Journal of Machine Learning Research 24 (2023) 1-63 Submitted 12/21; Revised 1/23; Published 2/23

Fast Objective & Duality Gap Convergence for Non-Convex Strongly-Concave Min-Max Problems with PL Condition

Zhishuai Guo[†] ZHISHGUO@TAMU.EDU
Yan Yan[‡] YAN.YAN1@WSU.EDU
Zhuoning Yuan[§] ZHUONING-YUAN@UIOWA.EDU
Tianbao Yang[†] TIANBAO-YANG@TAMU.EDU
[†] *Department of Computer Science and Engineering, Texas A&M University*
[‡] *School of Electrical Engineering and Computer Science, Washington State University*
[§] *Department of Computer Science, The University of Iowa*

Editor: Francesco Orabona

Abstract

This paper focuses on stochastic methods for solving smooth non-convex strongly-concave min-max problems, which have received increasing attention due to their potential applications in deep learning (e.g., deep AUC maximization, distributionally robust optimization). However, most of the existing algorithms are slow in practice, and their analysis revolves around the convergence to a nearly stationary point. We consider leveraging the Polyak-Łojasiewicz (PL) condition to design faster stochastic algorithms with stronger convergence guarantee. Although PL condition has been utilized for designing many stochastic minimization algorithms, their applications for non-convex min-max optimization remain rare. In this paper, we propose and analyze a generic framework of proximal stage-based method with many well-known stochastic updates embeddable. Fast convergence is established in terms of both **the primal objective gap and the duality gap**. Compared with existing studies, (i) our analysis is based on a novel Lyapunov function consisting of the primal objective gap and the duality gap of a regularized function, and (ii) the results are more comprehensive with improved rates that have better dependence on the condition number under different assumptions. We also conduct deep and non-deep learning experiments to verify the effectiveness of our methods.

Keywords: Min-Max Problems, Non-Convex Optimization, Stochastic Optimization, PL Condition, Proximal Stage-Based Method

1. Introduction

Min-max optimization has a broad range of applications in machine learning. In this paper, we consider a family of min-max optimization problems where the objective function is non-convex in terms of the min variable and is strongly concave in terms of the max variable. It covers a number of important applications in machine learning, such as deep AUC maximization (Ying et al., 2016; Liu et al., 2020b; Guo et al., 2020) and distributionally robust optimization (DRO) (Namkoong and Duchi, 2016, 2017; Rafique et al., 2018). In particular, we study stochastic gradient methods for solving the following **non-convex**

strongly-concave (NCSC) min-max problem:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathcal{Y}} f(x,y), \tag{1}$$

where $\mathcal{Y}\subseteq\mathbb{R}^{d'}$ is a convex closed set, $f(x,y)$ is smooth, non-convex in x and strongly concave in y. We assume the optimization is only through a stochastic gradient oracle that for any x,y returns unbiased stochastic gradient $(\mathcal{G}_x(x,y;\xi),\mathcal{G}_y(x,y;\xi))$, i.e., $\mathbb{E}[\mathcal{G}_x(x,y;\xi)]=\nabla f_x(x,y)$ and $\mathbb{E}[\mathcal{G}_y(x,y;\xi)]=\nabla f_y(x,y)$.

Stochastic algorithms for solving (1) have been studied in some recent papers (Lin et al., 2020a,b; Liu et al., 2020b; Rafique et al., 2018; Yan et al., 2020; Yang et al., 2020a). However, most of them are slow in practice by suffering from a high order of stochastic first-order oracle call complexity, while others hinge on a special structure of the objective function for constructing the update (Liu et al., 2020b). *How to improve the convergence for generic non-convex strongly-concave min-max problems remains an active research area.* There are two lines of work trying to reduce the stochastic first-order oracle call complexity of stochastic algorithms for NCSC min-max optimization. The first line is to leverage the geometrical structure of the objective function, in particular the Polyak-Łojasiewicz (PL) condition (Liu et al., 2020b; Yang et al., 2020a). The second line is leverage variance-reduction techniques (Luo et al., 2020; Yang et al., 2020a; Huang et al., 2022; Xu et al., 2020; Rafique et al., 2018).

In this paper, we conduct a comprehensive study to improve the convergence for NCSC min-max optimization by leveraging the Polyak-Łojasiewicz (PL) condition of the objective function. A smooth function $h(x)$ satisfies μ-PL condition on $\mathbb{R}^d$, if for any $x\in\mathbb{R}^d$ there exists $\mu>0$ such that $\|\nabla h(x)\|^2\geq 2\mu(h(x)-h(x_*))$, where x_* denotes a global minimum of h. Although the PL condition has been utilized extensively to improve the convergence for minimization problems (Allen-Zhu et al., 2019; Arora et al., 2019; Charles and Papailiopoulos, 2018; Du et al., 2019; Hardt and Ma, 2017; Karimi et al., 2016; Lei et al., 2017; Li and Liang, 2018; Li and Yuan, 2017; Li and Li, 2018; Nguyen et al., 2017; Polyak, 1963; Reddi et al., 2016; Wang et al., 2018; Zhou et al., 2018; Zhou and Liang, 2017), its application to non-convex min-max problems remains rare (Liu et al., 2020b; Nouiehed et al., 2019; Yang et al., 2020a). The key difference between the present work and these previous studies is that we focus on **improving the dependence of convergence rate on the condition number** (the ratio of smoothness parameter to the PL constant) for NCSC min-max optimization. Our contributions are summarized below.

- **Algorithms.** We analyze a generic framework of proximal stage-wise stochastic (PES) method, which in design is similar to practical stochastic gradient methods for deep learning. In particular, the step sizes are decreased geometrically in a stage-wise manner. Various stochastic updates can be leveraged as a plug-in in the PES framework, including stochastic optimistic gradient descent ascent (OGDA) update, stochastic gradient descent ascent (SGDA) update, and min-max adaptive stochastic gradient (AdaGrad) update, and min-max STORM update (a recursive variance reduced method).

- **Analysis.** We conduct novel analysis of the proposed stochastic methods by establishing fast convergence in terms of both *the primal objective gap* and *the duality gap* under different PL conditions. The analysis is based on a novel Lyapunov function that consists of the primal objective gap and the duality gap of a regularized problem. The convergence

Table 1: Comparison of sample complexities for achieving ϵ-Objective Gap and and ϵ-Duality Gap. $P(x)$ is L-smooth and is assumed to obey μ-PL condition; $f(x,y)$ is ℓ-smooth in terms of x and y, and is μ_y strongly concave in terms of y. For duality gap convergence, it requires a stronger assumption that $f(x,y)$ satisfies x-side μ_x-PL condition. * marks the results that are not available in the original work but are derived by us.

	Objective Gap		Duality Gap		Remarks on Conditions
	$L=\ell+\frac{\ell^2}{\mu_y}$	$L<\ell+\frac{\ell^2}{\mu_y}$	$L=\ell+\frac{\ell^2}{\mu_y}$	$L<\ell+\frac{\ell^2}{\mu_y}$	
Stoc-AGDA (Yang et al., 2020a)	$O\left(\frac{\ell^5}{\mu^2\mu_y^4\epsilon}\right)$	$O\left(\frac{\ell^5}{\mu^2\mu_y^4\epsilon}\right)$	$O\left(\frac{\ell^7}{\mu^2\mu_x\mu_y^5\epsilon}\right)^*$	$O\left(\frac{\ell^7}{\mu^2\mu_x\mu_y^5\epsilon}\right)^*$	w/o strong concavity
PES-OGDA PES-SGDA	$\widetilde{O}\left(\frac{\ell^4}{\mu^2\mu_y^3\epsilon}\right)$	$\widetilde{O}\left(\frac{(L+\ell)^2}{\mu^2\mu_y\epsilon}\right)$	$\widetilde{O}\left(\frac{\ell^5}{\mu^2\mu_x\mu_y^3\epsilon}\right)$	$\widetilde{O}\left(\frac{(L+\ell)^2\ell}{\mu^2\mu_x\mu_y\epsilon}\right)$	w/ strong concavity
PES-OGDA PES-SGDA	$\widetilde{O}\left(\frac{\ell}{\min\{\mu,\mu_y\}\epsilon}\right)$	$\widetilde{O}\left(\frac{\ell}{\min\{\mu,\mu_y\}\epsilon}\right)$	$\widetilde{O}\left(\frac{\mu\ell}{\mu_x\min\{\mu,\mu_y\}\epsilon}\right)$	$\widetilde{O}\left(\frac{\mu\ell}{\mu_x\min\{\mu,\mu_y\}\epsilon}\right)$	ρ-weakly Convex $\rho<O(\mu)$
PES-AdaGrad	$\widetilde{O}\left(\left(\frac{\ell^4}{\mu^2\mu_y^3\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$	$\widetilde{O}\left(\left(\frac{(L+\ell)^2}{\mu^2\mu_y\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$	$\widetilde{O}\left(\left(\frac{\ell^5}{\mu^2\mu_x\mu_y^3\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$	$\widetilde{O}\left(\left(\frac{(L+\ell)^2\ell}{\mu^2\mu_x\mu_y\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$	Slow SG Growth (growth rate $\alpha\in(0,1/2)$)
PES-STORM	$\widetilde{O}\left(\frac{\ell^2}{\mu\mu_y^2\epsilon}\right)$	$\widetilde{O}\left(\frac{\ell^2}{\mu\mu_y^2\epsilon}\right)$	$\widetilde{O}\left(\frac{\ell^4}{\mu\mu_x\mu_y^3\epsilon}\right)$	$\widetilde{O}\left(\frac{\ell^4}{\mu\mu_x\mu_y^3\epsilon}\right)$	Individual Smoothness

of the primal objective gap only requires a weaker PL condition defined on the primal objective. For the convergence of the duality gap, the objective function satisfying a pointwise PL condition in terms of x is assumed.

- **Improvements.** We make non-trivial improvements of the basic convergence rate by improving its dependence on the condition number under different conditions, include the almost-convexity condition with a small weak-convexity parameter, the slow growth condition of stochastic gradient for AdaGrad update, the individual smoothness condition for STORM update. The dependence on the condition number can be reduced from $O(\ell^4/\mu^2)$ to $O(\ell^2/\mu)$ and $O(\ell/\mu)$ under appropriate conditions. We summarize our convergence results on both objective gap and duality gap in Table 1.

Finally, we demonstrate the effectiveness of the proposed methods on non-convex AUC maximization with a square surrogate loss and non-convex distributionally robust optimization. It is also notable that the proposed method has been used in the literature for maximizing a robust objective for deep AUC maximization (Yuan et al., 2020), which further demonstrates the effectiveness of the proposed methods.

2. Related Work

2.1 Non-Convex Min-Max Optimization

Recently, there has been an increasing interest on non-convex min-max optimization (Rafique et al., 2018; Jin et al., 2019; Lin et al., 2018, 2020a; Liu et al., 2020a; Lu et al., 2020; Nouiehed et al., 2019; Sanjabi et al., 2018; Thekumparampil et al., 2019; Ostrovskii et al., 2020; Lin et al., 2020b; Yang et al., 2020a; Luo et al., 2020; Xu et al., 2020; Huang et al., 2022; Tran-Dinh et al., 2020; Lu et al., 2020; Boţ and Böhm, 2020; Zhao, 2020; Wang et al., 2020; Yang et al., 2020b; Zhang et al., 2021b; Qiu et al., 2020; Han et al., 2021; Tran-Dinh et al., 2020; Huang et al., 2021; Xian et al., 2021; Luo and Chen, 2021; Fiez et al., 2021; Xu et al.,

2021; Lei et al., 2021). Below, we focus on related works on stochastic optimization for non-convex concave min-max problems. Rafique et al. (2018) proposed stochastic algorithms for solving non-smooth weakly-convex and concave problems based on a proximal point method (Rockafellar, 1976). They established a convergence to a nearly stationary point of the primal objective function in the order of $O(1/\epsilon^6)$, where ϵ is the level for the first-order stationarity. When the objective function is strongly concave in terms of y and has certain special structure, they can reduce the stochastic first-order oracle call complexity to $O(1/\epsilon^4)$. The same order stochastic first-order oracle call complexity was achieved in (Yan et al., 2020) for weakly-convex strongly-concave problems without a special structure of the objective function. Lin et al. (2020a) analyzed a single-loop stochastic gradient descent ascent method for smooth non-convex (strongly)-concave min-max problems. Their analysis yields an stochastic first-order oracle call complexity of $O(1/\epsilon^8)$ for smooth non-convex concave problems and $O(1/\epsilon^4)$ for smooth non-convex strongly-concave problems. Recently, Boţ and Böhm (2020) extends the analysis to stochastic alternating (proximal) gradient descent ascent method. Improved first-order convergence for smooth problems has been established by leveraging variance-reduction techniques in (Luo et al., 2020; Yang et al., 2020a; Huang et al., 2022; Xu et al., 2020; Rafique et al., 2018). However, none of these works explicitly use the PL condition to improve the convergence. Directly applying PL condition to the first-order convergence result leads to a stochastic first-order oracle call complexity worse than $O(1/\epsilon)$ for the objective gap.

2.2 PL Games

PL conditions have been considered in min-max games. For example, Nouiehed et al. (2019) assumed that $h_x(y) = -f(x, y)$ satisfies PL condition for any x, which is referred to as **y-side PL condition**. The authors utilize the condition to design deterministic multi-step gradient descent ascent method for finding a first-order stationary point. In contrast, we consider the objective is strongly concave in terms of y, which is stronger than y-side pointwise PL condition. Recently, Liu et al. (2018) assume a PL condition for a NCSC formulation of deep AUC maximization, in which the PL condition is defined over the primal objective $P(x) = \max_{y\in\mathcal{Y}} f(x, y)$, which is referred to as **primal PL condition**. They established a stochastic first-order oracle call complexity of $O(1/\epsilon)$ for the primal objective gap convergence only. However, their algorithm and analysis are not applicable to a general NCSC problem without a special structure. In contrast, our algorithm is more generic and simpler as well, and we derive stronger convergence result in terms of the duality gap. In addition, our analysis is based on a novel Lyapunov function that consists of the primal objective gap and the duality gap of a regularized function, which allows us to establish the convergence of both the primal objective gap and the duality gap.

More recently, Yang et al. (2020a) considered a class of smooth non-convex non-concave problems, which satisfy both the y-side PL condition and x-side PL condition[1]. They proposed stochastic alternating gradient descent ascent (Stoc-AGDA) algorithms and established a global convergence for a Lyapunov function $P(x_t) - P_* + \lambda(P(x_t) - f(x_t, y_t))$ for a constant λ, which directly implies the convergence for the primal objective gap. After some manipulation, we can also derive the convergence for the duality gap under the assumption

1. We notice that the x-side PL condition can be replaced by the primal PL condition for their analysis.

that x-side PL condition holds. This work is different from (Yang et al., 2020a) in several perspectives: (i) their algorithm is based on alternating gradient descent ascent method with polynomially decreasing or very small step sizes, in contrast our algorithm is based on stage-wise stochastic methods with geometrically decreasing step sizes. This feature makes our algorithm more amenable to deep learning applications (Yuan et al., 2020); (ii) we make use of strong concavity of the objective function in terms of y and develop stronger convergence results. In particular, our stochastic first-order oracle call complexities have better dependence on condition numbers.

Finally, we note that there are a lot of research on deep learning to justify the PL condition. PL condition of a risk minimization problem has been shown to hold globally or locally on some networks with certain structures, activation or loss functions (Allen-Zhu et al., 2019; Arora et al., 2019; Charles and Papailiopoulos, 2018; Du et al., 2019; Hardt and Ma, 2017; Li and Liang, 2018; Li and Yuan, 2017; Zhou and Liang, 2017). For example, in (Du et al., 2019), they have shown that if the width of a two layer neural network is sufficiently large, PL condition holds within a ball centered at the initial solution and the global optimum would lie in this ball. Allen-Zhu et al. (2019) further shows that in overparameterized deep neural networks with ReLU activation, PL condition holds for a global optimum around a random initial solution.

3. Preliminaries

We denote by $\|\cdot\|$ the Euclidean norm of a vector. A function $h(x)$ is λ-strongly convex on $\mathcal{X}$ if for any $x, x' \in \mathcal{X}$, $\nabla h(x')^\top(x-x') + \frac{\lambda}{2}\|x-x'\|^2 \leq h(x) - h(x')$. A function $h(x)$ is ρ-weakly convex on $\mathcal{X}$ if for any $x, x' \in \mathcal{X}$, $\nabla h(x')^\top(x-x') - \frac{\rho}{2}\|x-x'\|^2 \leq h(x) - h(x')$. $h(x)$ is L-smooth if its gradient is L-Lipchitz continuous, i.e., $\|\nabla h(x) - \nabla h(x')\| \leq L\|x-x'\|, \forall x, x' \in \mathcal{X}$. An L-smooth function is also a L-weakly convex function. A smooth function $h(x)$ satisfies μ-PL condition on $\mathbb{R}^d$, if for any $x \in \mathbb{R}^d$ there exists $\mu > 0$ such that $\|\nabla h(x)\|^2 \geq 2\mu(h(x) - h(x_*))$, where x_* denotes a global minimum of h. Let $\hat{x}(y) = \arg\min_{x'} f(x', y)$ denote the set of optimal x for the fixed y and when the context is clear we abuse the notation $\hat{x}(y)$ to denote any point in that set. Let $\hat{y}(x) = \arg\max_{y' \in \mathcal{Y}} f(x, y')$ denote the optimal y for the fixed x.

For simplicity, we let $z = (x, y)^\top$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathcal{Y}$, $F(z) = (\nabla_x f(x,y), -\nabla_y f(x,y))^\top$ and $\mathcal{G}(z;\xi) = (\nabla_x f(x,y;\xi), -\nabla_y f(x,y;\xi))^\top \in \mathbb{R}^{d+d'}$. We abuse the notations $\|z\|^2 = \|x\|^2 + \|y\|^2$ and $\|F(z) - F(z')\|^2 = \|\nabla_x f(x,y) - \nabla_x f(x',y')\|^2 + \|\nabla_y f(x,y) - \nabla_y f(x',y')\|^2$. Let $P(x) = \max_{y \in \mathcal{Y}} f(x,y)$. The primal objective gap of a solution $x \in \mathcal{X}$ is defined as $P(x) - \min_{x \in \mathcal{X}} P(x)$. Below, we state some assumptions that will be used in our analysis.

Assumption 1 *(i) F is ℓ-Lipchitz continuous, i.e., $\|F(z) - F(z')\| \leq \ell\|z - z'\|$, for any $z, z' \in Z$ (ii) $f(x,y)$ is μ_y-strongly concave in y for any x; (iii) $P(x) = \max_{y \in \mathcal{Y}} f(x,y)$ is L-smooth and has a non-empty optimal set.*

Remark: Assumption 1(i) implies that $f(x,y)$ is ℓ-smooth in terms of x for any $y \in \mathcal{Y}$. Note that under Assumption 1(i) and (ii), we can derive that $P(x)$ is $(\ell + \ell^2/\mu_y)$-smooth (Lin et al., 2020a). However, we note that the smoothness parameter L could be much smaller than $(\ell + \ell^2/\mu_y)$, and hence we keep dependence on L, ℓ, μ_y explicitly. For example, consider $f(x,y) = x^\top y - \frac{\mu_y}{2}\|y\|^2 - (\frac{1}{2\mu_y} - \frac{L}{2})\|x\|^2$, $\mathcal{Y} = \mathbb{R}^{d'}$ with $L \ll 1 \ll 1/\mu_y$. Then we can see

that $F(z)$ is $\ell = (1 + \frac{1}{\mu_y} - L)$-Lipchitz continuous. However, $P(x) = \frac{L}{2}\|x\|^2$ is L-smooth function and L could be much smaller than $\ell + \ell^2/\mu_y$.

The following assumption is assumed regarding the stochastic gradients unless specified otherwise.

Assumption 2 *There exists $\sigma > 0$ such that $\mathbb{E}[\|\nabla_x f(x, y; \xi) - \nabla_x f(x, y)\|^2] \leq \sigma^2$ and $\mathbb{E}[\|\nabla_y f(x, y; \xi) - \nabla_y f(x, y)\|^2] \leq \sigma^2$.*

Remark: In order to use a simple stochastic gradient descent ascent update, we need to impose a different (non-typical) assumption on stochastic gradients for analysis, i.e., there exists $B > 0$ such that $\mathbb{E}[\|\nabla_x f(x, y; \xi)\|^2] \leq B^2$ and $\mathbb{E}[\|\nabla_y f(x, y; \xi)\|^2] \leq B^2$.

If $f(x, y)$ is ℓ-smooth, it is then weakly convex with a coefficient ρ no greater than ℓ, however, ρ can be much less than ℓ. In order to explore possibilities for deriving faster convergence, we could leverage the weak convexity of $f(x, y)$ in terms of x.

Assumption 3 *$f(x, y)$ is ρ-weakly convex in terms of x for any $y \in \mathcal{Y}$ with $0 < \rho \leq \ell$.*

For example, consider $f(x, y) = \ell x^\top y - \frac{\mu_y}{2}\|y\|^2 - \frac{\rho}{2}\|x\|^2$ with $\rho \leq \ell$. Then $F(z)$ is $(\ell + \max(\rho, \mu_y))$-Lipchitz continuous. However, $f(x, y)$ is ρ-weakly convex in terms of x for any y.

In the algorithms, let $\Pi_{\bar{z}}(\mathcal{G}) \in \mathcal{Z}$ and $\Pi^{\gamma}_{\bar{z}, x_0}(\mathcal{G}) \in \mathcal{Z}$ be defined as

$$\begin{aligned} \Pi_{\bar{z}}(\mathcal{G}) &= \arg\min_{z \in \mathcal{Z}} \mathcal{G}^\top z + \frac{1}{2}\|z - \bar{z}\|^2, \\ \Pi^{\gamma}_{\bar{z}, x_0}(\mathcal{G}) &= \arg\min_{z \in \mathcal{Z}} \mathcal{G}^\top z + \frac{1}{2}\|z - \bar{z}\|^2 + \frac{\gamma}{2}\|x - x_0\|^2. \end{aligned} \tag{2}$$

Let $\mathcal{P}_{\mathcal{Y}}(\cdot)$ denote an Euclidean projection to $\mathcal{Y}$.

4. PL-Strongly-Concave Problems and Applications in Machine Learning

Firstly, based on the definition of PL condition given in the last section, we define the different PL conditions for the min-max problem.

Definition 1 *$f(x, y)$ satisfies a primal μ-PL condition for some constant $\mu > 0$ if $P(x) = \max_{y \in \mathcal{Y}} f(x, y)$ satisfies μ-PL condition, i.e., $\|\nabla P(x)\|^2 \geq 2\mu(P(x) - \min_{x'} P(x'))$.*

Definition 2 *$f(x, y)$ satisfies a x-side μ_x-PL condition for some constant $\mu_x > 0$ if for any $y \in \mathcal{Y}$, $f(x, y)$ satisfies μ_x-PL condition, i.e., $\forall y \in \mathcal{Y}$, $\|\nabla_x f(x, y)\|^2 \geq 2\mu_x(f(x, y) - f(\hat{x}(y), y))$.*

We define almost PL conditions as follows.

Definition 3 *$f(x, y)$ satisfies an ϵ-almost primal μ-PL condition if for $P(x) = \max_{y \in \mathcal{Y}} f(x, y)$, there exists $\mu > 0$ such that $\|\nabla P(x)\|^2 \geq 2\mu(P(x) - \min_{x'} P(x') - \epsilon)$, where $\epsilon > 0$ is the accuracy level.*

Definition 4 *$f(x, y)$ satisfies an ϵ-almost x-side μ_x-PL condition if there exists $\mu_x > 0$ such that $\|\nabla_x f(x, y)\|^2 \geq 2\mu_x(f(x, y) - f(\hat{x}(y), y) - \epsilon)$, where $\epsilon > 0$ is the accuracy level.*

It is not hard to see that convergence rates under the ϵ-almost x-side PL condition or the ϵ-almost primal PL condition are identical to that under the x-side PL condition or the primal PL condition, respectively. Therefore, in the convergence analysis we focus on the x-side PL condition and the primal PL condition.

We define two kinds of PL-strongly-concave problems as follows.

Definition 5 *$f(x,y)$ is primal-PL-strongly-concave if $f(x,y)$ satisfies a primal μ-PL condition and is strong concave in y for any x.*

Definition 6 *$f(x,y)$ is x-side-PL-strongly-concave if $f(x,y)$ satisfies a x-side μ_x-PL condition and is strong concave in y for any x.*

It has been shown in Yang et al. (2020a) that the x-side μ_x-PL condition of $f(x,y)$ is stronger than μ-PL condition of $P(x)$ under strong concavity of $f(x,y)$ in terms of y.

Lemma 7 (Lemma A.3 of Yang et al. (2020a)) *If $f(x,y)$ satisfies x-side μ_x-PL condition on $\mathbb{R}^d$ and is strongly concave in y, then $P(x) = \max_{y\in\mathcal{Y}} f(x,y)$ satisfies μ-PL condition for some $\mu \geq \mu_x$.*

Here we show cases where the x-side μ_x-PL condition holds or does not hold. Fortunately, x-side PL condition (Assumption 6) is only needed in Section 6 to develop duality gap convergence. We can construct a function that does not obey a x-side μ_x-PL condition but satisfies a primal μ-PL condition. Let us consider $f(x,y) = xy - \frac{1}{2}y^2 - \frac{1}{4}x^2$ and $\mathcal{Y} = \mathbb{R}$. First, we show that μ_x-PL condition does not hold. To this end, fix $y = 1$, we can see that $|\nabla_x f(x,y)|^2 = (1 - x/2)^2$, and $\min_{x\in\mathcal{X}} f(x,1) = \min_x x(1 - x/4) - \frac{1}{2} = -\infty$. Hence, for $x = 2 + \epsilon$, we have $|\nabla_x f(x,y)|^2 = (\epsilon/2)^2$ and $f(x,1) - \min_{x\in\mathcal{X}} f(x,1) = \infty$. However, there exists no constant μ_x such that $|\nabla_x f(x,y)|^2 \geq \mu_x(f(x,1) - \min_{x\in\mathcal{X}} f(x,1))$ for $\epsilon \to 0$. Second, we can see that $P(x) = \max_y f(x,y) = \frac{x^2}{4}$ satisfies μ-PL condition with $\mu = 1/2$. This argument together with Theorem 10 implies that our result for the convergence of the primal objective gap only requires a weaker μ-PL condition other than the x-side μ_x-PL condition imposed in (Yang et al., 2020a). An example that satisfies both the x-side μ_x-PL condition and y-side strong concavity is $f(x,y) = \frac{1}{2}x^2 + \sin^2 x \sin^2 y - 2y^2$, which is verified in Lemma 44 in the Appendix.

Instead of imposing the x-side PL condition as in (Yang et al., 2020a), we use primal PL condition (Assumption 4) for proving the convergence of the primal objective gap, and use x-side PL condition (Assumption 6) only for proving the convergence of the duality gap. Yang et al. (2020a) also makes an extra assumption that there exists a saddle point, i.e., there exists (x_*, y_*) such that $f(x_*, y) \leq f(x_*, y_*) \leq f(x, y_*)$. However, we show in Lemma 8 that a saddle point (x_*, y_*) exists for the x-side-PL-strongly-concave problem.

Lemma 8 *Assume $f(x,y)$ satisfies a x-side μ_x-PL condition and is strongly concave in y and let $x_* = \arg\min_{x'} P(x')$ where $P(x) = \max_{y\in\mathcal{Y}} f(x,y)$. Then $(x_*, \hat{y}(x_*))$ is a saddle point of $f(x,y)$.*

It has been shown in Lemma 2.1 of (Yang et al., 2020a) that if the x-side μ_x-PL condition holds, then the saddle points, global min-max points, and stationary points are equivalent when $\mathcal{Y} = \mathbb{R}^{d'}$, where global min-max points, and stationary points are defined as

1. (x_*, y_*) is a global min-max point if for any (x, y): $f(x_*, y) \leq f(x_*, y_*) \leq \max_{y'} f(x, y')$.

