Optimal Epoch Stochastic Gradient Descent Ascent Methods for Min-Max Optimization

Yan Yan
School of EECS
Washington State University
yanyan.1@wsu.edu

Yi Xu
Machine Intelligence Technology
Alibaba Group US Inc
statxy@gmail.com

Qihang Lin
Department of Business Analytics
University of Iowa
qihang-lin@uiowa.edu

Wei Liu
Tencent AI Lab
wl2223@columbia.edu

Tianbao Yang
Department of CS
University of Iowa
tianbao-yang@uiowa.edu

Abstract

Epoch gradient descent method (a.k.a. Epoch-GD) proposed by [16] was deemed a breakthrough for stochastic strongly convex minimization, which achieves the optimal convergence rate of $O(1/T)$ with T iterative updates for the *objective gap*. However, its extension to solving stochastic min-max problems with strong convexity and strong concavity still remains open, and it is still unclear whether a fast rate of $O(1/T)$ for the *duality gap* is achievable for stochastic min-max optimization under strong convexity and strong concavity. Although some recent studies have proposed stochastic algorithms with fast convergence rates for min-max problems, they require additional assumptions about the problem, e.g., smoothness, bi-linear structure, etc. In this paper, we bridge this gap by providing a sharp analysis of epoch-wise stochastic gradient descent ascent method (referred to as Epoch-GDA) for solving strongly convex strongly concave (SCSC) min-max problems, without imposing any additional assumption about smoothness or the function's structure. To the best of our knowledge, our result is the first one that shows Epoch-GDA can achieve the optimal rate of $O(1/T)$ for the duality gap of general SCSC min-max problems. We emphasize that such generalization of Epoch-GD for strongly convex minimization problems to Epoch-GDA for SCSC min-max problems is non-trivial and requires novel technical analysis. Moreover, we notice that the key lemma can also be used for proving the convergence of Epoch-GDA for weakly-convex strongly-concave min-max problems, leading to a nearly optimal complexity without resorting to smoothness or other structural conditions.

1 Introduction

In this paper, we consider stochastic algorithms for solving the following min-max saddle-point problem with a general objective function f *without smoothness or any other special structure*:

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

where $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^n$ are closed convex sets and $f : X \times Y \to \mathbb{R}$ is continuous. It is of great interest to find a saddle-point solution to the above problem, which is defined as (x^*, y^*) such that $f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*), \forall x \in X, y \in Y$. Problem (1) covers a number of applications

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

in machine learning, including distributionally robust optimization (DRO) [31, 30], learning with non-decomposable loss functions [27, 11, 43, 26], and generative adversarial networks [13, 3].

In this work, we focus on two classes of the min-max problems: (i) strongly-convex strongly-concave (SCSC) problem where f is strongly convex in terms of x for any $y \in Y$ and is strongly concave in terms of y for any $x \in X$; (ii) weakly-convex strongly-concave (WCSC) problem, where there exists $\rho > 0$ such that $f(x, y) + \frac{\rho}{2}\|x\|^2$ is strongly convex in terms of x for any $y \in Y$ and is strongly concave in terms of y for any $x \in X$. Both classes have applications in machine learning [41, 36].

Although stochastic algorithms for convex-concave min-max problems have been studied extensively in the literature, their research is still far behind its counterpart for stochastic convex minimization problems. Below, we highlight some of these gaps to motivate the present work. For the sake of presentation, we first introduce some terminologies. The duality gap at (x, y) is defined as $\text{Gap}(x, y) := f(x, \hat{y}(x)) - f(\hat{x}(y), y)$, where $\hat{x}(y) := \arg\min_{x' \in X} f(x', y)$ and $\hat{y}(x) := \arg\max_{y' \in Y} f(x, y')$. If we denote by $P(x) := \max_{y' \in Y} f(x, y')$, then $P(x) - P(x^*)$ is the primal objective gap, where $x^* = \arg\min_{x \in X} P(x)$.

When f is convex in x and concave in y, many studies have designed and analyzed stochastic primal-dual algorithms for solving the min-max problems under different conditions of the problem (see references in next section). A standard result is provided by [32], which proves that primal-dual SGD suffers from a convergence rate of $O(1/\sqrt{T})$ for the duality gap without imposing any additional assumptions about the objective function. This is analogous to that for stochastic convex minimization [32]. However, the research of stochastic algorithms for SCSC problems lacks behind that for strongly convex minimization problems. A well-known result for stochastic strongly convex minimization is given by [16], which presents the first fast convergence rate $O(1/T)$ for stochastic strongly convex minimization by the Epoch-GD algorithm, which runs standard SGD in an epoch-wise manner by decreasing the step size geometrically. However, a fast rate of $O(1/T)$ for the *duality gap* of a stochastic algorithm **is still unknown for general SCSC problems**. We notice that there are extensive studies about stochastic algorithms with faster convergence rates than $O(1/\sqrt{T})$ for solving convex-concave min-max problems [46, 38, 37, 10, 6, 5, 35, 21, 41, 18, 47]. However, these works usually require additional assumptions about the objective function (e.g., smoothness, bilinear structure) or only prove the convergence in weaker measures (e.g., the primal objective gap, the distance of a solution to the saddle point).

We aim to bridge this gap by presenting the first optimal rate $O(1/T)$ of the duality gap for solving general SCSC problems. In particular, we propose an epoch-wise stochastic gradient descent ascent (Epoch-GDA) algorithm - a primal-dual variant of Epoch-GD that runs stochastic gradient descent update for the primal variable and stochastic gradient ascent update for the dual variable for solving (1). Although the algorithmic generalization is straightforward, the proof of convergence in terms of the duality gap for Epoch-GDA is not straightforward at all. We note that the key difference in the analysis of Epoch-GDA is that to upper bound the duality gap of a solution $(\bar{x}, \bar{y})$ we need to deal with the distance of an initial solution (x_0, y_0) to the reference solutions $(\hat{x}(\bar{y}), \hat{y}(\bar{x}))$, where $\hat{x}(\bar{y}) = \arg\min_{x' \in X} f(x', \bar{y})$ and $\hat{y}(\bar{x}) = \arg\max_{y' \in Y} f(\bar{x}, y')$ depend on $\bar{y}$ and $\bar{x}$, respectively. In contrast, in the analysis of the objective gap for Epoch-GD, one only needs to deal with the distance from an initial solution x_0 to the optimal solution x^*, i.e., $\|x_0 - x^*\|_2^2$, which by strong convexity can easily connects to the objective gap $P(x_0) - P(x^*)$, leading to the telescoping sum on the objective gap. Towards addressing the challenge caused by dealing with the duality gap, we present a key lemma that connects the distance measure $\|x_0 - \hat{x}(\bar{y})\|_2^2 + \|y_0 - \hat{y}(\bar{x})\|_2^2$ to the duality gap of (x_0, y_0) and $(\bar{x}, \bar{y})$. In addition, since we use the same technique as Epoch-GD for handling the variance of stochastic gradient by projecting onto a bounded ball with shrinking radius, we have to carefully prove that such restriction does not affect the duality gap for the original problem, which also needs to deal with bounding $\|x_0 - \hat{x}(\bar{y})\|_2^2$ and $\|y_0 - \hat{y}(\bar{x})\|_2^2$.

