v^t)\\&=\frac{r_1}{2}(\|x(y^{t+1},z^{t+1},v^t)-z^{t+1}\|^2-\|x(y^{t+1},z^{t+1},v^t)-z^t\|^2)\\&=\frac{r_1}{2}\langle z^{t+1}+z^t-2x(y^{t+1},z^{t+1},v^t),z^{t+1}-z^t\rangle.\end{aligned}$$

Finally, consider the third part,

$$\begin{aligned}③&\geq\langle\nabla_y d(y^t,z^t,v^t),y^{t+1}-y^t\rangle-\frac{L_d}{2}\|y^{t+1}-y^t\|^2\\&=\langle\nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t),y^{t+1}-y^t\rangle-\frac{L_d}{2}\|y^{t+1}-y^t\|^2.\end{aligned}$$

Combining above inequalities finishes the proof. ■

Lemma 7 (Proximal descent) *For all $t\geq 0$, the following inequality holds:*

$$\begin{aligned}p(z^t,v^t)\geq p(z^{t+1},v^{t+1})&-\frac{r_1}{2}\langle z^t+z^{t+1}-2x(y(z^{t+1},v^{t+1}),z^t,v^t),z^{t+1}-z^t\rangle+\\&\frac{r_2}{2}\langle v^t+v^{t+1}-2y(z^{t+1},v^{t+1}),v^{t+1}-v^t\rangle\end{aligned}$$

Proof From Sion's minmax theorem (Sion, 1958), we have

$$p(z,v)=\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}F(x,y,z,v)=\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}F(x,y,z,v)=\max_{y\in\mathcal{Y}}d(y,z,v),$$

Thus, it follows that

$$\begin{aligned}p(z^t,v^t)-p(z^{t+1},v^{t+1})&=d(y(z^t,v^t),z^t,v^t)-d(y(z^{t+1},v^{t+1}),z^{t+1},v^{t+1})\\&\geq d(y(z^{t+1},v^{t+1}),z^t,v^t)-d(y(z^{t+1},v^{t+1}),z^{t+1},v^{t+1})\\&\geq F(x(y(z^{t+1},v^{t+1}),z^t,v^t),y(z^{t+1},v^{t+1}),z^t,v^t)-\\&\quad F(x(y(z^{t+1},v^{t+1}),z^t,v^t),y(z^{t+1},v^{t+1}),z^{t+1},v^{t+1})\\&=\frac{r_1}{2}\langle z^t+z^{t+1}-2x(y(z^{t+1},v^{t+1}),z^t,v^t),z^t-z^{t+1}\rangle-\\&\quad\frac{r_2}{2}\langle v^t+v^{t+1}-2y(z^{t+1},v^{t+1}),v^t-v^{t+1}\rangle.\end{aligned}$$

The proof is complete. ■

Lemma 8 (Proximal ascent) *For all $t\geq 0$, the following inequality holds:*

$$g(v^{t+1})\geq g(v^t)+\frac{r_2}{2}\langle v^t+v^{t+1}-2y(z(v^{t+1}),v^t),v^t-v^{t+1}\rangle. \tag{11}$$

Proof From Sion's minmax theorem (Sion, 1958), we have

$$\begin{aligned} g(v) &= \min_z p(z,v) = \min_z \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x,y,z,v) \\ &= \min_z \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} F(x,y,z,v) = \min_z \max_{y\in\mathcal{Y}} d(y,z,v) \\ &= \min_z d(y(z,v),z,v) = \min_z F(x(y(z,v),z,v),y(z,v),z,v), \end{aligned}$$

Thus, it follows that

$$\begin{aligned} g(v^{t+1}) - g(v^t) &\geq d(y(z(v^{t+1}),v^{t+1}),z(v^{t+1}),v^{t+1}) - d(y(z(v^{t+1}),v^t),z(v^{t+1}),v^t) \\ &\geq d(y(z(v^{t+1}),v^t),z(v^{t+1}),v^{t+1}) - d(y(z(v^{t+1}),v^t),z(v^{t+1}),v^t) \\ &\geq F(x(y(z(v^{t+1}),v^t),z(v^{t+1}),v^{t+1}),y(z(v^{t+1}),v^t),z(v^{t+1}),v^{t+1}) - \\ &\quad F(x(y(z(v^{t+1}),v^t),z(v^{t+1}),v^{t+1}),y(z(v^{t+1}),v^t),z(v^{t+1}),v^t) \\ &= \frac{r_2}{2}\langle v^t + v^{t+1} - 2y(z(v^{t+1}),v^t), v^t - v^{t+1}\rangle. \end{aligned}$$

The proof is complete. ■

Appendix D. Proof of Theorem 1

From the results in Lemma 5, 6, 7 and 8, we know that

$$\begin{aligned} \Phi(x^t,y^t,z^t,v^t) \geq\ & \Phi(x^{t+1},y^{t+1},z^{t+1},v^{t+1}) + \left(\frac{1}{c} - \frac{L_x + r_1}{2}\right)\|x^{t+1} - x^t\|^2 - \\ & \left(\frac{L_y - r_2}{2} + L_d\right)\|y^{t+1} - y^t\|^2 + \frac{(2-\beta)r_1}{2\beta}\|z^{t+1} - z^t\|^2 + \\ & \frac{(2-\mu)r_2}{2\mu}\|v^{t+1} - v^t\|^2 + \underbrace{\langle \nabla_y F(x^{t+1},y^t,z^t,v^t), y^{t+1} - y^t\rangle}_{①} + \\ & \underbrace{2\langle \nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t) - \nabla_y F(x^{t+1},y^t,z^t,v^t), y^{t+1} - y^t\rangle}_{②} + \\ & \underbrace{2r_1\langle x(y(z^{t+1},v^{t+1}),z^t,v^t) - x(y^{t+1},z^{t+1},v^t), z^{t+1} - z^t\rangle}_{③} + \\ & \underbrace{2r_2\langle y(z(v^{t+1}),v^t) - y(z^{t+1},v^{t+1}), v^{t+1} - v^t\rangle}_{④} \end{aligned}$$

For the part ①, using projection update of y^{t+1}, we have

$$\langle \nabla_y F(x^{t+1},y^t,z^t,v^t), y^{t+1} - y^t\rangle \geq \frac{1}{\alpha}\|y^t - y^{t+1}\|^2. \tag{12}$$

The part ② is due to Lipschitz gradient property (see Assumption 1) and error bounds in Lemma 2:

$$\begin{aligned} & 2\langle \nabla_y F(x(y^t,z^t,v^t),y^t,z^t,v^t) - \nabla_y F(x^{t+1},y^t,z^t,v^t), y^{t+1} - y^t\rangle \\ \geq & -2L_y\|x(y^t,z^t,v^t) - x^{t+1}\|\|y^{t+1} - y^t\| \\ \geq & -L_y\sigma_6^2\|y^{t+1} - y^t\|^2 - L_y\sigma_6^{-2}\|x(y^t,z^t,v^t) - x^{t+1}\|^2 \\ \geq & -L_y\sigma_6^2\|y^{t+1} - y^t\|^2 - L_y\|x^{t+1} - x^t\|^2. \end{aligned} \tag{13}$$

