Constrained Smooth Minimax Optimization: Enhanced Convergence via Scaling Reduction

Abstract

In this paper, we present advanced algorithms for solving convex-concave minimax problems, leveraging *scaling reduction techniques* to enhance efficiency. Traditional approaches to minimax optimization often suffer from slow convergence rates and high computational costs due to poor condition number dependence. Our proposed methods significantly improve these aspects by incorporating proximal best response strategies, which streamline the optimization process. We demonstrate that our algorithms achieve accelerated convergence rates, making them highly effective for large-scale optimization tasks. These findings contribute to the field of mathematical programming by providing robust tools for efficient minimax optimization.

Keywords: Convex-Concave Minimax Optimization, Adversarial Training, Scaling Reduction, Proximal Best Response, Condition Number Dependence

1 Introduction

Convex-concave minimax problems are central to various fields such as game theory, machine learning, and optimization. These problems often arise in scenarios where two opposing objectives must be simultaneously optimized, leading to complex solution landscapes. Traditional methods for addressing convex-concave minimax problems typically involve iterative optimization techniques that can suffer from slow convergence rates and computational inefficiencies, particularly when dealing with large-scale problems or poorly conditioned systems. To tackle these challenges, recent advancements have focused on developing more efficient algorithms that leverage *scaling reduction techniques*. Scaling reduction aims to improve the condition number of the optimization problem, thereby enhancing the convergence rate and reducing the overall computational burden. By integrating these techniques with proximal best response strategies, we can further streamline the optimization process, leading to significant performance improvements.

Despite the said advancements, there remains a need for robust and scalable algorithms that can efficiently handle convex-concave minimax problems. Our research seeks to address this gap by introducing novel algorithms that build upon the principles of scaling reduction and proximal best response. Through comprehensive theoretical analysis and extensive numerical experiments, we demonstrate the efficacy of our proposed methods in achieving faster convergence and greater computational efficiency.

Mathematically, we study the minimax optimization problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; f(x,y) \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$ are convex sets, and $f(x,y)$ is μ_p-strongly convex in x, μ_d-strongly concave in y, and (ℓ_p, ℓ_H, ℓ_d)-smooth (c.f., Definition 4). This formulation also arises in many machine learning applications, including adversarial training [MMS+18, SND18], prediction and regression problems [XNLS04, TLJJ05], reinforcement learning [DCL+17, DSL+18, NCDL19] and generative adversarial networks [GPAM+14, ACB17].

We study the fundamental setting where f is smooth, strongly convex with respect to x and strongly concave with respect to y. In particular, we consider the function class $\mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$, where μ_p is the strong convexity modulus, μ_d is the strong concavity modulus, ℓ_p and ℓ_d characterize the smoothness with respect to x and y respectively, and ℓ_H characterizes the interaction between x and y (c.f., Definition 4). The reason to consider such a function class is twofold. First, the strongly convex-strongly concave setting is fundamental. Via reduction [LJJ20b], an efficient algorithm for this setting implies efficient algorithms for other settings, including strongly convex-concave, convex-concave, and non-convex-concave settings. Second, [ZHZ22]'s lower bound recently proved a gradient complexity lower bound $\Omega\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$, which naturally depends on the above parameters.[1][2]

In this setting, classic algorithms such as Gradient Descent-Ascent and ExtraGradient [Kor76] can achieve linear convergence [Tse95, ZHZ22]; however, their dependence on the condition number is far from optimal. Recently, [LJJ20b] showed an upper bound of $\tilde{O}\left(\sqrt{\frac{(\ell_p \vee \ell_d \vee \ell_H)^2}{\mu_p \mu_d}} \log^3\left(\frac{1}{\epsilon}\right)\right)$, which has a much tighter dependence on the condition number. In particular, when $\ell_H > \max\{\ell_p, \ell_d\}$, the dependence on the condition number matches the lower bound. However, when $\ell_H \ll \max\{\ell_p, \ell_d\}$, this dependence would no longer be tight. In particular, we note that, when x and y are completely decoupled (i.e., $\ell_H = 0$), the optimal gradient complexity bound is $\Theta\left(\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell_d}{\mu_d}} \cdot \log\left(\frac{1}{\epsilon}\right)\right)$ (the upper bound can be obtained by simply optimizing x and y separately). Moreover, [LJJ20b]'s result does not enjoy a linear rate, which may be undesirable if a high precision solution is needed. Further progress in this domain include [KG22] that improved the polylogarithmic factor, achieving the optimal linear rate in the coarse-grained setting, and [WL20] which achieved $\tilde{O}\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{(\ell_p \vee \ell_d \vee \ell_H) \cdot \ell_H}{\mu_p \mu_d}}\right) \log\left(\frac{1}{\epsilon}\right)\right)$, achieving improved complexity when $\ell_H \ll (\ell_p \vee \ell_d \vee \ell_H)$ (i.e., when the interaction component between x and y is weaker than the individual component).

In general constrained setting, we propose new algorithms in order to address these new issues. In special we uses the techniques of scaling reduction in §3 and apply to recent works [LJJ20b, WL20]. Our contribution can be summarized as follows.

(i) For general functions in $\mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$, we design an algorithm called *Proximal Best Response* (Algorithm 3), and prove a gradient complexity upper bound of

$$\tilde{O}\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \left(\frac{\ell_p \ell_d}{\mu_p \mu_d} \cdot \frac{\ell_H^2}{\mu_p \mu_d}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right) \tag{2}$$

This complexity bound achieves a linear convergence rate and enhanced dependency on condition numbers in certain regimes (e.g., when ℓ_H is small) [§4.1, Theorem 1]. In significance,

[1] This lower bound is also proved by Ibrahim et al. [IAGM20]. Although their result is stated for a narrower class of algorithms, their proof actually works for the broader class of algorithms considered in [ZHZ22].

[2] Throughout this work we denote for two reals $a \vee b := \max\{a, b\}$ and also $a \wedge b := \min\{a, b\}$.

such a complexity improves over the best known upper bound

$$\tilde{O}\left(\sqrt{\frac{(\ell_p \vee \ell_d \vee \ell_H)^2}{\mu_p \mu_d}} \log\left(\frac{1}{\epsilon}\right)\right)$$

by [KG22, LJJ20b] and

$$\tilde{O}\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{(\ell_p \vee \ell_d \vee \ell_H) \cdot \ell_H}{\mu_p \mu_d}}\right) \log\left(\frac{1}{\epsilon}\right)\right)$$

by [WL20]. Also, this matches [ZHZ22]'s lower bound

$$\Omega\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}}\right) \log\left(\frac{1}{\epsilon}\right)\right)$$

of the gradient complexity for any first-order method in a larger regime, achieving the state-of-the-art for that function class.

(ii) Via standard blackbox reduction, the previous complexity from the *strongly convex-strongly concave* indicates tighter upper bounds for the *strongly-convex-concave* minimax optimization [§5, Theorem 3]

$$\tilde{O}\left(\sqrt{\frac{\ell_d}{\epsilon}} + \sqrt{\frac{(\sqrt{\ell_p \ell_d} \vee \ell_H)\ell_H}{\mu_p \epsilon}}\right)$$

and the general *convex-concave* minimax optimization [§5, Theorem 4]

$$\tilde{O}\left(\sqrt{\frac{\ell_p \vee \ell_d}{\epsilon}} + \frac{\sqrt{(\sqrt{\ell_p \ell_d} \vee \ell_H)\ell_H}}{\epsilon}\right)$$

Such complexities achieve the state-of-the-art.

1.1 More Related Work

There is a long line of work on the convex-concave saddle point problem. Apart from GDA and ExtraGradient [Kor76, Tse95, Nem04, GBV+19], other algorithms with theoretical guarantees include OGDA [RS13, DISZ18, MOP20, AMLJG20], Hamiltonian Gradient Descent [ALW21] and Consensus Optimization [MNG17, ALW21, AMLJG20]. For the convex-concave case and strongly-convex-concave case, lower bounds have been proved by [OX21]. For the strongly-convex-strongly-concave case, the lower bound has been proved by [IAGM20] and [ZHZ22]. Some authors have studied the special case where the interaction between x and y is bilinear [CP11, CLO14, DH19] and variance reduction algorithms for finite sum objectives [CJST19, PB16].

The convex-concave saddle point problem can also be seen as a special case of variational inequalities with Lipschitz monotone operators [Nem04, KS80, GBV+19, HIMM19, Tse95]. Some existing algorithms for the saddle point problem, such as ExtraGradient, achieve the optimal rate in this more general setting as well [Nem04, Tse95].

Going beyond the convex-concave setting, some researchers have also studied the nonconvex-concave case recently [LJJ20a, TJNO19, RLLY22, LJJ20b, LTHC20, NSH+19, OLR21], with the goal being finding a stationary point of the nonconvex function $\phi(x) := \max_y f(x, y)$. By reducing to the strongly convex-strongly concave setting, [LJJ20b, ZXSL20, YOLH22] has achieved improved results for nonconvex-concave problems.

2 Preliminaries

In this work we are interested in strongly-convex strongly-concave smooth problems. We first review some standard definitions of strong convexity and smoothness.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is L-Lipschitz if for any $x, x' \in \mathbb{R}^n$*

$$\|f(x) - f(x')\| \leq L\|x - x'\|$$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is L-smooth if ∇f is L-Lipschitz.

Definition 2. *A differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-strongly convex if for any $x, x' \in \mathbb{R}^n$*

$$\phi(x') \geq \phi(x) + (x' - x)^\top \nabla\phi(x) + \frac{\mu}{2}\|x' - x\|^2$$

If $m = 0$, we recover the definition of convexity. If $-\phi$ is μ-strongly convex, ϕ is said to be μ-strongly concave. For a function $f(x, y)$, if for any y, $f(\cdot, y)$ is strongly convex, and for any x, $f(x, \cdot)$ is strongly concave, then f is said to be strongly convex-strongly concave.

Definition 3. *A differentiable function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is said to be (ℓ_p, ℓ_H, ℓ_d)-smooth if*

(i) For any y, $\nabla_x f(\cdot, y)$ is ℓ_p-Lipschitz

(ii) For any x, $\nabla_y f(x, \cdot)$ is ℓ_d-Lipschitz

(iii) For any x, $\nabla_x f(x, \cdot)$ is ℓ_H-Lipschitz

(iv) For any y, $\nabla_y f(\cdot, y)$ is ℓ_H-Lipschitz

In this work, we are interested in function that are strongly convex-strongly concave and smooth. Specifically, we study the following function class.

Definition 4. *The function class $\mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$ contains differentiable functions from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}$ such that:*

(i) For any y, $f(\cdot, y)$ is μ_p-strongly convex

(ii) For any x, $f(x, \cdot)$ is μ_d-strongly concave

(iii) f is (ℓ_p, ℓ_H, ℓ_d)-smooth

We assume

Assumption 1. $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$.

The optimal solution of the convex-concave minimax optimization problem (1) is the saddle point (x^*, y^*) defined as follows.

Definition 5. *(x^*, y^*) is a saddle point of $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ if for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$*

$$f(x, y^*) \geq f(x^*, y^*) \geq f(x^*, y)$$

For *strongly convex-strongly concave* (SCSC) functions, it is well known that such a saddle point exists and is unique. Meanwhile, the saddle point is a stationary point, i.e. $\nabla f(x^*, y^*) = 0$, and is the minimizer of $\phi(x) := \max_y f(x, y)$. For the design of numerical algorithms, we are satisfied with a close enough approximate of the saddle point, called *ϵ-saddle points*.

Definition 6. *$(\widehat{x}, \widehat{y})$ is an ϵ-saddle point of $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ if*

$$\max_{y \in \mathcal{Y}} f(\widehat{x}, y) - \min_{x \in \mathcal{X}} f(x, \widehat{y}) \leq \epsilon$$

Alternatively, we can also characterize optimality with the distance to the saddle point. In particular, let $z^* := [x^*; y^*]$, $\widehat{z} := [\widehat{x}; \widehat{y}]$, then one may require $\|\widehat{z} - z^*\| \leq \epsilon$. This implies that[3]

$$\max_{y} f(\widehat{x}, y) - \min_{x} f(x, \widehat{y}) \leq \frac{L^2}{\mu_p \wedge \mu_d} \epsilon^2$$

In this work we focus on first-order methods, that is, algorithms that only access f through gradient evaluations. The complexity of algorithms is measured through the gradient complexity: the number of gradient evaluations required to find an ϵ-saddle point (or get to $\|\widehat{z} - z^*\| \leq \epsilon$).

3 A Scaling Reduction Argument

En route our algorithm upper bounds analysis, we illustrate a scaling reduction argument; that is, we reparametrize the variables of the objective function by equalizing the two individual smoothness parameters. To do so, one of the convenient choices is by setting $\widetilde{\mathcal{X}} = \mathcal{X}$, $\widetilde{\mathcal{Y}} = \sqrt{\frac{\ell_d}{\ell_p}}\mathcal{Y} = \left\{\sqrt{\frac{\ell_d}{\ell_p}} y : y \in \mathcal{Y}\right\}$ and

$$\widetilde{x} = x \qquad \widetilde{y} = \sqrt{\frac{\ell_d}{\ell_p}} y \qquad \widetilde{F}(\widetilde{x}, \widetilde{y}) := F\left(\widetilde{x}, \sqrt{\frac{\ell_p}{\ell_d}} \widetilde{y}\right) = F(x, y)$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ or equivalently, $\widetilde{x} \in \widetilde{\mathcal{X}}$, $\widetilde{y} \in \widetilde{\mathcal{Y}}$. We consider

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y) \quad \longrightarrow \quad \min_{\widetilde{x} \in \widetilde{\mathcal{X}}} \max_{\widetilde{y} \in \widetilde{\mathcal{Y}}} \widetilde{F}(\widetilde{x}, \widetilde{y}) \tag{4}$$

Hence, we can make the following assumption without loss of generality.