2. (x_*, y_*) is a stationary point if $\nabla_x f(x_*, y_*) = \mathbf{0}$ and $\nabla_y f(x_*, y_*) = \mathbf{0}$.

Next we show two concrete application examples of PL-strongly-concave problems in machine learning.

Deep AUC Maximization The area under the ROC curve (AUC) on a population level for a scoring function $h : \mathcal{X} \to \mathbb{R}$ is defined as

$$AUC(h) = \Pr(h(\mathbf{a}) \geq h(\mathbf{a}')|b = 1, b' = -1), \tag{3}$$

where $\mathbf{a}, \mathbf{a}' \in \mathbf{R}^{d_0}$ are data features, $b, b' \in \{-1, 1\}$ are the labels, $\mathbf{z} = (\mathbf{a}, b)$ and $\mathbf{z}' = (\mathbf{a}', b')$ are drawn independently from $\mathbb{P}$. By employing the squared loss as the surrogate for the indicator function which is commonly used by previous studies (Ying et al., 2016; Liu et al., 2018, 2020b), the deep AUC maximization problem can be formulated as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{z},\mathbf{z}'} \left[(1 - h(\mathbf{w}; \mathbf{a}) + h(\mathbf{w}; \mathbf{a}'))^2 | b = 1, b' = -1\right], \tag{4}$$

where $h(\mathbf{w}; \mathbf{a})$ denotes the prediction score for a data sample $\mathbf{a}$ made by a deep neural network parameterized by $\mathbf{w}$. It was shown in (Ying et al., 2016) that the above problem is equivalent to the following min-max problem:

$$\min_{(\mathbf{w},s,r)} \max_{y\in\mathbb{R}} f(\mathbf{w}, s, r, y) = \mathbb{E}_{\mathbf{z}}[F(\mathbf{w}, s, r, y, \mathbf{z})], \tag{5}$$

where

$$\begin{aligned} F(\mathbf{w}, s, r, y; \mathbf{z}) =& (1 - p)(h(\mathbf{w}; \mathbf{a}) - s)^2 \mathbb{I}_{[b=1]} + p(h(\mathbf{w}; \mathbf{a}) - r)^2 \mathbb{I}_{[b=-1]} \\ &+ 2(1 + y)(ph(\mathbf{w}; \mathbf{a})\mathbb{I}_{[b=-1]} - (1 - p)h(\mathbf{w}; \mathbf{a})\mathbb{I}_{[b=1]}) - p(1 - p)y^2, \end{aligned} \tag{6}$$

where $p = \Pr(b = 1)$ denotes the prior probability that an example belongs to the positive class, and $\mathbb{I}$ denotes an indicator function whose output is 1 when the condition holds and 0 otherwise. We denote the primal variable by $x = (\mathbf{w}, s, r)$.

Obviously, the problem (5) is strongly concave on dual variable y for any primal variable x. Also, in the next lemma we show that $f(x, y)$ satisfies an ϵ-almost μ-PL condition with a high probability following the theory of over-parameterized deep learning for minimization problems in Theorem 1, 2, 3, 5 of (Allen-Zhu et al., 2019). We put all the proof in the appendix.

Lemma 9 *Assume that input data $\{(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_n, b_n)\}$, where $\mathbf{a}_i \in \mathbb{R}^{d_0}, b_i \in \{-1, 1\}$, satisfies $\|\mathbf{a}_i\| = 1$ and $\|\mathbf{a}_i - \mathbf{a}_j\| \geq \delta$. Consider a deep neural network with $h_{i,0} = \phi(A\mathbf{a}_i), h_{i,l} = \phi(W_l h_{i,l-1}), l = 1, \ldots, \tilde{L}, \hat{b}_i = B^T h_{i,\tilde{L}}$ where $A \in \mathbb{R}^{m\times d_0}, W_l \in \mathbb{R}^{m\times m}, B \in \mathbb{R}^m$ are randomly initialized, and ϕ is the ReLU activation function. Let $\mathbf{w}$ denote the vectorization of $(\mathbf{W}_1, \cdots, \mathbf{W}_{\tilde{L}})$ and $x = (\mathbf{w}, s, r)$ denote the primal variable. $h(\mathbf{w}; a_i) = \hat{b}_i$ be the output logit for the i-th data. Take $m = \widetilde{\Omega}(poly(n, \tilde{L}, \delta^{-1}, \epsilon))$, then with a high probability over randomness of W_0, A, B for every x with $\|\mathbf{w} - \mathbf{w}_0\| \leq O(\frac{\log m}{\sqrt{m}})$, $f(x, y)$ satisfies an ϵ-almost primal μ-PL condition.*

Distributionally Robust Optimization (DRO) DRO problem (Namkoong and Duchi, 2017; Rafique et al., 2018) has a min-max formulation of

$$\min_x \max_{y\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^n y_i f_i(x) - r(y), \tag{7}$$

where $f_i(x)$ can be a loss function on the i-th data using a neural network backbone parameterized by x, and $r(y)$ is a reguralization function. The spirit of this formulation is to put more weights to the data points with high losses, thus to increase the robustness of models. It would be strongly concave on y for any x if $r(y)$ is a strongly convex function. It has been shown in proof of Lemma 2 of (Qi et al., 2021) that $f(x, y)$ satisfies an ϵ-almost x-side μ_x-PL condition with a high probability for a similar network structure as in the above Lemma 9.

5. Algorithms and Objective Gap Convergence

In this section, we make the assumption of the primal PL condition.

Assumption 4 *$P(x) = \max_{y\in\mathcal{Y}} f(x, y)$ satisfies μ-PL condition.*

We present the proposed stochastic method in Algorithm 1. We would like to point out that our method follows the proximal point framework analyzed in (Liu et al., 2020b; Rafique et al., 2018; Yan et al., 2020). In particular, the proposed method includes multiple consecutive stages. In each stage, we employ a stochastic algorithm to solve the following proximal problem approximately:

$$f_k(x, y) = f(x, y) + \frac{\gamma}{2}\|x - x_0^k\|^2, \tag{8}$$

where γ is an appropriate regularization parameter to make f_k to be strongly convex and strongly concave. The reference point $x_0^k = \bar{x}_{k-1}$ is updated after each stage, i.e., after each inner loop. Let $\hat{x}_k(y) = \arg\min_{x'} f_k(x', y)$ denote the optimal x for the fixed y and $\hat{y}_k(x) = \arg\max_{y'\in\mathcal{Y}} f_k(x, y')$ denote the optimal y for the fixed x.

However, there are some key differences between the proposed method from that are analyzed in (Liu et al., 2020b; Rafique et al., 2018; Yan et al., 2020). We highlight the differences below. First, our method explicitly leverages the PL condition of the objective function by decreasing $\eta_k, 1/T_k$ geometrically (e.g, $e^{-\alpha k}$ for some $\alpha > 0$). In contrast, Rafique et al. (2018) and Yan et al. (2020) proposed to decrease $\eta_k, 1/T_k$ polynomially (e.g., $1/k$). Second, the restating point and the reference point $(\bar{x}_{k-1}, \bar{y}_{k-1})$ is simply the averaged or sampled solution of stochastic updates in our employed stochastic algorithm $\mathcal{A}$. In contrast, Liu et al. (2020b) and Rafique et al. (2018) assumed a special structure of the objective function and leverage its structure to compute a restarted solution for y. This makes our method much simpler to be implemented but makes the analysis more involved.

For stochastic algorithm $\mathcal{A}$, one can employ many stochastic primal-dual methods to solve $\min_x \max_y f_k(x, y)$. We consider four well-known methods with different stochastic updates. Stochastic gradient descent ascent (SGDA) update (option I) and min-max adaptive stochastic gradient (MinMax-AdaGrad) update (option III) are mostly interesting

Algorithm 1 Proximal Stage Stochastic Method: PES-$\mathcal{A}$

1: Initialization: $\bar{x}_0 \in \mathbb{R}^d, \bar{y}_0 \in \mathcal{Y}, \gamma, T_1, \eta_1, a$.
2: Option III: $\bar{u}_0 = \nabla_x f(\bar{x}_0, \bar{y}_0; \bar{\xi}), \bar{v}_0 = \nabla_y f(\bar{x}_0, \bar{y}_0; \bar{\xi})$.
3: **for** $k = 1, 2, ..., K$ **do**
4: $x_0^k = \bar{x}_{k-1}$, $y_0^k = \bar{y}_{k-1}$;
5: Option I$\sim$ III: $(\bar{x}_k, \bar{y}_k) = \mathcal{A}(f, x_0^k, y_0^k, \eta_k, T_k, \gamma)$;
6: Option IV: $(\bar{x}_k, \bar{y}_k, \bar{u}_k, \bar{v}_k) = \mathcal{A}(f, x_0^k, y_0^k, \eta_k, T_k, \gamma, \bar{u}_{k-1}, \bar{v}_{k-1})$;
7: $\eta_{k+1} = \eta_k/a$, $\eta_{k+1}^y = \eta_k^y/a$, $T_{k+1} = aT_k$;
8: **end for**
9: **return** $(\bar{x}_K, \bar{y}_K)$.

to practitioners. Stochastic optimistic gradient descent ascent (OGDA) update (option II) yields an algorithm with provable convergence result under standard assumptions for smooth problems that is more interesting to theoreticians, which was originated from stochastic mirror prox method proposed by (Juditsky et al., 2011). Min-max stochastic update based on the recursive variance reduced estimator STORM (Cutkosky and Orabona, 2019) (option IV) can lead to an improved rate without using large mini-batch.

5.1 Basic Results

Below, we present the basic convergence results of Algorithm 1 by employing stochastic OGDA update. Let $\text{Gap}(x, y) = \max_{y' \in \mathcal{Y}} f(x, y') - \min_{x' \in \mathcal{X}} f(x', y)$ be the duality gap of (x, y) on f and $\text{Gap}_k(x, y) = \max_{y' \in \mathcal{Y}} f_k(x, y') - \min_{x' \in \mathcal{X}} f_k(x', y)$ be the duality gap of (x, y) on f_k.

Theorem 10 *Consider Algorithm 1 that uses Option II: OGDA update in subroutine Algorithm 2. Suppose Assumption 1, 2, 3, 4 hold. Take $\gamma = 2\rho$ and denote $\hat{L} = L + 2\rho$ and $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$. Define $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\text{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \text{Gap}(\bar{x}_0, \bar{y}_0)$. Then we set $\eta_k = \eta_0 \exp(-(k-1)\frac{2\mu}{c+2\mu}) \leq \frac{1}{2\sqrt{2}\ell}$, $T_k = \left\lceil \frac{212}{\eta_0 \min\{\rho, \mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right) \right\rceil$. After $K = \left\lceil \max\left\{ \frac{c+2\mu}{2\mu} \log \frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu} \log \frac{208\eta_0 \hat{L} K \sigma^2}{(c+2\mu)\epsilon} \right\} \right\rceil$ stages, we have $\mathbb{E}[\Delta_{K+1}] \leq \epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\max\left\{ \frac{\ell(L+\rho)\epsilon_0}{\mu \min\{\rho, \mu_y\}\epsilon}, \frac{(L+\rho)^2\sigma^2}{\mu^2 \min\{\rho, \mu_y\}\epsilon} \right\} \right)$.*

Remark. This result would imply that it takes $\widetilde{O}\left(\frac{(L+\ell)^2}{\mu^2\mu_y\epsilon}\right)$ stochastic first-order oracle calls to reach an ϵ-level objective gap by setting $\rho = \ell$ (i.e., $f(x, y)$ is ℓ-weakly convex in terms of x under Assumption 1). With the worse-case value of $L = \ell + \frac{\ell^2}{\mu_y}$ (i.e., the ℓ-smoothness of $f(x, y)$ and μ_y-strongly concavity can imply the $\ell + \ell^2/\mu_y$-smoothness of $P(x)$ (Nouiehed et al., 2019)) , the total stochastic first-order oracle call complexity would be no greater than $\widetilde{O}\left(\frac{\ell^4}{\mu^2\mu_y^3\epsilon}\right)$. This is better than the stochastic first-order oracle call complexity of stochastic AGDA method in the order of $O\left(\frac{\ell^5}{\mu^2\mu_y^4\epsilon}\right)$ (Yang et al., 2020a).

The above result is achieved by analysis based on a novel Lyapunov function that consists of the primal objective gap $P(x_0^k) - P(x_*)$ and the duality gap of the proximal function $f_k(x, y)$. As a result, we can induce the convergence of duality gap in next section of the original problem with some extra assumptions.

Algorithm 2 Stochastic Algorithm for Each Stage

Option I∼III: $\mathcal{A}(f, x_0, y_0, \eta, T, \gamma)$,
Option IV: $\mathcal{A}(f, x_0, y_0, \eta, T, \gamma, u_0, v_0)$

Initialization: $\tilde{z}_0 = z_0 = (x_0, y_0)$, Option III: $g_{1:0} = []$

Let $\{\xi_0, \xi_1, \ldots, \xi_T\}$ be independent random variables, and $\mathcal{G}_\gamma(z; \xi) = \begin{pmatrix} \nabla_x f(x, y; \xi) + \gamma(x - x_0) \\ -\nabla_y f(x, y; \xi) \end{pmatrix}$.

for $t = 1, ..., T$ **do**

Option I: SGDA update:

$z_t = \Pi^{\gamma}_{z_{t-1}, x_0}(\eta \mathcal{G}(z_{t-1}; \xi_{t-1}))$;

Option II: OGDA update:

$$\begin{aligned} z_t &= \Pi_{\tilde{z}_{t-1}}(\eta \mathcal{G}_\gamma(z_{t-1}; \xi_{t-1})); \\ \tilde{z}_t &= \Pi_{\tilde{z}_{t-1}}(\eta \mathcal{G}_\gamma(z_t; \xi_t)); \end{aligned}$$

Option III: Min-Max AdaGrad update:

$$\begin{aligned} &g_{1:t} = [g_{1:t-1}, \mathcal{G}_\gamma(z_t; \xi_t)], \text{ and } s_{t,i} = \|g_{1:t,i}\|_2; \\ &\text{Set } H_t = \delta I + \operatorname{diag}(s_t), \psi_t(z) = \tfrac{1}{2}\langle z - z_0, H_t(z - z_0)\rangle; \\ &z_{t+1} = \arg\min_{z \in \mathcal{Z}} \eta z^T \left(\frac{1}{t} \sum_{\tau=1}^{t} \mathcal{G}_\gamma(z_\tau; \xi_\tau) \right) + \frac{1}{t}\psi_t(z); \end{aligned}$$

Option IV: Min-Max STORM update:

$$\begin{aligned} x_t &= x_{t-1} - \eta^x u_{t-1}, \\ y_t &= y_{t-1} + \eta^y(\mathcal{P}_{\mathcal{Y}}(y_{t-1} + \lambda v_{t-1}) - y_{t-1})); \\ u_t &= (1 - a_x)u_{t-1} + \nabla_x f(x_t, y_t; \xi_t) - (1 - a_x)\nabla_x f(x_{t-1}, y_{t-1}; \xi_t), \\ v_t &= (1 - a_y)v_{t-1} + \nabla_y f(x_t, y_t; \xi_t) - (1 - a_y)\nabla_y f(x_{t-1}, y_{t-1}; \xi_t); \end{aligned}$$

end for

Option I∼III: **return** $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t, \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$.

Option IV: **return** $(x_\tau, y_\tau, u_\tau, v_\tau)$ with a random index $\tau \in \{1, \ldots, T\}$.

The convergence results of using SGDA update are similar to the results presented above except that σ^2 is replaced by the upper bound B^2 of stochastic gradients, i.e., there exists $B > 0$ such that $\mathbb{E}[\|\nabla_x f(x, y; \xi)\|^2] \leq B^2$ and $\mathbb{E}[\|\nabla_y f(x, y; \xi)\|^2] \leq B^2$. This is a more restrictive assumption but holds in many practical applications (Hazan and Kale, 2014; Duchi et al., 2011).

Note that the number of iterations in k-th stage (i.e. T_k) does not depend on the initial solution $(\bar{x}_{k-1}, \bar{y}_{k-1})$. In each stage, we do not expect to solve the sub-problem accurately, i.e, to some ϵ-accurate level. Instead, each stage just optimizes the sub-problem in order to make the upper bound of Lyaponov function $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ decrease by a constant factor. And as k grows, $(\bar{x}_{k-1}, \bar{y}_{k-1})$ becomes a better and better solution to the original problem.

Below we highlight the proof sketch. For details of proof, please refer to that of Theorem 27 in the Appendix. What we need from the sub-problem solver is that it can provide a

convergence bound as

$$\mathbb{E}[\text{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{C_1}{\eta_k T_k}\mathbb{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \|\hat{y}_k(\bar{x}_k) - y_0^k\|^2] + \eta_k C_2,$$

which has Lemma 24 as an instantiation of Option II: OGDA subroutine. It is notable that the above upper bound depends on the initial solution (x_0^k, y_0^k) of this stage. To achieve this the number of iterations T_k does not need to depend on (x_0^k, y_0^k) and the constants C_1 and C_2 do not depend on the initial solution and do not depend on the stage index k. In Lemma 24, we can see $C_1 = 1$ and $C_2 = 13\sigma^2$, independent of the initial solution $(\bar{x}_{k-1}, \bar{y}_{k-1})$ and stage index k.

Then by setting $\eta_k = \eta_0 \exp(-(k-1)\frac{2\mu}{c+2\mu})$, $T_k = \left\lceil \frac{212C_1}{\eta_0 \min\{\rho, \mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$, both independent of the initial solution, we can guarantee that

$$\mathbb{E}[\text{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{\min\{\rho, \mu_y\}}{212}\mathbb{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \|\hat{y}_k(\bar{x}_k) - y_0^k\|^2] + \eta_k C_2, \tag{9}$$

which then by some theoretical deduction can lead to

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_k] + \frac{8\eta_k \hat{L} C_2}{c+2\mu}. \tag{10}$$

As η_k decreases exponentially as k increases, we can then guarantee the convergence of the Δ_{k+1} and therefore the convergence of the original problem.

5.2 Improved Rates when $\rho < O(\mu)$

Our first improved rate is for almost convex function, whose weak convexity parameter ρ is small enough. Such a condition has been considered in the literature for improving the convergence of non-convex minimization problem (Yuan et al., 2019; Chen et al., 2019a; Lan and Yang, 2019). In particular, we consider ρ is smaller than $O(\mu)$.

Theorem 11 *Suppose Assumption 1, 2, 3 , 4 hold and* $0 < \rho \leq \frac{\mu}{8}$. *Take* $\gamma = \frac{\mu}{4}$. *Define* $\Delta_k = 475(P(x_0^k) - P(x_*)) + 57\text{Gap}_k(x_0^k, y_0^k)$ *and* $\epsilon_0 = \text{Gap}(\bar{x}_0, \bar{y}_0)$. *Then we can set* $\eta_k = \eta_0 \exp(-\frac{k-1}{16}) \leq \frac{1}{2\sqrt{2}\ell}$, $T_k = \left\lceil \frac{384}{\eta_0 \min\{\mu/8, \mu_y\}} \exp\left(\frac{k-1}{16}\right)\right\rceil$. *After* $K = \left\lceil \max\left\{16 \log \frac{1200\epsilon_0}{\epsilon}, 16 \log \frac{15600\eta_0 K\sigma^2}{\epsilon}\right\}\right\rceil$ *stages, we can have* $\mathbb{E}[\Delta_{K+1}] \leq \epsilon$. *The total stochastic first-order oracle call complexity is* $\widetilde{O}\left(\frac{\max\{\ell\epsilon_0, \sigma^2\}}{\min\{\mu, \mu_y\}\epsilon}\right)$.

5.3 Improved Rates of Using Min-Max AdaGrad

Similar to the literature of AdaGrad for improving convergence of convex and non-convex minimization problems (Duchi et al., 2011; Chen et al., 2019b, 2018), we can also improve the convergence of NCSC min-max optimization by leveraging Min-Max AdaGrad update. In particular, the dependence on $1/\epsilon$ can be further reduced if the growth rate of the stochastic gradients is slow. In particular, we have the following theorem regarding Min-Max AdaGrad.

Theorem 12 *(Informal) Suppose Assumption 1, 3, 4 hold. Let $g^k_{1:T_k}$ denote the cumulative matrix of gradients in k-th stage. Suppose $\|g^k_{1:T_k,i}\|_2 \leq \delta T_k^\alpha$ and with $\alpha \in (0, 1/2]$. Then by setting parameters appropriately, PES-AdaGrad has the total stochastic first-order oracle call complexity of $\widetilde{O}\left(\left(\frac{\delta^2(L+\rho)^2(d+d')}{\mu^2 \min\{\rho,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$ in order to have $\mathbb{E}[\Delta_{K+1}] \leq \epsilon$, where Δ_k is defined as in Theorem 10.*

Remark: First let us justify the slow growth condition $\|g^k_{1:T_k,i}\|_2 \leq \delta T_k^\alpha$. Supposing the stochastic gradients are bounded, it is clear that $\|g^k_{1:T_k,i}\|_2 \leq O(T_k^{1/2})$. But the $\|g^k_{1:T_k,i}\|_2$ can actually grow in a slower order than that because as the algorithm goes on, more and more data would become easy for the model and thus generate small gradients. For example, in deep learning where the models are able to memorize a lot of the training data. We verify this assumption in the Experiment section (Figure 3) and similar phenomena has been reported in previous research on min-max optimization (see Figure 2 of Liu et al. (2020a)). Indeed it has been observed that overparameterized deep neural networks exhibit interpolation phenomenon, meaning that the model will have zero gradient at every example in the limit (Zhang et al., 2021a). Nevertheless, it is still nontrivial to prove this condition rigorously and we leave it as an open problem. The improvements lies at when the stochastic gradient grows slowly, the sample complexity has a better dependence than $1/\epsilon$, i.e. $O(1/\epsilon^{1/(2(1-\alpha))}) \leq O(1/\epsilon)$ when $\alpha \in (0, 1/2)$.

5.4 Improved Rates of Using Min-Max STORM

Our last improved rate is by leveraging the recursive variance reduced stochastic gradient estimator called STORM (Cutkosky and Orabona, 2019). This estimator has been used for non-convex min-max optimization (Huang et al., 2022). However, to the best of our knowledge, an improved rate under a PL condition for a NCSC optimization problem has not been established before. We make an additional assumption about the problem (1).

Assumption 5 *$f(x, y; \xi)$ is ℓ-smooth in terms of x and y in expectation, i.e., $\mathbb{E}_\xi[\|G(z;\xi) - G(z';\xi)\|^2] \leq \ell\|z - z'\|^2$.*

Theorem 13 *(Informal) Suppose Assumption 1, 2, 4, 5 hold. By setting parameters appropriately, PES-STORM has the total stochastic first-order oracle call complexity of $\widetilde{O}\left(\frac{\ell^2}{\mu\mu_y^2\epsilon}\right)$ in order to have $\mathbb{E}[P(\bar{x}_K) - P(x_*)] \leq \epsilon$.*

Remark: Compared to the complexity of PES-OGDA as implied by Theorem 10, the complexity of PES-STORM has a better dependence on the PL constant μ, which is usually small in practice.

6. Duality Gap Convergence

In this section, we provide a stronger guarantee by analyzing the duality gap convergence utilizing some extra assumptions. Similar to (Yang et al., 2020a), we make the following assumption. However, a difference is that we can prove the existence of a saddle point instead of imposing it.

Assumption 6 *$h_y(x) = f(x,y)$ satisfies x-side μ_x-PL condition for any $y \in \mathcal{Y}$, i.e., $\|\nabla_x f(x,y)\|^2 \geq 2\mu_x(f(x,y) - \min_x f(x,y))$, for any $x, y \in \mathcal{Y}$.*

Using Theorem 10 and Assumption 6, we have

Corollary 14 *Under the same setting as in Theorem 10, and suppose Assumption 6 holds as well. To achieve $\mathbb{E}[\text{Gap}(\bar{x}_K, \bar{y}_K)] \leq \epsilon$, the total number of stochastic first-order oracle call is $\widetilde{O}\left(\max\left\{\frac{(\rho/\mu_x+1)\ell(L+\rho)\epsilon_0}{\mu\min\{\rho,\mu_y\}\epsilon}, \frac{(\rho/\mu_x+1)(L+\rho)^2\sigma^2}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right\}\right)$.*

Remark. Note that compared with the stochastic first-order oracle call complexity of the primal objective gap convergence, the stochastic first-order oracle call complexity of the duality gap convergence is worse by a factor of $\rho/\mu_x + 1$. When $f(x,y)$ is ℓ-weakly convex with $\rho = \ell$, the stochastic first-order oracle call complexity for having ϵ-level duality gap is $\widetilde{O}\left(\frac{(L+\ell)^2\ell}{\mu^2\mu_x\mu_y\epsilon}\right)$, which reduces to $\widetilde{O}\left(\frac{\ell^5}{\mu^2\mu_x\mu_y^3\epsilon}\right)$ for the worst-case value of L. This result is better than the stochastic first-order oracle call complexity of stochastic AGDA method in the order of $O\left(\frac{\ell^7}{\mu^2\mu_x\mu_y^5\epsilon}\right)$ that is derived by us based on the result of (Yang et al., 2020a) (c.f. Lemma 17 in the Supplement). In addition, when $f(x,y)$ is ρ-weakly convex with $\mu_x < \rho < \mu_y$, the stochastic first-order oracle call complexity of PES-OGDA for having ϵ-level duality gap is $\widetilde{O}\left(\frac{(L+\ell)^2}{\mu^2\mu_x\epsilon}\right)$. Further, when $\rho < \mu_x$, we can set $\rho = \mu_x$ and then the stochastic first-order oracle call complexity for having ϵ-level duality gap is $\widetilde{O}\left(\frac{(L+\ell)^2}{\mu^2\min\{\mu_x,\mu_y\}\epsilon}\right)$.

Using Theorem 11 and Assumption 6, we have

Corollary 15 *Under the same setting as in Theorem 11 and suppose Assumption 6 holds as well. To achieve $\mathbb{E}[\text{Gap}(\bar{x}_K, \bar{y}_K)] \leq \epsilon$, the total number of stochastic first-order oracle calls is $\widetilde{O}\left(\frac{\mu\max\{\ell\epsilon_0,\sigma^2\}}{\mu_x\min\{\mu,\mu_y\}\epsilon}\right)$.*

Remark: Compared with results in Theorem 10 and Corollary 14, the sample complexities in Theorem 11 and Corollary 15 have better dependence on μ, μ_y. In addition, by setting $\mu = \mu_x$, the rate in Corollary 11 becomes $\widetilde{O}(\frac{1}{\min\{\mu_x,\mu_y\}\epsilon})$, which matches that established optimal rate in (Yan et al., 2020) for μ_x-strongly convex and μ_y-strongly concave problems up to a logarithmic factor. But we only require x-side μ_x-PL condition instead of μ_x-strongly convex in terms of x.

7. Experiments

In this section, we show some empirical results to verify the effectiveness of the proposed algorithms for deep and non-deep learning tasks.