As for the part ③, for any $\kappa > 0$ it follows that

$$\begin{aligned}
&2r_1\langle x(y(z^{t+1}, v^{t+1}), z^t, v^t) - x(y^{t+1}, z^{t+1}, v^t), z^{t+1} - z^t\rangle \\
=& 2r_1\langle x(y(z^{t+1}, v^{t+1}), z^t, v^t) - x(y(z^{t+1}, v^{t+1}), z^{t+1}, v^t), z^{t+1} - z^t\rangle + \\
&2r_1\langle x(y(z^{t+1}, v^{t+1}), z^{t+1}, v^t) - x(y^{t+1}, z^{t+1}, v^t), z^{t+1} - z^t\rangle \\
\geq& - 2r_1\sigma_2\|z^{t+1} - z^t\|^2 - \frac{r_1}{\kappa}\|z^{t+1} - z^t\|^2 - r_1\kappa\|x(y(z^{t+1}, v^{t+1}), z^{t+1}, v^t) - x(y^{t+1}, z^{t+1}, v^t)\|^2,
\end{aligned} \tag{14}$$

where the inequality is from the Cauchy-Schwarz inequality and AM-GM inequality. Similarly, we provide a lower bound for the part ④ with any $\kappa_1 > 0$ as follows:

$$\begin{aligned}
&2r_2\langle y(z(v^{t+1}), v^t) - y(z^{t+1}, v^{t+1}), v^{t+1} - v^t\rangle \\
=& 2r_2\langle y(z(v^{t+1}), v^t) - y(z(v^{t+1}), v^{t+1}), v^{t+1} - v^t\rangle + \\
&2r_2\langle y(z(v^{t+1}), v^{t+1}) - y(z^{t+1}, v^{t+1}), v^{t+1} - v^t\rangle \\
\geq& \ - 2r_2\sigma_5\|v^{t+1} - v^t\|^2 - \frac{r_2}{\kappa_1}\|v^{t+1} - v^t\|^2 - r_2\kappa_1\|y(z(v^{t+1}), v^{t+1}) - y(z^{t+1}, v^{t+1})\|^2.
\end{aligned} \tag{15}$$

Hence,

$$\begin{aligned}
\Phi(x^t, y^t, z^t, v^t) \geq& \Phi(x^{t+1}, y^{t+1}, z^{t+1}, v^{t+1}) + \left(\frac{1}{c} - \frac{L_x + r_1}{2} - L_y\right)\|x^{t+1} - x^t\|^2 + \\
&\left(\frac{1}{\alpha} - \frac{L_y - r_2}{2} - L_d - L_y\sigma_6^2\right)\|y^{t+1} - y^t\|^2 + r_1\left(\frac{2-\beta}{2\beta} - 2\sigma_2 - \frac{1}{\kappa}\right)\|z^{t+1} - z^t\|^2 \\
&+ r_2\left(\frac{2-\mu}{2\mu} - 2\sigma_5 - \frac{1}{\kappa_1}\right)\|v^{t+1} - v^t\|^2 \\
&- \underbrace{r_1\kappa\|x(y(z^{t+1}, v^{t+1}), z^{t+1}, v^t) - x(y^{t+1}, z^{t+1}, v^t)\|^2}_{⑤} \\
&- \underbrace{r_2\kappa_1\|y(z(v^{t+1}), v^{t+1}) - y(z^{t+1}, v^{t+1})\|^2}_{⑥}.
\end{aligned}$$

Next, we focus on the two negative terms. Following the fact $x(z, v) = x(y(z, v), z, v)$ and $x(y, z, v) = x(y, z, v')$, the inequality ⑤ is bounded as follow:

$$\begin{aligned}
&\|x(y(z^{t+1}, v^{t+1}), z^{t+1}, v^t) - x(y^{t+1}, z^{t+1}, v^t)\|^2 \\
=& \|x(z^{t+1}, v^{t+1}) - x(y^{t+1}, z^{t+1}, v^t)\|^2 \\
\leq& 4\|x(z^{t+1}, v^{t+1}) - x(z^t, v^t)\|^2 + 4\|x(z^t, v^t) - x(y_+^t(z^t, v^t), z^t, v^t)\|^2 + \\
&4\|x(y_+^t(z^t, v^t), z^t, v^t) - x(y^{t+1}, z^t, v^t)\|^2 + 4\|x(y^{t+1}, z^t, v^t) - x(y^{t+1}, z^{t+1}, v^t)\|^2 \\
\leq& 8\sigma_2^2\|z^{t+1} - z^t\|^2 + 8\sigma_1^2\|y(z^t, v^{t+1}) - y(z^t, v^t)\|^2 + 4\sigma_1^2\|y(z^t, v^t) - y_+^t(z^t, v^t)\|^2 + \\
&4\sigma_1^2\|y^{t+1} - y_+^t(z^t, v^t)\|^2 + 4\sigma_2^2\|z^t - z^{t+1}\|^2 \\
\leq& 12\sigma_2^2\|z^{t+1} - z^t\|^2 + 8\sigma_1^2\sigma_5^2\|v^t - v^{t+1}\|^2 + 4\sigma_1^2(1+\sigma_8)^2\|y^t - y_+^t(z^t, v^t)\|^2 + 4\sigma_1^2L_y^2\alpha^2\sigma_6^2\|x^{t+1} - x^t\|^2.
\end{aligned}$$

For the inequality ⑥, noting that $y(z(v),v)=y(v)$ and it follows that

$$
\begin{aligned}
&\|y(z(v^{t+1}),v^{t+1})-y(z^{t+1},v^{t+1})\|^2\\
&=\|y(v^{t+1})-y(z^{t+1},v^{t+1})\|^2\\
&\leq 4\|y(v^{t+1})-y(v^t)\|^2+4\|y(v^t)-y(z_+^t(v^t),v^t)\|^2+4\|y(z_+^t(v^t),v^t)-y(z^{t+1},v^t)\|^2+\\
&\quad 4\|y(z^{t+1},v^t)-y(z^{t+1},v^{t+1})\|^2\\
&\leq 4\sigma_5^2\|v^t-v^{t+1}\|^2+4\|y(v^t)-y(z_+^t(v^t),v^t)\|^2+4\sigma_3^2\|z_+^t(v^t)-z^{t+1}\|^2+4\sigma_5^2\|v^t-v^{t+1}\|^2\\
&\leq 8\sigma_5^2\|v^t-v^{t+1}\|^2+4\|y(v^t)-y(z_+^t(v^t),v^t)\|^2+12\sigma_3^2\beta^2\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2+\\
&\quad 12\sigma_3^2\beta^2\sigma_6^2\|x^t-x^{t+1}\|^2,
\end{aligned}
$$

where the third inequality comes from $z_+^t(v^t)=z^t+\beta(x(y(z^t,v^t),z^t,v^t)-z^t)$ and $z^{t+1}=z^t+\beta(x^{t+1}-z^t)$, which leads to

$$
\begin{aligned}
&\|z^{t+1}-z_+^t(v^t)\|^2\\
&=\beta^2\|x(z^t,v^t)-x^{t+1}\|^2\\
&\leq 3\beta^2\|x(z^t,v^t)-x(y_+^t(z^t,v^t),z^t,v^t)\|^2+3\beta^2\|x(y^t,z^t,v^t)-x(y_+^t(z^t,v^t),z^t,v^t)\|^2+\\
&\quad 3\beta^2\|x^{t+1}-x(y^t,z^t,v^t)\|^2\\
&\leq 3\beta^2\sigma_1^2(\sigma_8+1)^2\|y^t-y_+^t(z^t,v^t)\|^2+3\beta^2\sigma_1^2\|y^t-y_+^t(z^t,v^t)\|^2+3\beta^2\sigma_6^2\|x^t-x^{t+1}\|^2\\
&=3\beta^2\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2+3\beta^2\sigma_6^2\|x^t-x^{t+1}\|^2.
\end{aligned}
$$