Recall we have the symbolic reparametrization $\widetilde{x} = x, \widetilde{y} = \sqrt{\frac{\ell_d}{\ell_p}} y$, $\widetilde{F}(\widetilde{x}, \widetilde{y}) = F(x, y)$, $\widetilde{f}(\widetilde{x}, \widetilde{y}) = f(x, y)$ and also their derivatives

$$\nabla_{\widetilde{y}} \widetilde{F}(\widetilde{x}, \widetilde{y}) = \sqrt{\frac{\ell_p}{\ell_d}} \nabla_y F(x, y)$$

and second-order derivatives are, if exist,

$$\nabla^2_{\widetilde{x}\widetilde{y}} \widetilde{F}(\widetilde{x}, \widetilde{y}) = \sqrt{\frac{\ell_p}{\ell_d}} \nabla^2_{xy} F(x, y) \quad \text{and} \quad \nabla^2_{\widetilde{y}\widetilde{y}} \widetilde{F}(\widetilde{x}, \widetilde{y}) = \frac{\ell_p}{\ell_d} \nabla^2_{yy} F(x, y)$$

$\widetilde{F}(\widetilde{x}, \widetilde{y})$ is arguably $\left(\ell_p, \sqrt{\frac{\ell_p}{\ell_d}} \ell_H, \ell_p\right)$-smooth and $\left(\mu_p, \frac{\ell_p}{\ell_d} \mu_d\right)$-strongly-convex-strongly-concave. We thereby conclude

[3] See Fact 6 in §A.1 for proof.

Algorithm 1 AGD(g, x_0, T)

Require: Initial point x_0, smoothness constant L, strongly-convex modulus μ, number of iterations T ▷ [Nes13, (2.2.63)]

$\eta \leftarrow \frac{1}{L}$, $\kappa \leftarrow \frac{L}{\mu}$, $\theta \leftarrow \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$
$x_1 \leftarrow P_{\mathcal{X}}\left[x_0 - \eta \nabla g(x_0)\right]$, $\tilde{x}_1 \leftarrow x_1$
for $t = 2, \cdots, T+1$ **do**
    $x_t \leftarrow P_{\mathcal{X}}\left[\tilde{x}_{t-1} - \eta \nabla g(\tilde{x}_{t-1})\right]$
    $\tilde{x}_t \leftarrow x_t + \theta(x_t - x_{t-1})$
end for

Fact 1. *$F \in \mathscr{F}\left(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d\right)$ if and only if $\widetilde{F} \in \mathscr{F}\left(\mu_p, \frac{\ell_p}{\ell_d}\mu_d, \ell_p, \sqrt{\frac{\ell_p}{\ell_d}}\ell_H, \ell_p\right)$-strongly-convex-strongly-concave.*

Informal Version of Theorem 1. *For general functions in $\mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$, we design an algorithm called* Proximal Best Response *(Algorithm 3), and prove an iteration complexity of*

$$\tilde{O}\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \left(\frac{\ell_p \ell_d}{\mu_p \mu_d} \cdot \frac{\ell_H^2}{\mu_p \mu_d}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$$

to obtain an ϵ-optimal solution.

We argue that we only need to prove for the case of $\ell_p = \ell_d$—once this case is proved, the resulting sample complexity is, omitting a polylogarithmic factor of polylog $\left(\frac{1}{\epsilon}\right)$,

$$\begin{aligned}
&\sqrt{\frac{\widetilde{\ell}_p}{\widetilde{\mu}_p} \vee \frac{\widetilde{\ell}_d}{\widetilde{\mu}_d}} + \sqrt{\frac{(\widetilde{\ell}_p \vee \widetilde{\ell}_d \vee \widetilde{\ell}_H) \cdot \widetilde{\ell}_H}{\widetilde{\mu}_p \widetilde{\mu}_d}} + \sqrt{\frac{\widetilde{\ell}_H^2}{\widetilde{\mu}_p \widetilde{\mu}_d}} \\
&= \sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_p}{\frac{\ell_p}{\ell_d}\mu_d}} + \sqrt{\frac{(\ell_p \vee \ell_p \vee (\sqrt{\frac{\ell_p}{\ell_d}}\ell_H)) \cdot \sqrt{\frac{\ell_p}{\ell_d}}\ell_H}{\mu_p \frac{\ell_p}{\ell_d}\mu_d}} + \sqrt{\frac{(\sqrt{\frac{\ell_p}{\ell_d}}\ell_H)^2}{\mu_p \frac{\ell_p}{\ell_d}\mu_d}} \\
&= \sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\ell_p \cdot \sqrt{\frac{\ell_p}{\ell_d}}\ell_H}{\mu_p \frac{\ell_p}{\ell_d}\mu_d}} \vee \sqrt{\frac{\sqrt{\frac{\ell_p}{\ell_d}}\ell_H \cdot \sqrt{\frac{\ell_p}{\ell_d}}\ell_H}{\mu_p \frac{\ell_p}{\ell_d}\mu_d}} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}} \\
&= \sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\sqrt{\ell_p \ell_d} \cdot \ell_H}{\mu_p \mu_d}} \vee \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}} \\
&\asymp \sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \left(\frac{\ell_p \ell_d}{\mu_p \mu_d} \cdot \frac{\ell_H^2}{\mu_p \mu_d}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\mu_p \mu_d}}
\end{aligned}$$

which finishes the argument. **Seeing this we can safely assume $\ell_p = \ell_d$ till the rest of this paper.**

4 Enhanced Convergence Analysis

4.1 Unconstrained Settings

In this subsection, we design an efficient algorithm for general strongly convex-strongly concave functions in the unconstrained settings, i.e., $\mathcal{X} \equiv \mathbb{R}^n$, $\mathcal{Y} \equiv \mathbb{R}^m$. We first describe *alternating best response* (ABR, Algorithm 4), which is optimal as long as the interaction term ℓ_H is sufficiently small. Then we use the *accelerated proximal point algorithm* (APPA) twice to reduce a general problem into one solvable by Alternating Best Response. The final algorithm is achieved by combining the two algorithmic components in place.

To start with, let us consider a strongly-convex-strongly-concave function $f(x, y)$, and apply Algorithm 5 for f with proximal parameter $\beta = \ell_H$. We will show the algorithm can converge in $\tilde{O}\left(\sqrt{\frac{\ell_H}{\mu_p}}\right)$ iterations [Proposition 2], while in each iteration we need to solve a regularized minimax problem

$$\min_x \max_y \left\{ f(x, y) + \beta \|x - \widehat{x}_{t-1}\|^2 \right\}$$

This is equivalent to[4]

$$\min_y \max_x \left\{ -f(x, y) - \beta \|x - \widehat{x}_{t-1}\|^2 \right\}$$

so we can apply Algorithm 5 once more to this problem with parameter $\beta = \ell_H$, which requires $\tilde{O}\left(\sqrt{\frac{\ell_H}{\mu_d}}\right)$ iterations, in each of which one need to solve a minimax problem of the form

$$\begin{aligned} &\min_y \max_x \left\{ -f(x, y) - \beta \|x - \widehat{x}_{t-1}\|^2 + \beta \|y - \widehat{y}_{t'-1}\|^2 \right\} \\ &= -\min_x \max_y \left\{ f(x, y) + \beta \|x - \widehat{x}_{t-1}\|^2 - \beta \|y - \widehat{y}_{t'-1}\|^2 \right\} \end{aligned}$$

Hence, we reduced the original problem to a problem that is 2β-strongly convex with respect to x and 2β-strongly concave with respect to y. Now the interaction between x and y is (relatively) much weaker and one can easily see that $\ell_H \le \frac{1}{2}\sqrt{2\beta \cdot 2\beta}$. Consequently the final problem can be solved in $\tilde{O}\left(\frac{\ell_p}{\ell_H}\right)$ gradient evaluations using the Alternating Best Response algorithm. We first consider the case where $\ell_H > \max\{\mu_p, \mu_d\}$. The total gradient complexity would thus be

$$\tilde{O}\left(\sqrt{\frac{\ell_H}{\mu_p}}\right) \cdot \tilde{O}\left(\sqrt{\frac{\ell_H}{\mu_d}}\right) \cdot \tilde{O}\left(\sqrt{\frac{\ell}{\ell_H}}\right) = \tilde{O}\left(\sqrt{\frac{\ell \cdot \ell_H}{\mu_p \mu_d}}\right)$$

where $\ell \equiv (\ell_p \vee \ell_d \vee \ell_H)$. In order to deal with the case where $\ell_H < \max\{\mu_p, \mu_d\}$, we shall choose $\beta_1 = \max\{\ell_H, \mu_p\}$ for the first level of proximal point, and $\beta_2 = \max\{\ell_H, \mu_d\}$ for the second level of proximal point. In this case, the total gradient complexity bound can be shown to be

$$\tilde{O}\left(\sqrt{\frac{\beta_1}{\mu_p}}\right) \cdot \tilde{O}\left(\sqrt{\frac{\beta_2}{\mu_d}}\right) \cdot \tilde{O}\left(\sqrt{\frac{\ell}{\beta_1} + \frac{\ell}{\beta_2}}\right) = \tilde{O}\left(\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p \mu_d} + \frac{\ell_d}{\mu_d}}\right)$$

A formal description of the algorithm is provided in Algorithm 3, and a formal statement of the complexity upper bound is provided in Theorem 1. The proof utilizes the forthcoming Proposition 1 for APPA and is deferred to §4.1.3.

[4] Although Sion's Theorem does not apply here as we considered unconstrained problem, we can still exchange the order since the function is strongly-convex-strongly-concave [Har82].

Algorithm 2 APPA-ABR

Require: $g(\cdot,\cdot)$, Initial point $z_0 = [x_0; y_0]$, precision parameter M_1

$\beta_2 \leftarrow \max\{\mu_d, \ell_H\}$, $M_2 \leftarrow \frac{96\ell^{2.5}}{\mu_p \mu_d^{1.5}}$

$\widehat{y}_0 \leftarrow y_0$, $\kappa \leftarrow \frac{\beta_2}{\mu_d}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$, $t \leftarrow 0$

repeat

  $t \leftarrow t+1$

  $(x_t, y_t) \leftarrow$ABR$(g(x,y) - \beta_2\|y - \widehat{y}_{t-1}\|^2, [x_{t-1}; y_{t-1}], \frac{1}{M_2}, 2\beta_1, 2\beta_2, 3\ell, 3\ell)$

  $\widehat{y}_t \leftarrow y_t + \theta(y_t - y_{t-1}) + \tau(y_t - \widehat{y}_{t-1})$

until $\|\nabla g(x_t, y_t)\| \le \frac{\mu_p \wedge \mu_d}{9LM_1}\|\nabla g(x_0, y_0)\|$

Algorithm 3 Proximal Best Response

Require: Initial point $z_0 = [x_0; y_0]$

$\beta_1 \leftarrow \max\{\mu_p, \ell_H\}$, $M_1 \leftarrow \frac{80\ell^3}{\mu_p^{1.5}\mu_d^{1.5}}$

$\widehat{x}_0 \leftarrow x_0$, $\kappa \leftarrow \frac{\beta_1}{\mu_p}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$

for $t = 1, \cdots, T$ **do**

  $(x_t, y_t) \leftarrow$ APPA-ABR$(f(x,y) + \beta_1\|x - \widehat{x}_{t-1}\|^2, [x_{t-1}, y_{t-1}], M_1)$

  $\widehat{x}_t \leftarrow x_t + \theta(x_t - x_{t-1}) + \tau(x_t - \widehat{x}_{t-1})$

end for

Theorem 1. *Assume that $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$. In Algorithm 3, the gradient complexity to produce (x_T, y_T) such that $\|z_T - z^*\| \le \epsilon$ is*

$$O\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\ell \cdot \ell_H}{\mu_p \mu_d}}\right) \cdot \log^3\left(\frac{\ell^2}{\mu_p \mu_d}\right) \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon}\right)\right)$$

Note a scaling reduction argument gives the complexity of (2) which achieves the state-of-the-art.

4.1.1 Alternating Best Response

Let us first consider the extreme case where $\ell_H = 0$. In this case, there is no interaction between x and y, and $f(x,y)$ can be simply written as $h_1(x) - h_2(y)$, where h_1 and h_2 are strongly convex functions. Thus in this case, the following trivial algorithm solves the problem

$$x^* \leftarrow \arg\min_x f(x, y_0) \qquad y^* \leftarrow \arg\max_y f(x^*, y)$$

In other words, the equilibrium can be found by directly playing the best response to each other once.

Now, let us consider the case where ℓ_H is nonzero but small. In this case, would the best response dynamics converge to the saddle point? Specifically, consider the following procedure:

$$\begin{cases} x_{t+1} & \leftarrow \arg\min_x\{f(x, y_t)\} \\ y_{t+1} & \leftarrow \arg\max_y\{f(x_{t+1}, y)\} \end{cases} \tag{3}$$

Let us define

$$y^*(x) := \arg\max_y f(x,y) \qquad x^*(y) := \arg\min_x f(x,y)$$

Algorithm 4 Alternating Best Response (ABR)

Require: $g(\cdot,\cdot)$, Initial point $z_0 = [x_0; y_0]$, precision ϵ, parameters μ_p, μ_d, ℓ_p, ℓ_d

1: $\kappa_x \leftarrow \frac{\ell_p}{\mu_p}$, $\kappa_y \leftarrow \frac{\ell_d}{\mu_d}$, $T \leftarrow \left\lceil \log_2 \left(\frac{4\sqrt{\kappa_x + \kappa_y}}{\epsilon} \right) \right\rceil$

2: **for** $t = 0, \cdots, T$ **do**

3: $\quad x_{t+1} \leftarrow \mathrm{AGD}(g(\cdot, y_t), x_t, 2\sqrt{\kappa_x}\log(24\kappa_x))$
$\quad\quad \triangleright$ Run AGD on $g(\cdot, y_t)$ from x_t for $\Theta(\sqrt{\kappa_x}\log(\kappa_x))$ steps to get x_{t+1}

4: $\quad y_{t+1} \leftarrow \mathrm{AGD}(-g(x_{t+1}, \cdot), y_t, 2\sqrt{\kappa_y}\log(24\kappa_y))$
$\quad\quad \triangleright$ Run AGD on $-g(x_{t+1}, \cdot)$ from y_t for $\Theta(\sqrt{\kappa_y}\log(\kappa_y))$ steps to get y_{t+1}

5: **end for**

Because $y^*(x)$ is $\frac{\ell_H}{\mu_d}$-Lipschitz and $x^*(y)$ is $\frac{\ell_H}{\mu_p}$-Lipschitz,[5]

$$\|x_{t+1} - x^*\| = \|x^*(y_t) - x^*(y^*)\| \le \frac{\ell_H}{\mu_p}\|y_t - y^*\| = \frac{\ell_H}{\mu_p}\|y^*(x_t) - y^*(x^*)\| \le \frac{\ell_H^2}{\mu_p\mu_d}\|x_t - x^*\|$$

Thus when $\ell_H^2 < \mu_p\mu_d$, (3) is indeed a contraction. In fact, we can further replace the exact solution of the inner optimization problems with Nesterov's *Accelerated Gradient Descent* (AGD) for constant number of steps, as precisely described in Algorithm 4.