Non-convex Distributionally Robust Optimization. This task has been considered in (Rafique et al., 2018). The problem is formulated as:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathcal{S}}\sum_{i=1}^{n} y_i\phi(\log(1+\exp(-b_i\mathbf{a_i}^T x))) - \frac{\theta}{2}\|y - \frac{\mathbf{1}}{n}\|^2 \tag{11}$$

where $(\mathbf{a}_i, b_i)$ denotes feature label pair, $b_i \in \{-1, 1\}$, $\phi(s) = \log(1 + s/2)$ is a non-convex truncation function used to tackle outliers and noisy data, and $\mathcal{S}$ is a simplex. In experiments

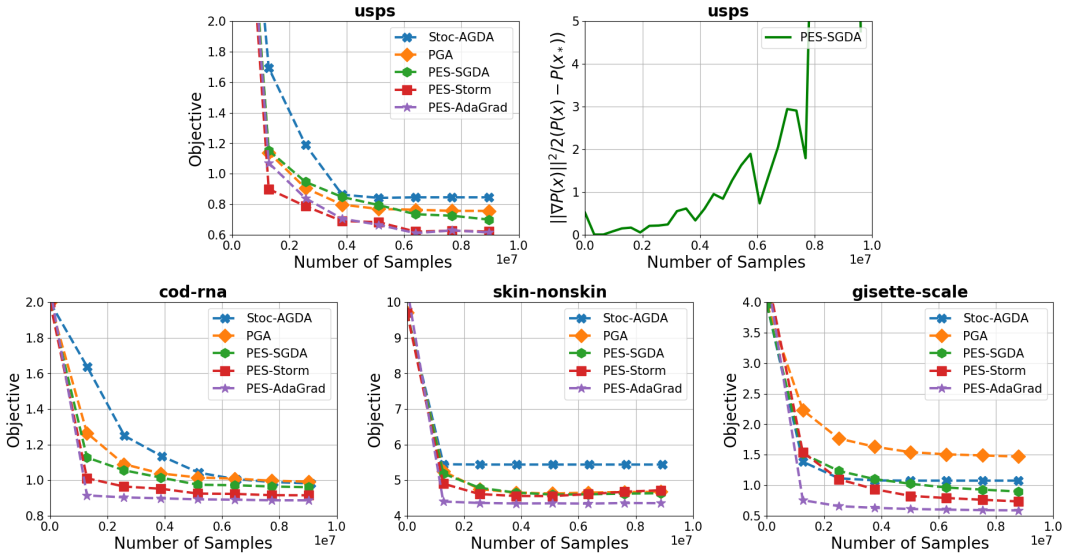

Figure 1: Results for Non-convex DRO.

the simplex constraint is handled by a projection algorithm in Duchi et al. (2008). We conduct experiments on four datasets from LibSVM website (Chang and Lin, 2011), i.e., gisette-scale, cod-rna, skin-nonskin and usps. For skin-nonskin, we randomly partition the dataset into training set and testing set of equal size. For other data sets we use the provided training/testing split. For usps, we make the first class to be the positive class and merge the other 9 classes into the negative class.

We first verify the PL condition of primal problem $P(x)$ of (11) empirically. We plot $\|\nabla P(x)\|^2/2(P(x) - P(x_*))$ in the second figure of the Figure 1.

We compare three variants of our method (PES-SGDA, PES-STORM, PES-AdaGrad) with two baselines Stoc-AGDA (Yang et al., 2020a), PGA (algorithm 1 (Rafique et al., 2018)). For all algorithms, we set $\theta = 10$. For Stoc-AGDA, the step sizes for x and y are set to be $\frac{\tau_1}{\lambda+t}$ and $\frac{\tau_1}{\lambda+t}$, respectively. τ_1 and τ_2 are tuned in $[1 \sim 1e3]$. γ is tuned in $[1 \sim 1e4]$. For PES-SGDA, PES-STORM, and PES-AdaGrad, we set $T_k = T_0 2^k$ and $\eta_k = \eta_0/2^k$, where T_0 and η_0 are tuned in $[500 \sim 5000]$, $[0.1, 0.05, 0.01, 0.001]$. γ is tuned in $[1 \sim 2000]$. The results are plotted in Figure 1. We can see that the proposed algorithms PES-SGDA, PES-STORM and PES-AdaGrad converge faster than the baselines in most cases. PES-STORM and PES-AdaGrad perform better than PES-SGDA on this task, which shows the potential to improve the performance by using STORM type variance techniques or adaptive methods when a task satisfies corresponding assumptions.

Deep AUC maximization. This task is similar to that considered in (Liu et al., 2020b). Deep AUC maximization with a square surrogate loss function is formulated as a NCSC min-max problem which has been introduced in the Section 4. We compare our algorithms, PES-SGDA (Option I), PES-OGDA (Option II), PES-AdaGrad (Option III), with five baseline methods, including stochastic gradient method (SGD) for solving a standard minimization formulation with cross-entropy loss, Stoc-AGDA (Yang et al., 2020a), PGA (algorithm 1 (Rafique et al., 2018)), PPD-SG and PPD-AdaGrad (Liu et al., 2020b)

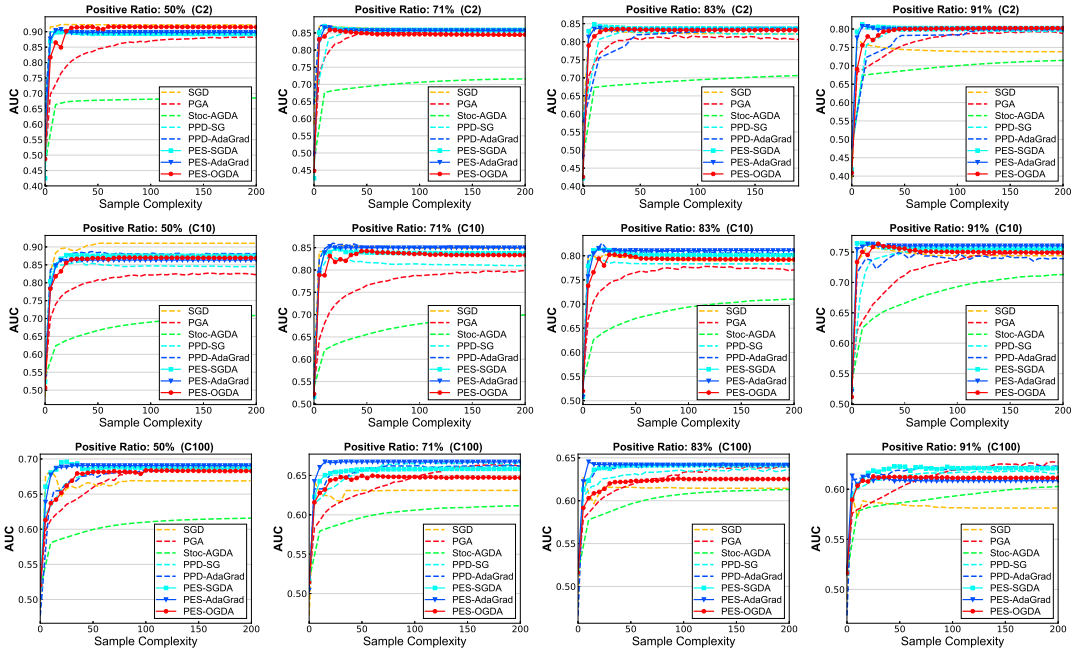

Figure 2: Comparison of testing AUC on Cat&Dog, CIFAR10, CIFAR100.

for solving the same AUC maximization problem. We learn ResNet-20 (He et al., 2016) with an ELU activation function.

For the parameter settings, we use a common stage-wise stepsize for SGD, i.e., the initial stepsize is divided by 10 at 40K, 60K of stochastic first-order oracle calls. For PPD-SG and PPD-AdaGrad, we follow the instructions in their works, i.e., $T_k = T_0 3^k, \eta_k = \eta_0/3^k$ and T_0, η_0 and γ are tuned in $[500 \sim 2000]$, $[0.1, 0.05, 0.01, 0.001]$, and $[100 \sim 2000]$, respectively. For Stoc-AGDA, the stepsize strategy follows $\frac{\tau_1}{\lambda+t}$, $\frac{\tau_2}{\lambda+t}$ for the dual and primal variables, respectively, where $\tau_1 \ll \tau_2$. The initial values τ_1, τ_2, λ are tuned in $[1, 5, 10, 15]$, $[5, 10, 15, 20]$, and $[1e3, 1e4]$, respectively. For our methods, we adopt the same strategy as PPD-SG and PPD-AdaGrad to tune the parameters.

We compare on three benchmark datasets: Cat&Dog (C2) (Elson et al., 2007), CIFAR10 (C10), CIFAR100 (C100) (Krizhevsky et al., 2009) which have 2, 10, 100 classes, respectively. To fit our task, we convert them into imbalanced datasets following the instructions in (Liu et al., 2020b). We firstly construct the binary dataset by splitting the original dataset into two portions with equal size (50% positive: 50% negative) and then we randomly remove $90\%, 80\%, 60\%$ data from negative samples on training data, which generate the imbalanced datasets with a positive:negative ratio of 91/9, 83/17, 71/29, respectively. We keep the testing data unchanged. We set the batch size to 128 for all datasets.

The testing AUC curve of all algorithms are reported in Figure 2, where the sample complexity indicates the number of samples used in the training up to 80K of stochastic first-order oracle calls. From the results, we can see that SGD works better (or similar to) than AUC-based methods on the balanced data (50%). However, PES-SGDA and PES-AdaGrad generally outperform SGD when the data is imbalanced, and outperforms PGA and Stoc-AGDA in almost all cases. In addition, the proposed methods performs similarly sometimes better than PPD-SG/PPD-AdaGrad except on C100 (91% positive ratio). This

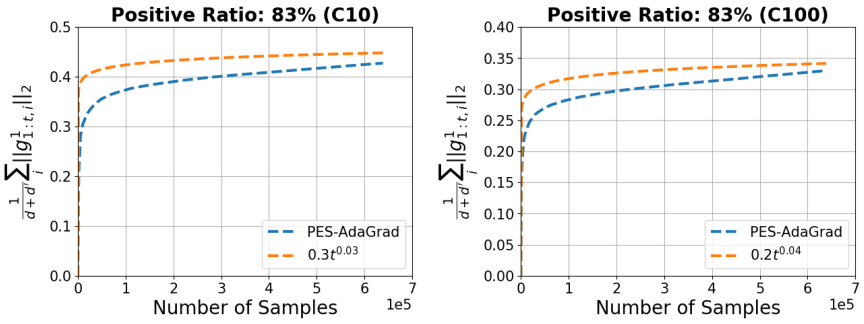

Figure 3: Verification of the Slow Growth Condition

is not surprising since PPD-SG/PPD-AdaGrad are designed for AUC maximization under the same PL condition by leveraging its structure and extra data samples for computing a restarted dual solution. In contrast, our algorithms directly use averaged dual solution for restarting. When the positive ratio is 91% on C100, we observe that PPD-AdaGrad performs better than our algorithms, showing that using the extra data samples may help in the extreme imbalanced cases. We also observe that the Stoc-AGDA performs worst in all cases with $O(\frac{1}{t})$ stepsize. For our methods, PES-SGDA and PES-AdaGrad perform generally better than PES-OGDA. In Figure 3, we verify the slow growth condition, i.e. $\|g^k_{1,T_k,i}\| \leq \delta T_k^\alpha$ used in the analysis of AdaGrad based algorithms, by plotting the $\frac{1}{d+d'}\sum_i \|g^1_{1:t,i}\|_2$ versus the sample complexity. We can seen that the growth of the aggregate of stochastic gradients is slower than the order of $O(\sqrt{T})$.

8. Conclusion

In this paper, we have presented generic stochastic algorithms for solving non-convex and strongly concave min-max optimization problems. We established convergence for both the objective gap and the duality gap under PL conditions of the objective function for different stochastic updates. The experiments on deep and non-deep learning tasks have demonstrated the effectiveness of our methods.

Acknowledgments

The feedback provided by the anonymous reviewers is greatly valued. We also wish to acknowledge the support received from the NSF Career Award #1844403, NSF Program #2110545, and NSF-Amazon Joint Program #2147253 for this work.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 242–252, 2019.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Pierre Bernhard and Alain Rapaport. On a theorem of danskin with an application to a theorem of von neumann-sion. *Nonlinear Analysis: Theory, Methods & Applications*, 24 (8):1163–1181, 1995.

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

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Zachary B. Charles and Dimitris S. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 744–753, 2018.

Zaiyi Chen, Yi Xu, Enhong Chen, and Tianbao Yang. Sadagrad: Strongly adaptive stochastic gradient methods. In *International Conference on Machine Learning*, pages 913–921. PMLR, 2018.

Zaiyi Chen, Yi Xu, Haoyuan Hu, and Tianbao Yang. Katalyst: Boosting convex katayusha for non-convex problems with a large condition number. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1102–1111, 2019a.

Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *7th International Conference on Learning Representations (ICLR)*, 2019b.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15210–15219, 2019.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations (ICLR)*, 2019.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159, 2011.

Jeremy Elson, John R. Douceur, Jon Howell, and Jared Saul. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In Peng Ning, Sabrina De Capitani di Vimercati, and Paul F. Syverson, editors, *Proceedings of the 2007 ACM Conference on Computer and Communications Security (CCS)*, pages 366–374, 2007.

Tanner Fiez, Chi Jin, Praneeth Netrapalli, and Lillian J Ratliff. Minimax optimization with smooth algorithmic adversaries. *arXiv preprint arXiv:2106.01488*, 2021.

Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 3864–3874, 2020.

Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *5th International Conference on Learning Representations (ICLR)*, 2017.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15 (71):2489–2512, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645, 2016.

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, 2019.

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

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal Machine Learning Research*, 23:36:1–36:70, 2022.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European*

Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pages 795–811, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Guanghui Lan and Yu Yang. Accelerated stochastic algorithms for nonconvex finite-sum and multiblock optimization. *SIAM Journal of Optimization*, 29(4):2753–2784, 2019.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2348–2358, 2017.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. *arXiv preprint arXiv:2105.03793*, 2021.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8157–8166, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 597–607, 2017.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5564–5574, 2018.

Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 6083–6093, 2020a.

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

Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with $O(1/n)$-convergence rate. In *International Conference on Machine Learning (ICML)*, pages 3189–3197, 2018.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2020a.

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. In *8th International Conference on Learning Representations (ICLR)*, 2020b.

Songtao Lu, Ioannis C. 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.

Luo Luo and Cheng Chen. Finding second-order stationary point for nonconvex-strongly-concave minimax problem. *arXiv preprint arXiv:2110.04814*, 2021.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2208–2216, 2016.

Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2971–2980, 2017.

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.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14905–14916, 2019.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

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

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2021.

Shuang Qiu, Zhuoran Yang, Xiaohan Wei, Jieping Ye, and Zhaoran Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *arXiv preprint arXiv:2008.10103*, 2020.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 314–323, 2016.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Maziar Sanjabi, Meisam Razaviyayn, and Jason D Lee. Solving non-convex non-concave min-max games under polyak-l ojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 12659–12670, 2019.

Quoc Tran-Dinh, Deyi Liu, and Lam M. Nguyen. Hybrid variance-reduced SGD algorithms for minimax problems with nonconvex-linear function. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Quoc Tran-Dinh, Deyi Liu, and Lam M Nguyen. Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. In *NeurIPS*, 2020.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021.

Tengyu Xu, Zhe Wang, Yingbin Liang, and H Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *arXiv e-prints*, pages arXiv–2006, 2020.

Zi Xu, Jingjing Shen, Ziqi Wang, and Yuhong Dai. Zeroth-order alternating randomized gradient projection algorithms for general nonconvex-concave minimax problems. *arXiv preprint arXiv:2108.00473*, 2021.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020a.

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

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*, pages 451–459, 2016.

Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2604–2614, 2019.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2103.15888*, 2021b.

Renbo Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv preprint arXiv:2003.04375*, 2020.

Renbo Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.

Renbo Zhao, William B Haskell, and Vincent YF Tan. An optimal algorithm for stochastic three-composite optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 428–437. PMLR, 2019.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3921–3932, 2018.

Yi Zhou and Yingbin Liang. Characterization of gradient dominance and regularity conditions for neural networks. *arXiv preprint arXiv:1710.06910*, 2017.

Appendix A. Convergence of Duality Gap by Stoc-AGDA Algorithm

To compare our algorithm with Stoc-AGDA in terms of convergence of duality gap, we derive Lemma 17 based on Theorem 3.3 of (Yang et al., 2020a). We first present an auxiliary lemma which is an extension of the Danskin's theorem.

Lemma 16 (Corollary of Theorem 1 of (Bernhard and Rapaport, 1995)) *In the min-max problem, when $f(x,y)$ is strong concave in y for any x then the gradient of the function $P(x) = \max_{y\in\mathcal{Y}}$ is $\nabla P(x) = \nabla_x f(x, \hat{y}(x))$ where $\hat{y}(x) = \arg\max_{y\in\mathcal{Y}} f(x,y)$.*

Then the convergence of duality gap by Stoc-AGDA algorithm is given in next lemma.

Lemma 17 *Supposes Assumption 1, 2, 4 and 6 hold. Stoc-AGDA would reach a ϵ-duality gap by a stochastic first-order oracle call complexity of $O\left(\frac{\ell^7}{\mu^2\mu_x\mu_y^5\epsilon}\right)$.*

Proof Yang et al. (2020a) defines the measure as following potential function,

$$P_t = E[P(x_t) - P(x_*)] + \frac{1}{10}E[P(x_t) - f(x_t, y_t)]. \tag{12}$$

By Theorem 3.3 of (Yang et al., 2020a), in Stoc-AGDA, $P_t \le \hat{\epsilon}$ after $O\left(\frac{\ell^5}{\mu^2\mu_y^4\hat{\epsilon}}\right)$ stochastic first-order oracle calls. It directly follows that the objective gap will be less than $\hat{\epsilon}$ after $O\left(\frac{\ell^5}{\mu^2\mu_y^4\hat{\epsilon}}\right)$ stochastic first-order oracle calls, i.e.,

$$P(x_t) - P(x_*) \le P_t \le \hat{\epsilon}. \tag{13}$$

Besides, after $O\left(\frac{\ell^5}{\mu^2\mu_y^4\hat{\epsilon}}\right)$ stochastic first-order oracle calls, we also have

$$f(x_t, \hat{y}(x_t)) - f(x_t, y_t) = P(x_t) - f(x_t, y_t) \le 10\hat{\epsilon}, \tag{14}$$

where the equality holds by the Lemma 16. By the μ_y-strong concavity of $f(x,\cdot)$, we have

$$\|y_t - \hat{y}(x_t)\|^2 \le \frac{f(x_t, \hat{y}(x_t)) - f(x_t, y_t)}{2\mu_y} \le \frac{5\hat{\epsilon}}{\mu_y}, \tag{15}$$

and

$$\begin{aligned}
\|\hat{y}(x_t) - y_*\|^2 &\le \frac{f(x_t, \hat{y}(x_t)) - f(x_t, y_*)}{2\mu_y} \\
&\le \frac{f(x_t, \hat{y}(x_t)) - f(x_*, y_*) + f(x_*, y_*) - f(x_t, y_*)}{2\mu_y} \\
&\le \frac{f(x_t, \hat{y}(x_t)) - f(x_*, y_*)}{2\mu_y} = \frac{P(x_t) - P(x_*)}{2\mu_y} \le \frac{\hat{\epsilon}}{2\mu_y}.
\end{aligned} \tag{16}$$

Thus,

$$\|y_t - y_*\|^2 \overset{(a)}{\le} 2\|y_t - \hat{y}(x_t)\|^2 + 2\|\hat{y}(x_t) - y_*\|^2 \le \frac{11\hat{\epsilon}}{\mu_y}, \tag{17}$$

where (a) holds since $\|\mathbf{a}-\mathbf{b}\|^2 = \|\mathbf{a}-\mathbf{c}+\mathbf{c}-\mathbf{b}\|^2 \leq 2\|\mathbf{a}-\mathbf{c}\|^2 + 2\|\mathbf{c}-\mathbf{b}\|^2$. Since $f(\cdot,\cdot)$ is ℓ-smooth and $f(\cdot, y)$ satisfies μ_x-PL condition for any y, we know $D(y) = \min_{x'} f(x', y)$ is smooth with coefficient $\ell + \frac{\ell^2}{\mu_x} \leq \frac{2\ell^2}{\mu_x}$ (Nouiehed et al., 2019; Yang et al., 2020a). Thus,

$$f(x_*, y_*) - f(\hat{x}(y_t), y_t) = D(y_*) - D(y_t) \leq \frac{2\ell^2}{2\mu_x}\|y_t - y_*\|^2 \leq \frac{11\ell^2\hat{\epsilon}}{\mu_x\mu_y}, \tag{18}$$

where the first equality holds by Lemma A.5 of (Nouiehed et al., 2019).

Then we know the duality gap is

$$\begin{aligned} f(x_t, \hat{y}(x_t)) - f(\hat{x}(y_t), y_t) &= f(x_t, \hat{y}(x_t)) - f(x_*, y_*) + f(x_*, y_*) - f(\hat{x}(y_t), y_t) \\ &\leq \hat{\epsilon} + \frac{11\ell^2\hat{\epsilon}}{\mu_x\mu_y}. \end{aligned} \tag{19}$$

To make the duality gap less than ϵ, we need $\hat{\epsilon} \leq O\left(\frac{\mu_x\mu_y\epsilon}{\ell^2}\right)$. Therefore, it takes $O\left(\frac{\ell^7}{\mu^2\mu_x\mu_y^5\epsilon}\right)$ stochastic first-order oracle calls to have a ϵ-duality gap for the Algorithm Stoc-AGDA that has been proposed in (Yang et al., 2020a). ■

Appendix B. Convergence Analysis of PES-SGDA

We present the convergence rate of primal gap and duality gap if SGDA update is used in Algorithm 2. Since the proof is similar to the version with Option II: OGDA as update, we include the proof in later sections together with the version using OGDA update.

Theorem 18 *Consider Algorithm 1 that uses Option I: SGDA update in subroutine Algorithm 2. Suppose Assumption 1, 3, 4 hold. Assume* $\mathbb{E}\|\nabla_x f(x,y;\xi)\|^2 \leq B^2$ *and* $\mathbb{E}\|\nabla_y f(x,y;\xi)\|^2 \leq B^2$. *Take* $\gamma = 2\rho$ *and denote* $\hat{L} = L + 2\rho$ *and* $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$. *Define* $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ *and* $\epsilon_0 = \mathrm{Gap}(\bar{x}_0, \bar{y}_0)$. *Then we can set* $\eta_k = \eta_0\exp(-(k-1)\frac{2\mu}{c+2\mu}) \leq \frac{1}{\rho}$, $T_k = \left\lceil \frac{212C_1}{\eta_0\min\{\rho,\mu_y\}}\exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$. *After* $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{80\eta_0\hat{L}KB_2}{(c+2\mu)\epsilon}\right\}\right\rceil$ *stages, we can have* $\Delta_{K+1} \leq \epsilon$. *The total stochastic first-order oracle call complexity is* $\widetilde{O}\left(\frac{(L+\rho)^2B^2}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)$.

Remark. The bounded stochastic gradient assumption i.e., $\mathbb{E}[\|\nabla_x f(x,y;\xi)\|^2] \leq B^2$ and $\mathbb{E}[\|\nabla_y f(x,y;\xi)\|^2] \leq B^2$ is only used for the analysis of our algorithm employing the SGDA update (Option I), and it is not used for other updates. It is notable that in min-max optimization it is an open question to get rid of the bounded stochastic gradient assumption for the vanilla SGDA updates in order to establish convergence bound for the duality gap. To the best of our knowledge, in the existing works over the gap convergence of stochastic min-max optimization that can achieve state-of-the-art complexity, they either use this bounded stochastic gradient assumption Nemirovski et al. (2009); Yan et al. (2020), or use some extra steps other than simple SGDA (Juditsky et al., 2011; Zhao, 2022; Hsieh et al., 2019; Zhao et al., 2019; Yang et al., 2020a).

Corollary 19 *Under the same setting as in Theorem 18 and suppose Assumption 6 holds aw well. To reach an ϵ-duality gap, it takes a total stochastic first-order oracle call complexity of $\widetilde{O}\left(\frac{(L+\rho)^2(\rho/\mu_x+1)B^2}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)$.*

Theorem 20 *Suppose Assumption 1, 6, 3 hold and $0<\rho\leq\frac{\mu}{8}$. Assume $\mathbb{E}\|\nabla_x f(x,y;\xi)\|^2\leq B^2$ and $\mathbb{E}\|\nabla_y f(x,y;\xi)\|^2\leq B^2$. Take $\gamma=2\rho$. Define $\Delta_k=475(P(x_0^k)-P(x_*))+57\mathrm{Gap}_k(x_0^k,y_0^k)$ and $\epsilon_0=\mathrm{Gap}(\bar{x}_0,\bar{y}_0)$. Then we can set $\eta_k=\eta_0\exp(-\frac{k-1}{16})\leq\frac{1}{\rho}$, $T_k=\left\lceil\frac{768}{\eta_0\min\{\mu/8,\mu_y\}}\exp\left(\frac{k-1}{16}\right)\right\rceil$. After $K=\left\lceil\max\left\{16\log\frac{1200\epsilon_0}{\epsilon},16\log\frac{6000\eta_0KB^2}{\epsilon}\right\}\right\rceil$ stages, we can have $\Delta_{K+1}\leq\epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\frac{B^2}{\min\{\mu,\mu_y\}\epsilon}\right)$.*

Corollary 21 *Under the same setting as in Theorem 20 and suppose Assumption 6 holds as well. To reach an ϵ-duality gap, it takes total stochastic first-order oracle call complexity of $\widetilde{O}\left(\frac{(\mu/\mu_x+1)B^2}{\min\{\mu,\mu_y\}\epsilon}\right)$.*

Appendix C. One Stage Analysis of PES-OGDA

We need the following lemmas from (Nemirovski, 2004).

Lemma 22 (Lemma 3.1 of (Nemirovski, 2004)) *For $z_0\in\mathcal{Z}$, let $w_1=\Pi_{z_0}(\zeta_1)$, $w_2=\Pi_{z_0}(\zeta_2)$. For any $z\in\mathcal{Z}$,*

$$\langle\zeta_2,w_1-z\rangle\leq\frac{1}{2}\|z-z_0\|^2-\frac{1}{2}\|w_2-z\|^2-\frac{1}{2}\|w_1-z_0\|^2-\frac{1}{2}\|w_1-w_2\|^2+\|\zeta_1-\zeta_2\|^2. \quad (20)$$

Lemma 23 (Corollary 2 of (Juditsky et al., 2011)) *Let $\zeta_1,\zeta_2,...$ be a sequence, we define a corresponding sequence $\{v_t\in\mathcal{Z}\}_{t=0}^T$ as*

$$v_t=\Pi_{v_{t-1}}(\zeta_t),v_0\in\mathcal{Z}, \quad (21)$$

we have for any $u\in\mathcal{Z}$,

$$\sum_{t=1}^T\langle\zeta_t,v_{t-1}-u\rangle\leq\frac{1}{2}\|v_0-u\|^2+\frac{1}{2}\sum_{t=1}^T\|\zeta_t\|^2. \quad (22)$$

Next we present the lemma that guarantees the converge of one call of Algorithm 2 with Option II: OGDA update.