Moreover, to make terms in the upper bound being consistent, we derive that

$$
\begin{aligned}
\|y^{t+1}-y^t\|^2&\geq\frac{1}{2}\|y^t-y_+^t(z^t,v^t)\|^2-\|y^{t+1}-y_+^t(z^t,v^t)\|^2\\
&\geq\frac{1}{2}\|y^t-y_+^t(z^t,v^t)\|^2-L_y^2\alpha^2\sigma_6^2\|x^t-x^{t+1}\|^2,
\end{aligned}
$$

and also

$$
\begin{aligned}
\|z^{t+1}-z^t\|^2&\geq\frac{1}{2}\|z^t-z_+^t(v^t)\|^2-\|z^{t+1}-z_+^t(v^t)\|^2\\
&\geq\frac{1}{2}\|z^t-z_+^t(v^t)\|^2-3\beta^2\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2-3\beta^2\sigma_6^2\|x^t-x^{t+1}\|^2.
\end{aligned}
$$

Summing the above inequalities up and let $s_1:=\frac{1}{c}-\frac{L_x+r_1}{2}-L_y$, $s_2:=\frac{1}{\alpha}-\frac{L_y-r_2}{2}-L_d-L_y\sigma_6^2$, $s_3:=r_1\left(\frac{2-\beta}{2\beta}-2\sigma_2-\frac{1}{\kappa}\right)$ and $s_4:=r_2\left(\frac{2-\mu}{2\mu}-2\sigma_5-\frac{1}{\kappa_1}\right)$, then we have

$$
\begin{aligned}
&\Phi_{r_1,r_2}(x^t,y^t,z^t,v^t)-\Phi_{r_1,r_2}(x^{t+1},y^{t+1},z^{t+1},v^{t+1})\\
&\geq s_1\|x^{t+1}-x^t\|^2+s_2\|y^{t+1}-y^t\|^2+s_3\|z^{t+1}-z^t\|^2+s_4\|v^{t+1}-v^t\|^2-12r_1\kappa\sigma_2^2\|z^{t+1}-z^t\|^2-\\
&\quad 8r_1\kappa\sigma_1^2\sigma_5^2\|v^t-v^{t+1}\|^2-4r_1\kappa\sigma_1^2(1+\sigma_8)^2\|y^t-y_+^t(z^t,v^t)\|^2-4r_1\kappa L_y^2\alpha^2\sigma_1^2\sigma_6^2\|x^t-x^{t+1}\|^2-\\
&\quad 8r_2\kappa_1\sigma_5^2\|v^t-v^{t+1}\|^2-4r_2\kappa_1\|y(v^t)-y(z_+^t(v^t),v^t)\|^2-12r_2\kappa_1\beta^2\sigma_3^2\sigma_6^2\|x^t-x^{t+1}\|^2-\\
&\quad 12r_2\kappa_1\beta^2\sigma_3^2\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2\\
&\geq(s_1-12r_2\kappa_1\beta^2\sigma_3^2\sigma_6^2-4r_1\kappa L_y^2\alpha^2\sigma_1^2\sigma_6^2-s_2L_y^2\alpha^2\sigma_6^2-3\beta^2\sigma_6^2s_3+36r_1\kappa\beta^2\sigma_6^2\sigma_2^2))\|x^{t+1}-x^t\|^2+\\
&\quad(\frac{s_2}{2}-4r_1\kappa\sigma_1^2(1+\sigma_8)^2-(12r_2\kappa_1\beta^2\sigma_3^2\sigma_1^2+3\beta^2\sigma_1^2s_3-36r_1\kappa\beta^2\sigma_1^2\sigma_2^2)(2+\sigma_8^2+2\sigma_8))\|y^t-y_+^t(z^t,v^t)\|^2+\\
&\quad\frac{s_3-12r_1\kappa\sigma_2^2}{2}\|z^t-z_+^t(v^t)\|^2+(s_4-8r_1\kappa\sigma_1^2\sigma_5^2-8r_2\kappa_1\sigma_5^2)\|v^t-v^{t+1}\|^2-4r_2\kappa_1\|y(v^t)-y(z_+^t(v^t),v^t)\|^2.
\end{aligned}
$$

Next, we will simplify the coefficients by the assumptions. Recall that $L_y = tL_x$ and then we have the following results.

- As $3(L_x+L_y) \le r_1 \le 4(L_x+L_y)$, then $\frac{t}{3+4t}+1 \le \sigma_1 \le \frac{t}{2+3t}+1$ and $\frac{3(1+t)}{3+4t} \le \sigma_2 \le \frac{4(1+t)}{2+3t}$. With these bounds and set $\kappa := (2+3t)\beta$ with $0 < \beta \le \frac{1}{500(1+t)}$, we derive that

$$
\begin{aligned}
s_3 - 12r_1\kappa\sigma_2^2 &= r_1\left(\frac{1}{\beta} - \frac{1}{2} - 2\sigma_2 - \frac{1}{\kappa} - 12\kappa\sigma_2^2\right) \\
&\ge r_1\left(\frac{1}{\beta} - \frac{1}{2} - \frac{8(1+t)}{2+3t} - \frac{1}{\kappa} - 192\kappa\frac{(1+t)^2}{(2+3t)^2}\right) \\
&\ge r_1\left(\frac{1+3t}{(2+3t)\beta} - \frac{18+19t}{2(2+3t)} - \frac{192\beta(1+t)^2}{2+3t}\right) \\
&\ge \frac{r_1}{(2+3t)\beta}\left(1+3t - \frac{18+19t}{1000(1+t)} - \frac{12}{15625}\right) \\
&\ge \frac{r_1}{(2+3t)\beta}\left(1+3t - \frac{19}{1000} - \frac{12}{15625}\right) \\
&= \frac{r_1(\frac{49}{50}+3t)}{\beta(2+3t)} \ge \frac{49r_1}{100\beta}
\end{aligned}
$$

and

$$
\begin{aligned}
s_3 - 12r_1\kappa\sigma_2^2 &= r_1\left(\frac{1}{\beta} - \frac{1}{2} - 2\sigma_2 - \frac{1}{\kappa} - 12\kappa\sigma_2^2\right) \\
&\le r_1\left(\frac{1}{\beta} - \frac{1}{2} - \frac{6(1+t)}{3+4t} - \frac{1}{\kappa} - 108\kappa\frac{(1+t)^2}{(3+4t)^2}\right) \\
&= \frac{r_1}{\beta}\left(\frac{1+3t}{2+3t} - \frac{(15+16t)\beta}{2(3+4t)} - 108(2+3t)\beta^2\frac{(1+t)^2}{(3+4t)^2}\right) \le \frac{r_1(1+3t)}{\beta(2+3t)}.
\end{aligned}
$$