The following theorem holds for the Alternating Best Response algorithm. The proof of the theorem, as well as a detailed version of Algorithm 4 can be found in §4.1.4.

Proposition 1. *If $g \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$ and $\ell_H \le \frac{1}{2}\sqrt{\mu_p\mu_d}$, Alternating Best Response returns (x_T, y_T) such that*

$$\|x_T - x^*\| + \|y_T - y^*\| \le \epsilon\left(\|x_0 - x^*\| + \|y_0 - y^*\|\right)$$

and the number of gradient evaluations is bounded by (with $\kappa_x = \frac{\ell_p}{\mu_p}$, $\kappa_y = \frac{\ell_d}{\mu_d}$)

$$O\left(\left(\sqrt{\kappa_x + \kappa_y}\right) \cdot \log\left(\kappa_x\kappa_y\right) \log\left(\frac{\kappa_x\kappa_y}{\epsilon}\right)\right)$$

Note that when ℓ_H is small, [ZHZ22]'s lower bound can be written as $\Omega\left(\sqrt{\kappa_x + \kappa_y}\log\left(\frac{1}{\epsilon}\right)\right)$. Thus Alternating Best Response matches this lower bound up to logarithmic factors.

4.1.2 Accelerated Proximal Point for Minimax Optimization

In the previous subsection, we showed that Alternating Best Response matches the lower bound when the interaction term ℓ_H is sufficiently small. However, in order to apply the algorithm to functions with $\ell_H > \frac{1}{2}\sqrt{\mu_p\mu_d}$, we need another algorithmic component, namely the accelerated proximal point algorithm [Gül92, LJJ20b]. However, the algorithm does not imply in the case where $\ell_H > \sqrt{\mu_p\mu_d}$. In order to deal with the general case, we would need to combine alternating best response with other algorithmic components, especially accelerated proximal point.

For a minimax optimization problem $\min_x \max_y f(x, y)$, define $\phi(x) := \max_y f(x, y)$. Suppose that we run the accelerated proximal point algorithm on $\phi(x)$ with proximal parameter β: then the number of iterations can be easily bounded, while in each iteration one needs to solve a proximal problem $\min_x \left\{\phi(x) + \beta\|x - \widehat{x}_t\|^2\right\}$. The key observation is that, this is equivalent to solving a

[5] See Fact 3 in §A.1 for proof.

Algorithm 5 Accelerated Proximal Point Algorithm

Require: Initial point $z_0 = [x_0; y_0]$, proximal parameter β, strongly-convex modulus μ_p

$\widehat{x}_0 \leftarrow x_0$, $\kappa \leftarrow \frac{\beta}{\mu_p}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$

for $t = 1, \cdots, T$ **do**

Suppose $(x_t^*, y_t^*) = \min_x \max_y f(x,y) + \beta\|x - \widehat{x}_{t-1}\|^2$. Find (x_t, y_t) such that

$$\|x_t - x_t^*\| + \|y_t - y_t^*\| \leq \frac{1}{M}\left(\|x_{t-1} - x_t^*\| + \|y_{t-1} - y_t^*\|\right)$$

$\widehat{x}_t \leftarrow x_t + \theta(x_t - x_{t-1}) + \tau(x_t - \widehat{x}_{t-1})$

end for

minimax optimization problem $\min_x \max_y \left\{f(x,y) + \beta\|x - \widehat{x}_t\|^2\right\}$. Thus via accelerated proximal point, we are able to reduce solving $\min_x \max_y f(x,y)$ to solving

$$\min_x \max_y \left\{f(x,y) + \beta\|x - \widehat{x}_t\|^2\right\}$$

This is exactly the idea behind Algorithm 5 (the idea was also used in [LJJ20b]). In the algorithm, M is a positive constant characterizing the precision of solving the subproblem, where we require $M \geq \text{poly}(\frac{\ell}{\mu_p}, \frac{\ell}{\mu_d}, \frac{\beta}{\mu_p})$. If $M \to \infty$, the algorithm exactly becomes an instance of accelerated proximal point on $\phi(x) = \max_y f(x,y)$.

The following theorem can be shown for Algorithm 5. The proof can be found in §4.1.5, and is based on the proof of Theorem 4.1 in [LJJ20b].

Proposition 2. *The number of iterations needed by Algorithm 5 to produce (x_T, y_T) such that*

$$\|x_T - x^*\| + \|y_T - y^*\| \leq \epsilon\left(\|x_0 - x^*\| + \|y_0 - y^*\|\right)$$

is at most

$$\widehat{T} = 8\sqrt{\kappa} \cdot \log\left(\frac{28\kappa^2\ell}{\mu_d}\sqrt{\frac{\ell^2}{\mu_p\mu_d}} \cdot \frac{1}{\epsilon}\right) \tag{4}$$

where $\kappa := \frac{\beta}{\mu_p}$.

4.1.3 Proof of Theorem 1

Proof of Theorem 1. We start the proof by verifying $f(x,y) + \beta_1\|x - \widehat{x}\|^2 - \beta_2\|y - \widehat{y}\|^2$ can indeed be solved by calling $\text{ABR}(\cdot, [x_0; y_0], \frac{1}{M_2}, 2\beta_1, 2\beta_2, 3\ell, 3\ell)$. Observe that $\ell_H \leq \beta_1, \beta_2 \leq \ell$. Since $f(x,y) + \beta_1\|x - \widehat{x}\|^2 - \beta_2\|y - \widehat{y}\|^2$ is $2\beta_1$-strongly convex in x and $2\beta_2$-strongly concave in y, we can see that $\frac{1}{2}\sqrt{2\beta_1 \cdot 2\beta_2} \geq \ell_H$. We can also verify that $f(x,y) + \beta_1\|x - \widehat{x}\|^2 - \beta_2\|y - \widehat{y}\|^2$ is 3ℓ-smooth, which follows from the fact that $\ell + \max\{2\beta_1, 2\beta_2\} \leq 3\ell$.

Therefore, we can apply Proposition 1 and conclude that at line 5 of Algorithm 2

$$\|x_t - x_t^*\| + \|y_t - y_t^*\| \leq \frac{1}{M_2}\left(\|x_{t-1} - x_t^*\| + \|y_{t-1} - y_t^*\|\right)$$

where

$$(x_t^*, y_t^*) := \min_x \max_y \{g(x,y) - \beta_2\|y - y_{t-1}\|^2\}$$

Here $g(x,y)$ refers to the argument passed to Algorithm 2, which in our case has the form $f(x,y)+\beta\|x-\widehat{x}_{t'-1}\|^2$. and such (x_t,y_t) is found in a gradient complexity of

$$O\left(\sqrt{\frac{\ell}{\beta_1}+\frac{\ell}{\beta_2}}\cdot\log\left(\frac{\ell^2}{\beta_1\beta_2}\right)\log\left(\frac{\ell^2}{\beta_1\beta_2}\cdot M_2\right)\right)=O\left(\sqrt{\frac{\ell}{\beta_1}+\frac{\ell}{\beta_2}}\cdot\log^2\left(\frac{\ell^2}{\mu_p\mu_d}\right)\right) \tag{5}$$

Next, we verify that Algorithm 2 is an instance of Algorithm 5 on the function $\widehat{g}(x,y):=-g(y,x)$. Notice that

$$\min_y\max_x\left\{-g(x,y)+\beta\|y-\widehat{y}\|^2\right\}=-\min_x\max_y\left\{g(x,y)-\beta\|y-\widehat{y}\|^2\right\}$$

That is, $\min_x\max_y\left\{g(x,y)-\|y-\widehat{y}\|^2\right\}$ has identical minimax point to $-g(x,y)+\beta\|y-\widehat{y}\|^2$. Thus we only need to verify that

$$M_2\geq 20\cdot\frac{\beta_2}{\mu_d'}\left(1+\frac{\ell'}{\mu_p'}\right)\sqrt{\frac{2\beta_2}{\mu_d'}+\frac{\ell'}{\mu_d'}+\frac{\ell_H^2}{\mu_p\mu_d}} \tag{6}$$

where $(\mu_p',\mu_d',\ell_p',\ell_H,\ell_d')$ are parameters for $f(x,y)+\beta_1\|x\|^2$, and $\ell'=\max\{\ell_H,\ell_p',\ell_d'\}$. Note that $\mu_p'\geq\mu_p+2\beta_1$, $\mu_d'=\mu_d$, $\ell_p'=\ell_d'\leq\ell+2\beta_1$, $\ell_H\leq\beta_1,\beta_2\leq\ell$. Thus

$$\begin{aligned}\text{RHS of (6)}&\leq 20\cdot\frac{\beta_2}{\mu_d}\sqrt{\frac{2\beta_2}{\mu_d'}+\frac{\ell+2\beta_1}{\mu_d}+\frac{\ell_H^2}{\mu_d(\mu_p+2\beta_1)}}\cdot\left(1+\frac{\ell+2\beta_1}{\mu_p+2\beta_1}\right)\\&\leq 20\cdot\frac{\ell}{\mu_d}\sqrt{\frac{2\ell}{\mu_d}+\frac{3\ell}{\mu_d}+\frac{\ell_H}{2\mu_d}}\left(1+\frac{\ell}{\mu_p}\right)\leq\frac{96\ell^{2.5}}{\mu_p\mu_d^{1.5}}=M_2\end{aligned}$$

Therefore, Algorithm 2 is indeed an instance of Inexact APPA (Algorithm 6). Notice that by the stopping condition of Algorithm 2, we have by applying Facts 4 and 5

$$\begin{aligned}&\|x_t-x^*\|+\|y_t-y^*\|\leq\frac{\sqrt{2}}{\mu_p\wedge\mu_d}\|\nabla g(x_t,y_t)\|\\&\leq\frac{\sqrt{2}}{\mu_p\wedge\mu_d}\cdot\frac{\mu_p\wedge\mu_d}{9LM_1}\|\nabla g(x_0,y_0)\|\\&\leq\frac{\sqrt{2}}{\mu_p\wedge\mu_d}\cdot\frac{\mu_p\wedge\mu_d}{9LM_1}\cdot 6L\left(\|x_0-x^*\|+\|y_0-y^*\|\right)\leq\frac{1}{M_1}\left(\|x_0-x^*\|+\|y_0-y^*\|\right)\end{aligned}$$

Thus when Algorithm 2 returns

$$\|x_t-x^*\|+\|y_t-y^*\|\leq\frac{1}{M_1}\left(\|x_0-x^*\|+\|y_0-y^*\|\right) \tag{7}$$

On the other hand, suppose that

$$\|x_t-x^*\|+\|y_t-y^*\|\leq\frac{1}{M_1}\frac{\mu_p\wedge\mu_d}{12\ell}\cdot\left(\|x_0-x^*\|+\|y_0-y^*\|\right)$$

we can show that

$$\|\nabla g(x_t,y_t)\|\leq 6L\left(\|x_t-x^*\|+\|y_t-y^*\|\right)$$

$$\leq \frac{\mu_p \wedge \mu_d}{2M_1} \left(\|x_0 - x^*\| + \|y_0 - y^*\| \right) \leq \frac{1}{M_1} \|\nabla g(x_0, y_0)\|$$

Thus in this case Algorithm 2 must return. By Proposition 2, we can see that Algorithm 2 always returns in at most

$$O\left(\sqrt{\frac{\beta_2}{\mu_d}} \cdot \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{12\ell}{\mu_p \wedge \mu_d} M_1 \right) \right) = O\left(\sqrt{\frac{\beta_2}{\mu_d}} \cdot \log\left(\frac{\ell^2}{\mu_p \mu_d} \right) \right) \tag{8}$$

iterations.

Finally, we verify that Algorithm 3 is an instance of Algorithm 5 on $f(x, y)$ with parameter β_1. Note that by (7), we only need to verify that

$$M_1 = \frac{80\ell^3}{\mu_p^{1.5} \mu_d^{1.5}} \geq 20 \cdot \frac{\beta_1}{\mu_p} \sqrt{\frac{2\beta_1}{\mu_p} + \frac{\ell}{\mu_p} + \frac{\ell_H^2}{\mu_p \mu_d}} \left(1 + \frac{\ell}{\mu_d} \right)$$

Observe that

$$20 \cdot \frac{\beta_1}{\mu_p} \sqrt{\frac{2\beta_1}{\mu_p} + \frac{\ell}{\mu_p} + \frac{\ell_H^2}{\mu_p \mu_d}} \left(1 + \frac{\ell}{\mu_d} \right) \leq 20 \cdot \frac{\ell}{\mu_p} \sqrt{\frac{2\ell}{\mu_p} + \frac{\ell}{\mu_p} + \frac{\ell^2}{\mu_p \mu_d}} \cdot \frac{2\ell}{\mu_d} \leq 20 \cdot \frac{\ell}{\mu_p} \cdot \sqrt{\frac{4\ell^2}{\mu_p \mu_d}} \cdot \frac{2\ell}{\mu_d} = M_1$$

Therefore Algorithm 3 is indeed an instance of Algorithm 5 on $f(x, y)$. As a result, by Proposition 2, the number of iterations needed such that $\|z_T - z^*\| \leq \epsilon$ is

$$O\left(\sqrt{\frac{\beta_1}{\mu_p}} \cdot \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon} \right) \right) \tag{9}$$

We now compute the total gradient complexity. Recall that $\beta_1 = \max\{\mu_p, \ell_H\}$, while $\beta_2 = \max\{\mu_d, \ell_H\}$. By (9), (8) and (5), the total gradient complexity of Algorithm 3 to reach $\|z_T - z^*\| \leq \epsilon$ is

$$\begin{aligned} &O\left(\sqrt{\frac{\beta_1}{\mu_p}} \cdot \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon} \right) \cdot \sqrt{\frac{\beta_2}{\mu_d}} \cdot \log\left(\frac{\ell^2}{\mu_p \mu_d} \right) \cdot \sqrt{\frac{\ell}{\beta_1} + \frac{\ell}{\beta_2}} \cdot \log^2\left(\frac{\ell^2}{\mu_p \mu_d} \right) \right) \\ &= O\left(\sqrt{\frac{\ell(\beta_1 + \beta_2)}{\mu_p \mu_d}} \cdot \log^3\left(\frac{\ell^2}{\mu_p \mu_d} \right) \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon} \right) \right) \end{aligned}$$

If $\ell_H \geq \max\{\mu_p, \mu_d\}$, then $\beta_1 = \beta_2 = \ell_H$, so

$$\sqrt{\frac{\ell(\beta_1 + \beta_2)}{\mu_p \mu_d}} = \sqrt{\frac{2\ell \cdot \ell_H}{\mu_p \mu_d}} \leq 2\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p \mu_d} + \frac{\ell_d}{\mu_d}}$$

Now consider the case where $\ell_H < \max\{\mu_p, \mu_d\}$. Without loss of generality, assume that $\mu_p \leq \mu_d$. Suppose that $\ell_H < \mu_d$, then $\ell = \ell_p$, $\beta_2 = \mu_d$, while $\beta_1 \leq \mu_d$. Hence

$$\sqrt{\frac{\ell(\beta_1 + \beta_2)}{\mu_p \mu_d}} \leq \sqrt{\frac{\ell_p \cdot 2\mu_d}{\mu_p \mu_d}} = \sqrt{\frac{2\ell_p}{\mu_p}} \leq 2\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p \mu_d} + \frac{\ell_d}{\mu_d}}$$

Thus in either case, $\sqrt{\frac{\ell(\beta_1+\beta_2)}{\mu_p \mu_d}} = O\left(\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p \mu_d} + \frac{\ell_d}{\mu_d}} \right)$. We conclude that the total gradient complexity of Algorithm 3 to find a point $z_T = [x_T; y_T]$ such that $\|z_T - z^*\| \leq \epsilon$ is

$$O\left(\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p \mu_d} + \frac{\ell_d}{\mu_d}} \cdot \log^3\left(\frac{\ell^2}{\mu_p \mu_d} \right) \log\left(\frac{\ell^2}{\mu_p \mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon} \right) \right)$$

□

4.1.4 Proof of Proposition 1

We proceed to prove Proposition 1.