Moreover, we notice that the aforementioned key lemma and the telescoping technique based on the duality gap can also be used for proving the convergence of Epoch-GDA for **finding an approximate stationary solution of general WCSC problems**. The algorithmic framework is similar to that proposed by [36], i.e., by solving SCSC problems successively, but with a subtle difference in handling the dual variable. In particular, we do not need additional condition on the structure of the objective function and extra care for dealing with the dual variable for restart as done in [36]. This key difference is caused by our sharper analysis, i.e., we use the telescoping sum based on the duality gap instead of the primal objective gap as in [36]. As a result, our algorithm and analysis lead to a

Table 1: Summary of complexity results of this work and previous works for finding an ϵ-duality-gap solution for SCSC or an ϵ-stationary solution for WCSC min-max problems. We focus on comparison of existing results without assuming smoothness of the objective function. Restriction means whether an additional condition about the objective function's structure is imposed.

Setting	Works	Restriction	Convergence	Complexity
SCSC	[32]	No	Duality Gap	$O\left(1/\epsilon^2\right)$
	[41]	Yes	Primal Gap	$O\left(1/\epsilon + n\log(1/\epsilon)\right)$
	This paper	**No**	**Duality Gap**	$\boldsymbol{O\left(1/\epsilon\right)}$
WCSC	[36]	No	Nearly Stationary	$\widetilde{O}\left(1/\epsilon^6\right)$
	[36]	Yes	Nearly Stationary	$\widetilde{O}\left(1/\epsilon^4 + n/\epsilon^2\right)$
	This paper	**No**	**Nearly Stationary**	$\boldsymbol{\widetilde{O}\left(1/\epsilon^4\right)}$

nearly optimal complexity for solving WCSC problems without the smoothness assumption on the objective [2] [1]. Finally, we summarize our results and the comparison with existing results in Table 1.

2 Related Work

Below, we provide an overview of related results in this area and the review is not necessarily exhaustive. In addition, we focus on the stochastic algorithms, and leave deterministic algorithms [4, 33, 42, 12, 34, 19, 14, 20, 28, 15] out of our discussion.

[32] is one of the early works that studies stochastic primal-dual gradient methods for convex-concave min-max problems, which establishes a convergence rate of $O(1/\sqrt{T})$ for the duality gap of general convex-concave problems. Following this work, many studies have tried to improve the algorithm and the analysis for a certain class of problems by exploring the smoothness condition of some component functions [23, 47, 21] or bilinear structure of the objective function [5, 6]. For example, [47] considers a family of min-max problems whose objective is $f(x) + g(x) + \phi(x, y) - J(y)$, where the smoothness condition is imposed on f and ϕ and strong convexity is imposed on f if necessary, and establishes optimal or nearly optimal complexity of a stochastic primal-dual hybrid algorithm. Although the dependence on each problem parameter of interest is made (nearly) optimal, the worst case complexity is still $O(1/\sqrt{T})$. [21] considers single-call stochastic extra-gradient and establishes $O(1/T)$ rate for smooth and strongly monotone variational inequalities in terms of the square distance from the returned solution to the saddle point. [44] also considers variational inequalities with a smoothing technique, so that it handles nonsmooth problems, but they derive the convergence of the square distance from the returned solution to the saddle point, as in [21]. The present work is complementary to these developments by making no assumption on smoothness or the structure of the objective but considers strong (weak) convexity and strong concavity of the objective function. It has applications in robust learning with non-smooth loss functions [41, 36].

In the machine learning community, many works have considered stochastic primal-dual algorithms for solving regularized loss minimization problems, whose min-max formulation usually exhibits bi-linear structure [46, 37, 39, 10, 35]. For example, [46] designs a stochastic primal-dual coordinate (SPDC) method for SCSC problems with bilinear structure, which enjoys a linear convergence for the duality gap. Similarly, in [45, 38], different variants of SPDC are proposed and analyzed for problems with the bilinear structure. [35] proposes stochastic variance reduction methods for a family of saddle-point problems with special structure that yields a linear convergence rate. An exception that makes no smoothness assumption and imposes no bilinear structure is a recent work [41]. It considers a family of functions $f(x, y) = y^\top \ell(x) - \phi^*(y) + g(x)$ and proposes a stochastic primal-dual algorithm similar to Epoch-GDA. The key difference is that [41] designs a particular scheme that computes a restarting dual solution based on $\nabla\phi(\ell(\bar{x}))$, where $\bar{x}$ is a restarting primal solution in order to derive a fast rate of $O(1/T)$ under strong convexity and strong concavity. Additionally, their

[1] Although [2] only concerns the lower bound of finding a stationary point of smooth non-convex problems $\min_x f(x)$ through stochastic first-order oracle, it is a special case of the WCSC problem.

fast rate $O(1/T)$ is in terms of the primal objective gap, which is weaker than our convergence result in terms of the duality gap.