- As $(\frac{t}{2+3t}+4)L_y + 4L_x \le r_2 \le 6L_y + 4L_x$,then $\sigma_5 \le \frac{(3t+2)^2}{5t^2+9t+4} \le \frac{3(3t+2)}{5t+4}$ and $\sigma_3 \le \frac{2(2t+5)}{2+3t}$. If we set $\kappa_1 = (5t+4)\mu$ and $\mu \le \frac{1}{180(3t+2)}$, then

$$
\begin{aligned}
&s_4 - 8r_1\kappa\sigma_1^2\sigma_5^2 - 8r_2\kappa_1\sigma_5^2 \\
&\ge r_2\left(\frac{1}{\mu} - \frac{1}{2} - 2\sigma_5 - \frac{1}{\kappa_1}\right) - 8r_2\sigma_5^2(\kappa\sigma_1^2 + \kappa_1) \\
&\ge r_2\left(\frac{1}{\mu} - \frac{1}{2} - \frac{6(3t+2)}{5t+4} - \frac{1}{\kappa_1}\right) - 8r_2\frac{9(3t+2)^2}{(5t+4)^2}\left(\frac{4\beta(2t+1)^2}{2+3t} + \kappa_1\right) \\
&\ge r_2\left(\frac{5t+3}{(5t+4)\mu} - \frac{41t+28}{2(5t+4)} - \frac{72(3t+2)(2t+1)^2}{125(1+t)(5t+4)^2} - \frac{72(3t+2)^2\mu}{(5t+4)}\right) \\
&\ge \frac{r_2}{(5t+4)\mu}\left(5t+3 - \frac{(41t+28)\mu}{2} - \frac{72(3t+2)(2t+1)^2\mu}{125(1+t)(5t+4)} - 72(3t+2)^2\mu^2\right) \\
&\ge \frac{r_2}{(5t+4)\mu}\left(5t+3 - \frac{41t+28}{360(3t+2)} - \frac{2(2t+1)^2}{625(1+t)(5t+4)} - \frac{1}{450}\right) \\
&\ge \frac{r_2(5t+2)}{(5t+4)\mu} \ge \frac{r_2}{2\mu},
\end{aligned}
$$

$$4r_2\kappa_1\sigma_3^2 + s_3 - 12r_1\kappa\sigma_2^2 \le r_2\left(\frac{16(2t+5)^2(5t+4)}{180(2+3t)^3} + \frac{1+3t}{\beta(2+3t)}\right) \le r_2\left(\frac{4(2t+5)}{9(2+3t)} + \frac{1+3t}{\beta(2+3t)}\right),$$

and

$$4r_2\kappa_1\sigma_3^2 + s_3 - 12r_1\kappa\sigma_2^2 \le r_1\left(\frac{16(2t+5)^2(5t+4)}{90(2+3t)^3} + \frac{1+3t}{\beta(2+3t)}\right) \le r_1\left(\frac{8(2t+5)}{9(2+3t)} + \frac{1+3t}{\beta(2+3t)}\right).$$

- Let $\frac{1}{c} - \frac{L_x+r_1}{2} \ge \frac{1}{2c}$ and $\frac{1}{2c} - L_y \ge \frac{1}{4c}$. Then we have $\frac{1}{c} \ge \max(L_x + r_1, 4L_y)$ which implies that $s_1 \ge \frac{1}{4c}$ and $\sigma_6 = \frac{2r_1+\frac{1}{c}}{r_1-L_x} \ge \frac{6L_x+6L_y+\frac{1}{c}}{3L_x+4L_y} \ge \max(\frac{10+9t}{3+4t}, \frac{6+10t}{3+4t})$. Suppose further that $\frac{1}{c} \le 4L_x + 4L_y$, then $\sigma_6 = \frac{2r_1+\frac{1}{c}}{r_1-L_x} \le \frac{8L_x+8L_y+\frac{1}{c}}{2L_x+3L_y} \le \frac{12(1+t)}{2+3t} \le 6$.
- Let $\frac{1}{\alpha} - L_d \ge \frac{2}{3\alpha}$ and $\frac{2}{3\alpha} - L_y\sigma_6^2 \ge \frac{1}{3\alpha}$. Then we have $\frac{1}{\alpha} \ge 3L_d$ and $s_2 \ge \frac{1}{3\alpha} + \frac{r_2-L_y}{2} \ge L_d + \frac{r_2-L_y}{2}$. If we set $\frac{1}{\alpha} \le 3L_d + L_y$, then $\sigma_8 = \frac{\frac{1}{\alpha}+L_d}{r_2-L_y} \le \frac{4L_d+L_y}{L_d+5L_y} \le \frac{103t^2+114t+32}{40t^2+38t+8} \le 4$.

With these bounds, we have

$$\begin{aligned}
&\frac{s_2}{2} - 4r_1\kappa\sigma_1(1+\sigma_8)^2 - 3\beta^2\sigma_1^2(4r_2\kappa_1\sigma_3^2 + s_3 - 12r_1\kappa\sigma_2^2)(2+\sigma_8^2+2\sigma_8)\\
\ge\ & \frac{L_d}{2} + \frac{r_2-L_y}{4} - \frac{2r_2(2t+1)(1+\sigma_8)^2}{125(1+t)} - \left(\frac{104r_2\beta^2\sigma_1^2(2t+5)}{3(2+3t)} + \frac{78r_2\beta\sigma_1^2(1+3t)}{2+3t}\right)\\
\ge\ & \frac{L_d}{2} + \frac{r_2-L_y}{4} - \frac{143r_2(\sigma_8+1)}{1250} - \left(\frac{26r_2(2t+5)(2t+1)^2}{46875(t+1)^2(3t+2)^3} + \frac{78r_2(3t+1)(2t+1)^2}{125(t+1)(3t+2)^3}\right)\\
\ge\ & \frac{r_2}{2} + \frac{(\sigma_1+1)L_y}{2} + \frac{r_2-L_y}{4} - \frac{143r_2}{250} - \frac{49r_2}{500}\\
\ge\ & \frac{2r_2}{25} + \frac{(\sigma_1+1)L_y}{4} \ge \frac{2r_2}{25}
\end{aligned}$$

and

$$\begin{aligned}
&s_1 - 3\beta^2\sigma_6^2(4r_2\kappa_1\sigma_3^2 + s_3 - 12r_1\kappa\sigma_2^2) - L_y^2\alpha^2\sigma_6^2(4r_1\kappa\sigma_1^2 + s_2)\\
\ge\ & \frac{L_x+r_1}{4} - \frac{432r_1\beta^2(1+t)^2}{(2+3t)^2}\left(\frac{8(2t+5)}{9(2+3t)} + \frac{1+3t}{\beta(2+3t)}\right) - \frac{4(2+3t)r_1\beta\sigma_1^2}{9\sigma_6^2} - \frac{L_y^2\sigma_6^2\alpha}{2}\\
\ge\ & \frac{r_1}{4} + \frac{r_1}{16(1+t)} - \frac{24(2t+5)r_1}{15625(2+3t)^3} - \frac{108r_1(1+3t)(1+t)}{125(2+3t)^3} - \frac{4r_1(1+2t)^2(3+4t)^2}{1125(2+3t)(1+t)(10+9t)^2} - \frac{L_y}{6}\\
\ge\ & \frac{r_1}{4} + \frac{r_1}{16(1+t)} - \frac{3r_1}{3125(1+t)} - \frac{18r_1}{125} - \frac{3r_1}{3125} - \frac{r_1}{18}\\
\ge\ & \frac{6r_1}{125} + \frac{3077r_1}{50000(1+t)} \ge \frac{6r_1}{125},
\end{aligned}$$

which is from

$$s_2 \le \frac{1}{3\alpha} + \frac{r_2-L_y}{2} + \frac{L_y}{2} = \frac{1}{3\alpha} + \frac{r_2}{2} \le \frac{1}{3\alpha} + (3+\frac{2}{t})L_y \le \frac{1}{3\alpha} + \frac{L_d}{2} \le \frac{1}{2\alpha}.$$