Proof of Proposition 1. Define $\tilde{x}_{t+1} := \arg\min_x f(x, y_t)$. Let us define $y^*(x) := \arg\max_y f(x, y)$, $x^*(y) := \arg\min_x f(x, y)$ and $\phi(x) := \max_y f(x, y)$. Also define $\tilde{x}_{t+1} := \arg\min_x f(x, y^*(x_t))$ and $\widehat{x}_{t+1} := \arg\min_x f(x, y_t)$.

The basic idea is the following. Because $y^*(\cdot)$ is $\frac{\ell_H}{\mu_d}$-Lipschitz and $x^*(\cdot)$ is $\frac{\ell_H}{\mu_p}$-Lipschitz (Fact 3)

$$\|x^*(y_t) - x^*\| = \|x^*(y_t) - x^*(y^*)\| \leq \frac{\ell_H}{\mu_p}\|y_t - y^*\|$$
$$\|y^*(x_{t+1}) - y^*\| = \|y^*(x_{t+1}) - y^*(x^*)\| \leq \frac{\ell_H}{\mu_d}\|x_{t+1} - x^*\|$$

By a standard analysis of accelerated gradient descent (Lemma 6), since $\widehat{x}_{t+1} = x^*(y_t)$ is the minimum of $f(\cdot, y_t)$ and x_t is the initial point

$$\|x_{t+1} - \widehat{x}_{t+1}\|^2 \leq (\kappa_x + 1)\|x_t - \widehat{x}_{t+1}\|^2 \cdot \left(1 - \frac{1}{\sqrt{\kappa_x}}\right)^{2\sqrt{\kappa_x}\log(24\kappa_x)}$$
$$\leq \|x_t - \widehat{x}_{t+1}\|^2 \cdot (\kappa_x + 1) \cdot \exp\{-2\log(24\kappa_x)\} \leq \frac{1}{256}\|x_t - \widehat{x}_{t+1}\|^2$$

That is

$$\|x_{t+1} - x^*(y_t)\| \leq \frac{1}{16}\|x_t - x^*(y_t)\| \leq \frac{1}{16}\left(\|x_t - x^*\| + \|x^*(y_t) - x^*\|\right)$$

Thus

$$\|x_{t+1} - x^*\| \leq \|x_{t+1} - x^*(y_t)\| + \|x^*(y_t) - x^*\| \leq \frac{17}{16} \cdot \frac{\ell_H}{\mu_p}\|y_t - y^*\| + \frac{1}{16}\|x_t - x^*\| \tag{10}$$

Similarly

$$\|y_{t+1} - y^*(x_{t+1})\| \leq \frac{1}{16}\|y_t - y^*(x_{t+1})\| \leq \frac{1}{16}\left(\|y_t - y^*\| + \|y^*(x_{t+1}) - y^*\|\right)$$

Thus

$$\|y_{t+1} - y^*\| \leq \|y_{t+1} - y^*(x_{t+1})\| + \|y^*(x_{t+1}) - y^*\| \leq \frac{17}{16} \cdot \frac{\ell_H}{\mu_d}\|x_{t+1} - x^*\| + \frac{1}{16}\|y_t - y^*\|$$
$$\leq \left(\frac{17^2}{16^2} \cdot \frac{\ell_H^2}{\mu_p\mu_d} + \frac{1}{16}\right)\|y_t - y^*\| + \frac{17\ell_H}{256\mu_d}\|x_t - x^*\| \leq 0.35\|y_t - y^*\| + \frac{17\ell_H}{256\mu_d}\|x_t - x^*\| \tag{11}$$

Define $C := 4\sqrt{\frac{\mu_d}{\mu_p}}$. By adding (10) and C times (11), one gets

$$\|x_{t+1} - x^*\| + C\|y_{t+1} - y^*\|$$
$$\leq \left(\frac{1}{16} + \frac{17\ell_H}{64\sqrt{\mu_p\mu_d}}\right)\|x_t - x^*\| + \left(0.35C + \frac{17}{16} \cdot \frac{\ell_H}{\mu_p}\right)\|y_t - y^*\|$$
$$\leq \frac{1}{2}\|x_t - x^*\| + \left(0.35 + \frac{17\ell_H}{64\sqrt{\mu_p\mu_d}}\right)C\|y_t - y^*\| \leq \frac{1}{2}\left(\|x_t - x^*\| + C\|y_t - y^*\|\right)$$

Algorithm 6 Inexact Accelerated Proximal Point Algorithm (Inexact APPA)

Require: Initial point x_0, proximal parameter β, strongly convex module μ

$\widehat{x}_0 \leftarrow x_0$, $\kappa \leftarrow \frac{\beta}{\mu}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$

for $t = 1, \cdots, T$ **do**

Find x_t such that

$$g(x_t) + \beta\|x_t - \widehat{x}_{t-1}\|^2 \leq \min_x\{g(x) + \beta\|x - \widehat{x}_{t-1}\|^2\} + \delta_t$$

$\widehat{x}_t \leftarrow x_t + \theta(x_t - x_{t-1}) + \tau(x_t - \widehat{x}_{t-1})$

end for

It follows that

$$\|x_T - x^*\| + C\|y_T - y^*\| \leq 2^{-T}\left(\|x_0 - x^*\| + C\|y_0 - y^*\|\right)$$

If $C \geq 1$, then

$$\|x_T - x^*\| + \|y_T - y^*\| \leq 4\sqrt{\frac{\mu_d}{\mu_p}} \cdot 2^{-T} \cdot (\|x_0 - x^*\| + \|y_0 - y^*\|)$$

On the other hand, if $C < 1$, then

$$\|x_T - x^*\| + \|y_T - y^*\| \leq \frac{2^{-T}}{C}\left(\|x_0 - x^*\| + \|y_0 - y^*\|\right) = \sqrt{\frac{\mu_p}{\mu_d}} \cdot 2^{-T-1}\left(\|x_0 - x^*\| + \|y_0 - y^*\|\right)$$

Since $\max\left\{\frac{\mu_p}{\mu_d}, \frac{\mu_d}{\mu_p}\right\} \leq \frac{\ell_p}{\mu_p \wedge \mu_d}$

$$\|x_T - x^*\| + \|y_T - y^*\| \leq 4\sqrt{\frac{\ell_p}{\mu_p \wedge \mu_d}} \cdot 2^{-T} \cdot (\|x_0 - x^*\| + \|y_0 - y^*\|) \tag{12}$$

The theorem follows from this inequality. □

4.1.5 Proof of Proposition 2

Before proving the theorem, we would first state the inexact accelerated proximal point algorithm [LJJ20b], which is the basis of Algorithm 5.

The following two lemmas about the inexact APPA algorithm follow from the proof of [LJJ20b, Theorem 4.1] in an earlier version of the paper. Here we provide their proofs for completeness. We state it without proving it.

Lemma 1. *Suppose that $\{(x_t, \widehat{x}_t)\}_{t\geq 0}$ are generated by running the inexact APPA algorithm on $g(\cdot)$. Then for any $t \geq 1$, x*

$$g(x) \geq g(x_t) - 2\beta(x - x_t)^\top(x_t - \widehat{x}_{t-1}) + \frac{\mu}{4}\|x - x_t\|^2 - 7\kappa\delta_t$$

Proof of Lemma 1. By definition

$$g(x_t) + \beta\|x_t - \widehat{x}_{t-1}\| \leq \min_x\{g(x) + \beta\|x - \widehat{x}_{t-1}\|^2\} + \delta_t$$

Define

$$x_t^* := \arg\min_x \{g(x) + \beta\|x - \widehat{x}_{t-1}\|^2\}$$

By the μ-strong convexity of $g(\cdot)$, we have for any x

$$g(x) + \beta\|x - \widehat{x}_{t-1}\|^2 \geq g(x_t^*) + \beta\|x_t^* - \widehat{x}_{t-1}\|^2 + \left(\frac{\mu}{2} + \beta\right)\|x - x_t^*\|^2$$

Equivalently

$$\begin{aligned} g(x) &\geq g(x_t) + \beta\|x_t - \widehat{x}_{t-1}\|^2 - \beta\|x - \widehat{x}_{t-1}\|^2 + \left(\beta + \frac{\mu}{2}\right)\|x - x_t^*\|^2 - \delta_t \\ &= g(x_t) - 2\beta(x - x_t)^\top(x_t - \widehat{x}_{t-1}) - \beta\|x - x_t\|^2 + \left(\beta + \frac{\mu}{2}\right)\|x - x_t^*\|^2 - \delta_t \end{aligned}$$

On the other hand, we have

$$\left(\beta + \frac{\mu}{2}\right)\|x - x_t^*\|^2 - \beta\|x - x_t\|^2 = \frac{\mu}{2}\|x - x_t\|^2 + (2\beta + \mu)(x - x_t)^\top(x_t - x_t^*) + \left(\beta + \frac{\mu}{2}\right)\|x_t - x_t^*\|^2$$

By Cauchy-Schwarz Inequality

$$(x - x_t)^\top(x_t - x_t^*) \geq -\frac{\mu}{4(2\beta + \mu)}\|x - x_t\|^2 - (1 + 2\kappa)\|x_t - x_t^*\|^2$$

Putting the pieces together yields

$$\begin{aligned} g(x) &\geq g(x_t) - 2\beta(x - x_t)^\top(x_t - \widehat{x}_{t-1}) + \frac{\mu}{2}\|x - x_t\|^2 + \left(\beta + \frac{\mu}{2}\right)\|x_t - x_t^*\|^2 \\ &\quad - \frac{\mu}{4}\|x - x_t\|^2 - (1 + 2\kappa)(2\beta + \mu)\|x_t - x_t^*\|^2 - \delta_t \\ &= g(x_t) - 2\beta(x - x_t)^\top(x_t - \widehat{x}_{t-1}) + \frac{\mu}{4}\|x - x_t\|^2 - (2\beta + \mu)\left(\frac{1}{2} + 2\kappa\right)\|x_t - x_t^*\|^2 - \delta_t \end{aligned}$$

Also, since $g(x) + \beta\|x - \widehat{x}_{t-1}\|^2$ is $(2\beta + \mu)$-strongly convex

$$\|x_t - x_t^*\|^2 \leq \frac{2}{2\beta + \mu}\left(g(x_t) + \beta\|x_t - \widehat{x}_{t-1}\|^2 - \min_x\{g(x) + \beta\|x - \widehat{x}_{t-1}\|^2\}\right) \leq \frac{2\delta_t}{2\beta + \mu}$$

Thus

$$g(x) \geq g(x_t) - 2\beta(x - x_t)^\top(x_t - \widehat{x}_{t-1}) + \frac{\mu}{4}\|x - x_t\|^2 - 7\kappa\delta_t$$

□

Lemma 2. *Suppose that $\{x_t\}_{t\geq 0}$ is generated by running the inexact APPA algorithm on $g(\cdot)$. There exists a sequence $\{\Lambda_t\}_{t\geq 0}$ such that*

(i) $\Lambda_0 - g(x^*) \leq 2(g(x_0) - g(x^*))$

(ii) $\Lambda_t \geq g(x_t)$

(iii) $\Lambda_{t+1} - g(x^*) \leq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)(\Lambda_t - g(x^*)) + 11\kappa\delta_{t+1}$

Proof of Lemma 2. Let us slightly abuse notation, and define a sequence of functions $\{\Lambda(x)\}_{t\geq 0}$ first:

$$\begin{aligned}\Lambda_0(x) &:= g(x_0) + \frac{\mu}{4}\|x - x_0\|^2 \\ \Lambda_{t+1}(x) &:= \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\Lambda_t(x) \\ &\quad + \frac{1}{2\sqrt{\kappa}}\left(g(x_{t+1}) + 2\beta(\widehat{x}_t - x_{t+1})^\top(x - x_{t+1}) + \frac{\mu}{4}\|x - x_{t+1}\|^2 + 14\kappa^{3/2}\delta_{t+1}\right)\end{aligned}$$

The sequence $\{\Lambda_t\}_{t\geq 0}$ in the lemma is then defined as $\Lambda_t := \Lambda_t(x^*)$. Note that later we do not need to make use of the explicit definition of Λ_t.