There is also increasing interest in stochastic primal-dual algorithms for solving WCSC min-max problems. To the best of our knowledge, [36] is probably the first work that comprehensively studies stochastic algorithms for solving WCSC min-max problems. To find a nearly ϵ-stationary point, their algorithms suffer from an $O(1/\epsilon^6)$ iteration complexity without strong concavity and an $O(1/\epsilon^4 + n/\epsilon^2)$ complexity with strong concavity and a special structure of the objective function that is similar to that imposed in [41]. Some recent works are trying to improve the complexity for solving WCSC min-max problems by exploring other conditions (e.g., smoothness) [25, 29, 40, 22]. For example, [25] establishes an $O(1/\epsilon^4)$ complexity for a single-loop stochastic gradient descent ascent method, while [29, 40, 22] make use of variance reduction or momentum to achieve $O(1/\epsilon^3)$ complexity. However, their analysis requires the smoothness condition and some of their algorithms need to use a large mini-batch size in the order $O(1/\epsilon^2)$. In contrast, we impose neither assumption about smoothness nor special structure of the objective function. The complexity of our algorithm is $\widetilde{O}(1/\epsilon^4)$ for finding a nearly ϵ-stationary point, which is the state of the art result for the considered non-smooth WCSC problem.

3 Preliminaries

This section provides some notations and assumptions used in the paper. We let $\|\cdot\|$ denote the Euclidean norm of a vector. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote the Fréchet subgradients and limiting Fréchet gradients by $\hat{\partial} f$ and ∂f , respectively, i.e., at x, $\hat{\partial} f(x) = \{v \in \mathbb{R}^d : \lim_{x \to x'} \inf \frac{f(x) - f(x') - v^\top (x - x')}{\|x - x'\|} \geq 0\}$, and $\partial f(x) = \{v \in \mathbb{R}^d : \exists x_k \xrightarrow{f} x, v_k \in \hat{\partial} f(x_k), v_k \to v, v \in \hat{\partial} f(x)\}$. Here $x_k \xrightarrow{f} x$ represents $x_k \to x$ with $f(x_k) \to f(x)$. A function $f(x)$ is μ-strongly convex on X if for any $x, x' \in X$, $\partial f(x')^\top (x - x') + \frac{\mu}{2}\|x - x'\|^2 \leq f(x) - f(x')$. A function $f(x)$ is ρ-weakly convex on X for any $x, x' \in X$ $\partial f(x')^\top (x - x') - \frac{\rho}{2}\|x - x'\|^2 \leq f(x) - f(x')$. Let $\mathcal{G}_x \in \partial_x f(x, y; \xi)$ denote a stochastic subgradient of f at x given y, where ξ is used to denote the random variable. Similarly, let $\mathcal{G}_y \in \partial_y f(x, y; \xi)$ denote a stochastic sugradient of f at y given x. Let $\Pi_\Omega[\cdot]$ denote the projection onto the set Ω, and let $\mathcal{B}(x, R)$ denote an Euclidean ball centered at x with a radius R. Denote by $dist(x, X)$ the distance between x and the set X, i.e., $dist(x, X) = \min_{v \in X} \|x - v\|$. Let $\tilde{O}(\cdot)$ hide some logarithmic factors.

For a WCSC min-max problem, it is generally a hard problem to find a saddle point. Hence, we use *nearly ϵ-stationarity* as the measure of convergence for solving WCSC problems [36], which is defined as follows.

Definition 1. *A solution x is a nearly ϵ-stationary point of $\min_x \psi(x)$ if there exist z and a constant $c > 0$ such that $\|z - x\| \leq c\epsilon$ and $dist(0, \partial\psi(z)) \leq \epsilon$.*

For a ρ-weakly convex function $\psi(x)$, let $z = \arg\min_{x \in \mathbb{R}^d} \psi(x) + \frac{\gamma}{2}\|x - \tilde{x}\|^2$ where $\gamma > \rho$ and $\tilde{x} \in \mathbb{R}^d$ is a reference point. Due to the strong convexity of the above problem, z is unique and $0 \in \partial\psi(z) + \gamma(z - \tilde{x})$, which results in $\gamma(\tilde{x} - z) \in \partial\psi(z)$, so that $dist(0, \partial\psi(z)) \leq \gamma\|\tilde{x} - z\|$. According to [8, 7, 9], we can find a nearly ϵ-stationary point $\tilde{x}$ as long as $\gamma\|\tilde{x} - z\| \leq \epsilon$.

Before ending this section, we present some assumptions that will be imposed in our analysis.

Assumption 1. *X and Y are closed convex sets. There exist initial solutions $x_0 \in X, y_0 \in Y$ and $\epsilon_0 > 0$ such that $Gap(x_0, y_0) \leq \epsilon_0$.*

Assumption 2. *(1) $f(x, y)$ is μ-strongly convex in x for any $y \in Y$ and λ-strongly concave in y for any $x \in X$. (2) There exist $B_1, B_2 > 0$ such that $\mathrm{E}[\exp(\frac{\|\mathcal{G}_x\|^2}{B_1^2})] \leq \exp(1)$ and $\mathrm{E}[\exp(\frac{\|\mathcal{G}_y\|^2}{B_2^2})] \leq \exp(1)$.*

Assumption 3. *(1) $f(x, y)$ is ρ-weakly convex in x for any $y \in Y$ and is λ-strongly concave in y for any $x \in X$. (2) $\mathrm{E}[\|\mathcal{G}_x\|^2] \leq M_1^2$ and $\mathrm{E}[\|\mathcal{G}_y\|^2] \leq M_2^2$.*

Remark: When $f(x, y)$ is smooth in x and y, the second condition in the above assumption can be replaced by the bounded variance condition.

Algorithm 1 Epoch-GDA for SCSC Min-Max Problems

1: Init.: $x_0^1 = x_0 \in X$, $y_0^1 = y_0 \in Y$, $\eta_x^1, \eta_y^1, R_1, T_1$
2: **for** $k = 1, 2, ..., K$ **do**
3: **for** $t = 0, 1, 2, ..., T_k - 1$ **do**
4: Compute stochastic gradients $\mathcal{G}_{x,t}^k \in \partial_x f(x_t^k, y_t^k; \xi_t^k)$ and $\mathcal{G}_{y,t}^k \in \partial_y f(x_t^k, y_t^k; \xi_t^k)$.
5: $x_{t+1}^k = \Pi_{X \cap \mathcal{B}(x_0^k, R_k)}(x_t^k - \eta_x^k \mathcal{G}_{x,t}^k)$
6: $y_{t+1}^k = \Pi_{Y \cap \mathcal{B}(y_0^k, R_k)}(y_t^k + \eta_y^k \mathcal{G}_{y,t}^k)$
7: **end for**
8: $x_0^{k+1} = \bar{x}_k = \frac{1}{T_k} \sum_{t=0}^{T_k - 1} x_t^k$, $y_0^{k+1} = \bar{y}_k = \frac{1}{T_k} \sum_{t=0}^{T_k - 1} y_t^k$
9: $\eta_x^{k+1} = \frac{\eta_x^k}{2}$, $\eta_y^{k+1} = \frac{\eta_y^k}{2}$, $R_{k+1} = R_k / \sqrt{2}$, $T_{k+1} = 2T_k$.
10: **end for**
11: Return $(\bar{x}_K, \bar{y}_K)$.