Together all the pieces, we get

$$
\begin{aligned}
&\Phi(x^t, y^t, z^t, v^t) - \Phi(x^{t+1}, y^{t+1}, z^{t+1}, v^{t+1}) \\
\geq\ & \frac{6r_1}{125}\|x^{t+1} - x^t\|^2 + \frac{2r_2}{25}\|y^t - y_+^t(z^t, v^t)\|^2 + \frac{r_2}{2\mu}\|v^t - v^{t+1}\|^2 + \frac{49r_1}{100\beta}\|z_+^t(v^t) - z^t\|^2 - \\
& 4r_2(t+2)\mu\|y(v^t) - y(z_+^t(v^t), v^t)\|^2.
\end{aligned}
$$

The proof is complete.

Appendix E. Proof of Proposition 1

Recall that $z_+(v) = z + \beta(x(y(z, v), z, v) - z)$. Note that $d(\cdot, z, v) = \min_{x\in\mathcal{X}} F(x, \cdot, z, v)$ is $(r_2 - L_y)$-strongly concave, then

$$
d(y(z(v), v), z(v), v) - d(y(z_+(v), v), z(v), v) \geq \frac{r_2 - L_y}{2}\|y(z_+(v), v) - y(z(v), v)\|^2. \tag{16}
$$

On the other side,

$$
\begin{aligned}
& d(y(z(v), v), z(v), v) - d(y(z_+(v), v), z(v), v) \\
\leq\ & \langle \nabla_y d(y(z_+(v), v), z(v), v), y(z(v), v) - y(z_+(v), v)\rangle \\
\leq\ & \langle \nabla_y d(y(z_+(v), v), z(v), v) - \nabla_y d(y(z_+(v), v), z_+(v), v), y(z(v), v) - y(z_+(v), v)\rangle + \\
& \langle \nabla_y d(y(z_+(v), v), z_+(v), v), y(z(v), v) - y(z_+(v), v)\rangle \\
\leq\ & L_y\sigma_2\sigma_3\|z_+^t(v) - z\|\|z_+(v) - z(v)\| \leq L_y\sigma_3\sigma_2 \operatorname{diam}(\mathcal{X})\|z - z_+(v)\|.
\end{aligned} \tag{17}
$$

Here, the third inequality is because $\langle \nabla_y d(y(z_+(v), v), z_+(v), v), y(z(v), v) - y(z_+(v), v)\rangle \leq 0$. Combining (16) and (17), we have

$$
\|y(z_+(v), v) - y(z(v), v)\|^2 \leq \frac{2L_y\sigma_2\sigma_3 \operatorname{diam}(\mathcal{X})}{r_2 - L_y}\|z - z_+(v)\|.
$$

The proof is complete.

Appendix F. Proof of Theorem 2

Before presenting the proof of Theorem 2, we first introduce the following lemma.

Lemma 9 *Let $\epsilon \geq 0$. Suppose that*

$$
\max\{\|x^{t+1} - x^t\|, \|y^t - y_+^t(z^t, v^t)\|, \|y^{t+1} - v^t\|, \|x^{t+1} - z^t\|\} \leq \epsilon,
$$

then there exists a $\rho > 0$ such that (x^{t+1}, y^{t+1}) is a $\rho\epsilon$-GS.

Proof We first note that

$$
\begin{aligned}
\operatorname{dist}(\mathbf{0}, \nabla_x f(x, y) + \partial\mathbf{1}_{\mathcal{X}}(x)) &= \inf_{\zeta}\{\|\nabla_x f(x, y) + \zeta\| : \zeta \in \partial\mathbf{1}_{\mathcal{X}}(x)\} \\
&= \inf_{z}\{\|\nabla_x f(\operatorname{proj}_{\mathcal{X}}(z), y) - \operatorname{proj}_{\mathcal{X}}(z) + z\| : x = \operatorname{proj}_{\mathcal{X}}(z)\} \\
&\leq \|\nabla_x f(x(y, x, v), y) - x(y, x, v) + x(y, x, v)\| \\
&= \|\nabla_x f(x(y, x, v), y)\| \leq r_1\|x(y, x, v) - x\|,
\end{aligned}
$$

where the second equality follows from (Li and Pong, 2018, Lemma 4.1), and the second inequality is from the first optimality condition of $x(y,z,v) = \mathrm{argmin}_{x\in\mathcal{X}}\{f(x,y) + \frac{r_1}{2}\|x-z\|^2\}$, which implies $\mathbf{0} \in \nabla_x f(\mathrm{proj}_{\mathcal{X}}(x(y,x,v),y) + r_1(x(y,x,v)-x)$. Next, we would further bound $\|x(y^{t+1},x^{t+1},v^{t+1}) - x^{t+1}\|$

$$
\begin{aligned}
&\|x(y^{t+1},x^{t+1},v^{t+1}) - x^{t+1}\| \\
\le\ &\|x^{t+1} - x(y^t,z^t,v^t)\| + \|x(y^t,z^t,v^t) - x(y^{t+1},z^t,v^t)\| + \\
&\|x(y^{t+1},z^t,v^t) - x(y^{t+1},x^{t+1},v^t)\| \\
\le\ &\sigma_6\|x^t - x^{t+1}\| + \sigma_1\|y^t - y^{t+1}\| + \sigma_2\|z^t - x^{t+1}\| \\
\le\ &(\sigma_6 + +L_y\sigma_1\alpha\sigma_6)\|x^t - x^{t+1}\| + \sigma_1\|y^t - y_+^t(z^t,v^t)\| + \sigma_2\|z^t - x^{t+1}\| \\
\le\ &(\sigma_6 + L_y\sigma_1\alpha\sigma_6 + \sigma_1 + \sigma_2)\epsilon.
\end{aligned}
$$

As for y^{t+1}, similarly we can obtain that

$$\mathrm{dist}(\mathbf{0}, \nabla_y f(x^{t+1},y^{t+1}) + \partial\mathbf{1}_{\mathcal{Y}}(y^{t+1})) \le r_2(\sigma_7 + \sigma_4 + L_y\alpha\sigma_7\sigma_6 + \sigma_5)\epsilon.$$