Proof of (i) is straightforward from the definition, as

$$\Lambda_0 - g(x^*) = \frac{\mu}{4}\|x^* - x_0\|^2 + g(x_0) - g(x^*) \leq \frac{1}{2}(g(x_0) - g(x^*)) + g(x_0) - g(x^*)$$

Proof of (ii) Now, let us show $\Lambda_t \geq \min_x \Lambda_t(x) \geq g(x_t)$ using induction. Let $w_t := \arg\min_x \Lambda_t(x)$ and $\Lambda_t^* := \min_x \Lambda_t(x)$. Observe that $\Lambda_t(x)$ is always a quadratic function of the form $\Lambda_t(x) = \Lambda_t^* + \frac{\mu}{4}\|x - w_t\|^2$. Then the following recursions hold for w_t and Λ_t^*:

$$\begin{aligned}w_{t+1} &= \left(1 - \frac{1}{2\sqrt{\kappa}}\right)w_t + 2\sqrt{\kappa}(x_{t+1} - \widehat{x}_t) + \frac{x_{t+1}}{2\sqrt{\kappa}} \\ \Lambda_{t+1}^* &= \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\Lambda_t^* + \frac{1}{2\sqrt{\kappa}}\left(g(x_{t+1}) + 14\kappa^{3/2}\delta_{t+1}\right) \\ &\quad + \frac{1}{2\sqrt{\kappa}}\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\left[\frac{\mu}{4}\|x_{t+1} - w_t\|^2 + 2\beta(\widehat{x}_t - x_{t+1})^\top(w_t - x_{t+1})\right]\end{aligned}$$

The recursion for w_{t+1} can be derived by differentiating both sides in the recusion of $\Lambda_t(x)$, while the recursion for Λ_{t+1}^* can be derived by plugging the recursion for w_{t+1} into $\Lambda_{t+1}^* = \Lambda_{t+1}(w_{t+1})$.

Now, assume that $\Lambda_t^* \geq g(x_t)$ for $t \leq T - 1$. Then

$$\begin{aligned}\Lambda_T^* &\geq \left(1 - \frac{1}{2\sqrt{\kappa}}\right)g(x_{T-1}) + \frac{1}{2\sqrt{\kappa}}\left(g(x_T) + 14\kappa^{3/2}\delta_T\right) \\ &\quad + \frac{1}{2\sqrt{\kappa}}\left(1 - \frac{1}{2\sqrt{\kappa}}\right)\left[\frac{\mu}{4}\|x_T - w_{T-1}\|^2 + 2\beta(\widehat{x}_{T-1} - x_T)^\top(w_{T-1} - x_T)\right] \qquad (13)\end{aligned}$$

Applying Lemma 1 with $x = x_{T-1}$ yields

$$g(x_{T-1}) \geq g(x_T) + 2\beta(x_{T-1} - x_T)^\top(\widehat{x}_{T-1} - x_T) + \frac{\mu}{4}\|x_{T-1} - x_T\|^2 - 7\kappa\delta_T \qquad (14)$$

Summing (13) and (14) gives

$$\begin{aligned}\Lambda_T^* &\geq g(x_T) + 2\beta\left(1 - \frac{1}{2\sqrt{\kappa}}\right)(\widehat{x}_{T-1} - x_T)^\top\left[(x_{T-1} - x_T) + \frac{w_{T-1} - x_T}{2\sqrt{\kappa}}\right] \\ &\geq g(x_T) + 2\beta\left(1 - \frac{1}{2\sqrt{\kappa}}\right)(\widehat{x}_{T-1} - x_T)^\top\left[(x_{T-1} - \widehat{x}_{T-1}) + \frac{w_{T-1} - \widehat{x}_{T-1}}{2\sqrt{\kappa}}\right]\end{aligned}$$

The second inequality follows from

$$(\widehat{x}_{T-1} - x_T)^\top \left(x_T - \widehat{x}_{T-1} + \frac{x_T - \widehat{x}_{T-1}}{2\sqrt{\kappa}} \right) \leq 0$$

By the update formula

$$\widehat{x}_{t+1} = x_{t+1} + \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}(x_{t+1} - x_t) + \frac{1}{2\sqrt{\kappa}+4\kappa}(x_{t+1} - \widehat{x}_t)$$

and the recursive rule for w_t, we get

$$\begin{aligned}
&(x_{t+1} - \widehat{x}_{t+1}) + \frac{1}{2\sqrt{\kappa}}(w_{t+1} - \widehat{x}_{t+1}) \\
&= x_{t+1} + \frac{1}{2\sqrt{\kappa}} \left[1 - \frac{1}{2\sqrt{\kappa}} w_t + 2\sqrt{\kappa}(x_{t+1} - \widehat{x}_t) + \frac{x_{t+1}}{2\sqrt{\kappa}} \right] \\
&\quad - \left(1 + \frac{2}{\sqrt{\kappa}} \right) \left(x_{t+1} + \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}(x_{t+1} - x_t) + \frac{1}{2\sqrt{\kappa}+4\kappa}(x_{t+1} - \widehat{x}_t) \right) \\
&= \left(1 - \frac{2}{\sqrt{\kappa}} \right) \left[x_t + \frac{1}{2\sqrt{\kappa}} w_t - \left(1 + \frac{1}{2\sqrt{\kappa}} \widehat{x}_t \right) \right]
\end{aligned}$$

Meanwhile, when $t = 0$, $x_t = \widehat{x}_t = w_t = x_0$. Thus by induction, we have for any t, $(x_t - \widehat{x}_t) + \frac{1}{2\sqrt{\kappa}}(w_t - \widehat{x}_t) = 0$. As a result $\Lambda_T^* \geq g(x_T)$. Again, by induction, this holds for all T.

Proof of (iii) Combining Lemma 1 and the recursion for $\Lambda_t(x)$

$$\Lambda_{t+1} = \Lambda_{t+1}(x^*) \leq \left(1 - \frac{1}{2\sqrt{\kappa}} \right) \Lambda_t + \frac{1}{2\sqrt{\kappa}} \left(g(x^*) + 14\kappa^{3/2}\delta_{t+1} + 7\kappa\delta_{t+1} \right)$$

It follows that

$$\Lambda_{t+1} - g(x^*) \leq \left(1 - \frac{1}{2\sqrt{\kappa}} \right) (\Lambda_t - g(x^*)) + 11\kappa\delta_{t+1}$$

This finished the entire proof. □

Now we are ready to prove Proposition 2.

Proof of Proposition 2. Define $\phi(x) := \max_y f(x, y)$ and $\widehat{\ell} := \ell + \frac{\ell_H^2}{\mu_d}$. Then $\phi(x)$ is μ_p-strongly convex and $\widehat{\ell}$-smooth. Observe that

$$x_t^* = \arg\min_x \left\{ \phi(x) + \beta \|x - \widehat{x}_{t-1}\|^2 \right\} \qquad y_t^* = \arg\max_y \ f(x_t^*, y)$$

Thus Algorithm 5 is an instance of the inexact APPA algorithm on $\phi(x)$ with proximal parameter β and strongly convex module μ_p, and with

$$\delta_t = \phi(x_t) + \beta \|x_t - \widehat{x}_{t-1}\|^2 - \min_x \left\{ \phi(x) + \beta \|x - \widehat{x}_{t-1}\|^2 \right\} \leq \frac{\widehat{\ell} + 2\beta}{2} \|x_t - x_t^*\|^2 \tag{15}$$

Here we used the fact that, for a L-smooth function $g(\cdot)$ whose minimum is x^*

$$g(x) - g(x^*) \le \frac{\ell}{2}\|x - x^*\|^2$$

Define

$$C_0 := 44\kappa^{1.5}\frac{\widehat{\ell} + 2\beta}{2}C_1^2 \qquad C_1 := \|x_0 - x^*\| + \|y_0 - y^*\|$$

Let us state the following induction hypothesis

$$\Delta_t := \Lambda_t - \phi(x^*) \le C_0\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^t \tag{16}$$

$$\epsilon_t := \|x_t - x_t^*\| + \|y_t - y_t^*\| \le C_1\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{t/2} \tag{17}$$

It is easy to verify that with our choice of C_0 and C_1, both (16) and (17) hold for $t = 0$.

Now, assume that (16) and (17) hold for $\tau = 1, 2, \cdots, t$. Define $y^*(\cdot) := \arg\max_y f(\cdot, y)$. By Fact 3, $y^*(\cdot)$ is $(\frac{\ell}{\mu_d})$-Lipschitz. Thus

$$\begin{aligned}\|y_t - y_{t+1}^*\| &\le \|y_t^* - y_{t+1}^*\| + \|y_t - y_t^*\| \le \|y^*(x_t^*) - y^*(x_{t+1}^*)\| + \epsilon_t \\ &\le \frac{\ell}{\mu_d}\cdot\left(\|x_t^* - x_t\| + \|x_t - x_{t+1}^*\|\right) + \epsilon_t \le \left(\frac{\ell}{\mu_d} + 1\right)\epsilon_t + \frac{\ell}{\mu_d}\|x_t - x_{t+1}^*\|\end{aligned}$$

It follows that

$$\epsilon_{t+1} \le \frac{1}{M}\left[\|x_t - x_{t+1}^*\| + \|y_t - y_{t+1}^*\|\right] \le \frac{1}{M}\left(1 + \frac{\ell}{\mu_d}\right)\left(\|x_t - x_{t+1}^*\| + \epsilon_t\right) \tag{18}$$

Note that by Lemma 2 and the induction hypothesis (16)

$$\phi(x_{t+1}^*) - \phi(x^*) \le \left(1 - \frac{1}{2\sqrt{\kappa}}\right)\Delta_t \le C_0\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^t$$

By the μ_p-strong convexity of $\phi(\cdot)$ (Fact 3)

$$\|x_{t+1}^* - x^*\| \le \sqrt{\frac{2}{\mu_p}\left(\phi(x_{t+1}^*) - \phi(x^*)\right)} \le \sqrt{\frac{2C_0}{\mu_p}}\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{t/2}$$

Meanwhile

$$\|x_t - x^*\| \le \sqrt{\frac{2}{\mu_p}\left(\phi(x_t) - \phi(x^*)\right)} \le \sqrt{\frac{2C_0}{\mu_p}}\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{t/2}$$

Therefore

$$\|x_t - x_{t+1}^*\| \le \|x_t - x^*\| + \|x_{t+1}^* - x^*\| \le 2\sqrt{\frac{2C_0}{\mu_p}}\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{t/2} \tag{19}$$

By (18), (17) and the fact that $M \ge 20\kappa\sqrt{2\kappa + \frac{\widehat{\ell}}{\mu_p}}\left(1 + \frac{\ell}{\mu_d}\right)$

$$\epsilon_{t+1} \le \frac{1}{M}\left(1 + \frac{\ell}{\mu_d}\right)\left(2\sqrt{\frac{2C_0}{\mu_p}} + C_1\right)\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{t/2}$$

$$\leq \frac{1}{M}\left(1+\frac{\ell}{\mu_d}\right)\cdot\left(1+2\sqrt{\frac{44\kappa^{1.5}(\widehat{\ell}+2\beta)}{\mu_p}}\right)C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{t/2} \qquad (C_0 = 44\kappa^{1.5}\frac{\widehat{\ell}+2\beta}{2}C_1^2)$$

$$\leq \frac{1+2\sqrt{44}\kappa\sqrt{\frac{\widehat{\ell}+2\beta}{\mu_p}}}{20\kappa\sqrt{2\kappa+\frac{\widehat{\ell}}{\mu_p}}}\cdot C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{t/2} \qquad (2\sqrt{44}+1<15)$$

$$\leq \frac{3}{4}C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{t/2} \leq C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{(t+1)/2}$$

Therefore (17) holds for $t+1$. Meanwhile, by (15) and Lemma 2

$$\Delta_{t+1} \leq \left(1-\frac{1}{2\sqrt{\kappa}}\right)\Delta_t + 11\kappa\cdot\frac{\widehat{\ell}+2\beta}{2}\epsilon_{t+1}^2$$
$$\leq \left(1-\frac{1}{2\sqrt{\kappa}}\right)C_0\left(1-\frac{1}{4\sqrt{\kappa}}\right)^t + 11\kappa\cdot\frac{\widehat{\ell}+2\beta}{2}\cdot C_1^2\left(1-\frac{1}{4\sqrt{\kappa}}\right)^t = C_0\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{t+1}$$

where we used the fact that

$$11\kappa\cdot\frac{\widehat{\ell}+2\beta}{2}\cdot C_1^2 = \frac{1}{4\sqrt{\kappa}}\cdot 44\kappa^{1.5}\frac{\widehat{\ell}+2\beta}{2}C_1^2 = \frac{C_0}{4\sqrt{\kappa}}$$

Thus (16) also holds for $t+1$. By induction on t, we can see that (16) and (17) both hold for all $t\geq 0$.

As a result

$$\|x_T - x^*\| \leq \sqrt{\frac{2}{\mu_p}\left[\phi(x_T)-\phi(x^*)\right]} \leq \sqrt{\frac{2}{\mu_p}\cdot 44\kappa^{1.5}\frac{\widehat{\ell}+2\beta}{2}C_1^2\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2}}$$
$$\leq C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2}\sqrt{88\kappa^{1.5}\cdot\left(\frac{\ell^2}{\mu_p\mu_d}+\kappa\right)}$$

Meanwhile

$$\|y_T - y^*\| \leq \|y_T - y^*(x_T)\| + \|y^* - y^*(x_T)\| \leq \epsilon_T + \frac{\ell_H}{\mu_d}\|x_T - x^*\|$$

Therefore

$$\|x_T - x^*\| + \|y_T - y^*\| \leq \epsilon_T + \left(\frac{\ell_H}{\mu_d}+1\right)\|x_T - x^*\|$$
$$\leq C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2} + \frac{2\ell}{\mu_d}\cdot C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2}\cdot\sqrt{88\kappa^{1.5}\cdot\left(\frac{\ell^2}{\mu_p\mu_d}+\kappa\right)}$$
$$\leq C_1\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2}\cdot\left[1+\frac{27\kappa^2\ell}{\mu_d}\sqrt{\frac{\ell^2}{\mu_p\mu_d}}\right]$$
$$\leq \frac{28\kappa^2\ell}{\mu_d}\sqrt{\frac{\ell^2}{\mu_p\mu_d}}\cdot\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{T/2}\cdot(\|x_0 - x^*\| + \|y_0 - y^*\|)$$

which proves the theorem.