4 Main Results

4.1 Strongly-Convex Strongly-Concave Min-Max Problems

In this subsection, we present the main result for solving SCSC problems. The proposed Epoch-GDA algorithm for SCSC min-max problems is shown in Algorithm 1. As illustrated, our algorithm consists of a series of epochs. In each epoch (Line 3 to 7), standard primal-dual updates are performed. After an epoch ends, in Line 8, the solutions $\bar{x}_k$ and $\bar{y}_k$ averaged over the epoch are returned as the initialization for the next epoch. In Line 9, step sizes $\eta_{x,k+1}$ and $\eta_{y,k+1}$, the radius R_{k+1} and the number of iterations T_{k+1} are also adjusted for the next epoch. The ball constraints $\mathcal{B}(x_0^k, R_k)$ and $\mathcal{B}(y_0^k, R_k)$ at each iteration are used for the convergence analysis in high probability as in [16, 17]. It is clear that Epoch-GDA can be considered as a primal-dual variant of Epoch-GD [16, 17].

The following theorem shows that the iteration complexity of Algorithm 1 to achieve an ϵ-duality gap for a general SCSC problem (1) is $O(1/\epsilon)$.

Theorem 1. *Suppose Assumption 1 and Assumption 2 hold and let $\delta \in (0, 1)$ be a failing probability and $\epsilon \in (0, 1)$ be the target accuracy level for the duality gap. Let $K = \lceil \log(\frac{\epsilon_0}{\epsilon}) \rceil$ and $\tilde{\delta} = \delta/K$, and the initial parameters are set by $R_1 \geq 2\sqrt{\frac{2\epsilon_0}{\min\{\mu,\lambda\}}}$, $\eta_x^1 = \frac{\min\{\mu,\lambda\} R_1^2}{40(5+3\log(1/\tilde{\delta}))B_1^2}$, $\eta_y^1 = \frac{\min\{\mu,\lambda\} R_1^2}{40(5+3\log(1/\tilde{\delta}))B_2^2}$ and*

$$T_1 \geq \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(5+3\log(1/\tilde{\delta}))\max\{B_1^2, B_2^2\}\right\}}{\min\{\mu,\lambda\}^2 R_1^2}.$$

Then the total number of iterations of Algorithm 1 to achieve an ϵ-duality gap, i.e., Gap$(\bar{x}_K, \bar{y}_K) \leq \epsilon$, with probability $1 - \delta$ is

$$T_{tot} = \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(5+3\log(1/\tilde{\delta}))\max\{B_1^2, B_2^2\}\right\}}{4\min\{\mu,\lambda\}\epsilon}.$$

Remark 1. *To the best of our knowledge, this is the first study that achieves a fast rate of $O(1/T)$ for the duality gap of a general SCSC min-max problem without any special structure assumption or smoothness of the objective function and an additional computational cost. In contrast, even if the algorithm in [41] attains the $O(1/T)$ rate of convergence, it i) only guarantees the convergence of the primal objective gap, rather than the duality gap, ii) additionally requires a special structure of the objective function, and iii) needs an extra $O(n)$ computational cost of the deterministic update at each outer loop to handle the maximization over y. In contrast, Algorithm 1 has stronger theoretical results with less restrictions of the problem structures and computational cost.*

Remark 2. *A lower bound of $O(1/T)$ for stochastic strongly convex minimization problems has been proven in [1, 17]. Due to Gap$(x, y) \geq P(x) - P(x^*)$, bounding the duality gap is more difficult than bounding the primal gap. This means that our convergence rate matches the lower bound and is therefore the best possible convergence rate without adding more assumptions.*

4.2 Weakly-Convex Strongly-Concave Problems

Algorithm 2 Epoch-GDA for WCSC Min-Max Problems

1: Init.: $x_0^1 = x_0 \in X$, $y_0^1 = y_0 \in Y$, $\gamma = 2\rho$.
2: **for** $k = 1, 2, ..., K$ **do**
3: Set $T_k = \frac{106(k+1)}{3}$, $\eta_x^k = \frac{4}{\rho(k+1)}$, $\eta_y^k = \frac{2}{\lambda(k+1)}$.
4: **for** $t = 1, 2, ..., T_k$ **do**
5: Compute $\mathcal{G}_{x,t}^k \in \partial_x f(x_t^k, y_t^k; \xi_t^k)$ and $\mathcal{G}_{y,t}^k \in \partial_y f(x_t^k, y_t^k; \xi_t^k)$.
6: $x_{t+1}^k = \arg\min_{x \in X} x^\top \mathcal{G}_{x,t}^k + \frac{1}{2\eta_x^k}\|x - x_t^k\|^2 + \frac{\gamma}{2}\|x - x_0^k\|^2$
7: $y_{t+1}^k = \arg\min_{y \in Y} -y^\top \mathcal{G}_{y,t}^k + \frac{1}{2\eta_y^k}\|y - y_t^k\|^2$
8: **end for**
9: $x_0^{k+1} = \bar{x}_k = \frac{1}{T}\sum_{t=0}^{T-1} x_t^k$, $y_0^{k+1} = \bar{y}_k = \frac{1}{T}\sum_{n=0}^{T-1} y_t^k$
10: **end for**
11: Return x_0^τ by τ randomly sampled from $\{1, ..., K\}$.