Lemma 24 *Suppose $f(x,y)$ is convex-concave and Assumption 2 holds. By running Algorithm 2 with OGDA update and input $(f,x_0,y_0,\eta\leq\frac{1}{4\sqrt{3}\ell},T)$, we have*

$$\mathbb{E}[f(\bar{x},\hat{y}(\bar{x}))-f(\hat{x}(\bar{y}),\bar{y})]\leq\frac{1}{\eta T}\mathbb{E}(\|\hat{x}(\bar{y})-x_0\|^2+\|\hat{y}(\bar{x})-y_0\|^2)+13\eta\sigma^2. \quad (23)$$

Proof [Proof of Lemma 24] Applying Lemma 22 with $z_0 = \tilde{z}_{t-1}$, $\zeta_1 = \eta\mathcal{G}(z_{t-1};\xi_{t-1})$, $\zeta_2 = \eta\mathcal{G}(z_t;\xi_t)$, and accordingly $w_1 = z_t$, $w_2 = \tilde{z}_t$, we get for any $z \in \mathcal{Z}$,

$$\begin{aligned}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle \leq & \frac{1}{2\eta}[\|z - \tilde{z}_{t-1}\|^2 - \|\tilde{z}_t - z\|^2] - \frac{1}{2\eta}[\|z_t - \tilde{z}_{t-1}\|^2 + \|z_t - \tilde{z}_t\|^2] \\ & + \eta\|\mathcal{G}(z_{t-1};\xi_{t-1}) - \mathcal{G}(z_t;\xi_t)\|^2.\end{aligned} \tag{24}$$

Taking average over $t = 1, ..., T$ and by the convexity of $f(x, y)$ in x, we have for any $x \in \mathcal{X}$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle \leq & \frac{\|z - \tilde{z}_0\|^2}{2\eta T} - \frac{1}{2\eta T}\sum_{t=1}^{T}(\|z_t - \tilde{z}_{t-1}\|^2 + \|z_t - \tilde{z}_t\|^2) \\ & + \frac{\eta}{T}\sum_{t=1}^{T}\|\mathcal{G}(z_{t-1};\xi_{t-1}) - \mathcal{G}(z_t;\xi_t)\|^2 \\ \leq & \frac{\|z - z_0\|^2}{2\eta T} - \frac{1}{2\eta T}\sum_{t=1}^{T}(\|z_t - \tilde{z}_{t-1}\|^2 + \|z_t - \tilde{z}_t\|^2) + \frac{3\eta}{T}\sum_{t=1}^{T}\|F(z_{t-1}) - F(z_t)\|^2 \\ & + \frac{3\eta}{T}\sum_{t=1}^{T}(\|\mathcal{G}(z_t;\xi_t) - F(z_t)\|^2 + \|\mathcal{G}(z_{t-1};\xi_{t-1}) - F(z_{t-1})\|^2),\end{aligned} \tag{25}$$

where the last inequality is due to $\left\|\sum_{k=1}^{K}\mathbf{a}_k\right\|^2 \leq K\sum_{k=1}^{K}\|\mathbf{a}_k\|^2$. Note that

$$\begin{aligned}&\sum_{t=1}^{T}(\|z_t - \tilde{z}_{t-1}\|^2 + \|z_t - \tilde{z}_t\|^2) = \sum_{t=0}^{T-1}\|z_{t+1} - \tilde{z}_t\|^2 + \sum_{t=1}^{T}\|z_t - \tilde{z}_t\|^2 \\ &= \sum_{t=1}^{T-1}\|z_{t+1} - \tilde{z}_t\|^2 + \|z_1 - \tilde{z}_0\|^2 + \sum_{t=1}^{T-1}\|z_t - \tilde{z}_t\|^2 \geq \frac{1}{2}\sum_{t=1}^{T-1}\|z_t - z_{t+1}\|^2 + \|z_1 - \tilde{z}_0\|^2 \\ &\geq \frac{1}{2}\sum_{t=0}^{T-1}\|z_t - z_{t+1}\|^2 = \frac{1}{2}\sum_{t=1}^{T}\|z_{t-1} - z_t\|^2.\end{aligned} \tag{26}$$

By the ℓ-smoothness of $f(x, y)$, we have

$$\begin{aligned}\|F(z_{t-1}) - F(z_t)\|^2 = & \|\nabla_x f(x_t, y_t) - \nabla_x f(x_{t-1}, y_{t-1})\|^2 + \|\nabla_y f(x_t, y_t) - \nabla_y f(x_{t-1}, y_{t-1})\|^2 \\ \leq & 2\|\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_{t-1})\|^2 + 2\|\nabla_x f(x_t, y_{t-1}) - \nabla_x f(x_{t-1}, y_{t-1})\|^2 \\ & + 2\|\nabla_y f(x_t, y_t) - \nabla_y f(x_t, y_{t-1})\|^2 + 2\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_{t-1}, y_{t-1})\|^2 \\ \leq & 2\ell^2\|y_t - y_{t-1}\|^2 + 2\ell^2\|x_t - x_{t-1}\|^2 + 2\ell^2\|y_t - y_{t-1}\|^2 + 2\ell^2\|x_t - x_{t-1}\|^2 \\ = & 4\ell^2\|z_{t-1} - z_t\|^2.\end{aligned}$$

Denote $\Theta_t = F(z_t) - \mathcal{G}(z_t;\xi_t)$. With the above two inequalities, (25) becomes

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle \\
&\le \frac{\|z-z_0\|^2}{2\eta T} - \frac{1}{4\eta T}\sum_{t=1}^{T}\|z_{t-1}-z_t\|^2 + \frac{12\eta\ell^2}{T}\sum_{t=1}^{T}\|z_{t-1}-z_t\|^2 + \frac{3\eta}{T}\sum_{t=1}^{T}(\|\Theta_t\|^2 + \|\Theta_{t-1}\|^2) \\
&\le \frac{\|z-z_0\|^2}{2\eta T} + \frac{3\eta}{T}\sum_{t=1}^{T}(\|\Theta_t\|^2 + \|\Theta_{t-1}\|^2),
\end{aligned} \tag{27}$$

where the last inequality holds because $\eta \le \frac{1}{4\sqrt{3}\ell}$.

Define a virtual sequence $\{\hat{z}_t \in \mathcal{X}\}_{t=0}^{T}$ as

$$\hat{z}_t = \Pi_{\hat{z}_{t-1}}(\eta\Theta_t), \hat{z}_0 = z_0. \tag{28}$$

Applying Lemma 23 with $\zeta_t = \eta\Theta_t = \eta(F(z_t) - \mathcal{G}(z_t;\xi_t)), v_t = \hat{z}_t$ and $u = z$, we have for any $z \in \mathcal{Z}$,

$$\frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, \hat{z}_{t-1} - z\rangle \le \frac{1}{2\eta T}\|z_0 - z\|^2 + \frac{\eta}{2T}\sum_{t=1}^{T}\|\Theta_t\|^2. \tag{29}$$

Using (27) and (29), we get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\langle F(z_t), z_t - z\rangle = \frac{1}{T}\sum_{t=1}^{T}[\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle + \langle \Theta_t, z_t - z\rangle] \\
&= \frac{1}{T}\sum_{t=1}^{T}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, z_t - \hat{z}_{t-1}\rangle + \frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, \hat{z}_{t-1} - z\rangle \\
&\le \frac{1}{\eta T}\|z_0 - z\|^2 + \frac{\eta}{T}\sum_{t=1}^{T}\left(\frac{7}{2}\|\Theta_{x,t}\|^2 + 3\|\Theta_{x,t-1}\|^2\right) + \frac{1}{T}\sum_{t=1}^{T}\langle \Theta_{x,t}, x_t - \hat{x}_{t-1}\rangle.
\end{aligned} \tag{30}$$

Note

$$\mathbb{E}[\langle \Theta_t, z_t - \hat{z}_{t-1}\rangle | z_t, \hat{z}_{t-1}, \Theta_{t-1}, ..., \Theta_0] = 0,$$

and by Assumption 2

$$\mathbb{E}[\|\Theta_t\|^2 | z_t, \tilde{z}_{t-1}, \Theta_{t-1}, ..., \Theta_0] \le 2\sigma^2.$$

Thus, taking expectation on both sides of (30), we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle F(z_t), z_t - z\rangle\right] \le \frac{1}{\eta T}\mathbb{E}\left[\|z_0 - z\|^2\right] + 13\eta\sigma^2. \tag{31}$$

By the fact $f(x,y)$ is convex-concave,

$$\begin{aligned}\mathbb{E}[f(\bar{x},y)-f(x,\bar{y})] \leq & \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f(x_t,y)-f(x,y_t))\right] \\ = & \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f(x_t,y)-f(x_t,y_t)+f(x_t,y_t)-f(x,y_t))\right] \\ \leq & \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\langle -\nabla_y f(x_t,y_t), y_t-y\rangle + \langle \nabla_x f(x_t,y_t), x_t-x\rangle)\right] \\ = & \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle F(z_t), z_t-z\rangle\right] \leq \frac{1}{\eta T}E[\|z_0-z\|^2]+13\eta\sigma^2.\end{aligned} \tag{32}$$

Then we can conclude by plugging in $z=(x,y)=(\hat{x}(\bar{y}),\hat{y}(\bar{x}))$. ■

Appendix D. Proof of Theorem 10 and Theorem 18

Before we prove these two theorems, we first present two lemmas from (Yan et al., 2020) and we introduce Theorem 27 that unifies the proof of Theorem 10 and Theorem 18.

Lemma 25 (Lemma 1 of (Yan et al., 2020)) *Suppose a function $h(x,y)$ is λ_1-strongly convex in x and λ_2-strongly concave in y. Consider the following problem*

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} h(x,y),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are convex sets. Denote $\hat{x}_h(y)=\arg\min_{x'\in\mathcal{X}} h(x',y)$ and $\hat{y}_h(x)=\arg\max_{y'\in\mathcal{Y}} h(x,y')$. Suppose we have two solutions (x_0,y_0) and (x_1,y_1). Then the following relation between variable distance and duality gap holds

$$\begin{aligned}\frac{\lambda_1}{4}\|\hat{x}_h(y_1)-x_0\|^2+\frac{\lambda_2}{4}\|\hat{y}_h(x_1)-y_0\|^2 \leq & \max_{y'\in\mathcal{Y}} h(x_0,y')-\min_{x'\in\mathcal{X}} h(x',y_0) \\ & +\max_{y'\in\mathcal{Y}} h(x_1,y')-\min_{x'\in\mathcal{X}} h(x',y_1).\end{aligned} \tag{33}$$

Lemma 26 (Lemma 5 of (Yan et al., 2020)) *We have the following lower bound for* $\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k)$

$$\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k)\geq\frac{3}{50}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1})+\frac{4}{5}(P(x_0^{k+1})-P(x_0^k)),$$

where $x_0^{k+1}=\bar{x}_k$ and $y_0^{k+1}=\bar{y}_k$.

We will introduce the following theorem that can unify the proof of Theorem 10 and Theorem 18 since their have pretty similar forms of bounds in solving the subproblem.

Theorem 27 *Suppose Assumption 1 and Assumption 4 hold. Assume we have a subroutine in the k-th stage of Algorithm 1 that can return $\bar{x}_k, \bar{y}_k$ such that*

$$\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{C_1}{\eta_k T_k}\mathbb{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \|\hat{y}_k(\bar{x}_k) - y_0^k\|^2] + \eta_k C_2, \tag{34}$$

where C_1 and C_2 are constants corresponding to the specific subroutine. Take $\gamma = 2\rho$ and denote $\hat{L} = L + 2\rho$ and $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$. Define $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \mathrm{Gap}(\bar{x}_0, \bar{y}_0)$. Then we can set $\eta_k = \eta_0 \exp(-(k-1)\frac{2\mu}{c+2\mu})$, $T_k = \left\lceil \frac{212C_1}{\eta_0 \min\{\rho,\mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$. After $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{16\eta_0\hat{L}KC_2}{(c+2\mu)\epsilon}\right\}\right\rceil$ stages, we can have $\Delta_{K+1} \leq \epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\max\left\{\frac{(L+\rho)C_1\epsilon_0}{\eta_0\mu\min\{\rho,\mu_y\}\epsilon}, \frac{(L+\rho)^2C_2}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right\}\right)$.

Proof [Proof of Theorem 27] Since $f(x,y)$ is ρ-weakly convex in x for any y, $P(x) = \max_{y' \in \mathcal{Y}} f(x,y')$ is also ρ-weakly convex. Taking $\gamma = 2\rho$, we have

$$\begin{aligned}
P(\bar{x}_{k-1}) &\geq P(\bar{x}_k) + \langle \nabla P(\bar{x}_k), \bar{x}_{k-1} - \bar{x}_k\rangle - \frac{\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&= P(\bar{x}_k) + \langle \nabla P(\bar{x}_k) + 2\rho(\bar{x}_k - \bar{x}_{k-1}), \bar{x}_{k-1} - \bar{x}_k\rangle + \frac{3\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&\overset{(a)}{=} P(\bar{x}_k) + \langle \nabla P_k(\bar{x}_k), \bar{x}_{k-1} - \bar{x}_k\rangle + \frac{3\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&\overset{(b)}{=} P(\bar{x}_k) - \frac{1}{2\rho}\langle \nabla P_k(\bar{x}_k), \nabla P_k(\bar{x}_k) - \nabla P(\bar{x}_k)\rangle + \frac{3}{8\rho}\|\nabla P_k(\bar{x}_k) - \nabla P(\bar{x}_k)\|^2 \\
&= P(\bar{x}_k) - \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{1}{4\rho}\langle \nabla P_k(\bar{x}_k), \nabla P(\bar{x}_k)\rangle + \frac{3}{8\rho}\|\nabla P(\bar{x}_k)\|^2,
\end{aligned} \tag{35}$$

where (a) and (b) hold by the definition of $P_k(x)$.

Rearranging the terms in (35) yields

$$\begin{aligned}
P(\bar{x}_k) - P(\bar{x}_{k-1}) &\leq \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 + \frac{1}{4\rho}\langle \nabla P_k(\bar{x}_k), \nabla P(\bar{x}_k)\rangle - \frac{3}{8\rho}\|\nabla P(\bar{x}_k)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 + \frac{1}{8\rho}(\|\nabla P_k(\bar{x}_k)\|^2 + \|\nabla P(\bar{x}_k)\|^2) - \frac{3}{8\rho}\|P(\bar{x}_k)\|^2 \\
&= \frac{1}{4\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{1}{4\rho}\|\nabla P(\bar{x}_k)\|^2 \\
&\overset{(b)}{\leq} \frac{1}{4\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{\mu}{2\rho}(P(\bar{x}_k) - P(x_*)),
\end{aligned} \tag{36}$$

where (a) holds by using $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$, and (b) holds by the μ-PL property of $P(x)$.

Thus, we have

$$(4\rho + 2\mu)\left(P(\bar{x}_k) - P(x_*)\right) - 4\rho(P(\bar{x}_{k-1}) - P(x_*)) \leq \|\nabla P_k(\bar{x}_k)\|^2. \tag{37}$$

Since $\gamma = 2\rho$, $f_k(x,y)$ is ρ-strongly convex in x and μ_y strong concave in y. Apply Lemma 25 to f_k, we know that

$$\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\mu_y}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2 \leq \mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k). \tag{38}$$

By the setting of $\eta_k = \eta_0 \exp\left(-(k-1)\frac{2\mu}{c+2\mu}\right)$, and $T_k = \left\lceil \frac{212C_1}{\eta_0 \min\{\rho,\mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$, we note that $\frac{C_1}{\eta_k T_k} \leq \frac{\min\{\rho,\mu_y\}}{212}$. Applying (34), we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \eta_k C_2 + \frac{1}{53}\mathbb{E}\left[\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\mu_y}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2\right] \\
&\leq \eta_k C_2 + \frac{1}{53}\mathbb{E}\left[\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right].
\end{aligned} \tag{39}$$

Since $P(x)$ is L-smooth and $\gamma = 2\rho$, then $P_k(x)$ is $\hat{L} = (L + 2\rho)$-smooth. According to Theorem 2.1.5 of (Nesterov, 2004), we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla P_k(\bar{x}_k)\|^2] \leq 2\hat{L}\mathbb{E}(P_k(\bar{x}_k) - \min_{x\in\mathbb{R}^d} P_k(x)) \leq 2\hat{L}\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&= 2\hat{L}\mathbb{E}[4\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&\leq 2\hat{L}\mathbb{E}\left[4\left(\eta_k C_2 + \frac{1}{53}\left(\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right)\right) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right] \\
&= 2\hat{L}\mathbb{E}\left[4\eta_k C_2 + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{155}{53}\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right].
\end{aligned} \tag{40}$$

Applying Lemma 26 to (40), we have

$$\begin{aligned}
\mathbb{E}[\|\nabla P_k(\bar{x}_k)\|^2] \leq 2\hat{L}\mathbb{E}\Bigg[&4\eta_k C_k + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) \\
&- \frac{155}{53}\left(\frac{3}{50}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k))\right)\Bigg] \\
= 2\hat{L}\mathbb{E}\Bigg[&4\eta_k C_2 + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{93}{530}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \frac{124}{53}(P(x_0^{k+1}) - P(x_0^k))\Bigg].
\end{aligned}$$

Combining this with (37), rearranging the terms, and defining a constant $c = 4\rho + \frac{248}{53}\hat{L} \in O(L + \rho)$, we get

$$\begin{aligned}
&(c + 2\mu)\,\mathbb{E}[P(x_0^{k+1}) - P(x_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1})] \\
&\leq \left(4\rho + \frac{248}{53}\hat{L}\right)\mathbb{E}[P(x_0^k) - P(x_*)] + \frac{8\hat{L}}{53}\mathbb{E}[\mathrm{Gap}_k(x_0^k, y_0^k)] + 8\eta_k\hat{L}C_2 \\
&\leq c\mathbb{E}\left[P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)\right] + 8\eta_k\hat{L}C_2.
\end{aligned} \tag{41}$$

Using the fact that $\hat{L} \geq \mu$,

$$(c + 2\mu)\frac{8\hat{L}}{53c} = \left(4\rho + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\rho + \frac{248}{53}\hat{L})} \leq \frac{8\hat{L}}{53} + \frac{16\mu\hat{L}}{248\hat{L}} \leq \frac{93}{265}\hat{L}. \tag{42}$$

Then, we have

$$\begin{aligned}&(c+2\mu)\mathbb{E}\left[P(x_0^{k+1})-P(x_*)+\frac{8\hat{L}}{53c}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1})\right]\\&\leq c\mathbb{E}\left[P(x_0^k)-P(x_*)+\frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k,y_0^k)\right]+8\eta_k\hat{L}C_2.\end{aligned}\tag{43}$$

Defining $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$, then

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_k] + \frac{8\eta_k\hat{L}C_2}{c+2\mu}.\tag{44}$$

Using this inequality recursively, it yields

$$E[\Delta_{K+1}] \leq \left(\frac{c}{c+2\mu}\right)^K E[\Delta_1] + \frac{8\hat{L}C_2}{c+2\mu}\sum_{k=1}^{K}\left(\eta_k\left(\frac{c}{c+2\mu}\right)^{K+1-k}\right).\tag{45}$$

By definition,

$$\begin{aligned}\Delta_1 &= P(x_0^1) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_1(x_0^1, y_0^1)\\&= P(\bar{x}_0) - P(x_*) + \left(f(\bar{x}_0, \hat{y}_1(\bar{x}_0)) + \frac{\gamma}{2}\|\bar{x}_0 - \bar{x}_0\|^2 - f(\hat{x}_1(\bar{y}_0), \bar{y}_0) - \frac{\gamma}{2}\|\hat{x}_1(\bar{y}_0) - \bar{x}_0\|^2\right)\\&\leq \epsilon_0 + f(\bar{x}_0, \hat{y}_1(\bar{x}_0)) - f(\hat{x}(\bar{y}_0), \bar{y}_0) \leq 2\epsilon_0.\end{aligned}$$

Using inequality $1 - x \leq \exp(-x)$, we have

$$\begin{aligned}\mathbb{E}[\Delta_{K+1}] &\leq \exp\left(\frac{-2\mu K}{c+2\mu}\right)\mathbb{E}[\Delta_1] + \frac{8\eta_0\hat{L}C_2}{c+2\mu}\sum_{k=1}^{K}\exp\left(-\frac{2\mu K}{c+2\mu}\right)\\&\leq 2\epsilon_0\exp\left(\frac{-2\mu K}{c+2\mu}\right) + \frac{8\eta_0\hat{L}C_2}{c+2\mu}K\exp\left(-\frac{2\mu K}{c+2\mu}\right).\end{aligned}$$

To make this less than ϵ, it suffices to make

$$\begin{aligned}&2\epsilon_0\exp\left(\frac{-2\mu K}{c+2\mu}\right) \leq \frac{\epsilon}{2},\\&\frac{8\eta_0\hat{L}C_2}{c+2\mu}K\exp\left(-\frac{2\mu K}{c+2\mu}\right) \leq \frac{\epsilon}{2}.\end{aligned}$$

Let K be the smallest value such that $\exp\left(\frac{-2\mu K}{c+2\mu}\right) \leq \min\{\frac{\epsilon}{4\epsilon_0}, \frac{(c+2\mu)\epsilon}{16\eta_0\hat{L}KC_2}\}$. We can set $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{16\eta_0\hat{L}KC_2}{(c+2\mu)\epsilon}\right\}\right\rceil$. Then, the total stochastic first-order oracle

call complexity is

$$\begin{aligned}\sum_{k=1}^{K} T_k &\leq O\left(\frac{212C_1}{\eta_0 \min\{\rho, \mu_y\}} \sum_{k=1}^{K} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right) \\ &\leq O\left(\frac{212C_1}{\eta_0 \min\{\rho, \mu_y\}} \frac{\exp(K\frac{2\mu}{c+2\mu}) - 1}{\exp(\frac{2\mu}{c+2\mu}) - 1}\right) \\ &\overset{(a)}{\leq} \widetilde{O}\left(\frac{cC_1}{\eta_0 \mu \min\{\rho, \mu_y\}} \max\left\{\frac{\epsilon_0}{\epsilon}, \frac{\eta_0 \hat{L} K C_2}{(c+2\mu)\epsilon}\right\}\right) \\ &\leq \widetilde{O}\left(\max\left\{\frac{(L+\rho)C_1\epsilon_0}{\eta_0 \mu \min\{\rho, \mu_y\}\epsilon}, \frac{(L+\rho)^2 C_2}{\mu^2 \min\{\rho, \mu_y\}\epsilon}\right\}\right),\end{aligned}$$

where (a) uses the setting of K and $\exp(x) - 1 \geq x$, and $\widetilde{O}$ suppresses logarithmic factors. ■

Proof [Proof of Theorem 10] With the above theorem, Theorem 10 directly follows. Noting Lemma 24, we can plug in $\eta_0 = \frac{1}{2\sqrt{2}\ell}$, $C_1 = 1$ and $C_2 = 13\sigma^2$ to Theorem 27. ■

Proof [Proof of Theorem 18] We need the following lemma to bound the convergence of the subproblem at each stage,

Lemma 28 (Lemma 4 of (Yan et al., 2020)) *Suppose Assumption 1 holds,* $\mathbb{E}\|\nabla_x f(x_t, y_t; \xi_t)\|^2 \leq B^2$ *and* $\mathbb{E}\|\nabla_y f(x_t, y_t; \xi_t)\|^2 \leq B^2$. *Set* $\gamma = 2\rho$. *By running Algorithm 1 with Option II: SGDA, it holds for* $k \geq 1$,

$$E[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq 5\eta_k B^2 + \frac{1}{T_k}\left\{\left(\frac{1}{\eta_k} + \frac{\rho}{2}\right) E[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{1}{\eta_k} E[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\right\}.$$

Using this lemma , we can set $\gamma = 2\rho$ and $\eta_0 = \frac{1}{\rho}$. Then it follows that

$$E[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq 5\eta_k B^2 + \frac{2}{\eta_k T_k}\left(E[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + E[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2\right).$$

We plug in $\eta_0 \leq \frac{1}{\rho}$, $C_1 = 2$ and $C_2 = 5B^2$ to Theorem 27 and the conclusion follows. ■

Appendix E. Proof of Theorem 11 and Theorem 20

We first present a lemma by plugging in Lemma 8 of (Yan et al., 2020). And then we a theorem that can unify the proof of Theorem 11 and Theorem 20. In the last, we prove Theorem 11 and Theorem 20.

Lemma 29 (Lemma 8 of (Yan et al., 2020)) *Suppose* $f(x, y)$ *is* $\frac{\mu}{8}$*-weakly convex in* x *for any* y *and set* $\gamma = \frac{\mu}{4}$. *Thus,* $f_k(x, y)$ *is* $\frac{\mu}{8}$*-strongly convex in* x. *Then* $\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)$ *can*

be lower bounded by the following inequalities

$$\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) \geq \left(3 - \frac{2}{\alpha}\right) \mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \frac{\mu\alpha}{8(1-\alpha)} \|x_0^{k+1} - x_0^k\|^2, (0 < \alpha \leq 1), \quad (46)$$

and

$$\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) \geq P(x_0^{k+1}) - P(x_0^k) + \frac{\mu}{8}\|\bar{x}_k - x_0^k\|^2, \textit{where } P(x) = \max_{y' \in \mathcal{Y}} f(x, y'). \quad (47)$$

Theorem 30 *Suppose $0 < \rho \leq \frac{\mu}{8}$ and suppose Assumption 1, 2, 3, 4 hold. Assume we have a subroutine in the k-th stage of Algorithm 1 that can return $\bar{x}_k, \bar{y}_k$ such that*

$$\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{C_1}{\eta_k T_k} \mathbb{E}[\|x - x_0^k\|^2 + \|y - y_0^k\|^2] + \eta_k C_2, \quad (48)$$

where C_1 and C_2 are constants corresponding to the specific subroutine. Take $\gamma = \frac{\mu}{4}$. Define $\Delta_k = 475(P(x_0^k) - P(x_*)) + 57\mathrm{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \mathrm{Gap}(\bar{x}_0, \bar{y}_0)$. Then we can set $\eta_k = \eta_0 \exp(-\frac{k-1}{16}) \leq \frac{1}{2\sqrt{2}\ell}$, $T_k = \left\lceil \frac{384C_1}{\eta_0 \min\{\mu/8, \mu_y\}} \exp\left(\frac{k-1}{16}\right) \right\rceil$. After $K = \left\lceil \max\left\{16 \log \frac{1200\epsilon_0}{\epsilon}, 16 \log \frac{1200\eta_0 K C_2}{\epsilon}\right\} \right\rceil$ stages, we can have $\Delta_{K+1} \leq \epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\max\left\{\frac{C_1\epsilon_0}{\eta_0 \min\{\mu, \mu_y\}\epsilon}, \frac{C_2}{\min\{\mu, \mu_y\}\epsilon}\right\}\right)$.