□

Algorithm 7 APPA-ABR (for Constrained Optimization)

Require: $g(\cdot,\cdot)$, Initial point $z_0 = [x_0; y_0]$, precision parameter M_1

1: $\beta_2 \leftarrow \max\{\mu_d, \ell_H\}$, $M_2 \leftarrow \frac{200\ell^3}{\mu_p \mu_d^2}$

2: $\widehat{y}_0 \leftarrow y_0$ $\kappa \leftarrow \frac{\beta_2}{\mu_d}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$, $T \leftarrow \left\lceil 8\sqrt{\kappa}\log\left(\frac{400\kappa^2\ell^2 M_1}{\mu_p\sqrt{\mu_p\mu_d}}\right)\right\rceil$

3: **for** $t = 1, \cdots, T$ **do**

4: $\quad (x_t', y_t') \leftarrow$ABR$(g(x,y) - \beta_2\|y - \widehat{y}_{t-1}\|^2, [x_{t-1}; y_{t-1}],\ \frac{1}{M_2},\ 2\beta_1,\ 2\beta_2,\ 3\ell,\ 3\ell)$

5: $\quad x_t \leftarrow P_{\mathcal{X}}\left[x_t' - \frac{1}{6\ell}\nabla_x g(x_t', y_t')\right]$

6: $\quad y_t \leftarrow P_{\mathcal{Y}}\left[y_t' + \frac{1}{6\ell}\left(\nabla_y g(x_t', y_t') - 2\beta_2(y_t' - \widehat{y}_{t-1})\right)\right]$

7: $\quad \widehat{y}_t \leftarrow y_t + \theta(y_t - y_{t-1}) + \tau(y_t - \widehat{y}_{t-1})$

8: **end for**

4.2 Algorithmic Modifications for Constrained Settings

In the constrained minimax optimization settings, x is constrained to a (possibly compact) convex set $\mathcal{X} \subseteq \mathbb{R}^n$ while y is constrained to a convex set $\mathcal{Y} \subseteq \mathbb{R}^m$. Assuming efficient projection operations, our algorithms can all be easily adapted to the constrained case. In particular for Algorithm 4, we only need to replace AGD with the constrained version; that is, set

$$x_t \leftarrow P_{\mathcal{X}}\left[\tilde{x}_{t-1} - \eta\nabla g(\tilde{x}_{t-1})\right]$$

where $P_{\mathcal{X}}[\cdot]$ denotes the projection onto convex set $\mathcal{X}$.

- For Algorithms 2 and 3, the modified versions are presented below. The only significant change is the addition of a projected gradient descent-ascent step in line 5-6 of Algorithm 2 and line 5-6 and 9-10 of Algorithm 3.

- For Algorithm 4, the only necessary modification is to *add projection steps to the Accelerated Gradient Descent Procedure*. The reason for the extra gradient step on line 2 is technical. From the original analysis [Nes13, Theorem 2.2.3], it only follows that

$$\|x_{T+1} - x^*\|^2 \le \left[\|x_1 - x^*\|^2 + \frac{2}{\mu}\left(f(x_1) - f(x^*)\right)\right] \cdot \left(1 - \frac{1}{\sqrt{\kappa}}\right)^T$$

  For constrained problems, $f(x_1) - f(x^*) \le \frac{\ell}{2}\|x_1 - x^*\|^2$ does not hold. However, with the initial projected gradient step, it can be shown that $\|x_1 - x^*\| \le \|x_0 - x^*\|$ and that $f(x_1) - f(x^*) \le \frac{\ell}{2}\|x_0 - x^*\|^2$ (see Lemma 7). Thus

$$\|x_{T+1} - x^*\|^2 \le (\kappa+1)\|x_0 - x^*\|^2 \left(1 - \frac{1}{\sqrt{\kappa}}\right)^T$$

  For Algorithm 2 and 3, the modified versions are presented below.

The most significant change is the addition of a projected gradient descent-ascent step in line 5-6 of Algorithm 2 and line 5-6 and 9-10 of Algorithm 3. The reason for this modification is very similar to that of the initial projected gradient descent step for AGD. For unconstrained problems, a small distance to the saddle point implies a small duality gap (Fact 6); however this may not be true for constrained problems, since the saddle point may no longer be a stationary point.

Algorithm 8 Proximal Best Response (for Constrained Optimization)

Require: Initial point $z_0 = [x_0; y_0]$

$\beta_1 \leftarrow \max\{\mu_p, \ell_H\}$, $M_1 \leftarrow \frac{120\ell^{3.5}}{\mu_p^2\mu_d^{1.5}}$

$\widehat{x}_0 \leftarrow x_0$, $\kappa \leftarrow \frac{\beta_1}{\mu_p}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$

for $t = 1, \cdots, T$ **do**

  $(x_t', y_t') \leftarrow \text{APPA-ABR}(f(x,y) + \beta_1\|x - \widehat{x}_{t-1}\|^2, [x_{t-1}, y_{t-1}], M_1)$

  $x_t \leftarrow P_{\mathcal{X}}\left[x_t' - \frac{1}{6\ell}\left(\nabla_x f(x_t', y_t') + 2\beta_1(x_t' - \widehat{x}_{t-1})\right)\right]$

  $y_t \leftarrow P_{\mathcal{Y}}\left[y_t' + \frac{1}{6\ell}\nabla_y f(x_t', y_t')\right]$

  $\widehat{x}_t \leftarrow x_t + \theta(x_t - x_{t-1}) + \tau(x_t - \widehat{x}_{t-1})$

end for

$\widehat{x} \leftarrow P_{\mathcal{X}}\left[x_T - \frac{1}{2\ell}\nabla_x f(x_T, y_T)\right]$

$\widehat{y} \leftarrow P_{\mathcal{Y}}\left[y_T + \frac{1}{2\ell}\nabla_y f(x_T, y_T)\right]$

This is also true for minimization: if $x^* = \arg\min_{x\in\mathcal{X}}\ g(x)$ where $g(x)$ is a ℓ-smooth function $g(x) - g(x^*) \le \frac{\ell}{2}\|x - x^*\|^2$ may not hold.

Fortunately there is a simple fix to this problem. By applying projected gradient descent-ascent once, we can assure that a small distance implies small duality gap. This is specified by the following lemma, which is the key reason why our result can be adapted to the constrained problem.

Lemma 3. *Suppose that $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$, (x^*, y^*) is a saddle point of f, $z_0 = (x_0, y_0)$ satisfies $\|z_0 - z^*\| \le \epsilon$. Let $\widehat{z} = (\widehat{x}, \widehat{y})$ be the result of one projected GDA update, i.e.*

$$\widehat{x} \leftarrow P_{\mathcal{X}}\left[x_0 - \frac{1}{2\ell}\nabla_x f(x_0, y_0)\right] \qquad \widehat{y} \leftarrow P_{\mathcal{Y}}\left[y_0 + \frac{1}{2\ell}\nabla_y f(x_0, y_0)\right]$$

Then $\|\widehat{z} - z^*\| \le \epsilon$, and

$$\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \min_{x\in\mathcal{X}} f(x, \widehat{y}) \le 2\left(1 + \frac{\ell_H^2}{(\mu_p \wedge \mu_d)^2}\right) L\epsilon^2$$

The proof of Lemma 3 is deferred to §A.3.

Because we would use Lemma 3 to replace (15) in the analysis of Algorithm 2 and 3, we would need to accordingly increase M_1 to $\frac{120\ell^{3.5}}{\mu_p^2\mu_d^{1.5}}$ and M_2 to $\frac{200\ell^3}{\mu_p\mu_d^2}$. Apart from this, another minor change in Algorithm 2 is that it would terminate after a fixed number of iterations instead of based on a termination criterion. The number of iterations is chosen such that $\|x_T - x^*\| + \|y_T - y^*\| \le \frac{1}{M_1}\left[\|x_0 - x^*\| + \|y_0 - y^*\|\right]$ is guaranteed.

4.3 Modification of Analysis for Constrained Settings

We now claim that after modifications to the algorithms, Theorem 1 holds for constrained cases.

Theorem 2 (Theorem 1, Constraint Settings)**.** *Assume that $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$. For Algorithm 3 tailored to constrained problems, the gradient complexity to find an ϵ-saddle point is*

$$O\left(\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \frac{\ell_d}{\mu_d}} + \sqrt{\frac{\ell \cdot \ell_H}{\mu_p\mu_d}}\right) \cdot \log^3\left(\frac{\ell^2}{\mu_p\mu_d}\right) \log\left(\frac{\ell^2}{\mu_p\mu_d} \cdot \frac{\ell\|z_0 - z^*\|^2}{\epsilon}\right)\right)$$

The proof of this theorem is, for the most part, the same as the unconstrained version. Hence, we only need to point out parts of the original proof that need to be modified for the constrained case.

(i) To start with, Proposition 1 holds in the constrained case. The proof of Proposition 1 only relies on the analysis of AGD and the Lipschitzness in Fact 3, and both still hold for constrained problems. (See [LJJ20b, Lemma B.2] for the proof of Fact 3 in constrained problems.)

(ii) As for Proposition 2, the key modification is about (15). As argued above, (15) uses the fact that $g(x) - g(x^*) \le \frac{\ell}{2}\|x - x^*\|^2$, which does not hold in constrained problems, since the optimum may not be a stationary point. Here, we would use Lemma 3 to derive a similar bound to replace (15). Note that originally (15) is only used to derive $\delta_t \le \frac{\widehat{\ell}+2\beta}{2}\epsilon_t^2$. Using Lemma 3, we can replace this with

$$\delta_t \le \max_{y\in\mathcal{Y}} \left\{ f(x_t, y) + \beta\|x_t - \widehat{x}_{t-1}\|^2 \right\} - \min_{x\in\mathcal{X}} \left\{ f(x, y_t) + \beta\|x - \widehat{x}_{t-1}\|^2 \right\} \le 2\left(1 + \frac{\ell_H^2}{\mu_p\mu_d}\right) L\epsilon_t^2$$

Accordingly, we can change C_0 to $44\kappa^{1.5} \cdot 2L\left(1 + \frac{\ell_H^2}{\mu_p\mu_d}\right) C_1^2$, and the assumption on M to $M \ge 20\kappa\sqrt{\frac{4\ell}{\mu_p}\left(1 + \frac{\ell_H^2}{\mu_p\mu_d}\right)}\left(1 + \frac{\ell}{\mu_d}\right)$. Then Proposition 2 would hold for the constrained case as well.

(iii) Finally, as for Theorem 1, we need to re-verify that M_1 and M_2 satisfy the new assumptions of M in order to apply Proposition 2. Observe that

$$20 \cdot \frac{\beta_2}{\mu_d} \cdot \sqrt{\frac{4(\ell + 2\beta_1)}{\mu_d} \cdot \left(1 + \frac{\ell_H^2}{2\beta_1 \cdot \mu_d}\right)} \cdot \left(1 + \frac{\ell}{\mu_p}\right) \le 20 \cdot \frac{\ell}{\mu_d} \cdot \sqrt{\frac{18\ell^3}{\mu_p\mu_d^2}} \cdot \frac{2\ell}{\mu_p} \le \frac{200\ell^3}{\mu_p\mu_d^2} = M_2$$

and that

$$20 \cdot \frac{\beta_1}{\mu_p} \cdot \sqrt{\frac{4\ell}{\mu_p} \cdot \frac{2\ell_H^2}{\mu_p\mu_d}} \cdot \frac{2\ell}{\mu_d} \le \frac{80\sqrt{2}\ell^{3.5}}{\mu_p^2\mu_d^{1.5}} \le M_1$$

It follows that the number of iterations needed to find $\|z_T - z^*\| \le \epsilon$ is

$$O\left(\sqrt{\frac{\ell_p}{\mu_p} + \frac{\ell \cdot \ell_H}{\mu_p\mu_d} + \frac{\ell_d}{\mu_d}} \cdot \log^3\left(\frac{\ell^2}{\mu_p\mu_d}\right) \log\left(\frac{\ell^2}{\mu_p\mu_d} \cdot \frac{\|z_0 - z^*\|}{\epsilon}\right)\right)$$

It follows from Lemma 3 that the duality gap of $(\widehat{x}, \widehat{y})$ is at most

$$\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \min_{x\in\mathcal{X}} f(x, \widehat{y}) \le 2\left(1 + \frac{\ell_H^2}{(\mu_p \wedge \mu_d)^2}\right) L\epsilon^2$$

Resetting ϵ to $\sqrt{\frac{\epsilon(\mu_p\wedge\mu_d)^2}{4\ell^3}}$ proves the theorem.

5 Scaling Reduction Meets Blackbox Reduction

As shown by [WL20, LJJ20b], convex-concave problems and strongly convex-concave problems can be reduced to strongly convex-strongly concave problems. Hence, Theorem 1 naturally implies improved algorithms for convex-concave and strongly convex-concave problems.

Theorem 3. *If $f(x,y)$ is (ℓ_p,ℓ_H,ℓ_d)-smooth and μ_p-strongly convex with respect to x, via reduction to Theorem 1, the gradient complexity of finding an ϵ-saddle point is*

$$\tilde{O}\left(\sqrt{\frac{\ell_d}{\epsilon}}+\sqrt{\frac{(\sqrt{\ell_p\ell_d}\vee\ell_H)\ell_H}{\mu_p\epsilon}}\right)$$

In comparison, [WL20]'s result in this setting is $\tilde{O}\left(\sqrt{\frac{\mu_p\cdot\ell_d+(\ell_p\vee\ell_d\vee\ell_H)\cdot\ell_H}{\mu_p\epsilon}}\right)$. Meanwhile a lower bound for this problem has been shown to be $\Omega\left(\sqrt{\frac{\ell_H^2}{\mu_p\epsilon}}\right)$ [ZHZ22]. In regimes where interaction is weak and $\ell_p \gg \ell_d$ (for instance $\epsilon \ll 1 \ll \mu_p \asymp \ell_H \asymp \ell_d \asymp \sqrt{\ell_p}$), our bound is a significant improvement over [WL20]'s result and is the state-of-the-art.

Theorem 4. *If $f(x,y)$ is (ℓ_p,ℓ_H,ℓ_d)-smooth and convex-concave, via reduction to Theorem 1, the gradient complexity to produce an ϵ-saddle point is*

$$\tilde{O}\left(\sqrt{\frac{\ell_p\vee\ell_d}{\epsilon}}+\frac{\sqrt{(\sqrt{\ell_p\ell_d}\vee\ell_H)\ell_H}}{\epsilon}\right)$$

The precise statement as well as the proofs can be found in §5.1. We remark that the reduction is for constrained minimax optimization, and (recall from §4.3) Theorem 1 holds for constrained problems after simple modifications to the algorithm.

In comparison, [WL20]'s result for this setting is $\tilde{O}\left(\sqrt{\frac{\ell_p\vee\ell_d}{\epsilon}}+\frac{\sqrt{(\ell_p\vee\ell_d\vee\ell_H)\cdot\ell_H}}{\epsilon}\right)$. Recall the classic [Nem04]'s ExtraGradient complexity is $O\left(\frac{\ell}{\epsilon}\right)$, and a lower bound for this setting has shown to be $\Omega\left(\sqrt{\frac{\ell_p\vee\ell_d}{\epsilon}}+\frac{\ell_H}{\epsilon}\right)$ [OX21]. In selective regimes where interaction is weak and ℓ_p, ℓ_d differ significantly in magnitude (e.g. $\epsilon \ll 1 \ll \ell_H \asymp \ell_d \asymp \sqrt{\ell_p}$), our result can be a significant improvement over [WL20]'s result and is closer to the lower bound.