In this subsection, we present the convergence results for solving WCSC problems, where the objective function $f(x, y)$ in (1) is ρ-weakly convex in x and λ-strongly concave in y. The proposed Epoch-GDA algorithm for WCSC min-max problems is summarized in Algorithm 2. As our Algorithm 1, Algorithm 2 consists of a number of epochs. As shown in Line 4 to Line 8, each epoch performs primal-dual updates on x and y. When updating x at the k-th stage, an additional regularizer $\frac{\gamma}{2}\|x - x_0^k\|^2$ is added, where the value $\gamma = 2\rho$. The added term is used to handle the weak convexity condition. After an epoch ends, average solutions of both x and y are restarted as the initial ones for the next epoch. The step sizes for updating x and y are set to $O(1/(\rho k))$ and $O(1/(\lambda k))$ at the k-th epoch, respectively. If we define $\hat{f}_k(x, y) = f(x, y) + \frac{\gamma}{2}\|x - x_0^k\|^2$, we can see that $\hat{f}_k(x, y)$ is ρ-strongly convex in x and λ-strongly concave in y, since $f(x, y)$ is ρ-weakly convex and $\gamma = 2\rho$. Indeed, for each inner loop of Algorithm 2, we actually work on the SCSC problem $\min_{x \in X} \max_{y \in Y} \hat{f}_k(x, y)$.

It is worth mentioning the key difference between our algorithm and the recently proposed stochastic algorithm PG-SMD [36] for WCSC problems with a special structural objective function. PG-SMD also consists of two loops. For each inner loop, it runs the same updates with the added regularizer on x as Algorithm 2. It restarts x by averaging the solutions over the inner loop, like our $\bar{x}_k$, but restarts y by taking the deterministic maximization of (1) over y given $\bar{x}_k$, leading to an additional $O(n)$ computational complexity per epoch. In addition, PG-SMD sets $\eta_y^k = O(1/(\gamma\lambda^2 k))$. Although Algorithm 2 shares similar updates to PG-SMD, our analysis yields stronger results under weaker assumptions — the same iteration complexity $\tilde{O}(1/\epsilon^4)$ without deterministic updates for y and special structure in the objective function. This is due to our sharper analysis that makes use of the telescoping sum based on the duality gap of $\hat{f}_k$ instead of the primal objective gap.

Let $\hat{P}(x) = P(x) + \mathbb{I}_X(x)$ where $\mathbb{I}_X(x)$ denotes the indicator function of the constraint set X at x. The convergence result of Algorithm 2 that achieves a nearly ϵ-stationary point with $\tilde{O}(1/\epsilon^4)$ iteration complexity is summarized below.

Theorem 2. *Suppose Assumption 3 holds. Algorithm 2 guarantees* $\mathrm{E}[dist(0, \partial\hat{P}(\hat{x}_\tau^*))^2] \leq \gamma^2 \mathrm{E}[\|\hat{x}_\tau^* - x_0^\tau\|^2] \leq \epsilon^2$ *after* $K = \max\left\{\frac{1696\gamma(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}\ln(\frac{1696\gamma(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}), \frac{1376\gamma\epsilon_0}{5\epsilon^2}\right\}$ *epochs, where* τ *is randomly sampled from* $\{1, ..., K\}$ *and* $(\hat{x}_k^*, \hat{y}_k^*)$ *is the saddle-point of* $\hat{f}_k(x, y)$. *The total number of iteration is* $\sum_{k=1}^K T_k = \tilde{O}(\frac{1}{\epsilon^4})$.

Remark 3. *Theorem 2 shows that the iteration complexity of Algorithm 2 to attain an ϵ-nearly stationary point is $\tilde{O}(1/\epsilon^4)$. It improves the result of [36] for WCSC problems in terms of two aspects. First, [36] requires a stronger condition on the structure of the objective function, while our analysis simply assumes a general objective function $f(x, y)$. Second, [36] requires to solve the maximization over y at each epoch, which may introduce an $O(n)$ computational complexity for $y \in \mathbb{R}^n$ [2]. In contrast, our algorithm restarts both the primal variable x and dual variable y at each epoch, which does not need an additional cost.*

[2] Although the exact maximization over y for restarting next epoch might be solved approximately, it still requires additional overhead.

Finally, we note that when $f(x, y)$ is smooth in x and y, we can use stochastic Mirror Prox algorithm [23] to replace the stochastic gradient descent ascent updates (Step 6 and Step 7) such that we can use a bounded variance assumption of the stochastic gradients instead of bounded second-order moments. It is a simple exercise to finish the proof by following our analysis of Theorem 2.

We prove the expectation result for WCSC in Theorem 2 for consistency with previous results [36]. In fact, we can also derive the high probability version. We provide a proof sketch at the end of the proof of Theorem 2 in the appendix and leave the details to the longer version.

5 Analysis

In this section, we present the proof of Theorem 1 and a proof sketch of Theorem 2. As we mentioned at the introduction, the key challenge in the analysis of Epoch-GDA lies in handling the variable distance measure $\|\hat{x}(y_1) - x_0\|^2 + \|\hat{y}(x_1) - y_0\|^2$ for any $(x_0, y_0) \in X \times Y$ and $(x_1, y_1) \in X \times Y$ and its connection to the duality gaps, where $\hat{x}(y_1) = \arg\min_{x' \in X} f(x', y_1)$ and $\hat{y}(x_1) = \arg\max_{y' \in Y} f(x_1, y')$. Hence, we first introduce the following key lemma that is useful in the analysis of Epoch-GDA for both SCSC and WCSC problems. It connects the variable distance measure $\|\hat{x}(y_1) - x_0\|^2 + \|\hat{y}(x_1) - y_0\|^2$ to the duality gaps at (x_0, y_0) and (x_1, y_1).