The proof is complete. ■

Proof of Theorem 2 Firstly, it is easy to check that $\Phi(x,y,z,v)$ is lower bounded by $\bar{F}$. Let

$$\zeta := \max\left\{\frac{6r_1}{125}\|x^{t+1}-x^t\|^2, \frac{2r_2}{25}\|y^t - y_+^t(z^t,v^t)\|^2, \frac{r_2}{2\mu}\|v^t - v^{t+1}\|^2, \frac{49r_1}{100\beta}\|z_+^t(v^t) - z^t\|^2\right\}.$$

Then we consider the following two cases separately:

- there exists $k \in \{0,1,\cdots,K-1\}$ such that
$$\frac{1}{2}\zeta \le 4r_2(t+2)\mu\|y(v^t) - y(z_+^t(v^t),v^t)\|^2; \tag{18}$$
- there exists $k \in \{0,1,\cdots,K-1\}$ such that
$$\frac{1}{2}\zeta \ge 4r_2(t+2)\mu\|y(v^t) - y(z_+^t(v^t),v^t)\|^2. \tag{19}$$

From Lemma 1, we know that

$$
\begin{aligned}
\|z^t - z_+^t(v^t)\|^2 &\le \frac{800(t+2)r_2\beta\mu}{49r_1}\|y(v^t) - y(z_+^t(v^t),v^t)\|^2 \\
&\le \frac{800(t+2)r_2\omega\beta\mu}{49r_1}\|z^t - z_+^t(v^t)\|.
\end{aligned}
\tag{20}
$$

Then, we have $\|z_+^t(v^t) - z^t\| \le \rho_1\beta\mu$, where $\rho_1 := \frac{800(t+2)r_2\omega}{49r_1}$. Armed with this, we can bound other terms as follows:

$$
\begin{aligned}
\|x^{t+1} - x^t\|^2 &\le \frac{250r_2(t+2)\mu}{3r_1}\|y(v^t) - y(z_+^t(v^t),v^t)\|^2 \\
&\le \frac{500r_2(t+2)\omega\mu}{3r_1}\|z_+^t(v^t) - z^t\| = \rho_2\beta\mu^2,
\end{aligned}
$$

$$
\begin{aligned}
\|y^{t+1}-v^t\|^2 &= \frac{1}{\mu^2}\|v^{t+1}-v^t\|^2 \\
&\le 16(t+2)\|y(v^t)-y(z_+^t(v^t),v^t)\|^2 \le 16(t+2)\omega\|z^t-z_+^t(v^t)\| = \rho_3\beta\mu, \\
\|y^t-y_+^t(z^t,v^t)\|^2 &\le 100(t+2)\mu\|y(v^t)-y(z_+^t(v^t),v^t)\|^2 \\
&\le 100(t+2)\omega\mu\|z^t-z_+^t(v^t)\| = \rho_4\beta\mu^2, \\
\|x^{t+1}-z^t\|^2 &= \frac{1}{\beta^2}\|z^{t+1}-z^t\|^2 \\
&\le \frac{2}{\beta^2}\|z^t-z_+^t(v^t)\|^2 + 6\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2 + 6\sigma_6^2\|x^t-x^{t+1}\|^2 \\
&\le 2\rho_1^2\mu^2 + 6\sigma_1^2(2+\sigma_8^2+2\sigma_8)\rho_4\beta\mu^2 + 6\sigma_6^2\rho_2\beta\mu^2 = 2\rho_1^2\mu^2 + \rho_5\beta\mu^2.
\end{aligned}
$$

where $\rho_2 := \frac{500r_2(t+2)\omega\rho_1}{3r_1}$, $\rho_3 := 16(t+2)\omega\rho_1$, $\rho_4 := 100(t+2)\omega\rho_1$ and $\rho_5 := 6\sigma_1^2(2+\sigma_8^2+2\sigma_8)\rho_4 + 6\sigma_6^2\rho_2$. According to Lemma 9, there exists $\rho > 0$ such that (x^{t+1}, y^{t+1}) is a $\rho\epsilon$-GS, where $\epsilon = \max\{\sqrt{\rho_2}\beta^{\frac{1}{2}}\mu, \sqrt{\rho_4}\beta^{\frac{1}{2}}\mu, \sqrt{\rho_3}\beta^{\frac{1}{2}}\mu^{\frac{1}{2}}, \sqrt{2}\rho\mu + \sqrt{\rho_5}\beta^{\frac{1}{2}}\mu\}$. Now, we consider the second phase. Since

$$
\begin{aligned}
&\Phi(x^t,y^t,z^t,v^t) - \Phi(x^{t+1},y^{t+1},z^{t+1},v^{t+1}) \\
\ge& \frac{3r_1}{125}\|x^{t+1}-x^t\|^2 + \frac{r_2}{25}\|y^t-y_+^t(z^t,v^t)\|^2 + \frac{r_2}{4\mu}\|v^t-v^{t+1}\|^2 + \frac{49r_1}{200\beta}\|z_+^t(v^t)-z^t\|^2
\end{aligned}
$$

holds for $t \in \{0,1,\cdots,T-1\}$, we know that

$$
\begin{aligned}
&\Phi(x^0,y^0,z^0,v^0) - \bar{F} \\
\ge& \sum_{T=0}^{T-1} \frac{3r_1}{125}\|x^{t+1}-x^t\|^2 + \frac{r_2}{25}\|y^t-y_+^t(z^t,v^t)\|^2 + \frac{r_2}{4\mu}\|v^t-v^{t+1}\|^2 + \frac{49r_1}{200\beta}\|z_+^t(v^t)-z^t\|^2 \\
\ge& T\min\left\{\frac{3r_1}{125}, \frac{r_2}{25}, \frac{r_2}{4}, \frac{49r_1}{200}\right\}\left(\|x^{t+1}-x^t\|^2 + \|y^t-y_+^t(z^t,v^t)\|^2\right) + \\
& T\min\left\{\frac{3r_1}{125}, \frac{r_2}{25}, \frac{r_2}{4}, \frac{49r_1}{200}\right\}\left(\frac{1}{\mu}\|v^t-v^{t+1}\|^2 + \frac{1}{\beta}\|z_+^t(v^t)-z^t\|^2\right)
\end{aligned}
$$

Since $\Phi(x,y,z,v) \ge \bar{F}$, therefore, there exists $k \in \{0,1,\cdots,K-1\}$ such that

$$
\begin{aligned}
&\max\left\{\|x^{t+1}-x^t\|^2, \|y^t-y_+^t(z^t,v^t)\|^2, \frac{1}{\mu}\|v^t-v^{t+1}\|^2, \frac{1}{\beta}\|z_+^t(v^t)-z^t\|^2\right\} \\
\le& \frac{\Phi_{r_1,r_2}(x^0,y^0,z^0,v^0)-\bar{F}}{T\min\{\frac{3r_1}{125}, \frac{r_2}{25}, \frac{r_2}{4}, \frac{49r_1}{200}\}} =: \frac{\eta}{T}.
\end{aligned}
$$