Proof [Proof of Theorem 30] We have the following relation between $P(x_0^k) - P(x_*)$ and $\mathrm{Gap}_k(x_0^k, y_0^k)$,

$$\begin{aligned}
&P(x_0^k) - P(x_*) = f(x_0^k, \hat{y}(x_0^k)) - f(x_*, y_*) \leq f(x_0^k, \hat{y}(x_0^k)) - f(x_*, y_0^k) \\
&= f(x_0^k, \hat{y}(x_0^k)) + \frac{\gamma}{2}\|x_0^k - x_0^k\|^2 - f(x_*, y_0^k) - \frac{\gamma}{2}\|x_* - x_0^k\|^2 + \frac{\gamma}{2}\|x_* - x_0^k\|^2 \\
&= f_k(x_0^k, \hat{y}(x_0^k)) - f_k(x_*, y_0^k) + \frac{\gamma}{2}\|x_* - x_0^k\|^2 \\
&\leq \hat{f}_k(x_0^k, \hat{y}_k(x_0^k)) - f_k(\hat{x}_k(y_0^k), y_0^k) + \frac{\gamma}{2}\|x_* - x_0^k\|^2 \\
&= \mathrm{Gap}_k(x_0^k, y_0^k) + \frac{\gamma}{2}\|x_* - x_0^k\|^2 \\
&\leq \mathrm{Gap}_k(x_0^k, y_0^k) + \frac{\gamma}{4\mu}(P(x_0^k) - P(x_*)),
\end{aligned} \quad (49)$$

where the first inequality holds by the Lemma 16, and the last inequality due to the μ-PL condition of $P(x)$. Since we take $\gamma = \frac{\mu}{4}$, we know that $1 - \frac{\gamma}{4\mu} = \frac{15}{16}$. Then it follows that

$$P(x_0^k) - P(x_*) \leq \frac{16}{15}\mathrm{Gap}_k(x_0^k, y_0^k). \quad (50)$$

Since $\rho < \frac{\mu}{8}$ and $\gamma = \frac{\mu}{4}$, we know that $f_k(x, y)$ is $\lambda_x = \frac{\mu}{8}$-strongly convex in x. By the setting $\eta_k = \eta_0 \exp\left(-\frac{k-1}{16}\right), T_k = \left\lceil \frac{384C_1}{\eta_0 \min\{\lambda_x, \mu_y\}} \exp\left(\frac{k-1}{16}\right)\} \right\rceil$, we note that $\frac{C_1}{\eta_k T_k} \leq \frac{\min\{\lambda_x, \mu_y\}}{384}$. Applying 48, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] &\leq \eta_k C_2 + \frac{1}{96}\left(\frac{\lambda_x}{4}\mathbb{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{\mu_y}{4}\mathbb{E}[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\right) \\
&\leq \eta_k C_2 + \frac{1}{96}E[\mathrm{Gap}_k(x_0^k, y_0^k)] + \frac{1}{96}E[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)],
\end{aligned} \quad (51)$$

where the last inequality follows from Lemma 25. Rearranging the terms, we have

$$\frac{95}{96}\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k)] \le \eta_k C_2 + \frac{1}{96}\mathbb{E}[\mathrm{Gap}_k(x_0^k,y_0^k)]. \tag{52}$$

Since $\rho \le \frac{\mu}{8}$, $f(x,y)$ is also $\frac{\mu}{8}$-weakly convex in x. Then we use Lemma 29 to lower bound the LHS of (52) with $\alpha = \frac{5}{6}$,

$$\begin{aligned}
&\frac{95}{96}\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k) = \frac{95}{576}\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k) + \frac{475}{576}\mathrm{Gap}_k(\bar{x}_k,\bar{y}_k)\\
&\overset{(a)}{\ge} \frac{95}{576}\left(\frac{3}{5}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1}) - \frac{5}{8}\mu\|x_0^{k+1}-x_0^k\|^2\right)\\
&\quad + \frac{475}{576}(P(x_0^{k+1})-P(x_*)) + \frac{475}{576}(P(x_*)-P(x_0^k)) + \frac{475}{576}\frac{\mu}{8}\|x_0^k - x_0^{k+1}\|^2\\
&= \frac{57}{576}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1}) + \frac{475}{576}(P(x_0^{k+1})-P(x_*))\\
&\quad - \frac{475}{576}\cdot\frac{15}{16}(P(x_0^k)-P(x_*)) - \frac{475}{576}\left(1-\frac{15}{16}\right)(P(x_0^k)-P(x_*))\\
&\overset{(b)}{\ge} \frac{57}{576}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1}) + \frac{475}{576}(P(x_0^{k+1})-P(x_*))\\
&\quad - \frac{475}{576}\cdot\frac{15}{16}(P(x_0^k)-P(x_*)) - \frac{475}{576}\cdot\frac{1}{15}\mathrm{Gap}_k(x_0^k,y_0^k),
\end{aligned} \tag{53}$$

where (a) uses Lemma 29 and (b) uses (50). Combining (52) and (53), we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{475}{576}(P(x_0^{k+1})-P(x_*)) + \frac{57}{576}\mathrm{Gap}_{k+1}(x_0^{k+1},y_0^{k+1})\right]\\
&\le \mathbb{E}\left[\eta_k C_2 + \frac{475}{576}\cdot\frac{15}{16}(P(x_0^k)-P(x_*)) + \frac{475}{576}\cdot\frac{1}{15}\mathrm{Gap}_k(x_0^k,y_0^k) + \frac{1}{96}\mathrm{Gap}_k(x_0^k,y_0^k)\right]\\
&\le \eta_k C_2 + \frac{15}{16}\mathbb{E}\left[\frac{475}{576}(P(x_0^k)-P(x_*)) + \frac{57}{576}\mathrm{Gap}_k(x_0^k,y_0^k)\right].
\end{aligned} \tag{54}$$

Defining $\Delta_k = 475(P(x_0^k)-P(x_*)) + 57\mathrm{Gap}_k(x_0^k,y_0^k)$, we have

$$\mathbb{E}[\Delta_{k+1}] \le 600\eta_k C_2 + \frac{15}{16}\mathbb{E}[\Delta_k] \le \exp(-1/16)\,\mathbb{E}[\Delta_k] + 600\eta_k C_2, \tag{55}$$

and

$$\begin{aligned}
&\Delta_1 = 475(P(x_0^1)-P(x_*)) + 57\mathrm{Gap}_1(x_0^1,y_0^1)\\
&= 475(P(\bar{x}_0)-P(x_*)) + 57\left(f(\bar{x}_0,\hat{y}_1(\bar{x}_0)) + \frac{\gamma}{2}\|\bar{x}_0-\bar{x}_0\|^2 - f(\hat{x}_1(\bar{y}_0),\bar{y}_0) - \|\hat{x}_1(\bar{y}_0)-\bar{x}_0\|^2\right)\\
&\le 475\epsilon_0 + 57\left(f(\bar{x}_0,\hat{y}_1(\bar{x}_0)) - f(\hat{x}(\bar{y}_0),\bar{y}_0)\right) \le 600\epsilon_0.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[\Delta_{K+1}] &\le \exp(-K/16)\,\Delta_1 + 600C_2\sum_{k=1}^{K}\eta_k\exp(-(K+1-k)/(16))\\
&= \exp(-K/16)\,\Delta_1 + 600C_2\sum_{k=1}^{K}(\eta_0\exp(-K/16))\\
&\le 600\epsilon_0\exp(-K/16) + 600\eta_0 C_2 K\exp(-K/16).
\end{aligned} \tag{56}$$

To make this less than ϵ, we just need to make

$$600\epsilon_0 \exp\left(-K/16\right) \leq \frac{\epsilon}{2},$$
$$600\eta_0 C_2 K \exp\left(-K/16\right) \leq \frac{\epsilon}{2}.$$

Let K be the smallest value such that $\exp\left(\frac{-K}{16}\right) \leq \min\{\frac{\epsilon}{1200\epsilon_0}, \frac{\epsilon}{1200\eta_0 C_2 K}\}$. We can set $K = \left\lceil \max\left\{16\log\left(\frac{1200\epsilon_0}{\epsilon}\right), 16\log\left(\frac{1200\eta_0 C_2 K}{\epsilon}\right)\right\}\right\rceil$. Then the total stochastic first-order oracle call complexity is

$$\begin{aligned}
\sum_{k=1}^{K} T_k &\leq O\left(\frac{384C_1}{\eta_0 \min\{\lambda_x, \mu_y\}} \sum_{k=1}^{K} \exp\left(\frac{k-1}{16}\right)\right) \\
&\leq O\left(\frac{384C_1}{\eta_0 \min\{\lambda_x, \mu_y\}} \frac{\exp\left(\frac{K}{16}\right) - 1}{\exp\left(\frac{1}{16}\right) - 1}\right) \\
&\leq \widetilde{O}\left(\max\left\{\frac{C_1\epsilon_0}{\eta_0 \min\{\mu, \mu_y\}\epsilon}, \frac{KC_2}{\min\{\mu, \mu_y\}\epsilon}\right\}\right) \\
&\leq \widetilde{O}\left(\max\left\{\frac{C_1\epsilon_0}{\eta_0 \min\{\mu, \mu_y\}\epsilon}, \frac{C_2}{\min\{\mu, \mu_y\}\epsilon}\right\}\right).
\end{aligned} \tag{57}$$

■

Proof [Proof of Theorem 11] Plugging in Theorem 30 with $\eta_0 = \frac{1}{2\sqrt{2}\ell}$, $C_1 = 1$ and $C_2 = 5B^2$, we get the conclusion. ■

Proof [Proof of Theorem 20] We can plug in $\eta_0 = \frac{1}{\rho}$, $C_1 = 2$ and $C_2 = 5B^2$ to Theorem 27. And the conclusion follows. ■

Appendix F. Analysis of PES-AdaGrad

In this section, we analyze AdaGrad in solving the strongly convex-strongly concave problem. Define $\|u\|_H = \sqrt{u^T H u}$, $\psi_0(z) = 0$, ψ_T^* to be the conjugate of $\frac{1}{\eta}\psi_T$, i.e., $\psi_t^*(z) = \sup_{z' \in \mathcal{Z}} \{\langle z, z'\rangle - \frac{1}{\eta}\psi_t(z')\}$. We first present a supporting lemma,

Lemma 31 *For a sequence $\zeta_1, \zeta_2, ...$, define a sequence $\{u_t \in \mathcal{Z}\}_{t=0}^{T+1}$ as*

$$u_{t+1} = \arg\min_{u \in \mathcal{X}} \frac{\eta}{t} \sum_{\tau=1}^{t} \langle \zeta_\tau, u\rangle + \frac{1}{t}\psi_t(u), u_0 = z_0, \tag{58}$$

where $\psi_t(\cdot)$ is defined in Algorithm 2 with Option III: AdaGrad. Then for any $u \in \mathcal{Z}$,

$$\sum_{t=1}^{T} \langle \zeta_t, u_t - u\rangle \leq \frac{1}{\eta}\psi_T(u) + \frac{\eta}{2}\sum_{t=1}^{T} \|\zeta_t\|_{\psi_{t-1}^*}^2. \tag{59}$$

Proof [Proof of Lemma 31]

$$\begin{aligned}\sum_{t=1}^{T}\langle\zeta_t, u_t - u\rangle &= \sum_{t=1}^{T}\langle\zeta_t, u_t\rangle - \sum_{t=1}^{T}\langle\zeta_t, u\rangle - \frac{1}{\eta}\psi_T(u) + \frac{1}{\eta}\psi_T(u)\\ &\le \frac{1}{\eta}\psi_T(x) + \sum_{t=1}^{T}\langle\zeta_t, u_t\rangle + \sup_{u\in\mathcal{Z}}\left\{\left\langle -\sum_{t=1}^{T}\zeta_t, u\right\rangle - \frac{1}{\eta}\psi_T(u)\right\} \qquad (60)\\ &= \frac{1}{\eta}\psi_T(u) + \sum_{t=1}^{T}\langle\zeta_t, u_t\rangle + \psi_T^*\left(-\sum_{t=1}^{T}\zeta_t\right).\end{aligned}$$

Note that

$$\begin{aligned}\psi_T^*\left(-\sum_{t=1}^{T}\zeta_t\right) &\overset{(a)}{=} \left\langle -\sum_{t=1}^{T}\zeta_t, u_{T+1}\right\rangle - \frac{1}{\eta}\psi_T(u_{T+1}) \overset{(b)}{\le} \left\langle -\sum_{t=1}^{T}\zeta_t, u_{T+1}\right\rangle - \frac{1}{\eta}\psi_{T-1}(u_{T+1})\\ &\le \sup_{u\in\mathcal{Z}}\left\{\left\langle -\sum_{t=1}^{T}\zeta_t, u\right\rangle - \frac{1}{\eta}\psi_{T-1}(u)\right\} = \psi_{T-1}^*\left(-\sum_{t=1}^{T}\zeta_t\right) \qquad (61)\\ &\overset{(c)}{\le} \psi_{T-1}^*\left(-\sum_{t=1}^{T-1}\zeta_t\right) + \left\langle -\zeta_T, \nabla\psi_{T-1}^*\left(-\sum_{t=1}^{T-1}\zeta_t\right)\right\rangle + \frac{\eta}{2}\|\zeta_T\|_{\psi_{T-1}^*}^2,\end{aligned}$$

where (a) holds due to the updating rule, (b) holds since $\psi_{t+1}(u) \ge \psi_t(u)$, (c) uses the fact that $\psi_t(u)$ is 1-strongly convex w.r.t. $\|\cdot\|_{\psi_t} = \|\cdot\|_{H_t}$ and hence $\psi_t^*(\cdot)$ is η-smooth w.r.t. $\|\cdot\|_{\psi_t^*} = \|\cdot\|_{(H_t)^{-1}}$.

Noting $\nabla\psi_{T-1}^*\left(-\sum_{t=1}^{T-1}\zeta_t\right) = u_T$ and adding $\sum_{t=1}^{T}\langle\zeta_t, u_t\rangle$ to both sides of (61),

$$\sum_{t=1}^{T}\langle\zeta_t, u_t\rangle + \psi_T^*\left(-\sum_{t=1}^{T}\zeta_t\right) \le \sum_{t=1}^{T-1}\langle\zeta_t, u_t\rangle + \psi_{T-1}^*\left(-\sum_{t=1}^{T-1}\zeta_t\right) + \frac{\eta}{2}\|\zeta_T\|_{\psi_{T-1}^*}^2. \qquad (62)$$

Using (62) recursively and noting that $\psi_0(u) = 0$, we have

$$\sum_{t=1}^{T}\langle\zeta_t, u_t\rangle + \psi_{x,T}^*\left(-\sum_{t=1}^{T}\zeta_t\right) \le \frac{\eta}{2}\sum_{t=1}^{T}\|\zeta_t\|_{\psi_{t-1}^*}^2. \qquad (63)$$

Combining (60) and (63), we have

$$\sum_{t=1}^{T}\langle\zeta_t, u_t - u\rangle \le \frac{1}{\eta}\psi_T(u) + \frac{\eta}{2}\sum_{t=1}^{T}\|\zeta_t\|_{\psi_{t-1}^*}^2.$$

■

Lemma 32 *Suppose $f(x, y)$ is convex-concave. And also assume $\|\mathcal{G}_t\|_\infty \le \delta$.*
Set $T = M\left\lceil \max\{\frac{\delta+\max_i\|g_{1:T,i}\|}{m}, m\sum_{i=1}^{d+d'}\|g_{1:T,i}\|\}\right\rceil$. By running Algorithm 2 with Option III:

AdaGrad, with input (f, x_0, y_0, η, T), we have

$$E[\text{Gap}(\bar{x}, \bar{y})] \leq \frac{m}{\eta M}(\|z - z_0\|^2) + \frac{4\eta}{mM}. \tag{64}$$

Proof Applying Lemma 31 with $\zeta_t = \mathcal{G}_t$ and $u_t = z_t$, for any $z \in \mathcal{Z}$,

$$\sum_{t=1}^{T} \langle \mathcal{G}_t, z_t - z \rangle \leq \frac{1}{\eta}\psi_T(z) + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathcal{G}_t\|^2_{\psi^*_{t-1}}. \tag{65}$$

By Lemma 4 of (Duchi et al., 2011), we know that $\sum_{t=1}^{T} \|\mathcal{G}_t\|^2_{\psi^*_{t-1}} \leq 2\sum_{i=1}^{d+d'} \|g_{1:T,i}\|$. Hence, for any $z \in \mathcal{Z}$

$$\begin{aligned}
\sum_{t=1}^{T} \langle \mathcal{G}_t, z_t - z \rangle &\leq \frac{1}{\eta}\psi_T(z) + \eta \sum_{i=1}^{d+d'} \|g_{1:T,i}\|_2 \\
&= \frac{\delta\|z_0 - z\|^2}{2\eta} + \frac{\langle z_0 - z,\ \text{diag}\ (s_T)(z_0 - z)\rangle}{2\eta} + \eta \sum_{i=1}^{d+d'} \|g_{1:T,i}\| \\
&\leq \frac{\delta + \max_i \|g_{1:T,i}\|}{2\eta}\|z_0 - z\|^2 + \eta \sum_{i=1}^{d+d'} \|g_{1:T,i}\|.
\end{aligned} \tag{66}$$

Then, we define the following auxiliary sequence $\{\hat{z}_t \in \mathcal{Z}\}_{t=0}^{T}$,

$$\hat{z}_{t+1} = \arg\min_{z \in \mathcal{Z}} \frac{\eta}{t}\sum_{\tau=1}^{t} \langle F(z_t) - \mathcal{G}(z_t; \xi_t)), z\rangle + \frac{1}{t}\psi_t(z), \hat{z}_0 = z_0. \tag{67}$$

Denote $\Theta_t = F(z_t) - \mathcal{G}(z_t; \xi_t)$. Applying Lemma 31 with $\zeta_t = \Theta_t$ and $u_t = \hat{z}_t$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \Theta_t, \hat{z}_t - z \rangle &\leq \frac{1}{\eta}\psi_T(z) + \frac{\eta}{2}\sum_{t=1}^{T} \|\Theta_t\|^2_{\psi^*_{t-1}} \\
&\leq \frac{\delta + \max_i \|g_{1:T,i}\|}{2\eta}\|z_0 - z\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\Theta_t\|^2_{\psi^*_{t-1}}.
\end{aligned} \tag{68}$$

To deal with the last term in the above inequality, we have in expectation that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \|\Theta_t\|^2_{\psi^*_{t-1}}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\|\Theta_t\|^2_{\psi^*_{t-1}}\right] \\
&= \sum_{t=1}^{T} (\mathbb{E}\left[\|\mathcal{G}(z_t; \xi_t)\|^2_{\psi^*_{t-1}}\right] - \|F(z_t)\|^2_{\psi^*_{t-1}}) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \|\mathcal{G}(z_t; \xi_t)\|^2_{\psi^*_{t-1}}\right] \leq 2\mathbb{E}\left[\sum_{i=1}^{d+d'} \|g_{1:T,i}\|\right],
\end{aligned} \tag{69}$$

where the second equality uses the fact that $\mathbb{E}[\mathcal{G}(z_t;\xi_t)] = F(z_t)$ and the last inequality uses Lemma 4 of (Duchi et al., 2011).

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle F(z_t), z_t - z\rangle\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, z_t - z\rangle\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \mathcal{G}(z_t;\xi_t), z_t - z\rangle\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, z_t - \hat{z}_t\rangle\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, \hat{z}_t - z\rangle\right] \\
&\overset{(a)}{\leq} E\left[\frac{\delta + \max_i \|g_{1:T,i}\|}{\eta T}\|z_0 - z\|^2\right] + 2\frac{\eta}{T}\mathbb{E}\left[\sum_{i=1}^{d}\|g_{1:T,i}\|\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle \Theta_t, z_t - \hat{z}_t\rangle\right] \\
&= \mathbb{E}\left[\frac{\delta + \max_i \|g_{1:T,i}\|}{\eta T}\|z_0 - z\|^2\right] + 2\frac{\eta}{T}\mathbb{E}\left[\sum_{i=1}^{d+d'}\|g_{1:T,i}\|\right],
\end{aligned} \tag{70}$$

where the last equality holds because $\mathbb{E}[\langle \Theta_t, z_t - \hat{z}_t\rangle | z_t, \hat{z}_t, \Theta_{t-1}, ..., \Theta_0] = 0$, and (a) uses (66), (68) and (69). Then for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{E}[f(\bar{x}, y) - f(x, \bar{y})] \leq& \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f(x_t, y) - f(x, y_t))\right] \\
=& \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f(x_t, y) - f(x_t, y_t) + f(x_t, y_t) - f(x, y_t))\right] \\
\leq& \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\langle -\nabla_y f(x_t, y_t), y_t - y\rangle + \langle \nabla_x f(x_t, y_t), x_t - x\rangle)\right] \\
=& \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\langle F(z_t), z_t - z\rangle\right] \\
\overset{(a)}{\leq}& \mathbb{E}\left[\frac{\delta + \max_i \|g_{1:T,i}\|}{\eta T}\|z_0 - z\|^2\right] + 2\frac{\eta}{T}\mathbb{E}\left[\sum_{i=1}^{d+d'}\|g_{1:T,i}\|\right] \\
\overset{(b)}{\leq}& \frac{m}{\eta M}\mathbb{E}[\|x_0 - x\|^2 + \|y_0 - y\|^2] + \frac{4\eta}{mM},
\end{aligned} \tag{71}$$

where (a) uses (70), and the last inequality is due to $T = M\left\lceil \max\{\frac{\delta + \max_i \|g_{1:T,i}\|}{m}, m\sum_{i=1}^{d+d'}\|g_{1:T,i}\|\}\right\rceil$. Then we can conclude by plugging in $(x, y) = (\hat{x}(\bar{y}), \hat{y}(\bar{x}))$. ■

Now we formally restate the Theorem 12 as:

Theorem 33 (Formal version of Theorem 12) *Suppose Assumption 1, 3, 4 hold. Let $g^k_{1:T_k}$ denote the cumulative matrix of gradients in k-th stage. Suppose $\|g^k_{1:T_k,i}\|_2 \leq \delta T_k^{\alpha}$ and with $\alpha \in (0, 1/2]$. Then by setting parameters appropriately, $\gamma = 2\rho$, $m = 1/\sqrt{d + d'}$, $\eta_k =$*

$2\eta_0 \exp\left(-\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)$, $M_k = \frac{212m}{\eta_0 \min(\ell,\mu_y)} \exp\left(\frac{k-1}{2}\frac{2\mu}{c+2\mu}\right)$, *and* $T_k = \left\lceil M_k \max\left\{\frac{\delta+\max_i \|g^k_{1:T_k,i}\|_2}{2m}, m\sum_{i=1}^{d+d'} \|g^k_{1:T_k,i}\|_2\right\}\right\rceil$, *and after* $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\left(\frac{4\epsilon_0}{\epsilon}\right), \frac{c+2\mu}{2\mu}\log\left(\frac{16\eta_0^2\hat{L}\min(\rho,\mu_y)K}{53m^2(c+2\mu)\epsilon}\right)\right\}\right\rceil$ *stages, we have PES-AdaGrad has the total stochastic first-order oracle call complexity of* $\widetilde{O}\left(\left(\frac{\delta^2(L+\rho)^2(d+d')}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$ *in order to have* $\mathbb{E}[\Delta_{K+1}] \le \epsilon$, *where* Δ_k *is defined as in Theorem 10.*

Proof By analysis in proof of Theorem 10, we have the following inequalities that do not depend on the optimization algorithm

$$\left(1+\frac{\mu}{2\rho}\right)(P(\bar{x}_k) - P(x_*)) - (P(\bar{x}_{k-1}) - P(x_*)) \le \frac{1}{4\rho}\|\nabla P_k(\bar{x}_k)\|^2, \tag{72}$$

$$\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) \ge \frac{3}{50}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)), \tag{73}$$

and

$$\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\mu_y}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2 \le \mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k). \tag{74}$$

Set $m = 1/\sqrt{d+d'}$, $\eta_k = \eta_0 \exp\left(-\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)$, $M_k = \left\lceil \frac{212m}{\eta_0 \min\{\rho,\mu_y\}} \exp\left(\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)\right\rceil$. Note that,

$$T_k = \left\lceil M_k \max\left\{\frac{\delta + \max_i \|g_{1:T,i}\|}{m}, m\sum_{i=1}^{d+d'} \|g_{1:T,i}\|\right\}\right\rceil \le 2\sqrt{d+d'}\delta M_k T_k^{\alpha}. \tag{75}$$

Thus, $T_k \le (2\sqrt{d+d'}\delta M_k)^{\frac{1}{1-\alpha}}$. Noting $\frac{m}{\eta_k M_k} \le \frac{\min\{\rho,\mu_y\}}{212}$, we can plug in Lemma 32 as

$$\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \le \mathbb{E}\left[\frac{4\eta_k}{mM_k}\right] + \frac{1}{53}\mathbb{E}\left[\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right]. \tag{76}$$

$$\begin{aligned}
&\mathbb{E}[\|\nabla P_k(\bar{x}_k)\|^2] \le 2\hat{L}\mathbb{E}[P_k(\bar{x}_k) - \min_{x\in\mathbb{R}^d} P_k(x)] \le 2\hat{L}\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&= 2\hat{L}\mathbb{E}[4\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&\le 2\hat{L}\mathbb{E}\left[4\left(\frac{4\eta_k}{mM_k} + \frac{1}{53}\left(\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right)\right) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right] \\
&= 2\hat{L}\mathbb{E}\left[16\frac{\eta_k}{mM_k} + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{155}{53}\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right] \\
&\le 2\hat{L}\mathbb{E}\left[16\frac{\eta_k}{mM_k} + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{93}{530}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \frac{124}{53}(P(x_0^{k+1}) - P(x_0^k))\right],
\end{aligned}$$

where the last inequality uses (73). Combining this with (72) and arranging terms, with a constant $c = 4\rho + \frac{248}{53}\hat{L}$, we have

$$\begin{aligned}&(c+2\mu)\mathbb{E}[P(x_0^{k+1}) - P(x_*)] + \frac{93\hat{L}}{265}\mathbb{E}[\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1})] \\ &\leq c\mathbb{E}[P(x_0^k) - P(x_*)] + \frac{8\hat{L}}{53}\mathbb{E}[\mathrm{Gap}_k(x_0^k, y_0^k)] + \frac{32\eta_k\hat{L}}{mM_k}.\end{aligned} \tag{77}$$

Then using the fact that $\hat{L} \geq \mu$, by similar analysis as in proof of Theorem 10, we have

$$\begin{aligned}&(c+2\mu)\mathbb{E}\left[P(x_0^{k+1}) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_{k+1}(x_0^k, y_0^k)\right] \\ &\leq c\mathbb{E}\left[P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)\right] + \frac{32\eta_k\hat{L}}{mM_k}.\end{aligned} \tag{78}$$

Defining $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \mathrm{Gap}(x_0, y_0)$, then

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_k] + \frac{32\eta_k\hat{L}}{(c+2\mu)mM_k}. \tag{79}$$

Noting $\Delta_1 \leq 2\epsilon_0$ and $(1-x) \leq \exp(-x)$,

$$\begin{aligned}\mathbb{E}[\Delta_{k+1}] &\leq \left(\frac{c}{c+2\mu}\right)^K \mathbb{E}[\Delta_1] + \frac{32\hat{L}}{(c+2\mu)m}\sum_{k=1}^K \frac{\eta_k}{M_k}\left(\frac{c}{c+2\mu}\right)^{K+1-k} \\ &\leq 2\epsilon_0 \exp\left(\frac{-2\mu K}{c+2\mu}\right) + \frac{32\hat{L}\eta_0^2\min(\rho,\mu_y)}{212m^2(c+2\mu)}\sum_{k=1}^K \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\exp\left(-\frac{2\mu(K+1-k)}{c+2\mu}\right) \\ &\leq 2\epsilon_0 \exp\left(\frac{-2\mu K}{c+2\mu}\right) + \frac{8\eta_0^2\hat{L}\min(\rho,\mu_y)}{53m^2(c+2\mu)}K\exp\left(-\frac{2\mu K}{c+2\mu}\right).\end{aligned} \tag{80}$$

To make this less than ϵ, we just need to make

$$\begin{aligned}&2\epsilon_0 \exp\left(\frac{-2\mu K}{c+2\mu}\right) \leq \frac{\epsilon}{2}, \\ &\frac{8\eta_0^2\hat{L}\min(\rho,\mu_y)}{53m^2(c+2\mu)}K\exp\left(-\frac{2\mu K}{c+2\mu}\right) \leq \frac{\epsilon}{2}.\end{aligned} \tag{81}$$

Let K be the smallest integer such that $\exp\left(\frac{-2\mu K}{c+2\mu}\right) \leq \min\{\frac{\epsilon}{4\epsilon_0}, \frac{53m^2(c+2\mu)\epsilon}{16\eta_0^2\hat{L}\min(\rho,\mu_y)K}\}$. We can set $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\left(\frac{4\epsilon_0}{\epsilon}\right), \frac{c+2\mu}{2\mu}\log\left(\frac{16\eta_0^2\hat{L}\min(\rho,\mu_y)K}{53m^2(c+2\mu)\epsilon}\right)\right\}\right\rceil$. Recall

$$T_k \leq (2\sqrt{d+d'}\delta M_k)^{\frac{1}{1-\alpha}} \leq \left[\frac{424\delta}{\eta_0\min\{\rho,\mu_y\}}\exp\left(\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)\right]^{\frac{1}{1-\alpha}}. \tag{82}$$

Then the total number of stochastic first-order oracle calls is

$$\begin{aligned}
&\sum_{k=1}^{K} T_k \le O\left(\sum_{k=1}^{K}\left[\frac{\delta}{\eta_0\min\{\rho,\mu_y\}}\exp\left(\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)\right]^{\frac{1}{1-\alpha}}\right)\\
&\le O\left(\sum_{k=1}^{K}\left(\frac{\delta}{\eta_0\min\{\rho,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\exp\left(\frac{k-1}{2(1-\alpha)}\frac{2\mu}{c+2\mu}\right)\right)\\
&\le O\left(\left(\frac{\delta}{\eta_0\min\{\rho,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\frac{\exp\left(K\frac{2\mu}{2(1-\alpha)(c+2\mu)}-1\right)}{\exp\left(\frac{2\mu}{2(1-\alpha)(c+2\mu)}\right)-1}\right)\\
&\overset{(a)}{\le} O\left(\left(\frac{\delta}{\eta_0\min\{\rho,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\left(\frac{c+2\mu}{2\mu}\right)^{\frac{1}{2(1-\alpha)}}\left(\max\left\{\frac{4\epsilon_0}{\epsilon},\frac{16\eta_0^2\hat{L}\min(\rho,\mu_y)K}{53\epsilon m^2(c+\mu)}\right\}\right)^{\frac{1}{2(1-\alpha)}}\right)\\
&\le \widetilde{O}\left(\left(\max\left\{\frac{\delta^2 c}{\eta_0^2\mu(\min\{\rho,\mu_y\})^2},\frac{\delta^2\hat{L}c(d+d')}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right\}\right)^{\frac{1}{2(1-\alpha)}}\right)\\
&\le \widetilde{O}\left(\left(\frac{\delta^2(L+\rho)^2(d+d')}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right),
\end{aligned}$$

where (a) uses the inequality that $\exp(ax)-1\ge x^a$ for any $0<a<1$ and $x>0$, noting that $0<\frac{2\mu}{c+2\mu}<1$ and $\frac{1}{2(1-\alpha)}>0$. ■

Appendix G. More Analysis on PES-AdaGrad

We have already shown in Theorem 12 about the convergence of primal gap for our Algorithm with Option III: Adagrad update. In this section, we show a corollary about the convergence of duality gap based on Theorem 12. What is more, in parallel with our analysis on Option II: OGDA update, we show some convergence results under the condition that $\rho\le\frac{\mu}{8}$.