5.1 Proof Details

In this subsection, we discuss how Theorem 1 implies improved bounds for strongly convex-concave problems and convex-concave problems via reductions established in [LJJ20b].

Let us consider minimax optimization problem $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y)$, where $f(x,y)$ is μ_p-strongly convex with respect to x, concave with respect to y, and (ℓ_p,ℓ_H,ℓ_d)-smooth. Here, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are bounded sets, with diameters $D_x = \max_{x,x'\in\mathcal{X}}\|x-x'\|$ and $D_y = \max_{y,y'\in\mathcal{Y}}\|y-y'\|$. Recall from Definition 6 that $(\widehat{x},\widehat{y})$ is an ϵ-saddle point of f if

$$\max_{y\in\mathcal{Y}} f(\widehat{x},y)-\min_{x\in\mathcal{X}} f(x,\widehat{y})\le\epsilon$$

We first have

Fact 2. *Let $\widehat{z} = [\widehat{x}; \widehat{y}]$. Then $\|\widehat{z} - z^*\| \le \epsilon$ implies*

$$\max_y f(\widehat{x}, y) - \min_x f(x, \widehat{y}) \le \frac{1}{2}\left(\ell + \frac{\ell_H^2}{\mu_p \wedge \mu_d}\right)\epsilon^2 \le \frac{\ell^2}{\mu_p \wedge \mu_d}\epsilon^2$$

Following [LJJ20b], we show that

Lemma 4. *Let us consider the function*[6]

$$f_{\epsilon,y}(x, y) := f(x, y) - \frac{\epsilon}{2D_y^2}\|y - y_0\|^2$$

Then an $\frac{\epsilon}{2}$-saddle point of $f_{\epsilon,y}$ is an ϵ-saddle point of f.

Proof of Lemma 4. Let $x^*(\cdot) := \arg\min_{x\in\mathcal{X}} f(x, \cdot)$ and $y^*(\cdot) := \arg\max_{y\in\mathcal{Y}} f(\cdot, y)$. Obviously, for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$,

$$f(x, y) - \frac{\epsilon}{2} \le f_{\epsilon,y}(x, y) \le f(x, y)$$

Thus if $(\widehat{x}, \widehat{y})$ is a $\frac{\epsilon}{2}$-saddle point of $f_{\epsilon,y}$, then

$$f(\widehat{x}, y^*(\widehat{x})) \le f_{\epsilon,y}(\widehat{x}, y^*(\widehat{x})) + \frac{\epsilon}{2} \le \max_{y\in\mathcal{Y}} f_{\epsilon,y}(\widehat{x}, y) + \frac{\epsilon}{2}$$
$$f(x^*(\widehat{y}), \widehat{y}) \ge f_{\epsilon,y}(x^*(\widehat{y}), \widehat{y}) \ge \min_{x\in\mathcal{X}} f_{\epsilon,y}(x, \widehat{y})$$

It immediately follows that

$$\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \min_{x\in\mathcal{X}} f(x, \widehat{y}) \le \frac{\epsilon}{2} + \max_{y\in\mathcal{Y}} f_{\epsilon,y}(\widehat{x}, y) - \min_{x\in\mathcal{X}} f_{\epsilon,y}(x, \widehat{y}) \le \epsilon$$

□

Thus to find an ϵ-saddle point of f, we only need to find an $\frac{\epsilon}{2}$-saddle point of $f_{\epsilon,y}$. We can now prove Theorem 3 by reducing to (the constrained version of) Theorem 1.

Observe that $f_{\epsilon,y}$ belongs to $\mathcal{F}(\mu_p, \frac{\epsilon}{D_y^2}, \ell_p, \ell_H, \ell_d + \frac{\epsilon}{D_y^2})$. Thus by Theorem 1, the gradient complexity of finding a $\frac{\epsilon}{2}$-saddle point in $f_{\epsilon,y}$ is[7]

$$\tilde{O}\left(\sqrt{\frac{\ell_p}{\mu_p} \vee \left(\ell_d \cdot \frac{D_y^2}{\epsilon}\right)} + \left(\frac{\ell_p}{\mu_p}\left(\ell_d \cdot \frac{D_y^2}{\epsilon}\right) \cdot \frac{\ell_H^2}{\mu_p \frac{\epsilon}{D_y^2}}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\mu_p \frac{\epsilon}{D_y^2}}}\right) = \tilde{O}\left(\sqrt{\frac{\ell_d}{\epsilon}} + \sqrt{\frac{(\sqrt{\ell_p \ell_d} \vee \ell_H)\ell_H}{\mu_p \epsilon}}\right)$$

which proves Theorem 3.

Similarly, if $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex with respect to x, concave with respect to y and (ℓ_p, ℓ_H, ℓ_d)-smooth, we have

Lemma 5. *Let us consider the function*

$$f_\epsilon(x, y) := f(x, y) + \frac{\epsilon}{4D_x^2}\|x - x_0\|^2 - \frac{\epsilon}{4D_y^2}\|y - y_0\|^2$$

Then if $(\widehat{x}, \widehat{y})$ is an $\frac{\epsilon}{2}$-saddle point of f_ϵ, it is an ϵ-saddle point of f.

[6] Such a function is invariant under scaling reduction.

[7] Here it is assumed that ϵ is sufficiently small, i.e. $\epsilon \le \max\{\ell_H, \mu_p\}D_y^2$.

Proof of Lemma 5. It can be shown that for any $\widehat{x} \in \mathcal{X}$

$$\max_{y\in\mathcal{Y}} \left\{ f(\widehat{x}, y) + \frac{\epsilon}{4D_x^2}\|\widehat{x} - x_0\|^2 - \frac{\epsilon}{4D_y^2}\|y - y_0\|^2 \right\} \geq \max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \frac{\epsilon}{4}$$

Similarly, for any $\widehat{y} \in \mathcal{Y}$

$$\min_{x\in\mathcal{X}} \left\{ f(x, \widehat{y}) + \frac{\epsilon}{4D_x^2}\|x - x_0\|^2 - \frac{\epsilon}{4D_y^2}\|\widehat{y} - y_0\|^2 \right\} \leq \min_{x\in\mathcal{Y}} f(x, \widehat{y}) + \frac{\epsilon}{4}$$

Therefore, if $(\widehat{x}, \widehat{y})$ is an $\frac{\epsilon}{2}$-saddle point of f_ϵ, it is an ϵ-saddle point of f, as

$$\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \min_{x\in\mathcal{X}} f(x, \widehat{y}) \leq \frac{\epsilon}{2} + \max_{y\in\mathcal{Y}} f_\epsilon(\widehat{x}, y) - \min_{x\in\mathcal{X}} f_\epsilon(x, \widehat{y}) \leq \epsilon$$

□

Observe that f_ϵ belongs to $\mathcal{F}(\frac{\epsilon}{2D_x^2}, \frac{\epsilon}{2D_y^2}, \ell_p + \frac{\epsilon}{2D_x^2}, \ell_H, \ell_d + \frac{\epsilon}{2D_y^2})$. Thus by Theorem 1, the gradient complexity of finding an $\frac{\epsilon}{2}$-saddle point of f_ϵ is

$$\begin{aligned}
&\tilde{O}\left(\left(\sqrt{\frac{\ell_p + \frac{\epsilon}{2D_x^2}}{\frac{\epsilon}{2D_x^2}} \vee \frac{\ell_d + \frac{\epsilon}{2D_y^2}}{\frac{\epsilon}{2D_y^2}}} + \left(\frac{\ell_p + \frac{\epsilon}{2D_x^2}}{\frac{\epsilon}{2D_x^2}} \frac{\ell_d + \frac{\epsilon}{2D_y^2}}{\frac{\epsilon}{2D_y^2}} \cdot \frac{\ell_H^2}{\frac{\epsilon}{2D_x^2}\frac{\epsilon}{2D_y^2}}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\frac{\epsilon}{2D_x^2}\frac{\epsilon}{2D_y^2}}}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right) \\
&= \tilde{O}\left(\sqrt{\frac{\ell_p \vee \ell_d}{\epsilon}} + \left(\frac{\ell_p}{\epsilon}\frac{\ell_d}{\epsilon} \cdot \frac{\ell_H^2}{\epsilon^2}\right)^{1/4} + \sqrt{\frac{\ell_H^2}{\epsilon^2}}\right) = \tilde{O}\left(\sqrt{\frac{\ell_p \vee \ell_d}{\epsilon}} + \frac{\sqrt{(\sqrt{\ell_p \ell_d} \vee \ell_H)\ell_H}}{\epsilon}\right)
\end{aligned}$$

which proves Theorem 4.

6 Conclusion

In this work, we studied convex-concave minimax optimization problems. For general strongly convex-strongly concave problems, our Proximal Best Response algorithm achieves linear convergence with improved constants in certain regimes (e.g. when the interaction parameter ℓ_H is small and ℓ_p and ℓ_d are relatively unbalanced). Via known blackbox reductions [WL20, LJJ20b], this result implies better upper bounds for strongly convex-concave and convex-concave problems.

References

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[ALW21] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In *Algorithmic Learning Theory*, pages 3–47. PMLR, 2021.

[AMLJG20] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020.

[CJST19] Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. *Advances in Neural Information Processing Systems*, 32:11377–11388, 2019.

[CLO14] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[DCL+17] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[DH19] Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 196–205. PMLR, 2019.

[DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[DSL+18] 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.

[GBV+19] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[GPAM+14] 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*, 27:2672–2680, 2014.

[Gül92] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[Har82] Joachim Hartung. An extension of Sion's minimax theorem with an application to a method for constrained games. *Pacific Journal of Mathematics*, 103(2):401–408, 1982.

[HIMM19] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32:6938–6948, 2019.

[IAGM20] Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

[KG22] Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[Kor76] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[KS80] David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and Their Applications*, volume 31. Society for Industrial and Applied Mathematics, 1980.

[LJJ20a] 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.

[LJJ20b] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

[LTHC20] 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.

[MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[MNG17] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. 30:1825–1835, 2017.

[MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[NCDL19] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32:2315–2325, 2019.

[Nem04] 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.

[Nes83] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[Nes13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[NSH+19] 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*, 32:14934–14942, 2019.

[OLR21] 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.

[OX21] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185:1–35, 2021.

[PB16] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29:1416–1424, 2016.

[RLLY22] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

[RS13] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[SND18] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[TJNO19] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32:12680–12691, 2019.

[TLJJ05] Ben Taskar, Simon Lacoste-Julien, and Michael I Jordan. Structured prediction via the extra-gradient method. *Advances in Neural Information Processing Systems*, 18:1345–1352, 2005.

[Tse95] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[WL20] Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.

[XNLS04] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. *Advances in Neural Information Processing Systems*, 17:1537–1544, 2004.

[YOLH22] 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.

[ZHZ22] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194:901–935, 2022.

[ZXSL20] 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.

A Miscellaneous

A.1 Useful Facts

In this section we collect some useful facts of functions in $\mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$. Some of the facts are known (see, e.g., [LJJ20b, ZHZ22]) and we provide the proofs for completeness.

Fact 3. *Suppose $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$. Let us define $y^*(x) := \arg\max_y f(x,y)$, $x^*(y) := \arg\min_x f(x,y)$, $\phi(x) := \max_y f(x,y)$ and $\psi(y) := \min_x f(x,y)$. Then, we have that*

(i) y^* is $\frac{\ell_H}{\mu_d}$-Lipschitz; x^* is $\frac{\ell_H}{\mu_p}$-Lipschitz

(ii) $\phi(x)$ is μ_p-strongly convex and $\left(\ell_p + \frac{\ell_H^2}{\mu_d}\right)$-smooth; $\psi(y)$ is μ_d-strongly concave and $\left(\ell_d + \frac{\ell_H^2}{\mu_p}\right)$-smooth

Proof. (i) Consider arbitrary x and x'. By definition, $\nabla_y f(x, y^*(x)) = \nabla_y f(x', y^*(x')) = 0$. By the definition of (ℓ_p, ℓ_H, ℓ_d)-smoothness, $\|\nabla_y f(x', y^*(x))\| \le \ell_H \|x - x'\|$. Thus

$$\mu_d \|y^*(x) - y^*(x')\| \le \|\nabla_y f(x', y^*(x))\| \le \ell_H \|x - x'\|$$

This proves that $y^*(\cdot)$ is $\frac{\ell_H}{\mu_d}$-Lipschitz. Similarly $x^*(\cdot)$ is $\frac{\ell_H}{\mu_p}$-Lipschitz.

(ii) By Danskin's Theorem, $\nabla\phi(x) = \nabla_x f(x, y^*(x))$. Thus for any x, x'

$$\begin{aligned}
&\|\nabla\phi(x) - \nabla\phi(x')\| = \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x'))\| \\
&\le \|\nabla_x f(x, y^*(x)) - \nabla_x f(x, y^*(x'))\| + \|\nabla_x f(x, y^*(x')) - \nabla_x f(x', y^*(x'))\| \\
&\le \ell_H \cdot \|y^*(x) - y^*(x')\| + \ell_p \|x - x'\| \le \left(\ell_p + \frac{\ell_H^2}{\mu_d}\right) \|x - x'\|
\end{aligned}$$

On the other hand, for any x, x'

$$\begin{aligned}
&\phi(x') - \phi(x) - (x' - x)^\top \nabla\phi(x) = f(x', y^*(x')) - f(x, y^*(x)) - (x' - x)^\top \nabla_x f(x, y^*(x)) \\
&\ge f(x', y^*(x)) - f(x, y^*(x)) - (x' - x)^\top \nabla_x f(x, y^*(x)) \ge \frac{\mu_p}{2} \|x' - x\|^2
\end{aligned}$$

Thus $\phi(x)$ is μ_p-strongly convex and $\left(\ell_p + \frac{\ell_H^2}{\mu_d}\right)$-smooth. By symmetric arguments, one can show that $\psi(y)$ is μ_d-strongly concave and $\left(\ell_d + \frac{\ell_H^2}{\mu_p}\right)$-smooth.