Note that $\|y^{t+1}-v^t\|^2 = \frac{1}{\mu^2}\|v^{t+1}-v^t\|^2 \le \frac{\eta}{\mu T}$ and

$$
\begin{aligned}
\|x^{t+1}-z^t\|^2 &= \frac{1}{\beta^2}\|z^{t+1}-z^t\|^2 \\
&\le \frac{2}{\beta^2}\|z^t-z_+^t(v^t)\|^2 + 6\sigma_1^2(2+\sigma_8^2+2\sigma_8)\|y^t-y_+^t(z^t,v^t)\|^2 + 6\sigma_6^2\|x^t-x^{t+1}\|^2
\end{aligned}
$$

$$\leq \frac{2\eta}{\beta T} + \frac{\eta(6\sigma_1^2(2+\sigma_8^2+2\sigma_8)+6\sigma_6^2)}{T}.$$

Then there exists $\rho > 0$ such that (x^{t+1}, y^{t+1}) is a $\rho\epsilon$-GS, where $\epsilon = \max\left\{\sqrt{\frac{\eta}{T}}, \sqrt{\frac{\eta}{T}}, \sqrt{\frac{\eta}{\mu T}}, \sqrt{\frac{2\eta}{\beta T}} + \sqrt{\frac{\eta(6\sigma_1^2(2+\sigma_8^2+2\sigma_8)+6\sigma_6^2)}{T}}\right\}$. If we choose $\beta = T^{-\gamma_1}, \mu = T^{-\gamma_2}$ with $\gamma_1, \gamma_2 \in (0,1)$, then

- In the first phase, it is an $\mathcal{O}(T^{-\min(\frac{\gamma_1+\gamma_2}{2}, \gamma_2)})$-GS;
- In the second phase, it is an $\mathcal{O}(T^{-\min(\frac{1-\gamma_1}{2}, \frac{1-\gamma_2}{2})})$-GS.

For the simple case $\gamma_1 = \gamma_2 = \frac{1}{2}$, our algorithm achieves an $\mathcal{O}(T^{-\frac{1}{4}})$-GS.

Appendix G. Relationship Between Different Notions of Stationary Points

In this section, we illustrate quantitative relationship among several notions of stationary measure.

Definition 2 *The point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is said to be a*

- *ϵ-proximal game stationary point* (PGS) *if*

$$\|\nabla_x d(y, x, v)\| \leq \epsilon \quad \textit{and} \quad \|\nabla_y h(x, z, y)\| \leq \epsilon.$$

- *ϵ-minimax game stationary point* (MGS) *if*

$$\|\nabla_x p(x, y)\| \leq \epsilon \quad \textit{and} \quad \|\nabla_y p(x, y)\| \leq \epsilon.$$

Proposition 2 (GS $\Longleftrightarrow$ PGS) *Suppose that the pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is ϵ-*GS *(resp. ϵ-*PGS*). Then, (x, y) is $\mathcal{O}(\epsilon)$-*PGS *(resp. $\mathcal{O}(\epsilon)$-*GS*).*

Proof Firstly, note that $\nabla_x d(y, x, v) = r_1(x - x(y, x, v))$, we just consider the equivalence of $\|x - x(y, x, v)\|$ and $\operatorname{dist}(\mathbf{0}, \nabla_x f(x, y) + \partial \mathbf{1}_{\mathcal{X}}(x))$. We find that

$$\begin{aligned}
\operatorname{dist}(\mathbf{0}, \nabla_x f(x, y) + \partial \mathbf{1}_{\mathcal{X}}(x)) &= \inf_{\zeta}\{\|\nabla_x f(x, y) + \zeta\| : \zeta \in \partial \mathbf{1}_{\mathcal{X}}(x)\} \\
&= \inf_{z}\{\|\nabla_x f(\operatorname{proj}_{\mathcal{X}}(z), y) - \operatorname{proj}_{\mathcal{X}}(z) + z\| : x = \operatorname{proj}_{\mathcal{X}}(z)\} \\
&\leq \|\nabla_x f(x(y, x, v), y) - x(y, x, v) + x(y, x, v)\| \\
&= \|\nabla_x f(x(y, x, v), y)\| \leq r_1 \|x(y, x, v) - x\|,
\end{aligned}$$

where the second equality follows from (Li and Pong, 2018, Lemma 4.1), and the second inequality is from the first optimality condition of $x(y, z, v) = \operatorname{argmin}_{x \in \mathcal{X}}\{f(x, y) + \frac{r_1}{2}\|x - z\|^2\}$, which implies $\mathbf{0} \in \nabla_x f(\operatorname{proj}_{\mathcal{X}}(x(y, x, v), y) + r_1(x(y, x, v) - x)$. On the other hand, let $x_+(y, x, v) := \operatorname{proj}_{\mathcal{X}}(x - c\nabla_x F(x, y, x, v))$, then from the primal error bound (see Pang (1987)) we know that

$$\|x - x(y, x, v)\| \leq \frac{cL_x + cr_1 + 1}{cr_1 - cL_x}\|x - x_+(y, x, v)\|.$$

Moreover, since $\nabla_x F(x, y, x, v) = \nabla_x f(x, y)$, it follows from (Li and Pong, 2018, Lemma 4.1) that

$$\begin{aligned}\|x - x(y, x, v)\| &\leq \frac{cL_x + cr_1 + 1}{cr_1 - cL_x}\|x - x_+(y, x, v)\| \\ &= \frac{cL_x + cr_1 + 1}{cr_1 - cL_x}\|x - \mathrm{proj}_{\mathcal{X}}(x - c\nabla_x f(x, y))\| \\ &\leq \frac{cL_x + cr_1 + 1}{c^2 r_1 - c^2 L_x}\,\mathrm{dist}(\mathbf{0}, \nabla_x f(x, y) + \partial \mathbf{1}_{\mathcal{X}}(x)).\end{aligned}$$

The similar analysis can be applied to derive the bounds for $\|y - y(x, z, y)\|$ and $\mathrm{dist}(\mathbf{0}, -\nabla_y f(x, y) + \mathbf{1}_{\mathcal{Y}}(x))$. The proof is complete. ■

Proposition 3 (MGS $\Longleftrightarrow$ PGS) *Suppose that the pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a ϵ-*MGS *(resp. ϵ-*PGS*), then it is also a $\mathcal{O}(\epsilon)$-*PGS *(resp. $\mathcal{O}(\epsilon)$-*MGS*).*

Proof From the Definition 1, if $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a ϵ-MGS, then

$$\|\nabla_x p(x, y)\| = r_1\|x - x(x, y)\| \leq \epsilon \quad \text{and} \quad \|\nabla_y p(x, y)\| = r_2\|y - y(x, y)\| \leq \epsilon.$$

Noting that $x(z, v) = x(y(z, v), z, v)$ and $y(z, v) = y(x(z, v), z, v)$, we further derive that

$$\begin{aligned}\|\nabla_x d(y, x, v)\| &= r_1\|x - x(y, x, v)\| \\ &\leq r_1(\|x - x(x, y)\| + \|x(x, y) - x(y, x, v)\|) \\ &\leq r_1(\|x - x(x, y)\| + \sigma_1\|y - y(x, y)\|) \\ &\leq \left(\frac{r_1\sigma_1}{r_2} + 1\right)\epsilon,\end{aligned}$$

and

$$\begin{aligned}\|\nabla_y h(x, z, y)\| &= r_2\|y - y(x, z, y)\| \\ &\leq r_2\|y - y(x, y)\| + r_2\|y(x, y) - y(x, z, y)\| \\ &= r_2\|y - y(x, y)\| + r_2\|y(x(x, y), x, y) - y(x, z, y)\| \\ &\leq r_2\|y - y(x, y)\| + r_2\sigma_4\|x(x, y) - x\| \\ &\leq \left(\frac{r_2\sigma_4}{r_1} + 1\right)\epsilon,\end{aligned}$$

which implies (x, y) is a $\mathcal{O}(\epsilon)$-PGS.