Corollary 34 *Under same setting as in Theorem 12 and suppose Assumption 6 holds as well. To reach an ϵ-duality gap, the total stochastic first-order oracle call complexity is* $\widetilde{O}\left(\left(\left(\frac{\rho}{\mu_x}+1\right)\frac{\delta^2(L+\rho)^2(d+d')}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$.

Theorem 35 *Suppose Assumption 1, 3, 6, hold and $\rho\le\frac{\mu}{8}$. Define a constant $c=4\rho+\frac{248}{53}\hat{L}\in O(L+\rho)$. $g^k_{1:T_k}$ denotes the cumulative matrix of gradients $g_{1:T}$ in k-th stage. Suppose $\|g^k_{1:T_k,i}\|_2\le\delta T_k^{\alpha}$ and with $\alpha\in(0,1/2]$. Then by setting $\gamma=2\rho$, $m=1/\sqrt{d+d'}$, $\eta_k=2\eta_0\exp\left(-\frac{(k-1)}{2}\frac{2\mu}{c+2\mu}\right)$, $M_k=\frac{212m}{\eta_0\min(\ell,\mu_y)}\exp\left(\frac{k-1}{2}\frac{2\mu}{c+2\mu}\right)$, and $T_k=\left\lceil M_k\max\left\{\frac{\delta+\max_i\|g^k_{1:\tau,i}\|_2}{2m},m\sum_{i=1}^{d+d'}\|g^k_{1:\tau,i}\|_2\right\}\right\rceil$, and after $K=\left\lceil\max\left\{\frac{c+2\mu}{2\mu}\log\left(\frac{4\epsilon_0}{\epsilon}\right),\frac{c+2\mu}{2\mu}\log\left(\frac{16\eta_0^2\hat{L}\min(\rho,\mu_y)K}{53m^2(c+2\mu)\epsilon}\right)\right\}\right\rceil$ stages, we have $\tilde{O}\left(\left(\frac{\delta^2(d+d')}{\min\{\mu,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$.*

Corollary 36 *Under same setting as in Theorem 35 and suppose Assumption 6 holds as well. To reach an ϵ-duality gap, the total stochastic first-order oracle call complexity is* $\tilde{O}\left(\left(\left(\frac{\mu}{\mu_x}+1\right)\frac{\delta^2(d+d')}{\min\{\mu,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right)$.

Proof [Proof of Theorem 35] By analysis in the Proof of Theorem 11, we know that when $\rho < \frac{\mu}{8}$ and $\gamma = \frac{\mu}{4}$,

$$P(x_0^k) - P(x_*) \leq \frac{16}{15}\text{Gap}_k(x_0^k, y_0^k) \tag{83}$$

and $f_k(x,y)$ is $\lambda_x = \frac{\mu}{8}$-strongly convex in x.

Set $m = 1/\sqrt{d+d'}$, $\eta_k = \eta_0 \exp\left(-\frac{(k-1)}{32}\right)$, $M_k = \frac{384m}{\eta_0 \min\{\lambda_x,\mu_y\}}\exp\left(\frac{(k-1)}{32}\right)$. Note that,

$$T_k = \left\lceil M_k \max\left\{\frac{\delta + \max_i \|g_{1:T,i}\|}{m}, m\sum_{i=1}^{d+d'}\|g_{1:T,i}\|\right\}\right\rceil \leq 2\sqrt{d+d'}\delta M_k T_k^{\alpha}. \tag{84}$$

Thus, $T_k \leq (2\sqrt{d+d'}\delta M_k)^{\frac{1}{1-\alpha}}$. Since $\frac{m}{\eta_k M_k} \leq \frac{\min\{\lambda_x,\mu_y\}}{384}$, we can apply Lemma 32 to get

$$\begin{aligned}\mathbb{E}[\text{Gap}_k(\bar{x}_k,\bar{y}_k)] &\leq \frac{4\eta_k}{mM_k} + \frac{1}{96}\left(\frac{\lambda_x}{4}\mathbb{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{\mu_y}{4}\mathbb{E}[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\right)\\ &\leq \frac{4\eta_k}{mM_k} + \frac{1}{96}\mathbb{E}[\text{Gap}_k(x_0^k, y_0^k)] + \frac{1}{96}\mathbb{E}[\text{Gap}_k(\bar{x}_k,\bar{y}_k)],\end{aligned} \tag{85}$$

where the last inequality follows from Lemma 25. Rearranging terms, we have

$$\frac{95}{96}\mathbb{E}[\text{Gap}_k(\bar{x}_k,\bar{y}_k)] \leq \frac{4\eta_k}{mM_k} + \frac{1}{96}\mathbb{E}[\text{Gap}_k(x_0^k, y_0^k)]. \tag{86}$$

Since $\rho \leq \frac{\mu}{8}$, $f(x,y)$ is also $\frac{\mu}{8}$-weakly convex in x. Then, similar to the analysis in proof of Theorem 11, we use Lemma 29 to lower bound the LHS of (86) with $\alpha = \frac{5}{6}$,

$$\begin{aligned}\frac{95}{96}\text{Gap}_k(\bar{x}_k,\bar{y}_k) \geq& \frac{57}{576}\text{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{475}{576}(P(x_0^{k+1}) - P(x_*))\\ &- \frac{475}{576}\cdot\frac{15}{16}(P(x_0^k) - P(x_*)) - \frac{475}{576}\cdot\frac{1}{15}\text{Gap}_k(x_0^k, y_0^k)\end{aligned} \tag{87}$$

Combining (86) and (87), we get

$$\begin{aligned}&\mathbb{E}\left[\frac{475}{576}(P(x_0^{k+1}) - P(x_*)) + \frac{57}{576}\text{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1})\right]\\ &\leq \frac{4\eta_k}{mM_k} + \frac{475}{576}\cdot\frac{15}{16}\mathbb{E}[P(x_0^k) - P(x_*)] + \frac{475}{576}\cdot\frac{1}{15}\mathbb{E}[\text{Gap}_k(x_0^k, y_0^k)] + \frac{1}{96}\mathbb{E}[\text{Gap}_k(x_0^k, y_0^k)]\\ &\leq \frac{4\eta_k}{mM_k} + \frac{15}{16}\mathbb{E}\left[\frac{475}{576}(P(x_0^k) - P(x_*)) + \frac{57}{576}\text{Gap}_k(x_0^k, y_0^k)\right].\end{aligned}$$

Defining $\Delta_k = 475(P(x_0^k) - P(x_*)) + 57\text{Gap}_k(x_0^k, y0^k)$ and $\epsilon_0 = \text{Gap}(x_0, y_0)$, we have

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{15}{16}\mathbb{E}[\Delta_k] + \frac{4\eta_k}{mM_k} \leq \exp\left(-\frac{1}{16}\right)\mathbb{E}[\Delta_k] + \frac{4\eta_k}{mM_k} \tag{88}$$

and $\Delta_1 \leq 600\epsilon_0$. Thus,

$$\begin{aligned}\mathbb{E}[\Delta_{K+1}] \leq & \exp\left(-\frac{K}{16}\right)\Delta_1 + \frac{4}{m}\sum_{k=1}^{K}\frac{\eta_k}{M_k}\exp\left(-\frac{K+1-k}{16}\right) \\ = & \exp\left(-\frac{K}{16}\right)\Delta_1 + \frac{\eta_0^2\min\{\lambda_x,\mu_y\}}{96m^2}\sum_{k=1}^{K}\exp\left(-\frac{K}{16}\right) \\ \leq & 600\epsilon_0\exp\left(-\frac{K}{16}\right) + \frac{\eta_0^2\min\{\lambda_x,\mu_y\}}{96m^2}K\exp\left(-\frac{K}{16}\right).\end{aligned} \tag{89}$$

To make this less than ϵ, we just need to make

$$\begin{aligned}&600\epsilon_0\exp\left(-\frac{K}{16}\right) \leq \frac{\epsilon}{2}, \\ &\frac{\eta_0^2\min\{\rho,\mu_y\}}{96m^2}K\exp\left(-\frac{K}{16}\right) \leq \frac{\epsilon}{2}.\end{aligned} \tag{90}$$

Let K be the smallest integer such that $\exp\left(\frac{-K}{16}\right) \leq \min\{\frac{\epsilon}{1200\epsilon_0}, \frac{48m^2\epsilon}{\eta_0^2\min\{\rho,\mu_y\}K}\}$. Recall $T_k \leq (2\sqrt{d+d'}\delta M_k)^{\frac{1}{1-\alpha}} = \left(\frac{768\delta}{\eta_0\min\{\lambda_x,\mu_y\}}\exp\left(\frac{k-1}{32}\right)\right)^{\frac{1}{1-\alpha}}$. Then the total stochastic first-order oracle call complexity is

$$\begin{aligned}\sum_{k=1}^{K}T_k \leq & O\left(\sum_{k=1}^{K}\left[\frac{\delta}{\eta_0\min\{\lambda_x,\mu_y\}}\exp\left(\frac{k-1}{32}\right)\right]^{\frac{1}{1-\alpha}}\right) \\ \leq & O\left(\sum_{k=1}^{K}\left(\frac{\delta}{\eta_0\min\{\lambda_x,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\exp\left(\frac{k-1}{32(1-\alpha)}\right)\right) \\ \leq & O\left(\left(\frac{\delta}{\eta_0\min\{\lambda_x,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\frac{\exp\left(\frac{K}{2(1-\alpha)\cdot 16}\right)-1}{\exp\left(\frac{1}{2(1-\alpha)\cdot 16}\right)-1}\right) \\ \leq & \tilde{O}\left(\left(\frac{\delta}{\eta_0\min\{\lambda_x,\mu_y\}}\right)^{\frac{1}{1-\alpha}}\left(\max\left\{\frac{\epsilon_0}{\epsilon},\frac{\eta_0^2\min\{\lambda_x,\mu_y\}K}{m^2\epsilon}\right\}\right)^{\frac{1}{2(1-\alpha)}}\right) \\ \leq & \tilde{O}\left(\left(\max\left\{\frac{\delta^2\epsilon_0}{\eta_0^2(\min\{\mu,\mu_y\})^2\epsilon},\frac{\delta^2(d+d')}{\min\{\mu,\mu_y\}\epsilon}\right\}\right)^{\frac{1}{2(1-\alpha)}}\right) \\ \leq & \tilde{O}\left(\left(\frac{\delta^2(d+d')}{\min\{\mu,\mu_y\}\epsilon}\right)^{\frac{1}{2(1-\alpha)}}\right).\end{aligned}$$

■

Appendix H. Proof of Corollary 14, 19, 34

Proof Let (x_*, y_*) denote a saddle point solution of $\min_{x\in\mathbb{R}^d}\max_{y\in\mathcal{Y}} f(x,y)$.

Note that $x_0^{K+1} = x_K, y_0^{K+1} = \bar{y}_K$. Suppose we have $\mathbb{E}[\text{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] \leq \hat{\epsilon}$ after K stages. Noting $\gamma = 2\rho$, $f_k(x, y)$ is ρ-strongly convex and μ_y-strongly concave. By Lemma 25, we know that

$$\mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \leq \frac{4}{\rho} 2\mathbb{E}[\text{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] \leq \frac{8\hat{\epsilon}}{\rho}. \tag{91}$$

Since $\nabla_x f_{K+1}(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) = \nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) + \gamma(\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}) = 0$, we have

$$\mathbb{E}[\|\nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1})\|^2] = \gamma^2 \mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \leq 32\rho\hat{\epsilon}$$

Using the μ_x-PL condition of $f(\cdot, y_0^{K+1})$ in x,

$$\mathbb{E}\left[f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})\right] \leq \mathbb{E}\left[\frac{\|\nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1})\|^2}{2\mu_x}\right] \leq \frac{16\rho\hat{\epsilon}}{\mu_x}.$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[\text{Gap}(x_0^{K+1}, y_0^{K+1})\right] = \mathbb{E}\left[f(x_0^{K+1}, \hat{y}(x_0^{K+1})) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})\right] \\
&= \mathbb{E}\Big\{f(x_0^{K+1}, \hat{y}(x_0^{K+1})) + \frac{\gamma}{2}\|x_0^{K+1} - x_0^{K+1}\|^2 - f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - \frac{\gamma}{2}\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2 \\
&\quad + f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1}) + \frac{\gamma}{2}\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2\Big\} \\
&= \mathbb{E}[\text{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] + \mathbb{E}[f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})] \\
&\quad + \frac{\gamma}{2}\mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \\
&\leq \hat{\epsilon} + \frac{16\rho\hat{\epsilon}}{\mu_x} + 8\hat{\epsilon} \leq O\left(\frac{\rho\hat{\epsilon}}{\mu_x} + \hat{\epsilon}\right).
\end{aligned}$$

To have $\text{Gap}(x_0^{K+1}, y_0^{K+1}) \leq \epsilon$, we need $\hat{\epsilon} \leq O\left(\left(\frac{\rho}{\mu_x} + 1\right)^{-1}\epsilon\right)$. Plug $\hat{\epsilon}$ into Theorem 10, 12, 18, we can prove Corollary 14, 34, 19, respectively. ■

Appendix I. Proof of Corollary 15, 21, 36

Proof Let (x_*, y_*) denote a saddle point solution of $\min_{x\in\mathbb{R}^d}\max_{y\in\mathcal{Y}} f(x, y)$ and $x_{K+1}^* = \min_{x\in\mathbb{R}^d} P_k(x)$. Note that $x_0^{K+1} = x_K, y_0^{K+1} = \bar{y}_K$. Suppose $\mathbb{E}[\text{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] \leq \hat{\epsilon}$ after K stages.

By the setting $\rho \leq \frac{\mu}{8}$ and $\gamma = \frac{\mu}{4}$, $f_k(x, y)$ is $\frac{\mu}{8}$-strongly convex and μ_y-strongly concave. By Lemma 25, we know that

$$\mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \leq \frac{64}{\mu}\mathbb{E}[\text{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] \leq \frac{64\hat{\epsilon}}{\mu}. \tag{92}$$

Since $\nabla_x f_{K+1}(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) = \nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) + \gamma(\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}) = 0$, we have

$$\mathbb{E}[\|\nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1})\|^2] = \gamma^2 \mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \leq 4\mu\hat{\epsilon}. \tag{93}$$

Using the μ_x-PL condition of $f(\cdot, y_0^{K+1})$ in x,

$$\mathbb{E}\left[f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})\right] \leq \mathbb{E}\left[\frac{\|\nabla_x f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1})\|^2}{2\mu_x}\right] \leq \frac{2\mu\hat{\epsilon}}{\mu_x}.$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Gap}(x_0^{K+1}, y_0^{K+1})\right] = \mathbb{E}\left[f(x_0^{K+1}, \hat{y}(x_0^{K+1})) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})\right] \\
&= \mathbb{E}\Bigg\{f(x_0^{K+1}, \hat{y}(x_0^{K+1})) + \frac{\gamma}{2}\|x_0^{K+1} - x_0^{K+1}\|^2 - f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - \frac{\gamma}{2}\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2 \\
&\quad + f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1}) + \frac{\gamma}{2}\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2\Bigg\} \\
&= \mathbb{E}[\mathrm{Gap}_{K+1}(x_0^{K+1}, y_0^{K+1})] + \mathbb{E}[f(\hat{x}_{K+1}(y_0^{K+1}), y_0^{K+1}) - f(\hat{x}(y_0^{K+1}), y_0^{K+1})] \\
&\quad + \frac{\gamma}{2}\mathbb{E}[\|\hat{x}_{K+1}(y_0^{K+1}) - x_0^{K+1}\|^2] \\
&\leq \hat{\epsilon} + \frac{2\mu\hat{\epsilon}}{\mu_x} + 8\hat{\epsilon} \leq O\left(\frac{\mu\hat{\epsilon}}{\mu_x} + \hat{\epsilon}\right).
\end{aligned}$$

To have $\mathrm{Gap}(x_0^{K+1}, y_0^{K+1}) \leq \epsilon$, we need $\hat{\epsilon} \leq O\left(\left(\frac{\mu}{\mu_x} + 1\right)^{-1}\epsilon\right)$. Plug $\hat{\epsilon}$ into Theorem11, 20, 35 we can prove Corollary 15, 21, 36, respectively. ■

Appendix J. Analysis of Option IV: PES-Storm

In this section, we present the formal version of Theorem 13 in the Theorem 41 and show its proof. Denote $d_t = (v_t, u_t)$, where the component v_t is corresponding to primal variable x and the component u_t is corresponding to dual variable. Also denote $\eta = (\eta^x, \eta^y)$, $a = (a_x, a_y)$.

J.1 Auxiliary Lemmas

In this subsection, we show some lemmas that are needed to prove Theorem 41.

Lemma 37 *In Algorithm 2 with Option IV: Storm. setting* $0 < \eta^x \leq \frac{1}{2L}$, *we have*

$$P(x_{t+1}) - P(x_t) \leq -\frac{\eta^x}{4}\|v_t\|^2 + \eta^x\ell^2\|\hat{y}(x_t) - y_t\|^2 + \eta^x\|\nabla_x f(x_t, y_t) - v_t\|^2. \tag{94}$$

Proof Using the L-smoothness of $P(x) = \max_{y' \in \mathcal{Y}} f(x, y')$,

$$\begin{aligned}
P(x_{t+1}) &\le P(x_t) + \langle \nabla P(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= P(x_t) + \langle \nabla P(x_t) - \nabla_x f(x_t, y_t), x_{t+1} - x_t \rangle + \langle \nabla_x f(x_t, y_t) - v_t, x_{t+1} - x_t \rangle \\
&\quad + \langle v_t, x_{t+1} - x_t \rangle + \frac{L(\eta^x)^2}{2}\|v_t\|^2 \\
&\le P(x_t) + \eta^x \|\nabla P(x_t) - \nabla_x f(x_t, y_t)\|^2 + \frac{1}{4\eta^x}\|x_{t+1} - x_t\|^2 \\
&\quad + \eta^x \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{4\eta^x}\|x_{t+1} - x_t\|^2 + \langle v_t, x_{t+1} - x_t \rangle + \frac{L(\eta^x)^2}{2}\|v_t\|^2 \\
&= P(x_t) + \eta^x \|\nabla P(x_t) - \nabla_x f(x_t, y_t)\|^2 + \frac{\eta^x}{4}\|v_t\|^2 + \eta^x \|\nabla_x f(x_t, y_t) - v_t\|^2 \\
&\quad + \frac{\eta^x}{4}\|v_t\|^2 - \eta^x \|v_t\|^2 + \frac{L(\eta^x)^2}{2}\|v_t\|^2 \\
&\le P(x_t) + \eta^x \ell^2 \|y_t - \hat{y}(x_t)\|^2 + \eta^x \|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\eta^x}{4}\|v_t\|^2,
\end{aligned}$$

where the last inequality uses the setting $\eta^x \le \frac{1}{2L}$. ■

Lemma 38 *In Algorithm 2 with Option IV: Storm, setting $0 < a_x, a_y < 1$, we have*

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\
&\le (1 - a_x)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + 8(1 - a_x)^2 \ell^2 (\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2) + 2a_x^2 \sigma^2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - u_{t+1}\|^2 \\
&\le (1 - a_y)\mathbb{E}\|\nabla_y f(x_t, y_t) - u_t\|^2 + 8(1 - a_y)^2 \ell^2 (\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2) + 2a_y^2 \sigma^2.
\end{aligned}$$

Proof By the update rule of v, we get

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\
&= \mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - a_x)v_t - (1 - a_x)\nabla_x f(x_t, y_t; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1})\|^2 \\
&\le \mathbb{E}\|(1 - a_x)(v_t - \nabla_x f(x_t, y_t)) + (1 - a_x)[\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_t; \xi_{t+1})] \\
&\qquad - [\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})]\|^2] \\
&= \mathbb{E}\|(1 - a_x)(v_t - \nabla_x f(x_t, y_t))\|^2 \\
&\quad + \mathbb{E}\|(1 - a_x)[\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_t; \xi_{t+1})] - [\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})]\|^2 \\
&= \mathbb{E}\|(1 - a_x)(v_t - \nabla_x f(x_t, y_t))\|^2 \\
&\quad + \mathbb{E}\|(1 - a_x)[\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_t; \xi_{t+1})] - (1 - a_x)[\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})] \\
&\qquad - a_x[\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})]\|^2
\end{aligned}$$

Then using $\mathbb{E}[\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_t; \xi_{t+1})] = 0$ and $\mathbb{E}[\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})] = 0$, we continue the above inequality as

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \le (1-a_x)\mathbb{E}\|v_t - \nabla_x f(x_t, y_t)\|^2 \\
&\quad + 2(1-a_x)^2\mathbb{E}\|[\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y_t; \xi_{t+1})] - [\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})] \\
&\quad + 2a_x^2\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})\|^2 \\
&\le (1-a_x)\mathbb{E}\|v_t - \nabla_x f(x_t, y_t)\|^2 + 4(1-a_x)^2\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla_x f(x_{t+1}, y_{t+1})\|^2 \\
&\quad + 4(1-a_x)^2\mathbb{E}\|\nabla_x f(x_t, y_t; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1})\|^2 + 2a_x^2\sigma^2 \\
&\le (1-a_x)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + 8(1-a_x)^2\ell^2(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2) + 2a_x^2\sigma^2.
\end{aligned}$$

By similar analysis on y-side, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - u_{t+1}\|^2 \\
&\le (1-a_y)\mathbb{E}\|\nabla_y f(x_t, y_t) - u_t\|^2 + 8(1-a_y)^2\ell^2(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2) + 2a_y^2\sigma^2.
\end{aligned}$$

■

The next lemma follows from Lemma 18 of (Huang et al., 2022). We include the proof for the sake of completeness.

Lemma 39 *In Algorithm 2 with Option IV, setting $\eta^y \le \min\{1, \frac{1}{6\ell}\}, \lambda = \frac{1}{6\ell}$, we have*

$$\begin{aligned}
\|y_{t+1} - \hat{y}(x_{t+1})\|^2 \le &(1 - \frac{\mu_y\eta^y\lambda}{4})\|y_t - \hat{y}(x_t)\|^2 - \frac{3\eta^y\lambda^2}{4}\|u_t\|^2 + \frac{5\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2 \\
&+ \frac{5\ell^2(\eta^x)^2}{\eta^y\lambda\mu_y^3}\|v_t\|^2.
\end{aligned}$$

Proof Using μ_y-strong concavity of $f(x, y)$ in y,

$$\begin{aligned}
f(x_t, y) &\le f(x_t, y_t) + \langle\nabla_y f(x_t, y_t), y - y_t\rangle - \frac{\mu_y}{2}\|y - y_t\|^2 \\
&= f(x_t, y_t) + \langle u_t, y - \tilde{y}_{t+1}\rangle + \langle\nabla_y f(x_t, y_t) - u_t, y - \tilde{y}_{t+1}\rangle \\
&\quad + \langle\nabla_y f(x_t, y_t), \tilde{y}_{t+1} - y_t\rangle - \frac{\mu_y}{2}\|y - y_t\|^2.
\end{aligned} \tag{95}$$

Using ℓ-smoothness of $f(x, y)$,

$$-f(x_t, \tilde{y}_{t+1}) \le -f(x_t, y_t) - \langle\nabla_y f(x_t, y_t), \tilde{y}_{t+1} - y_t\rangle + \frac{\ell}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{96}$$

Adding the above two inequalities, we get

$$\begin{aligned}
&f(x_t, y) - f(x_t, \tilde{y}_{t+1}) \le \\
&\langle u_t, y - \tilde{y}_{t+1}\rangle + \langle\nabla_y f(x_t, y_t) - u_t, y - \tilde{y}_{t+1}\rangle - \frac{\mu_y}{2}\|y - y_t\|^2 + \frac{\ell}{2}\|\tilde{y}_{t+1} - y_t\|^2.
\end{aligned} \tag{97}$$

Note that the update of y is

$$\begin{aligned}\tilde{y}_{t+1} &= \mathcal{P}_{\mathcal{Y}}(y_t + \lambda u_t),\\ y_{t+1} &= y_t + \eta^y(\tilde{y}_{t+1} - y_t),\end{aligned} \tag{98}$$

where $\lambda = \frac{1}{6\ell}$. Since $\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda u_t) = \arg\min_{y\in\mathcal{Y}} \frac{1}{2}\|y - y_t - \lambda u_t\|^2$ and $\frac{1}{2}\|y - y_t - \lambda u_t\|^2$ is convex in y, we have

$$\langle \tilde{y}_{t+1} - y_t - \lambda u_t, y - \tilde{y}_{t+1}\rangle \geq 0, y \in \mathcal{Y}. \tag{99}$$

Then we get

$$\begin{aligned}&\langle u_t, y - \tilde{y}_{t+1}\rangle \leq \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - \tilde{y}_{t+1}\rangle = \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - \tilde{y}_{t+1}\rangle + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle\\ &= -\frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle.\end{aligned}$$

Thus,

$$\begin{aligned}f(x_t, y) - f(x_t, \tilde{y}_{t+1}) \leq &-\left(\frac{1}{\lambda} - \frac{\ell}{2}\right)\|\tilde{y}_{t+1} - y_t\|^2 + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle\\ &+ \langle \nabla_y f(x_t, y_t) - u_t, y - \tilde{y}_{t+1}\rangle - \frac{\mu_y}{2}\|y - y_t\|^2.\end{aligned} \tag{100}$$

Plugging in $y = \hat{y}(x_t)$,

$$\begin{aligned}0 \leq f(x_t, \hat{y}(x_t)) - f(x_t, \tilde{y}_{t+1}) \leq &-\left(\frac{1}{\lambda} - \frac{\ell}{2}\right)\|\tilde{y}_{t+1} - y_t\|^2 + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, \hat{y}(x_t) - y_t\rangle\\ &+ \langle \nabla_y f(x_t, y_t) - u_t, \hat{y}(x_t) - \tilde{y}_{t+1}\rangle - \frac{\mu_y}{2}\|\hat{y}(x_t) - y_t\|^2.\end{aligned} \tag{101}$$

By $y_{t+1} = y_t + \eta^y(\tilde{y}_{t+1} - y_t)$, we have

$$\begin{aligned}&\|y_{t+1} - \hat{y}(x_t)\|^2 = \|y_t + \eta^y(\tilde{y}_{t+1} - y_t) - \hat{y}(x_t)\|^2\\ &= \|y_t - \hat{y}(x_t)\|^2 + 2\eta^y\langle \tilde{y}_{t+1} - y_t, y_t - \hat{y}(x_t)\rangle + (\eta^y)^2\|\tilde{y}_{t+1} - y_t\|^2\\ &\leq \|y_t - \hat{y}(x_t)\|^2 + (\eta^y)^2\|\tilde{y}_{t+1} - y_t\|^2 - \eta^y(2 - \ell\lambda)\|\tilde{y}_{t+1} - y_t\|^2\\ &\quad + 2\eta^y\lambda\langle \nabla_y f(x_t, y_t) - u_t, \hat{y}(x_t) - \tilde{y}_{t+1}\rangle - \mu_y\eta^y\lambda\|\hat{y}(x_t) - y_t\|^2\\ &\leq \|y_t - \hat{y}(x_t)\|^2 - (2\eta^y - (\eta^y)^2 - \ell\lambda\eta^y)\|\tilde{y}_{t+1} - y_t\|^2\\ &\quad + 2\eta^y\lambda\left[\frac{2}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2 + \frac{\mu_y}{8}\|\hat{y}(x_t) - y_{t+1}\|^2\right] - \mu_y\eta^y\lambda\|\hat{y}(x_t) - y_t\|^2\\ &\leq (1 - \mu_y\eta^y\lambda)\|y_t - \hat{y}(x_t)\|^2 - (2\eta^y - (\eta^y)^2 - \ell\lambda\eta^y)\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta^y\mu_y\lambda}{2}\|\hat{y}(x_t) - y_t\|^2\\ &\quad + \frac{\eta^y\mu_y\lambda}{2}\|y_t - \tilde{y}_{t+1}\|^2 + \frac{4\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2\\ &\leq (1 - \frac{\mu_y\eta^y\lambda}{2})\|y_t - \hat{y}(x_t)\|^2 - (2\eta^y - (\eta^y)^2 - \ell\lambda\eta^y - \frac{\eta^y\mu_y\lambda}{2})\|\tilde{y}_{t+1} - y_t\|^2 + \frac{4\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2\\ &\leq (1 - \frac{\mu_y\eta^y\lambda}{2})\|y_t - \hat{y}(x_t)\|^2 - \frac{3}{4}\eta^y\|\tilde{y}_{t+1} - y_t\|^2 + \frac{4\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2\\ &= (1 - \frac{\mu_y\eta^y\lambda}{2})\|y_t - \hat{y}(x_t)\|^2 - \frac{3\eta^y\lambda^2}{4}\|u_t\|^2 + \frac{4\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2,\end{aligned}$$

where the last inequality holds because $\mu_y \leq \ell, \eta^y \leq \min\{1, \frac{1}{6\ell}\}$. Using the above inequalities, we get

$$\begin{aligned}
&\|y_{t+1} - \hat{y}(x_{t+1})\|^2 = \|y_{t+1} - \hat{y}(x_t) + \hat{y}(x_t) - \hat{y}(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta^y \mu_y \lambda}{4})\|y_{t+1} - \hat{y}(x_t)\|^2 + (1 + \frac{4}{\eta^y \mu_y \lambda})\|\hat{y}(x_t) - \hat{y}(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta^y \mu_y \lambda}{4})\|y_{t+1} - \hat{y}(x_t)\|^2 + (1 + \frac{4}{\eta^y \mu_y \lambda})\frac{\ell^2}{\mu_y^2}\|x_{t+1} - x_t\|^2 \\
&\leq (1 - \frac{\mu_y \eta^y \lambda}{2})(1 + \frac{\eta^y \mu_y \lambda}{4})\|y_t - \hat{y}(x_t)\|^2 - \frac{3\eta^y \lambda^2}{4}\|u_t\|^2 \\
&\quad + (1 + \frac{\eta^y \mu_y \lambda}{4})\frac{4\eta^y \lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2 + (1 + \frac{4}{\eta^y \mu_y \lambda})\frac{\ell^2}{\mu_y^2}(\eta^x)^2\|v_t\|^2 \\
&\leq (1 - \frac{\mu_y \eta^y \lambda}{4})\|y_t - \hat{y}(x_t)\|^2 - \frac{3\eta^y \lambda^2}{4}\|u_t\|^2 + \frac{5\eta^y \lambda}{\mu_y}\|\nabla_y f(x_t, y_t) - u_t\|^2 + \frac{5\ell^2(\eta^x)^2}{\eta^y \lambda \mu_y^3}\|v_t\|^2,
\end{aligned}$$

where the second inequality is because $\hat{y}(\cdot)$ is $\frac{\ell}{\mu_y}$-Lipshitz (Lin et al., 2020a). ■

The following lemma analyze the convergence of one stage in PES-Storm.