Lemma 1. *Consider the following μ-strongly convex in x and λ-strongly concave problem $\min_{x \in \Omega_1} \max_{y \in \Omega_2} f(x, y)$. Let (x^*, y^*) denote the saddle point solution to this problem. Suppose we have two solutions $(x_0, y_0) \in \Omega_1 \times \Omega_2$ and $(x_1, y_1) \in \Omega_1 \times \Omega_2$. Then the following relation between variable distance and duality gaps holds*

$$\begin{aligned}\frac{\mu}{4}\|\hat{x}(y_1) - x_0\|^2 + \frac{\lambda}{4}\|\hat{y}(x_1) - y_0\|^2 \leq & \max_{y' \in \Omega_2} f(x_0, y') - \min_{x' \in \Omega_1} f(x', y_0) \\ & + \max_{y' \in \Omega_2} f(x_1, y') - \min_{x' \in \Omega_1} f(x', y_1). \end{aligned} \tag{2}$$

5.1 Proof of Theorem 1 for the SCSC setting

The key idea is to first show the convergence of the duality gap with respect to the ball constraints $\mathcal{B}(x_0^k, R_k)$ and $\mathcal{B}(y_0^k, R_k)$ in an epoch (Lemma 2). Then we investigate the condition to make $\hat{x}(\bar{y}_k) \in \mathcal{B}(x_0^k, R_k)$ and $\hat{y}(\bar{x}_k) \in \mathcal{B}(y_0^k, R_k)$ given the average solution $(\bar{x}_k, \bar{y}_k)$, which allows us to derive the duality gap $\text{Gap}(\bar{x}_k, \bar{y}_k)$ for the original problem. Finally, under such conditions, we show how the duality gap between two consecutive outer loops can be halved (Theorem 3), which implies the total iteration complexity (Theorem 1). Below, we omit superscript k when it applies to all epochs.

Lemma 2. *Suppose Assumption 2 holds. Let Line 3 to 7 of Algorithm 1 run for T iterations (omitting the k-index) by fixed step sizes η_x and η_y. Then with the probability at least $1 - \tilde{\delta}$ where $0 < \tilde{\delta} < 1$, for any $x \in X \cap \mathcal{B}(x_0, R)$ and $y \in Y \cap \mathcal{B}(y_0, R)$, $\bar{x} = \sum_{t=0}^{T-1} x_t/T$, $\bar{y} = \sum_{t=0}^{T-1} y_t/T$ satisfy*

$$\begin{aligned} f(\bar{x}, y) - f(x, \bar{y}) \leq & \frac{\|x - x_0\|^2}{\eta_x T} + \frac{\|y - y_0\|^2}{\eta_y T} + \frac{\eta_x B_1^2 + \eta_y B_2^2}{2}(5 + 3\log(1/\tilde{\delta})) \\ & + \frac{4(B_1 + B_2)R\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{T}}. \end{aligned} \tag{3}$$

Remark 4. *Lemma 2 is a standard analysis for an epoch of Algorithm 1. The difficulty arises when attempting to plug x and y into (3). In order to derive the duality gap on the LHS of (3), we have to plug in $x \leftarrow \hat{x}(\bar{y})$ and $y \leftarrow \hat{y}(\bar{x})$. Nevertheless, it is unclear whether $\hat{x}(\bar{y}) \in \mathcal{B}(x_0, R)$ and $\hat{y}(\bar{x}) \in \mathcal{B}(y_0, R)$, which is the requirement for x and y to be plugged into (3). In the following lemma, we investigate the condition to make $\hat{x}(\bar{y}) \in \mathcal{B}(x_0, R)$ and $\hat{y}(\bar{x}) \in \mathcal{B}(y_0, R)$ based on Lemma 1.*

Lemma 3. *Suppose Assumption 2 holds. Let $\hat{x}_R(y) := \arg\min_{x \in X \cap \mathcal{B}(x_0, R)} f(x, y)$ and $\hat{y}_R(x) := \arg\max_{y \in Y \cap \mathcal{B}(y_0, R)} f(x, y)$. Assume the initial duality gap $\text{Gap}(x_0, y_0) \leq \epsilon_0$. Let Lines 3 to 7 of Algorithm 1 run T iterations with $\tilde{\delta} \in (0, 1)$, $R \geq 2\sqrt{\frac{2\epsilon_0}{\min\{\mu, \lambda\}}}$, $\eta_x = \frac{\min\{\mu, \lambda\} R^2}{40(5 + 3\log(1/\tilde{\delta}))B_1^2}$,*

$\eta_y = \frac{\min\{\mu,\lambda\}R^2}{40(5+3\log(1/\tilde{\delta}))B_2^2}$ *and*

$$T \geq \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(5+3\log(1/\tilde{\delta}))\max\{B_1^2, B_2^2\}\right\}}{\mu^2 R^2}.$$

Then, with probability at least $1-\tilde{\delta}$*, it holds* $\|\hat{x}_R(\bar{y}) - x_0\| < R, \|\hat{y}_R(\bar{x}) - y_0\| < R$.

Remark 5. *Lemma 3 shows that if we properly set the values of* R*,* η_x*,* η_y *and* T*, then* $\hat{x}_R(\bar{y})$ *and* $\hat{y}_R(\bar{x})$ *are the interior points of* $\mathcal{B}(x_0, R)$ *and* $\mathcal{B}(y_0, R)$ *with high probability. Therefore, we conclude that* $\hat{x}(\bar{y}) = \hat{x}_R(\bar{y})$ *and* $\hat{y}(\bar{x}) = \hat{y}_R(\bar{x})$ *with probability* $1-\tilde{\delta}$ *under the conditions of Lemma 3, which allows us to derive the duality gap in LHS of (3) of Lemma 2.*

We would highlight that $\hat{x}(\bar{y}) \in \mathcal{B}(x_0, R)$ and $\hat{y}(\bar{x}) \in \mathcal{B}(y_0, R)$ have to be confirmed in high probability, rather than in expectation. If we show $\mathrm{E}[\|\hat{x}(\bar{y}) - x_0\|] < R$, it is still unclear $\hat{x}(\bar{y}) \in \mathcal{B}(x_0, R)$, as pointed in [47]. The following theorem gives the relation of duality gaps between two consecutive epochs of Algorithm 1 by using Lemma 2 and the conditions proven by Lemma 3.

Theorem 3. *Consider the* k*-th epoch of Algorithm 1 with an initial solution* (x_0^k, y_0^k) *and the ending averaged solution* $(\bar{x}_k, \bar{y}_k)$*. Suppose Assumption 2 holds and Gap*$(x_0^k, y_0^k) \leq \epsilon_{k-1}$*. Let* $R_k \geq 2\sqrt{\frac{2\epsilon_{k-1}}{\min\{\mu,\lambda\}}}$ *(i.e.* $\epsilon_{k-1} \leq \frac{\min\{\mu,\lambda\}R_k^2}{8}$*),* $\eta_x^k = \frac{\min\{\mu,\lambda\}R_k^2}{40(5+3\log(1/\tilde{\delta}))B_1^2}$*,* $\eta_y^k = \frac{\min\{\mu,\lambda\}R_k^2}{40(5+3\log(1/\tilde{\delta}))B_2^2}$ *and*

$$T_k \geq \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(5+3\log(1/\tilde{\delta}))\max\{B_1^2, B_2^2\}\right\}}{\min\{\mu,\lambda\}^2 R_k^2}.$$

Then we have with probability $1-\tilde{\delta}$*, Gap*$(\bar{x}_k, \bar{y}_k) \leq \frac{\min\{\mu,\lambda\}R_k^2}{16}$.