Now, let us consider when the pair (x, y) is a ϵ-PGS. We can see from the following two inequalities

$$\begin{aligned}\|x - x(z, v)\| &\leq \|x - x(y, z, v)\| + \|x(y, z, v) - x(z, v)\| \\ &\leq \|x - x(y, z, v)\| + \sigma_1\|y - y(z, v)\|,\end{aligned} \tag{21}$$

and

$$\begin{aligned}\|y - y(z, v)\| &\leq \|y - y(x, z, v)\| + \|y(x, z, v) - y(z, v)\| \\ &\leq \|y - y(x, z, v)\| + \sigma_4\|x - x(z, v)\|,\end{aligned} \tag{22}$$

that if $\|x - x(z, v)\| \leq \epsilon$, then there exists a constant $\rho > 0$ such that $\|y - y(z, v)\| \leq \rho\epsilon$ and vice versa. In equality (21), the term $\|y - y(z, v)\|$ can be further bounded.

$$
\begin{aligned}
\|y - y(z,v)\| &\leq \sigma_8 \|y - y_+(z,v)\| \\
&= \sigma_8 \|y - \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x(y,z,v), y, z, v))\| \\
&\leq \sigma_8 \|y - \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x,y,z,v))\| + \\
&\quad \sigma_8 \| \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x,y,z,v)) - \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x(y,z,v), y, z, v))\| \\
&\leq L_y \alpha \sigma_8 \|x - x(y,z,v)\| + \sigma_8 \|y - \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x,y,z,v))\| \\
&\leq L_y \alpha \sigma_8 \|x - x(y,z,v)\| + \|y - y(x,z,v)\| + \sigma_8 \|y(x,z,v) - \mathrm{proj}_{\mathcal{Y}}(y + \alpha \nabla_y F(x,y,z,v))\| \\
&\leq L_y \alpha \sigma_8 \|x - x(y,z,v)\| + \|y - y(x,z,v)\| + \\
&\quad \sigma_8 \|y(x,z,v) - y + \alpha(\nabla_y F(x, y(x,z,v), z, v) - \nabla_y F(x,y,z,v))\| \\
&\leq L_y \alpha \sigma_8 \|x - x(y,z,v)\| (2 + L_y + r_2) \sigma_8 \|y - y(x,z,v)\|.
\end{aligned}
$$

Combining them and we have

$$
\|x - x(z,v)\| \leq (1 + L_y \alpha \sigma_1 \sigma_8) \|x - x(y,z,v)\| + \sigma_1 \sigma_8 (2 + L_y + r_2) \|y - y(x,z,v)\|,
$$

which implies that (x, y) is also a $\mathcal{O}(\epsilon)$-MGS. The proof is complete. ■

Appendix H. Details about Examples in Figure 1

In this section, we will characterize the properties of two toy examples mentioned in Figure 1. The PŁ condition can imply quadratic growth (QG) condition under the assumption 1 (Karimi et al., 2016). It is obvious that the dual functions of these two examples do not satisfy QG condition globally. Therefore, we mainly discuss whether they satisfy "weak MVI" and "α-interaction dominant" conditions, which are two representative classes of conditions in the nonconvex-nonconcave setting.

H.1. Proof of Proposition for (Hsieh et al., 2021, Example 5.2)

This subsection considers the "Forsaken" example in (Hsieh et al., 2021, Example 5.2) on the constraint set $\mathcal{X} = \mathcal{Y} = \{z : -1.5 \leq z \leq 1.5\}$. In (Pethick et al., 2022), they have checked that "Forsaken" example violates "weak MVI" condition with $\rho < -\frac{1}{2L}$. Therefore, we only consider the α-interaction dominant condition here. By simple calculation, we get $\nabla^2_{xx} f(x, y) = \frac{1}{2} - 6x^2 + 5x^4$, $\nabla^2_{xy} f(x, y) = \nabla^2_{yx} f(x, y) = 1$, and $\nabla^2_{yy} f(x, y) = -\frac{1}{2} + 6y^2 - 5y^4$. Armed with these, α can be found globally by minimizing the following equation:

$$
\begin{aligned}
&\nabla^2_{xx} f(x,y) + \nabla^2_{xy} f(x,y)(\eta \mathbf{1} - \nabla^2_{yy} f(x,y))^{-1} \nabla^2_{yx} f(x,y) \\
&= \frac{1}{2} - 6x^2 + 5x^4 + (\eta + \frac{1}{2} - 6y^2 + 5y^4)^{-1}.
\end{aligned}
$$

It is less than zero when $(x, y) = (1, 0)$. That is, $\alpha < 0$ in the constraint set, which means α-interaction dominant condition is violated for primal variable x. Similar proof could be adapted for the dual variable. This rules out the convergence guarantees of damped PPM, which is validated in Figure 1(a).

H.2. Proof of Proposition for (Grimmer et al., 2020)

This "Bilinearly-Coupled Minimax" example is mentioned as a representative example where α is in the interaction moderate regime (Grimmer et al., 2020). Experiments also validate that the solution path will be globally trapped into a limit cycle (See Figure 1(*d*)). For this reason, we would only check the "weak MVI" condition. In this example, $\mathcal{X} = \mathcal{Y} = \{z : -4 \leq z \leq 4\}$, $G(u) = [\nabla_x f(x, y); -\nabla_y f(x, y)] = [4x^3 - 20x + 10y; 4y^3 - 20y - 10x]$ and $u^\star = [0; 0]$. Then ρ can be found by globally minimize $\rho(u) := \frac{\langle G(u), u - u^\star \rangle}{\|G(u)\|^2}$ for all $u \in \mathcal{X} \times \mathcal{Y}$. Notice that

$$\begin{aligned}
\frac{\langle G(u), u - u^\star \rangle}{\|G(u)\|^2} &= \frac{4x^4 - 20x^2 + 10xy + 4y^4 - 20y^2 - 10xy}{4((2x^3 - 10x + 5y)^2 + (2y^3 - 10y - 5x)^2)} \\
&= \frac{x^4 + y^4 - 5x^2 - 5y^2}{(2x^3 - 10x + 5y)^2 + (2y^3 - 10y - 5x)^2}.
\end{aligned}$$

We have $\rho(u) = \frac{\langle G(u), u - u^\star \rangle}{\|G(u)\|^2} = -\frac{4}{89}$ when $u = [x; y] = [0; 1]$, which implies that $\rho < -\frac{4}{89}$. Moreover, we find $L = 172$, so $\rho < -\frac{4}{89} < -\frac{1}{344} = -\frac{1}{2L}$. We conclude that this example does not satisfy the "weak MVI" condition and the limit cycle phenomena is actually observed in Figure 1(*e*).