Lemma 40 *By setting* $\eta^x = \frac{\mu_2^2}{1000\ell^2}\eta^y, a_x = \frac{800\ell}{\mu_y}(\eta^y)^2, a_y = \frac{800\ell}{\mu_y}(\eta^y)^2$, $\eta^y \leq O(\frac{1}{30}\sqrt{\frac{\mu_y}{\ell}})$ *to ensure* $0 < a_x, a_y < 1$, *one stage of Algorithm 2 with Option IV: Storm returns an solution* (x_τ, y_τ) *such that*

$$\begin{aligned}
&\mathbb{E}\|y_\tau - \hat{y}(x_\tau)\|^2 + \frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_x f(x_\tau, y_\tau) - v_\tau\|^2] + \frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_y f(x_\tau, y_\tau) - v_\tau\|^2] + \frac{1}{8}\mathbb{E}\|v_\tau\|^2] \\
&\leq \frac{\Gamma_1 - \Gamma_{T+1}}{\eta^x T} + \frac{4C\ell(\eta^y)^3\sigma^2}{\mu_y \eta^x},
\end{aligned}$$

where $C = 1600$ *and* τ *is sampled from* $1, ..., T$.

Proof Defining a Lyapunov function as in (Huang et al., 2022),

$$\Gamma_t = P(x_t) + \frac{\mu_y}{\ell}\left(9\ell^2\|y_t - \hat{y}(x_t)\|^2 + \frac{1}{\eta^y}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{\eta^y}\|\nabla_y f(x_t, y_t) - u_t\|^2\right).$$

Then we have

$$
\begin{aligned}
&\Gamma_{t+1}-\Gamma_t\\
&=P(x_{t+1})-P(x_t)+\frac{9\mu_y}{\ell}\ell^2(\|y_{t+1}-\hat{y}(x_{t+1})\|^2-\|y_t-\hat{y}(x_t)\|^2)\\
&\quad+\frac{\mu_y}{\ell}\left(\frac{1}{\eta^y}\|\nabla_x f(x_{t+1},y_{t+1})-v_{t+1}\|^2-\frac{1}{\eta^y}\|\nabla_x f(x_t,y_t)-v_t\|^2\right)\\
&\quad+\frac{\mu_y}{\ell}\left(\frac{1}{\eta^y}\|\nabla_y f(x_{t+1},y_{t+1})-u_{t+1}\|^2-\frac{1}{\eta}\|\nabla_y f(x_t,y_t)-u_t\|^2\right)\\
&\le-\frac{\eta^x}{4}\|v_t\|^2+\eta^x\ell^2\|\hat{y}(x_t)-y_t\|^2+\eta^x\|\nabla_x f(x_t,y_t)-v_t\|^2\\
&\quad+\frac{9\mu_y}{\ell}\ell^2\left(-\frac{\mu_y\eta^y\lambda}{4}\|y_t-\hat{y}(x_t)\|^2-\frac{3\eta^y\lambda^2}{4}\|u_t\|^2+\frac{5\eta^y\lambda}{\mu_y}\|\nabla_y f(x_t,y_t)-u_t\|^2+\frac{5\ell^2(\eta^x)^2}{\eta^y\lambda\mu_y^3}\|v_t\|^2\right)\\
&\quad-\frac{\mu_y a_x}{16\ell\eta^y}\mathbb{E}[\|\nabla_x f(x_t,y_t)-v_t\|^2]+\frac{\mu_y\ell^2}{2\ell\eta^y}((\eta^x)^2\mathbb{E}[\|v_t\|^2]+(\eta^y)^2\lambda^2\mathbb{E}[\|u_t\|^2])+\frac{\mu_y a_x^2\sigma^2}{8\ell\eta^y}\\
&\quad-\frac{\mu_y a_y}{16\ell\eta^y}\mathbb{E}[\|\nabla_y f(x_t,y_t)-u_t\|^2]+\frac{\mu_y\ell^2}{2\ell\eta^y}((\eta^x)^2\mathbb{E}[\|v_t\|^2]+(\eta^y)^2\lambda^2\mathbb{E}[\|u_t\|^2])+\frac{\mu_y a_y^2\sigma^2}{8\ell\eta^y}\\
&\le(-\frac{9\mu_y^2\ell\lambda\eta^y}{4}+\eta^x\ell^2)\|y_t-\hat{y}(x_t)\|^2\\
&+(\eta^x-\frac{\mu_y a_x}{16\ell\eta^y})\mathbb{E}[\|\nabla_x f(x_t,y_t)-v_t\|^2]+(45\eta^y\ell\lambda-\frac{\mu_y a_y}{16\ell\eta^y})\mathbb{E}[\|\nabla_y f(x_t,y_t)-u_t\|^2]\\
&-(\frac{\eta^x}{4}-\frac{\mu_y\ell(\eta^x)^2}{\eta^y}-\frac{45\ell^3(\eta^x)^2}{\eta^y\lambda\mu_y^2})\mathbb{E}\|v_t\|^2+(\mu_y\ell\eta^y-9\mu_y\ell\frac{3\eta^y}{4})\lambda^2\mathbb{E}[\|u_t\|^2]+\frac{2\mu_y a_x^2\sigma^2}{\ell\eta^y}+\frac{2\mu_y a_y^2\sigma^2}{\ell\eta^y},
\end{aligned}
$$

where the first inequality uses Lemma 37, Lemma 38, Lemma 39.

Taking $\eta^x=\frac{\mu_2^2}{1000\ell^2}\eta^y, a_x=\frac{800\ell}{\mu_y}(\eta^y)^2, a_y=\frac{800\ell}{\mu_y}(\eta^y)^2$, $\eta^y\le O(\sqrt{\frac{\mu_y}{\ell}})$ to ensure $0<a_x,a_y<1$, we get

$$
\begin{aligned}
\Gamma_{t+1}-\Gamma_t\le&-\eta^x\|y_t-\hat{y}(x_t)\|^2-\eta^y\mathbb{E}[\|\nabla_x f(x_t,y_t)-v_t\|^2]\\
&-\eta^y\mathbb{E}[\|\nabla_y f(x_t,y_t)-v_t\|^2]-\frac{1}{8}\eta^x\mathbb{E}\|v_t\|^2+\frac{4C\ell(\eta^y)^3\sigma^2}{\mu_y},
\end{aligned}
\tag{102}
$$

where $C=1600$.

Thus,

$$
\begin{aligned}
&\eta^x\|y_t-\hat{y}(x_t)\|^2+\eta^y\mathbb{E}[\|\nabla_x f(x_t,y_t)-v_t\|^2]+\eta^y\mathbb{E}[\|\nabla_y f(x_t,y_t)-v_t\|^2]+\frac{1}{8}\eta^x\mathbb{E}\|v_t\|^2\\
&\le\Gamma_t-\Gamma_{t+1}+\frac{4C\ell(\eta^y)^3\sigma^2}{\mu_y}.
\end{aligned}
\tag{103}
$$

Taking average over $t=1,..,T$,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}[\|y_t-\hat{y}(x_t)\|^2+\frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_x f(x_t,y_t)-v_t\|^2]+\frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_y f(x_t,y_t)-v_t\|^2]+\frac{1}{8}\mathbb{E}\|v_t\|^2]\\
&\le\frac{\Gamma_1-\Gamma_{T+1}}{\eta^x T}+\frac{4C\ell(\eta^y)^3\sigma^2}{\mu_y\eta^x}.
\end{aligned}
\tag{104}
$$

Randomly sample τ from $1, ..., T$, we obtain

$$\mathbb{E}\|y_\tau - \hat{y}(x_\tau)\|^2 + \frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_x f(x_\tau, y_\tau) - v_\tau\|^2] + \frac{\eta^y}{\eta^x}\mathbb{E}[\|\nabla_y f(x_\tau, y_\tau) - v_\tau\|^2] + \frac{1}{8}\mathbb{E}\|v_\tau\|^2]$$
$$\leq \frac{\Gamma_1 - \Gamma_{T+1}}{\eta^x T} + \frac{4C\ell(\eta^y)^3\sigma^2}{\mu_y \eta^x}.$$

■

Theorem 13 is formally restated as follows:

Theorem 41 (Formal version of Theorem 13) *Assume Assumption 1, 2, 4, 5 hold. Define a constant* $\epsilon_1 = \frac{C\ell^2\sigma^2}{2\mu\mu_y^2}$ *and* $\epsilon_k = \epsilon_1/2^k$, *where* $C = 1600$. *By setting* $\eta_k^y = \min\{\frac{1}{30}\sqrt{\frac{\mu_y}{\ell}}, \sqrt{\frac{\mu\mu_y^3\epsilon_k}{320C\ell^3\sigma^2}}\}$, $\eta_k^x = \frac{\mu_y^2}{1000\ell^2}\eta_k^y$, $T_k = O\left(\max\{\frac{1}{\mu\eta_k^x}, \frac{\mu_y^3}{\ell^3\eta_k^x\eta_k^y}\}\right)$, *after* $K = O(log(\epsilon_1/\epsilon))$ *stages,* $\mathbb{E}[P(\bar{x}_k) - P(x_*)] \leq \epsilon$. *The total stochastic first-order oracle call complexity is* $\widetilde{O}\left(\frac{\ell^{7/2}}{\mu^{3/2}\mu_y^{7/2}\epsilon^{1/2}} + \frac{\ell^2}{\mu\mu_y^2\epsilon}\right)$.

Proof Without loss of generality, let us assume that the initialization of the first stage $P(x_0^1) - P(x_*) = \epsilon_0 \geq \frac{C\ell^2\sigma^2}{2\mu\mu_y^2}$, i.e. $\sqrt{\frac{\mu\mu_y^3\epsilon_0}{320C\ell^3\sigma^2}} > \frac{1}{30}\sqrt{\frac{\mu_y}{\ell}}$. The case where $\epsilon_0 \leq \frac{320C\ell^2\sigma^2}{\mu\mu_y^2}$ can be simply covered by our proof. Then denote $\epsilon_1 = \frac{C\ell^2\sigma^2}{2\mu\mu_y^2}$ and $\epsilon_k = \epsilon_1/2^k$.

Let's consider the first stage, we have initialization such that $P(x_0) - P(x_*) = \epsilon_0$ and $\mathbb{E}[\|\nabla_x f(x_0, y_0) - v_0\|^2 + \|\nabla_y f(x_0, y_0) - u_0\|^2] \leq \sigma^2$.

We bound the error of the first stage's output as follows

$$\mathbb{E}\ell^2\|\bar{y}_1 - \hat{y}(\bar{x}_1)\|^2 + \frac{\eta_1^y}{\eta_1^x}\mathbb{E}[\|\nabla_x f(\bar{x}_1, \bar{y}_1) - \bar{v}_1\|^2] + \frac{\eta_1^y}{\eta_1^x}\mathbb{E}[\|\nabla_y f(\bar{x}_1, \bar{y}_1) - \bar{v}_1\|^2] + \frac{1}{8}\mathbb{E}\|\bar{v}_1\|^2]$$
$$\leq \frac{\mathbb{E}[P(x_0)] - P(x_*)}{\eta^x 1 T_1} + \frac{\mu_y\ell^2\|y_0 - \hat{y}(x_0)\|^2}{\ell\eta_1^x T_1} + \frac{\mu_y}{\ell\eta_1^y\eta_1^x T_1}\|\nabla_x f(\bar{x}_1, \bar{y}_1) - \bar{v}_1\|^2$$
$$+ \frac{\mu_y}{\ell\eta_1^y\eta_1^x T_1}\|\nabla_y f(\bar{x}_1, \bar{y}_1) - \bar{u}_1\|^2 + \frac{4C\ell(\eta_1^y)^3\sigma^2}{\mu_y\eta_1^x}$$
$$\leq \frac{\mu\epsilon_1}{16},$$

where the last inequality is by the setting $\eta_1^x = \frac{\mu_y^2}{1000\ell^2}\eta_1^y$, $\eta_1^y = \sqrt{\frac{\mu_y}{\ell}}$ and $T_1 = O\left(\max\{\frac{\epsilon_0}{\eta_1^x\mu\epsilon_1}, \frac{\mu_y\sigma^2}{\mu\eta_1^x\eta_1^y\epsilon_1}, \frac{\mu_y\ell D}{\eta_1^x\mu\epsilon_1}\}\right)$, which is in the order of a constant and where D denotes the diameter of $\mathcal{Y}$. This result implies that

$$\begin{aligned}
&\mathbb{E}\ell^2\|\bar{y}_1 - \hat{y}(\bar{x}_1)\|^2 \leq \frac{\mu\epsilon_1}{16}, \\
&\mathbb{E}[\|\nabla_x f(x_\tau, y_\tau) - v_\tau\|^2] + \mathbb{E}[\|\nabla_y f(x_\tau, y_\tau) - v_\tau\|^2] \leq \frac{\mu\mu_y^2\epsilon_1}{16\ell^2}, \\
&\mathbb{E}\|\bar{v}_1\|^2 \leq \frac{\mu\epsilon_1}{2}.
\end{aligned} \tag{105}$$

Using the μ-PL condition of $P(x)$,

$$\begin{aligned}P(\bar{x}_1) - P(x_*) &\leq \frac{1}{2\mu}\|\nabla P(\bar{x}_1)\|^2 = \frac{1}{2\mu}\|\nabla P(\bar{x}_1) - \nabla_x f(\bar{x}_1, \bar{y}_1) + \nabla_x f(\bar{x}_1, \bar{y}_1) - \bar{v}_1 + \bar{v}_1\|^2 \\ &\leq \frac{1}{2\mu}(3\ell^2\|\bar{y}_1 - \hat{y}(\bar{x}_1)\|^2 + 3\|\nabla_x f(x_t, y_t) - v_t\|^2 + 3\|v_t\|^2) \leq \epsilon_1,\end{aligned}$$

where the second inequality has used $\nabla P(x) = \nabla f(x, \hat{y}(x))$, which is by the Lemma 16.

Starting from the second stage, we will prove by induction. Suppose the initialization of k-th stage ($k \geq 2$) satisfies $\mathbb{E}[P(\bar{x}_{k-1}) - P(x_*)] \leq \epsilon_{k-1}$, $\mathbb{E}[\|\nabla_x f(\bar{x}_{k-1}, \bar{y}_{k-1}) - v_{k-1}\|^2 + \|\nabla_y f(\bar{x}_{k-1}, \bar{y}_{k-1}) - u_{k-1}\|^2] \leq \frac{\mu\mu_y^2\epsilon_{k-1}}{\ell^2}$. The error of the output of k-th stage can be bounded as

$$\begin{aligned}&\mathbb{E}\ell^2\|\bar{y}_k - \hat{y}(\bar{x}_k)\|^2 + \frac{\eta_k^y}{\eta_k^x}\mathbb{E}[\|\nabla_x f(\bar{x}_k, \bar{y}_k) - v_k\|^2] + \frac{\eta_k^y}{\eta_k^x}\mathbb{E}[\|\nabla_y f(\bar{x}_k, \bar{y}_k) - v_k\|^2] + \frac{1}{8}\mathbb{E}\|v_k\|^2] \\ &\leq \frac{\mathbb{E}[P(\bar{x}_{k-1})] - P(x_*)}{\eta_k^x T_k} + \frac{\mu_y\ell^2\|\bar{y}_{k-1} - \hat{y}(\bar{x}_{k-1})\|^2}{\ell\eta_k^x T_k} + \frac{\mu_y}{\ell\eta_k^y\eta_k^x T_k}\|\nabla_x f(\bar{x}_k, \bar{y}_k) - v_k\|^2 \\ &\quad + \frac{\mu_y}{\ell\eta_k^y\eta_k^x T_k}\|\nabla_y f(\bar{x}_k, \bar{y}_k) - u_k\|^2 + \frac{4C\ell(\eta_k^y)^3\sigma^2}{\mu_y\eta_k^x} \\ &\leq \frac{\epsilon_{k-1}}{\eta_k^x T_k} + \frac{\mu_y\ell^2\mu\epsilon_{k-1}}{\ell\eta_k^x T_k} + \frac{\mu\mu_y^3\epsilon_{k-1}}{\ell^3\eta_k^y\eta_k^x T_k} + \frac{\mu\mu_y^3\epsilon_{k-1}}{\ell^3\eta_k^y\eta_k^x T_k} + \frac{4C\ell(\eta_k^y)^3\sigma^2}{\mu_y\eta_k^x} \\ &\leq \frac{\mu\epsilon_k}{16},\end{aligned} \tag{106}$$

where the last inequality is due to the setting $\eta_k^y = \sqrt{\frac{\mu\mu_y^3\epsilon_k}{320C\ell^3\sigma^2}}$, $\eta_k^x = \frac{\mu_y^2}{1000\ell^2}\eta_k^y$, $T_k = O\left(\max\left\{\frac{1}{\mu\eta_k^x}, \frac{\mu_y^3}{\ell^3\eta_k^x\eta_k^y}\right\}\right)$.

Similar to in the analysis of first stage, this result implies that

$$\begin{aligned}&\mathbb{E}[\ell^2\|\bar{y}_k - \hat{y}(\bar{x}_k)\|^2] \leq \frac{\mu\epsilon_{k-1}}{16}, \\ &\mathbb{E}[\|\nabla_x f(\bar{x}_k, \bar{y}_k) - v_k\|^2] + \mathbb{E}[\|\nabla_y f(\bar{x}_k, \bar{y}_k) - v_k\|^2] \leq \frac{\mu\mu_y^2\epsilon_k}{16\ell^2}, \\ &\mathbb{E}\|v_k\|^2 \leq \frac{\mu\epsilon_k}{2}.\end{aligned} \tag{107}$$

Using the μ-PL condition of $P(x)$, we obtain

$$\begin{aligned}P(\bar{x}_k) - P(x_*) &\leq \frac{1}{2\mu}\|\nabla P(\bar{x}_k)\|^2 = \frac{1}{2\mu}\|\nabla P(\bar{x}_k) - \nabla_x f(\bar{x}_k, \bar{y}_k) + \nabla_x f(\bar{x}_k, \bar{y}_k) - v_k + v_k\|^2 \\ &\leq \frac{1}{2\mu}(3\ell^2\|\bar{y}_k - \hat{y}(\bar{x}_k)\|^2 + 3\|\nabla_x f(\bar{x}_k, \bar{y}_k) - v_k\|^2 + 3\|v_k\|^2) \leq \epsilon_k.\end{aligned} \tag{108}$$

By induction we know that after $K = 1 + \log(\epsilon_1/\epsilon)$ stages, $P(\bar{x}_K) - P(x_*) \leq 0$. Total complexity is

$$\begin{aligned}\sum_{k=1}^{K} T_k &= O\left(\sum_{k=2}^{K}\left(\frac{1}{\mu\eta_k^x} + \frac{\mu_y^3}{\ell^3\eta_k^x\eta_k^y}\right)\right) \\ &= O\left(\sum_{k=2}^{K}\left(\frac{\ell^2}{\mu\mu_y^2\sqrt{\mu\mu_y^3\epsilon_k/\ell^3}} + \frac{\mu_y^3\ell^2}{\mu_y^2\mu\mu_y^3\epsilon_k}\right)\right) \\ &= \widetilde{O}\left(\frac{\ell^{7/2}}{\mu^{3/2}\mu_y^{7/2}\epsilon^{1/2}} + \frac{\ell^2}{\mu\mu_y^2\epsilon}\right).\end{aligned} \tag{109}$$

■

In the following corollary, we analyze the convergence of duality gap by PES-Storm.

Corollary 42 *Under the same setting as in Theorem 41 and suppose Assumption 5 as well. To achieve* $\mathbb{E}[\mathrm{Gap}(\bar{x}_K, \bar{y}_K)] \leq \epsilon$, *the total number of stochastic first-order oracle calls is* $\widetilde{O}\left(\frac{\ell^{9/2}}{\mu^{3/2}\mu_x^{1/2}\mu_y^{9/2}\epsilon^{1/2}} + \frac{\ell^4}{\mu\mu_x\mu_y^3\epsilon}\right)$.

Proof Assume after K stages, we have the output $\bar{x}_K, \bar{y}_K$ such that

$$\begin{aligned}P(\bar{x}_K) - P(x_*) &\leq \frac{1}{2\mu}\|\nabla P(\bar{x}_K)\|^2 \\ &\leq \frac{1}{2\mu}\left[3\ell^2\|\bar{y}_K - \hat{y}(\bar{x}_K)\|^2 + 3\|\nabla_x f(\bar{x}_K, \bar{y}_K) - v_K\|^2 + 3\|v_K\|^2\right] \\ &\leq \hat{\epsilon},\end{aligned} \tag{110}$$

and

$$\mathbb{E}\|\bar{y}_K - \hat{y}(\bar{x}_K)\|^2 \leq \frac{\mu\hat{\epsilon}}{16\ell^2}. \tag{111}$$

Hence, by the strong concavity,

$$\begin{aligned}\|\hat{y}(\bar{x}_K) - y_*\|^2 &\leq \frac{f(\bar{x}_K, \hat{y}(\bar{x}_K)) - f(\bar{x}_K, \bar{y}_K)}{2\mu_y} \\ &\leq \frac{f(\bar{x}_K, \hat{y}(\bar{x}_K)) - f(x_*, y_*) + f(x_*, y_*) - f(\bar{x}_K, \bar{y}_K)}{2\mu_y} \\ &\leq \frac{f(\bar{x}_K, \hat{y}(\bar{x}_K)) - f(x_*, y_*)}{2\mu_y} \\ &\leq \frac{P(\bar{x}_K) - P(x_*)}{2\mu_y} \\ &\leq \frac{\hat{\epsilon}}{2\mu_y}.\end{aligned} \tag{112}$$

Thus,

$$\|\bar{y}_K - y_*\|^2 \leq 2\|\bar{y}_K - \hat{y}(\bar{x}_K)\|^2 + 2\|\hat{y}(\bar{x}_K) - y_*\|^2 \leq \frac{\hat{\epsilon}}{\mu_y}. \tag{113}$$

And the dual function $D(y) = \min_{x'} f(x', y)$ is $\ell + \frac{\ell^2}{\mu_x} \leq \frac{2\ell^2}{\mu_x}$, where μ_x is the x-side PL condition coefficient. Therefore, we have

$$f(x_*, y_*) - f(\hat{x}(\bar{y}_K), \bar{y}_K) = D(y_*) - D(\bar{y}_K) \leq \frac{2\ell^2}{\mu_x}\|\bar{y}_K - y_*\|^2 \leq \frac{\ell^2\hat{\epsilon}}{\mu_x\mu_y}. \tag{114}$$

Then we know the duality gap is

$$\begin{aligned} f(\bar{x}_K, \hat{y}(\bar{x}_K)) - f(\hat{x}(\bar{y}_K), \bar{y}_K) &= f(\bar{x}_K, \hat{y}(\bar{x}_K)) - f(x_*, y_*) + f(x_*, y_*) - f(\hat{x}(\bar{y}_K), \bar{y}_K) \\ &\leq \hat{\epsilon} + \frac{\ell^2\hat{\epsilon}}{\mu_x\mu_y}. \end{aligned} \tag{115}$$

To make the duality gap less than ϵ, we need $\hat{\epsilon} \leq O(\frac{\mu_x\mu_y\epsilon}{\ell^2})$. Therefore, it takes $\widetilde{O}\left(\frac{\ell^{9/2}}{\mu^{3/2}\mu_x^{1/2}\mu_y^{9/2}\epsilon^{1/2}} + \frac{\ell^4}{\mu\mu_x\mu_y^3\epsilon}\right)$ to have a ϵ-duality gap. ■

Appendix K. Justification of PL condition

In this section, we show the analysis of cases where the x-side PL condition can hold, and show the properties that follow from the x-side PL condition. We need to introduce a auxiliary lemma as follows.

Lemma 43 (Corallary 5.1 of (Li and Pong, 2018)) *Suppose $h(x) = g(Ax)$, where A is a matrix and $g(\cdot)$ is strongly convex, then $h(x)$ satisfies a μ-PL condition.*

K.1 Proof of Lemma 9

Here we prove the Lemma 9 which justifies the PL condition for the non-convex AUC maximization problem.