Remark 6. *Theorem 3 shows that after running* T_k *iterations at the* k*-th stage, the upper bound of the duality gap would be halved with high probability, i.e., from* $\frac{\min\{\mu,\lambda\}R_k^2}{8}$ *to* $\frac{\min\{\mu,\lambda\}R_k^2}{16}$*. Then, in order to make the duality gap of each outer loop of Algorithm 1 halved from the last epoch, we can simply set* $R_{k+1}^2 = \frac{R_k^2}{2}$*, and accordingly,* $\eta_{x,k+1} = \frac{\eta_x^k}{2}$*,* $\eta_{y,k+1} = \frac{\eta_y^k}{2}$ *and* $T_{k+1} = 2T_k$*.*

Proof. (of Theorem 3) For any $x \in \mathcal{B}(x_0^k, R_k)$ and $y \in \mathcal{B}(y_0^k, R_k)$, we have $\|x - x_0^k\| \leq R$ and $\|y - y_0^k\| \leq R$, so by (3) of Lemma 2, we have with probability $1-\tilde{\delta}$

$$f(\bar{x}_k, y) - f(x, \bar{y}_k) \overset{(a)}{\leq} \frac{R_k^2}{\eta_x^k T_k} + \frac{R_k^2}{\eta_y^k T_k} + \frac{\eta_x^k B_1^2}{2}(5+3\log(1/\tilde{\delta})) + \frac{\eta_y^k B_2^2}{2}(5+3\log(1/\tilde{\delta}))$$
$$+ \frac{4(B_1+B_2)R_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{T_k}} \overset{(b)}{\leq} \frac{\min\{\mu,\lambda\}R_k^2}{16}, \tag{4}$$

where inequality (a) is due to $x \in \mathcal{B}(x_0^k, R_k)$ and $y \in \mathcal{B}(y_0^k, R_k)$. Inequality (b) is due to the values of η_x^k, η_y^k and T_k. Recall the definitions $\hat{x}(\bar{y}_k) = \arg\min_{x\in X} f(x, \bar{y}_k)$ and $\hat{y}(\bar{x}_k) = \arg\max_{y\in Y} f(\bar{x}_k, y)$. By Lemma 3, we have $\hat{x}(\bar{y}_k) \in \mathcal{B}(x_0^k, R_k)$ and $\hat{y}(\bar{x}_k) \in \mathcal{B}(y_0^k, R_k)$ with probability $1-\tilde{\delta}$. Then from (4) we have

$$\mathrm{Gap}(\bar{x}_k, \bar{y}_k) = \max_{y\in Y} f(\bar{x}_k, y) - \min_{x\in X} f(x, \bar{y}_k) \leq \frac{\min\{\mu,\lambda\}R_k^2}{16}.$$

□

Given the condition $\mathrm{Gap}(x_0^k, y_0^k) \leq \epsilon_{k-1} \leq \frac{\min\{\mu,\lambda\}R_k^2}{8}$, we then conclude that running T_k iterations in an epoch of Algorithm 1 would halve the duality gap with high probability. As indicated in Theorem 3, the duality gap $\mathrm{Gap}(\bar{x}_k, \bar{y}_k)$ can be halved as long as the condition of Theorem 3 holds. Then Theorem 1 is implied (the detailed proof is in Supplementary Materials).

5.2 Proof Sketch of Theorem 2 for the WCSC setting

Due to limit of space, we only present a sketch here and present the full proof in the Supplement. Recall $\hat{f}_k(x,y) = f(x,y) + \frac{\gamma}{2}\|x - x_0^k\|^2$. Let us denote its duality gap by $\widehat{\mathrm{Gap}}_k(x,y) = \hat{f}_k(x, \hat{y}_k(x)) - \hat{f}_k(\hat{x}_k(y), y)$, where we define $\hat{y}_k(x) := \arg\max_{y'\in Y} \hat{f}_k(x, y')$ given $x \in X$ and

$\hat{x}_k(y) := \arg\min_{x' \in X} \hat{f}_k(x', y)$ given $y \in Y$. Its saddle point solution is denoted by $(\hat{x}_k^*, \hat{y}_k^*)$, i.e., $\hat{f}_k(\hat{x}_k^*, y) \leq \hat{f}_k(\hat{x}_k^*, \hat{y}_k^*) \leq \hat{f}_k(x, \hat{y}_k^*)$ for any $x \in X$ and $y \in Y$. The key idea of our analysis is to connect the duality gap $\widehat{\text{Gap}}_k(x_0^k, y_0^k)$ to $\gamma^2\|\hat{x}_k^* - x_0^k\|^2$, and then by making $\gamma^2\|\hat{x}_k^* - x_0^k\|^2 \leq \epsilon^2$, we can show that x_0^k is a nearly ϵ-stationary point. To this end we first establish a bound of the duality gap for the regularized problem $\hat{f}_k(x, y)$ for the k-th epoch (Lemma 4). Then we connect it to $\gamma\|\hat{x}_k^* - x_0^k\|^2$ (Lemma 5). Finally, we bound $\gamma\|\hat{x}_k^* - x_0^k\|^2$ by a telescoping sum of $\text{E}[\widehat{\text{Gap}}_k(x_0^k, y_0^k)] - \text{E}[\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})]$ and $\text{E}[P(x_0^k) - P(x_0^{k+1})]$.