□

Fact 4. *Let $z := [x; y]$ and $z^* := [x^*; y^*]$. Then*

$$\frac{1}{\sqrt{2}} \left(\|x - x^*\| + \|y - y^*\|\right) \le \|z - z^*\| \le \|x - x^*\| + \|y - y^*\|$$

Proof. This can be easily proved using the AM-GM inequality. □

Fact 5. *Let $z := [x; y] \in \mathbb{R}^{m+n}$, $z^* := [x^*; y^*]$. Then*

$$(\mu_p \wedge \mu_d) \|z - z^*\| \le \|\nabla f(x, y)\| \le 2L \|z - z^*\|$$

Proof. By strong convexity we have for any x, y [Nes13]

$$f(x, y^*(x)) - f(x, y) \leq \frac{1}{2\mu_d}\|\nabla_y f(x, y)\|^2$$

Similarly

$$f(x, y) - f(x^*(y), y) \leq \frac{1}{2\mu_p}\|\nabla_x f(x, y)\|^2$$

Thus

$$\|\nabla f(x, y)\|^2 = \|\nabla_x f(x, y)\|^2 + \|\nabla_y f(x, y)\|^2 \geq 2(\mu_p \wedge \mu_d)\left(\phi(x) - \psi(y)\right)$$

Here $\phi(\cdot) = \max_y f(\cdot, y)$, $\psi(\cdot) = \min_x f(x, \cdot)$. By Proposition 3, ϕ is μ_p-strongly convex while ψ is μ_d-strongly concave. Hence

$$\phi(x) - \psi(y) \geq \frac{\mu_p \wedge \mu_d}{2}\left(\|x - x^*\|^2 + \|y - y^*\|^2\right) = \frac{\mu_p \wedge \mu_d}{2}\|z - z^*\|^2$$

It follows that

$$\|\nabla f(x, y)\| \geq (\mu_p \wedge \mu_d)\|z - z^*\|$$

On the other hand

$$\|\nabla_x f(x, y)\| \leq \ell_H\|y - y^*\| + \ell_p\|x - x^*\| \qquad \|\nabla_y f(x, y)\| \leq \ell_H\|x - x^*\| + \ell_d\|y - y^*\|$$

As a result

$$\|\nabla f(x, y)\|^2 \leq \ell\left(\|x - x^*\| + \|y - y^*\|\right)^2 \leq 4\ell^2\|z - z^*\|^2$$

□

Fact 6. *Let $\widehat{z} = [\widehat{x}; \widehat{y}]$. Then $\|\widehat{z} - z^*\| \leq \epsilon$ implies*

$$\max_y f(\widehat{x}, y) - \min_x f(x, \widehat{y}) \leq \frac{1}{2}\left(\ell + \frac{\ell_H^2}{\mu_p \wedge \mu_d}\right)\epsilon^2$$

Proof. Define $\phi(x) = \max_y\ f(x, y)$ and $\psi(y) = \min_x\ f(x, y)$. Then

$$\max_y f(\widehat{x}, y) - \min_x f(x, \widehat{y}) = \phi(\widehat{x}) - \psi(\widehat{y})$$

By Fact 3, ϕ is $\left(\ell_p + \frac{\ell_H^2}{\mu_p}\right)$-smooth while ψ is $\left(\ell_d + \frac{\ell_H^2}{\mu_p}\right)$-smooth. Since $\phi(x^*) = \psi(y^*)$, $\nabla\phi(x^*) = 0$, $\nabla\psi(y^*) = 0$

$$\begin{aligned}
&\phi(\widehat{x}) - \psi(\widehat{y}) \leq \frac{1}{2}\left(\ell_p + \frac{\ell_H^2}{\mu_p}\right)\|\widehat{x} - x^*\|^2 + \frac{1}{2}\left(\ell_d + \frac{\ell_H^2}{\mu_d}\right)\|\widehat{y} - y^*\|^2 \\
&\leq \frac{1}{2}\left(\ell + \frac{\ell_H^2}{\mu_p \wedge \mu_d}\right)\left(\|\widehat{x} - x^*\|^2 + \|\widehat{y} - y^*\|^2\right)
\end{aligned}$$

□

A.2 Accelerated Gradient Descent

Nesterov's Accelerated Gradient Descent [Nes83] is an optimal first-order algorithm for smooth and convex functions. Here we present a version of AGD for minimizing an L-smooth and μ-strongly convex functions $g(\cdot)$. It is a crucial building block for the algorithms in this work.

The following classical theorem holds for AGD. It implies that the complexity is $O\left(\sqrt{\kappa}\log\left(\frac{1}{\epsilon}\right)\right)$, which improves over the $O\left(\kappa\log\left(\frac{1}{\epsilon}\right)\right)$ bound for gradient descent.

Lemma 6. *([Nes13, Theorem 2.2.3]) In the AGD algorithm,*

$$\|x_T - x^*\|^2 \le (\kappa+1)\left(1-\frac{1}{\sqrt{\kappa}}\right)^T \cdot \|x_0 - x^*\|^2$$

where $\kappa := \frac{L}{\mu}$.

A.3 Proof of Lemma 3

In this section we target to prove Lemma 3 which is repeated here

Lemma 3. *Suppose that* $f \in \mathcal{F}(\mu_p, \mu_d, \ell_p, \ell_H, \ell_d)$, (x^*, y^*) *is a saddle point of* f, $z_0 = (x_0, y_0)$ *satisfies* $\|z_0 - z^*\| \le \epsilon$. *Let* $\widehat{z} = (\widehat{x}, \widehat{y})$ *be the result of one projected GDA update, i.e.*

$$\widehat{x} \leftarrow P_{\mathcal{X}}\left[x_0 - \frac{1}{2\ell}\nabla_x f(x_0, y_0)\right] \qquad \widehat{y} \leftarrow P_{\mathcal{Y}}\left[y_0 + \frac{1}{2\ell}\nabla_y f(x_0, y_0)\right]$$

Then $\|\widehat{z} - z^*\| \le \epsilon$, *and*

$$\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \min_{x\in\mathcal{X}} f(x, \widehat{y}) \le 2\left(1 + \frac{\ell_H^2}{(\mu_p \wedge \mu_d)^2}\right) L\epsilon^2$$

First we prove the following properties of one-step projected gradient, $\widehat{x} = P_{\mathcal{X}}\left[x_0 - \frac{1}{L}\nabla g(x_0)\right]$:

Lemma 7. *If a univariate function* $g : \mathcal{X} \to \mathbb{R}$ *is* L*-smooth,* $x^* = \arg\min_{x\in\mathcal{X}}\ g(x)$, *then*

$$\|\widehat{x} - x^*\| \le \|x_0 - x^*\|$$

and

$$g(\widehat{x}) - g(x^*) \le \frac{L}{2}\|x_0 - x^*\|^2$$

Proof of Lemma 7. By Corollary 2.2.1 of [Nes13] we have

$$(x_0 - \widehat{x})^T(x_0 - x^*) \ge \frac{1}{2}\|\widehat{x} - x_0\|^2$$

Therefore

$$\begin{aligned}\|\widehat{x} - x^*\|^2 &= \|(x_0 - x^*) + (\widehat{x} - x_0)\|^2 \\ &= \|x_0 - x^*\|^2 + 2(x_0 - x^*)^\top(\widehat{x} - x_0) + \|\widehat{x} - x_0\|^2 \le \|x_0 - x^*\|^2\end{aligned}$$

Meanwhile, note that

$$\widehat{x} = \arg\min_{x\in\mathcal{X}}\left\{\nabla g(x_0)^\top x + \frac{L}{2}\|x - x_0\|^2\right\}$$

By the optimality condition and the L-strong convexity of $\nabla g(x_0)^\top x + \frac{L}{2}\|x - x_0\|^2$, we have

$$\nabla g(x_0)^\top \widehat{x} + \frac{L}{2}\|\widehat{x} - x_0\|^2 + \frac{L}{2}\|x_1 - x^*\|^2 \le \nabla g(x_0)^\top x^* + \frac{L}{2}\|x^* - x_0\|^2$$

Thus

$$\nabla g(x_0)^\top (\widehat{x} - x^*) \le \frac{L}{2}\left[\|x^* - x_0\|^2 - \|\widehat{x} - x_0\|^2 - \|\widehat{x} - x^*\|^2\right]$$

It follows that

$$\begin{aligned} g(\widehat{x}) - g(x^*) &\le \nabla g(\widehat{x})^\top (\widehat{x} - x^*) = \nabla g(x_0)^\top (\widehat{x} - x^*) + (\nabla g(\widehat{x}) - \nabla g(x_0))^T (\widehat{x} - x^*) \\ &\le \frac{L}{2}\|x^* - x_0\|^2 \underbrace{-\frac{L}{2}\|\widehat{x} - x_0\|^2 - \frac{L}{2}\|\widehat{x} - x^*\|^2 + L\|\widehat{x} - x_0\| \cdot \|\widehat{x} - x^*\|}_{\le 0} \le \frac{L}{2}\|x^* - x_0\|^2 \end{aligned}$$

□

We then prove Lemma 3.

Proof of Lemma 3. This can be seen as a special case of Proposition 2.2 [Nem04]. Define the monotone operator field to be $F(z) := \begin{bmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{bmatrix}$. Note that the $\widehat{z}$ can also be written as

$$\widehat{z} = \operatorname*{arg\,min}_{z \in \mathcal{X} \times \mathcal{Y}} \left\{ \ell \|z - z_0\|^2 + F(z_0)^\top z \right\}$$

Now, define $z' = (x', y')$ to be

$$x' \leftarrow P_{\mathcal{X}}\left[x_0 - \frac{1}{2\ell}\nabla_x f(\widehat{x}, \widehat{y})\right] \qquad y' \leftarrow P_{\mathcal{Y}}\left[y_0 + \frac{1}{2\ell}\nabla_y f(\widehat{x}, \widehat{y})\right]$$

In other words

$$z' = \operatorname*{arg\,min}_{z \in \mathcal{X} \times \mathcal{Y}} \left\{ \ell \|z - z_0\|^2 + F(\widehat{z})^\top z \right\}$$

By the optimality condition and 2ℓ-strong convexity of $\ell\|z - z_0\|^2 + F(\widehat{z})^\top z$, for any $z \in \mathcal{X} \times \mathcal{Y}$

$$\ell\|z' - z_0\|^2 + F(\widehat{z})^\top z' + \ell\|z' - z\|^2 \le \ell\|z - z_0\|^2 + F(\widehat{z})^\top z$$

Similarly, by optimality of $\widehat{z}$

$$\ell\|\widehat{z} - z_0\|^2 + F(z_0)^\top \widehat{z} + \ell\|z' - \widehat{z}\|^2 \le \ell\|z' - z_0\|^2 + F(z_0)^\top z'$$

Thus

$$\begin{aligned} &F(\widehat{z})^\top (\widehat{z} - z) = F(\widehat{z})^\top (z' - z) + F(\widehat{z})^\top (\widehat{z} - z') \\ &= F(\widehat{z})^\top (z' - z) + F(z_0)^\top (\widehat{z} - z') + (F(\widehat{z}) - F(z_0))^\top (\widehat{z} - z') \\ &\le \ell\left(\|z - z_0\|^2 - \|z' - z_0\|^2 - \|z' - z\|^2\right) \\ &\quad + (F(\widehat{z}) - F(z_0))^\top (\widehat{z} - z') + \ell\left(\|z' - z_0\|^2 - \|\widehat{z} - z_0\|^2 - \|z' - \widehat{z}\|^2\right) \\ &\le \ell\left(\|z - z_0\|^2 - \|z' - z\|^2\right) + \underbrace{2\ell\|\widehat{z} - z_0\| \cdot \|\widehat{z} - z'\| - \ell\|\widehat{z} - z'\|^2 - L\|\widehat{z} - z_0\|^2}_{\le 0} \end{aligned}$$

$$\leq \ell \left(\|z - z_0\|^2 - \|z' - z\|^2 \right)$$

Here we used the fact that for any z_1, z_2, $\|F(z_1) - F(z_2)\| \leq 2\ell \|z_1 - z_2\|$. Note that (by convexity and concavity)

$$\begin{aligned} &F(\widehat{z})^\top (\widehat{z} - z) = \nabla_x f(\widehat{x}, \widehat{y})^\top (\widehat{x} - x) - \nabla_y f(\widehat{x}, \widehat{y})^\top (\widehat{y} - y) \\ &\geq [f(\widehat{x}, \widehat{y}) - f(x, \widehat{y})] + [f(\widehat{x}, y) - f(\widehat{x}, \widehat{y})] \geq f(\widehat{x}, y) - f(x, \widehat{y}) \end{aligned}$$

If we choose x and y to be $x^*(\widehat{y})$ and $y^*(\widehat{x})$, we can see that

$$\begin{aligned} &\max_{y \in \mathcal{Y}} f(\widehat{x}, y) - \min_{x \in \mathcal{X}} f(x, \widehat{y}) \leq \ell \|z - z_0\|^2 \leq 2\ell \|z - z^*\|^2 + 2\ell \|z^* - z_0\|^2 \\ &\leq 2\ell \|x^*(\widehat{y}) - x^*\|^2 + 2\ell \|y^*(\widehat{x}) - y^*\|^2 + 2\ell \|z^* - z_0\|^2 \leq \frac{2\ell_H^2}{(\mu_p \wedge \mu_d)^2} \cdot \ell \|\widehat{z} - z^*\|^2 + 2\ell \|z^* - z_0\|^2 \end{aligned}$$

By Corollary 2.2.1 [Nes13], $(x_0 - \widehat{x})^\top (x_0 - x^*) \geq \frac{1}{2} \|\widehat{x} - x_0\|^2$. Therefore

$$\begin{aligned} &\|\widehat{x} - x^*\|^2 = \|(x_0 - x^*) + (\widehat{x} - x_0)\|^2 \\ &= \|x_0 - x^*\|^2 + 2(x_0 - x^*)^\top (\widehat{x} - x_0) + \|\widehat{x} - x_0\|^2 \leq \|x_0 - x^*\|^2 \end{aligned}$$

Similarly, $\|\widehat{y} - y^*\| \leq \|y_0 - y^*\|$. Thus

$$\|\widehat{z} - z^*\|^2 = \|\widehat{x} - x^*\|^2 + \|\widehat{y} - y^*\|^2 \leq \|z_0 - z^*\|^2 \leq \epsilon^2$$

It follows that

$$\max_{y \in \mathcal{Y}} f(\widehat{x}, y) - \min_{x \in \mathcal{X}} f(x, \widehat{y}) \leq 2\ell \cdot \left(\frac{\ell_H^2}{(\mu_p \wedge \mu_d)^2} + 1 \right) \epsilon^2$$

□