Proof [Proof of Lemma 9] Before diving into analyzing the min-max formulation of the AUC maximization problem, we investigate the property of the optimal solution to a AUC maximization problem. Reconstruct a data set $\{(\mathbf{a}_1, c_1), ..., (\mathbf{a}_i, c_i), ..., (\mathbf{a}_n, c_n)\}$ where $c_i = i$ if $b_i = 1$ and $c_i = 0$ if $b_i = -1$. Consider the problem

$$\min_{\mathbf{w}} F_1(\mathbf{w}) := \sum_{i=1}^{n}(h(\mathbf{w}; \mathbf{a}_i) - b_i)^2. \tag{116}$$

By Theorem 1 and Theorem 3 of Allen-Zhu et al. (2019), we know that for $\mathbf{w}_* = \arg\min F_1(\mathbf{w})$, $\|\mathbf{w}_* - \mathbf{w}_0\|_2 \leq \omega$ and $F_1(\mathbf{w}_*) = 0$ where $\omega = O(\frac{\log m}{\sqrt{m}})$ and $\mathbf{w}_0$ is a random initialization. Then we know that $\mathbf{w}_*$ is a optimal solution to problem (4) as well with the optimal objective to be 0. Therefore, $\mathbf{w}_*$ is also a optimal solution of the Problem (5).

Then let us consider the min-max formulation of the AUC maximization problem. For the n input data points, the problem (5) can be written as

$$\min_{(\mathbf{w},s,r)} \max_{\alpha\in\mathbb{R}} f(\mathbf{w},s,r,y) = \frac{1}{n}\sum_{i=1}^{n} F(\mathbf{w},s,r,y,\mathbf{z}_i). \tag{117}$$

From Section 12 and Section 13 of (Allen-Zhu et al., 2019), we know that $h(\mathbf{w};\mathbf{a}) \leq O(\log m)$. Then by a similar analysis of Lemma 7 and Lemma 8 of (Guo et al., 2020), it holds that $\max\{|s|,|r|,|y|\} \leq O(\log m)$.

By Theorem 5 of (Allen-Zhu et al., 2019), for $\|\mathbf{w}-\mathbf{w}_0\| \leq \omega$, with probability at least $1 - e^{-\widetilde{\Omega}(m\omega^{2/3}\tilde{L})}$,

$$h(\mathbf{w};\mathbf{a}) = h(\mathbf{w}_0;\mathbf{a}) + \langle \nabla h(\mathbf{w}_0;\mathbf{a}), \mathbf{w}-\mathbf{w}_0\rangle \pm \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}). \tag{118}$$

Then for any fixed y and for $\|\mathbf{w}-\mathbf{w}_0\| \leq \omega$, with probability at least $1 - e^{-\widetilde{\Omega}(m\omega^{2/3}\tilde{L})}$,

$$\begin{aligned}
f(x,y) &= f(\mathbf{w},s,r,y)\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[(1-p)(h(\mathbf{w};\mathbf{a}_i)-s)^2\mathbb{I}_{[b_i=1]} + p(h(\mathbf{w};\mathbf{a}_i)-r)^2\mathbb{I}_{[b_i=-1]}\\
&\qquad + 2(1+y)(ph(\mathbf{w};\mathbf{a}_i)\mathbb{I}_{[b_i=-1]} - (1-p)h(\mathbf{w};\mathbf{a}_i)\mathbb{I}_{[b_i=1]}) - p(1-p)y^2\Big]\\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[(1-p)(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}) - s)^2\mathbb{I}_{[b_i=1]}\\
&\qquad + p(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}) - r)^2\mathbb{I}_{[b_i=-1]}\\
&\qquad + 2(1+y)p(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}))\mathbb{I}_{[b_i=-1]}\\
&\qquad - 2(1+y)(1-p)(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}))\mathbb{I}_{[b_i=1]}\\
&\qquad - p(1-p)y^2\Big].
\end{aligned} \tag{119}$$

Then,

$$\begin{aligned}
\hat{y}(x) =& \frac{1}{(1-p)n}\sum_{i=1}^{n}(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle)\mathbb{I}_{[b_i=-1]}\\
&- \frac{1}{pn}\sum_{i=1}^{n}(h(\mathbf{w}_0;\mathbf{a}_i) + \langle\nabla h(\mathbf{w}_0;\mathbf{a}_i),\mathbf{w}-\mathbf{w}_0\rangle)\mathbb{I}_{[b_i=1]} + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}).
\end{aligned} \tag{120}$$

Thus,

$$
\begin{aligned}
P(x) = \max_y f(x,y) = \frac{1}{n}\sum_{i=1}^n \Big[& (1-p)(\langle \nabla h(\mathbf{w}_0;\mathbf{a}_i), \mathbf{w}-\mathbf{w}_0\rangle - s)^2 \mathbb{I}_{[b_i=1]} \\
& + p(\langle \nabla h(\mathbf{w}_0;\mathbf{a}_i), \mathbf{w}-\mathbf{w}_0\rangle - r)^2 \mathbb{I}_{[b_i=-1]} \\
& + \frac{1}{p(1-p)}\big(p\langle \frac{1}{n}\sum_{j=1}^n \nabla h(\mathbf{w}_0;\mathbf{a}_j), \mathbf{w}-\mathbf{w}_0\rangle \mathbb{I}_{[b_j=-1]} - (1-p)\langle \nabla \frac{1}{n}\sum_{j=1}^n h(\mathbf{w}_0;\mathbf{a}_j), \mathbf{w}-\mathbf{w}_0\rangle \mathbb{I}_{[b_j=1]}\big)^2 \Big] \\
& + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}) \\
= {} & \|Hx-c\|^2 + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m})
\end{aligned} \tag{121}
$$

where $H \in \mathbb{R}^{(n+1)\times 3}$ and $c \in \mathbb{R}^{n+1}$. For $0 \le i \le n$, the i-th row is $H_i = (\sqrt{1-p}\nabla h(\mathbf{w}_0;\mathbf{a}_i), -1, 0)$ if $b_i = 1$ and $H_i = (\sqrt{p}\nabla h(\mathbf{w}_0;\mathbf{a}_i), 0, -1)$ if $b_i = -1$; and the i-th element of c is $c_i = \sqrt{1-p}\langle \nabla h(\mathbf{w}_0;\mathbf{a}_i), \mathbf{w}_0\rangle$. The last row of H is $(\frac{1}{\sqrt{p(1-p)}}(p\frac{1}{n}\sum_{i=1}^n \nabla h(\mathbf{w}_0;\mathbf{a}_i)\mathbb{I}_{[b_i=-1]} - (1-p)\frac{1}{n}\sum_{i=1}^n \nabla h(\mathbf{w}_0;\mathbf{a}_i)\mathbb{I}_{[b_i=1]}), 0, 0)$ and the last element of c is $\frac{1}{\sqrt{p(1-p)}}\langle (p\frac{1}{n}\sum_{i=1}^n \nabla h(\mathbf{w}_0;\mathbf{a}_i)\mathbb{I}_{[b_i=-1]} - (1-p)\frac{1}{n}\sum_{i=1}^n \nabla h(\mathbf{w}_0;\mathbf{a}_i)\mathbb{I}_{[b_i=1]}), \mathbf{w}_0\rangle$.

With (119) and Lemma 43, we know that for some $\mu > 0$ and any y, i.e.,

$$
2\mu(P(x) - P(x_*)) \le \|\nabla P(x)\|^2 + \widetilde{O}(\tilde{L}^3\omega^{4/3}\sqrt{m}). \tag{122}
$$

Since $\omega = O(\frac{\log m}{\sqrt{m}})$, by the choice of m, we know that

$$
\|\nabla P(x)\|^2 \ge 2\mu(P(x) - P(x_*) - \epsilon). \tag{123}
$$

■

K.2 An Example of x-side-PL-Strongly-Concave Problem

Lemma 44 *Let $x, y \in \mathbb{R}$, $f(x,y) = \frac{1}{2}x^2 + \sin^2 x \sin^2 y - 2y^2$. We have that $f(x,y)$ satisfies a x-side $\frac{1}{12}$ PL condition, is 2-strongly concave in y and has a saddle point $(0,0)$.*

Proof For any x,

$$
\nabla_y^2 f(x,y) = 2\sin^2 x \cos(2y) - 4 \in [-6, -2]. \tag{124}
$$

Thus, $f(x,y)$ is 2-strongly concave in y.

For any y, we know that $\hat{x}(y) = 0$, and

$$
|\nabla_x^2 f(x,y)| = |1 + 2\cos(2x)\sin^2 y|^2 \le 3, \tag{125}
$$

which together with (124) implies that $f(x,y)$ is 6-smooth.

We also get

$$\frac{|\nabla_x f(x,y)|}{|x-\hat{x}(y)|} = \frac{|x+\sin(2x)\sin^2 y|}{|x|} \geq \frac{1}{2}, \tag{126}$$

which together with the 6-smoothness we know that $f(x,y)$ satisfies a x-side $\frac{1}{12}$-PL condition by Appendix A of (Karimi et al., 2016).

Also it is easy to verify that

$$f(0,y) \leq f(0,0) \leq f(x,0), \tag{127}$$

therefore $(0,0)$ is a saddle point.

■

K.3 Existence of a saddle point

Proof [Proof of Lemma 8] Since $x_* = \arg\min_{x'} P(x')$, $\nabla P(x_*) = \nabla_x f(x_*, \hat{y}(x_*)) = 0$, where the first equality holds by the Lemma 16. Then noting the x-side PL condition $2\mu_x(f(x_*,\hat{y}(x_*)) - \min_{x'} f(x',\hat{y}(x_*))) \leq \|\nabla_x f(x_*,\hat{y}(x_*))\|^2 = 0$, we have

$$x_* \in \hat{x}(\hat{y}(x_*)), \tag{128}$$

which is one of the optimal x corresponding to the $\hat{y}(x_*)$.

Then we can conclude that $(x_*, \hat{y}(x_*))$ is a saddle point of $f(x,y)$, i.e., for any x and $y \in \mathcal{Y}$,

$$f(x_*, y) \leq f(x_*, \hat{y}(x_*)) \leq f(x, \hat{y}(x_*)). \tag{129}$$

■

Appendix L. Using Different Step Sizes for Primal and Dual Variables

In previous sections, we used the same step size for for primal and dual Variables in order to simplify the analysis. Actually, step sizes for primal and dual variables can be set different. In this section, we provide an analysis and rewrite the algorithm in Algorithm 3, 4. Similar as before, we first provide a unified theorem.

Algorithm 3 Proximal Stage Stochastic Method: PES-$\mathcal{A}$

1: Initialization: $\bar{x}_0 \in \mathbb{R}^d, \bar{y}_0 \in \mathcal{Y}, \gamma, T_1, \eta_1^x, \eta_1^y, a$.
2: **for** $k = 1, 2, ..., K$ **do**
3: $\quad x_0^k = \bar{x}_{k-1}$, $y_0^k = \bar{y}_{k-1}$;
4: $\quad (\bar{x}_k, \bar{y}_k) = \mathcal{A}(f, x_0^k, y_0^k, \eta_k^x, \eta_k^y, T_k, \gamma)$;
5: $\quad \eta_{k+1}^x = \eta_k^x/a$,$\eta_{k+1}^y = \eta_k^y/a$, $T_{k+1} = aT_k$;
6: **end for**
7: **return** $(\bar{x}_K, \bar{y}_K)$.

Algorithm 4 Stochastic Algorithm $\mathcal{A}(f, x_0, y_0, \eta^x, \eta^y, T, \gamma, u_0, v_0)$

Initialization: (x_0, y_0)
Let $\{\xi_0, \xi_1, \ldots, \xi_T\}$ be independent random variables.
for $t = 1, ..., T$ **do**
$x_t = \Pi^{\gamma}_{x_{t-1}, x_0}(\eta^x \mathcal{G}(x_{t-1}; \xi_{t-1}))$;
$y_t = \Pi_{y_{t-1}}(\eta^y \mathcal{G}(y_{t-1}; \xi_{t-1}))$;
end for
return $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t, \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$.

Theorem 45 *Suppose Assumption 1 and Assumption 4 hold. Assume we have a subroutine in the k-th stage of Algorithm 3 that can return $\bar{x}_k, \bar{y}_k$ such that*

$$\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \mathbb{E}[\frac{C_1}{\eta_k^x T_k}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{C_1}{\eta_k^y T_k}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2] + (\eta_k^x + \eta_k^x)C_2, \quad (130)$$

where C_1 and C_2 are constants corresponding to the specific subroutine. Take $\gamma = 2\rho$ and denote $\hat{L} = L + 2\rho$ and $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$. Define $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \mathrm{Gap}(\bar{x}_0, \bar{y}_0)$. Then we can set $\eta_k^x = \eta_0^x \exp(-(k-1)\frac{2\mu}{c+2\mu})$, $\eta_k^y = \eta_0^y \exp(-(k-1)\frac{2\mu}{c+2\mu})$, $T_k = \left\lceil \frac{212C_1}{\min\{\eta_0^x\rho, \eta_0^y\mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$. After $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{16(\eta_0^x+\eta_0^y)\hat{L}KC_2}{(c+2\mu)\epsilon}\right\}\right\rceil$ stages, we can have $\Delta_{K+1} \leq \epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\max\left\{\frac{(L+\rho)C_1\epsilon_0}{\min\{\eta_0^x\rho, \eta_0^y\mu_y\}\mu\epsilon}, (\frac{1}{\rho} + \frac{\eta_0^x}{\eta_0^y\mu_y} + \frac{\eta_0^y}{\eta_0^x\rho} + \frac{1}{\mu_y})\frac{(L+\rho)^2C_2}{\mu^2\epsilon}\right\}\right)$.

Remark. As long as $O(\eta_0^x) \leq \eta_0^y \leq O(\frac{\eta_0^x}{\mu_y})$, $\eta_0^x \geq O(\mu\mu_y)$, and $\eta_0^y \geq O(\mu)$, then η_0^x, η_0^y can be separately tuned without harming the complexity bound.

Proof [Proof of Theorem 45] Since $f(x, y)$ is ρ-weakly convex in x for any y, $P(x) = \max_{y' \in \mathcal{Y}} f(x, y')$ is also ρ-weakly convex. Taking $\gamma = 2\rho$, we have

$$\begin{aligned}
P(\bar{x}_{k-1}) &\geq P(\bar{x}_k) + \langle \nabla P(\bar{x}_k), \bar{x}_{k-1} - \bar{x}_k\rangle - \frac{\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&= P(\bar{x}_k) + \langle \nabla P(\bar{x}_k) + 2\rho(\bar{x}_k - \bar{x}_{k-1}), \bar{x}_{k-1} - \bar{x}_k\rangle + \frac{3\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&\overset{(a)}{=} P(\bar{x}_k) + \langle \nabla P_k(\bar{x}_k), \bar{x}_{k-1} - \bar{x}_k\rangle + \frac{3\rho}{2}\|\bar{x}_{k-1} - \bar{x}_k\|^2 \\
&\overset{(b)}{=} P(\bar{x}_k) - \frac{1}{2\rho}\langle \nabla P_k(\bar{x}_k), \nabla P_k(\bar{x}_k) - \nabla P(\bar{x}_k)\rangle + \frac{3}{8\rho}\|\nabla P_k(\bar{x}_k) - \nabla P(\bar{x}_k)\|^2 \\
&= P(\bar{x}_k) - \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{1}{4\rho}\langle \nabla P_k(\bar{x}_k), \nabla P(\bar{x}_k)\rangle + \frac{3}{8\rho}\|\nabla P(\bar{x}_k)\|^2,
\end{aligned} \quad (131)$$

where (a) and (b) hold by the definition of $P_k(x)$.

Rearranging the terms in (131) yields

$$\begin{aligned}
P(\bar{x}_k) - P(\bar{x}_{k-1}) &\leq \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 + \frac{1}{4\rho}\langle \nabla P_k(\bar{x}_k), \nabla P(\bar{x}_k)\rangle - \frac{3}{8\rho}\|\nabla P(\bar{x}_k)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{8\rho}\|\nabla P_k(\bar{x}_k)\|^2 + \frac{1}{8\rho}(\|\nabla P_k(\bar{x}_k)\|^2 + \|\nabla P(\bar{x}_k)\|^2) - \frac{3}{8\rho}\|P(\bar{x}_k)\|^2 \\
&= \frac{1}{4\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{1}{4\rho}\|\nabla P(\bar{x}_k)\|^2 \\
&\overset{(b)}{\leq} \frac{1}{4\rho}\|\nabla P_k(\bar{x}_k)\|^2 - \frac{\mu}{2\rho}(P(\bar{x}_k) - P(x_*)),
\end{aligned} \tag{132}$$

where (a) holds by using $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$, and (b) holds by the μ-PL property of $P(x)$.

Thus, we have

$$(4\rho + 2\mu)\left(P(\bar{x}_k) - P(x_*)\right) - 4\rho(P(\bar{x}_{k-1}) - P(x_*)) \leq \|\nabla P_k(\bar{x}_k)\|^2. \tag{133}$$

Since $\gamma = 2\rho$, $f_k(x, y)$ is ρ-strongly convex in x and μ_y strong concave in y. Apply Lemma 25 to f_k, we know that

$$\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\mu_y}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2 \leq \mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k). \tag{134}$$

By the setting of $\eta_k^x = \eta_0^x \exp\left(-(k-1)\frac{2\mu}{c+2\mu}\right)$, $\eta_k^y = \eta_0^y \exp\left(-(k-1)\frac{2\mu}{c+2\mu}\right)$, and $T_k = \left\lceil \frac{212C_1}{\min\{\eta_0^x\rho, \eta_0^y\mu_y\}} \exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$, we note that $\frac{C_1}{\eta_k^x T_k} \leq \frac{\rho}{212}$ and $\frac{C_1}{\eta_k^y T_k} \leq \frac{\mu_y}{212}$. Applying (130), we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq (\eta_k^x + \eta_k^y)C_2 + \frac{1}{53}\mathbb{E}\left[\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\mu_y}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2\right] \\
&\leq (\eta_k^x + \eta_k^y)C_2 + \frac{1}{53}\mathbb{E}\left[\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right].
\end{aligned} \tag{135}$$

Since $P(x)$ is L-smooth and $\gamma = 2\rho$, then $P_k(x)$ is $\hat{L} = (L + 2\rho)$-smooth. According to Theorem 2.1.5 of (Nesterov, 2004), we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla P_k(\bar{x}_k)\|^2] \leq 2\hat{L}\mathbb{E}(P_k(\bar{x}_k) - \min_{x\in\mathbb{R}^d} P_k(x)) \leq 2\hat{L}\mathbb{E}[\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&= 2\hat{L}\mathbb{E}[4\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)] \\
&\leq 2\hat{L}\mathbb{E}\left[4\left((\eta_k^x + \eta_k^y)C_2 + \frac{1}{53}\left(\mathrm{Gap}_k(x_0^k, y_0^k) + \mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right)\right) - 3\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right] \\
&= 2\hat{L}\mathbb{E}\left[4(\eta_k^x + \eta_k^y)C_2 + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{155}{53}\mathrm{Gap}_k(\bar{x}_k, \bar{y}_k)\right].
\end{aligned} \tag{136}$$

Applying Lemma 26 to (136), we have

$$\begin{aligned}\mathbb{E}[\|\nabla P_k(\bar{x}_k)\|^2] \leq 2\hat{L}\mathbb{E}\Bigg[4(\eta_k^x+\eta_k^y)C_k + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) \\ - \frac{155}{53}\left(\frac{3}{50}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k))\right)\Bigg] \\ = 2\hat{L}\mathbb{E}\left[4(\eta_k^x+\eta_k^y)C_2 + \frac{4}{53}\mathrm{Gap}_k(x_0^k, y_0^k) - \frac{93}{530}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \frac{124}{53}(P(x_0^{k+1}) - P(x_0^k))\right].\end{aligned}$$

Combining this with (133), rearranging the terms, and defining a constant $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$, we get

$$\begin{aligned}&(c+2\mu)\,\mathbb{E}[P(x_0^{k+1}) - P(x_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1})] \\ &\leq \left(4\rho + \frac{248}{53}\hat{L}\right)\mathbb{E}[P(x_0^k) - P(x_*)] + \frac{8\hat{L}}{53}\mathbb{E}[\mathrm{Gap}_k(x_0^k, y_0^k)] + 8(\eta_k^x+\eta_k^y)\hat{L}C_2 \\ &\leq c\mathbb{E}\left[P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)\right] + 8(\eta_k^x+\eta_k^y)\hat{L}C_2.\end{aligned} \tag{137}$$

Using the fact that $\hat{L} \geq \mu$,

$$(c+2\mu)\frac{8\hat{L}}{53c} = \left(4\rho + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\rho + \frac{248}{53}\hat{L})} \leq \frac{8\hat{L}}{53} + \frac{16\mu\hat{L}}{248\hat{L}} \leq \frac{93}{265}\hat{L}. \tag{138}$$

Then, we have

$$\begin{aligned}&(c+2\mu)\mathbb{E}\left[P(x_0^{k+1}) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1})\right] \\ &\leq c\mathbb{E}\left[P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)\right] + 8(\eta_k^x+\eta_k^y)\hat{L}C_2.\end{aligned} \tag{139}$$

Defining $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$, then

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_k] + \frac{8(\eta_k^x+\eta_k^y)\hat{L}C_2}{c+2\mu}. \tag{140}$$

Using this inequality recursively, it yields

$$E[\Delta_{K+1}] \leq \left(\frac{c}{c+2\mu}\right)^K E[\Delta_1] + \frac{8\hat{L}C_2}{c+2\mu}\sum_{k=1}^{K}\left((\eta_k^x+\eta_k^y)\left(\frac{c}{c+2\mu}\right)^{K+1-k}\right). \tag{141}$$

By definition,

$$\begin{aligned}\Delta_1 &= P(x_0^1) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_1(x_0^1, y_0^1) \\ &= P(\bar{x}_0) - P(x_*) + \left(f(\bar{x}_0, \hat{y}_1(\bar{x}_0)) + \frac{\gamma}{2}\|\bar{x}_0 - \bar{x}_0\|^2 - f(\hat{x}_1(\bar{y}_0), \bar{y}_0) - \frac{\gamma}{2}\|\hat{x}_1(\bar{y}_0) - \bar{x}_0\|^2\right) \\ &\leq \epsilon_0 + f(\bar{x}_0, \hat{y}_1(\bar{x}_0)) - f(\hat{x}(\bar{y}_0), \bar{y}_0) \leq 2\epsilon_0.\end{aligned}$$

Using inequality $1-x \le \exp(-x)$, we have

$$\mathbb{E}[\Delta_{K+1}] \le \exp\left(\frac{-2\mu K}{c+2\mu}\right)\mathbb{E}[\Delta_1] + \frac{8(\eta_0^x+\eta_0^y)\hat{L}C_2}{c+2\mu}\sum_{k=1}^{K}\exp\left(-\frac{2\mu K}{c+2\mu}\right)$$
$$\le 2\epsilon_0 \exp\left(\frac{-2\mu K}{c+2\mu}\right) + \frac{8(\eta_0^x+\eta_0^y)\hat{L}C_2}{c+2\mu}K\exp\left(-\frac{2\mu K}{c+2\mu}\right).$$

To make this less than ϵ, it suffices to make

$$2\epsilon_0 \exp\left(\frac{-2\mu K}{c+2\mu}\right) \le \frac{\epsilon}{2},$$
$$\frac{8(\eta_0^x+\eta_0^y)\hat{L}C_2}{c+2\mu}K\exp\left(-\frac{2\mu K}{c+2\mu}\right) \le \frac{\epsilon}{2}.$$

Let K be the smallest value such that $\exp\left(\frac{-2\mu K}{c+2\mu}\right) \le \min\{\frac{\epsilon}{4\epsilon_0}, \frac{(c+2\mu)\epsilon}{16(\eta_0^x+\eta_0^y)\hat{L}KC_2}\}$. We can set $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{16(\eta_0^x+\eta_0^y)\hat{L}KC_2}{(c+2\mu)\epsilon}\right\}\right\rceil$. Then, the total stochastic first-order oracle call complexity is

$$\sum_{k=1}^{K} T_k \le O\left(\frac{212C_1}{\min\{\eta_0^x,\eta_0^y\}\min\{\rho,\mu_y\}}\sum_{k=1}^{K}\exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right)$$
$$\le O\left(\frac{212C_1}{\min\{\eta_0^x,\eta_0^y\}\min\{\rho,\mu_y\}}\frac{\exp(K\frac{2\mu}{c+2\mu})-1}{\exp(\frac{2\mu}{c+2\mu})-1}\right)$$
$$\overset{(a)}{\le} \widetilde{O}\left(\frac{cC_1}{\min\{\eta_0^x\rho,\eta_0^y\mu_y\}\mu}\max\left\{\frac{\epsilon_0}{\epsilon}, \frac{(\eta_0^x+\eta_0^y)\hat{L}KC_2}{(c+2\mu)\epsilon}\right\}\right)$$
$$\le \widetilde{O}\left(\max\left\{\frac{(L+\rho)C_1\epsilon_0}{\min\{\eta_0^x\rho,\eta_0^y\mu_y\}\mu\epsilon}, \frac{(\eta_0^x+\eta_0^y)(L+\rho)^2C_2}{\min\{\eta_0^x\rho,\eta_0^y\mu_y\}\mu^2\epsilon}\right\}\right)$$
$$\le \widetilde{O}\left(\max\left\{\frac{(L+\rho)C_1\epsilon_0}{\min\{\eta_0^x\rho,\eta_0^y\mu_y\}\mu\epsilon}, (\frac{1}{\rho}+\frac{\eta_0^x}{\eta_0^y\mu_y}+\frac{\eta_0^y}{\eta_0^x\rho}+\frac{1}{\mu_y})\frac{(L+\rho)^2C_2}{\mu^2\epsilon}\right\}\right),$$

where (a) uses the setting of K and $\exp(x)-1 \ge x$, and $\widetilde{O}$ suppresses logarithmic factors. ■

Theorem 46 *Consider Algorithm 3 that uses Algorithm 4 as a subroutine. Suppose Assumption 1, 3, 4 hold. Assume $\mathbb{E}\|\nabla_x f(x,y;\xi)\|^2 \le B^2$ and $\mathbb{E}\|\nabla_y f(x,y;\xi)\|^2 \le B^2$. Take $\gamma = 2\rho$ and denote $\hat{L} = L+2\rho$ and $c = 4\rho + \frac{248}{53}\hat{L} \in O(L+\rho)$. Define $\Delta_k = P(x_0^k) - P(x_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_k(x_0^k, y_0^k)$ and $\epsilon_0 = \mathrm{Gap}(\bar{x}_0, \bar{y}_0)$. Then we can set $\eta_k^x = \eta_0^x\exp(-(k-1)\frac{2\mu}{c+2\mu}) \le \frac{1}{\rho}$,$\eta_k^y = \eta_0^y\exp(-(k-1)\frac{2\mu}{c+2\mu})$, $T_k = \left\lceil \frac{212C_1}{\min\{\eta_0^x\rho,\eta_0^y\mu_y\}}\exp\left((k-1)\frac{2\mu}{c+2\mu}\right)\right\rceil$. After $K = \left\lceil \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{80\eta_0\hat{L}KB_2}{(c+2\mu)\epsilon}\right\}\right\rceil$ stages, we can have $\Delta_{K+1} \le \epsilon$. The total stochastic first-order oracle call complexity is $\widetilde{O}\left(\frac{(L+\rho)^2B^2}{\mu^2\min\{\rho,\mu_y\}\epsilon}\right)$.*

Proof [Proof of Theorem 46] Using Lemma 28, we can set $\gamma = 2\rho$ and $\eta_0 = \frac{1}{\rho}$. Then it follows that

$$E[\text{Gap}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{2}{\eta_k^x T_k} E[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{2}{\eta_k^y T_k} E[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2 + \frac{5(\eta_k^x + \eta_k^y)B^2}{2}.$$

We plug in $C_1 = 2$ and $C_2 = 5B^2/2$ to Theorem 45 and the conclusion follows. ■