6 Conclusions

In this paper, we filled the gap between stochastic min-max and minimization optimization problems. We proposed Epoch-GDA algorithms for general SCSC and general WCSC problems, which do not impose any additional assumptions on the smoothness or the structure of the objective function. Our key lemma provides sharp analysis of Epoch-GDA for both problems. For SCSC min-max problems, to the best of our knowledge, our result is the first one to show that Epoch-GDA achieves the optimal rate of $O(1/T)$ for the duality gap of general SCSC min-max problems. For WCSC min-max problems, our analysis allows us to derive the best complexity $\tilde{O}(1/\epsilon^4)$ of Epoch-GDA to reach a nearly ϵ-stationary point, which does not require smoothness, large mini-batch sizes or other structural conditions.

Broader Impact

A discussion about broader impact is not applicable since our work is very theoretical and currently has no particular application.

Acknowledgments and Disclosure of Funding

T. Yang is partially supported by National Science Foundation Career Award (NSF 1844403) and NSF grant #1933212. Most work of Y. Yan was done when he worked in the University of Iowa.

References

[1] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, International Convention Centre, Sydney, Australia, 2017.

[4] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, May 2011.

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

[6] Cong Dang and Guanghui Lan. Randomized first-order methods for saddle point optimization. *arXiv preprint arXiv:1409.8625*, 2014.

[7] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *CoRR*, abs/1803.06523, 2018.

[8] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $\text{O}(k^{-1/4})$ on weakly convex functions. *CoRR*, /abs/1802.02988, 2018.

[9] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

[10] Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *arXiv preprint arXiv:1802.01504*, 2018.

[11] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. In *Advances in Neural Information Processing Systems 30*, pages 497–505. 2017.

[12] Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. Frank-wolfe algorithms for saddle point problems. *arXiv preprint arXiv:1610.07797*, 2016.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[14] Davood Hajinezhad and Mingyi Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, 176(1-2):207–245, 2019.

[15] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

[16] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 421–436, 2011.

[17] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[18] Le Thi Khanh Hien, Renbo Zhao, and William B Haskell. An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. *arXiv preprint arXiv:1711.03669*, 2017.

[19] Mingyi Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv preprint arXiv:1604.00543*, 2016.

[20] Mingyi Hong, Jason D Lee, and Meisam Razaviyayn. Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization. *arXiv preprint arXiv:1802.08941*, 2018.

[21] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*, 2019.

[22] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020.

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

[24] Guanghui Lan, Arkadi Nemirovski, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

[25] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *CoRR*, abs/1906.00331, 2019.

[26] Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

[27] Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with $O(1/n)$-convergence rate. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[28] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *arXiv preprint arXiv:1902.08294*, 2019.

[29] Luo Luo, Haishan Ye, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *arXiv preprint arXiv:2001.03724*, 2020.

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

[31] Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2975–2984, 2017.

[32] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

[33] Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16:235–249, 01 2005.

[34] 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*, pages 14905–14916, 2019.

[35] Balamurugan Palaniappan and Francis R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1408–1416, 2016.

[36] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, abs/1810.02060, 2018.

[37] Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. *Journal of Machine Learning Research (JMLR)*, 2013.

[38] Conghui Tan, Tong Zhang, Shiqian Ma, and Ji Liu. Stochastic primal-dual method for empirical risk minimization with o (1) per-iteration complexity. In *Advances in Neural Information Processing Systems*, pages 8366–8375, 2018.

[39] Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3694–3702. JMLR. org, 2017.

[40] Tengyu Xu, Zhe Wang, Yingbin Liang, and H Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *arXiv preprint arXiv:2006.09361*, 2020.

[41] Yan Yan, Yi Xu, Qihang Lin, Lijun Zhang, and Tianbao Yang. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

[42] Tianbao Yang, Mehrdad Mahdavi, Rong Jin, and Shenghuo Zhu. An efficient primal dual prox method for non-smooth optimization. *Machine Learning*, 98(3):369–406, 2015.

[43] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in neural information processing systems*, pages 451–459, 2016.

[44] Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. Self-tuned stochastic approximation schemes for non-lipschitzian stochastic multi-user optimization and nash games. *IEEE Transactions on Automatic Control*, 61(7):1753–1766, 2015.

[45] Adams Wei Yu, Qihang Lin, and Tianbao Yang. Doubly stochastic primal-dual coordinate method for regularized empirical risk minimization with factorized data. *CoRR*, abs/1508.03390, 2015.

[46] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

[47] Renbo Zhao. Optimal algorithms for stochastic three-composite convex-concave saddle point problems. *arXiv preprint arXiv:1903.01687*, 2019.

A Proof of Theorem 2 for the WCSC setting

For the proof of Theorem 2, we need to merge the constraint set X into the objective function, so that we can derive the convergence of the near ϵ-stationary point. Recall originally $\hat{f}_k(x,y) = f(x,y) + \frac{\gamma}{2}\|x - x_0^k\|^2$. Now we re-define $\hat{f}_k(x,y) = f(x,y) + \frac{\gamma}{2}\|x - x_0^k\|^2 + \mathbb{I}_X(x)$, where $I_X(x)$ is the indicator function of the constraint set X at x. Similarly, we re-define $P(x) = \max_{y\in Y} f(x,y) + \mathbb{I}_X(x)$ by merging the constraint set X into the objective function. In the following, we can derive the convergence of $dist(0, \partial P(x))$.

Let us denote its duality gap by $\widehat{\text{Gap}}_k(x,y) = \hat{f}_k(x, \hat{y}_k(x)) - \hat{f}_k(\hat{x}_k(y), y)$, where we define $\hat{y}_k(x) := \arg\max_{y'\in Y} \hat{f}_k(x,y')$ given $x \in X$ and $\hat{x}_k(y) := \arg\min_{x'\in X} \hat{f}_k(x',y)$ given $y \in Y$. Its saddle point solution is denoted by $(\hat{x}_k^*, \hat{y}_k^*)$, i.e., $\hat{f}_k(\hat{x}_k^*, y) \le \hat{f}_k(\hat{x}_k^*, \hat{y}_k^*) \le \hat{f}_k(x, \hat{y}_k^*)$ for any $x \in X$ and $y \in Y$. The key idea of our analysis is to connect the duality gap $\widehat{\text{Gap}}_k(x_0^k, y_0^k)$ to $\gamma^2\|\hat{x}_k^* - x_0^k\|^2$, and then by making $\gamma^2\|\hat{x}_k^* - x_0^k\|^2 \le \epsilon^2$, we can show that x_0^k is a nearly ϵ-stationary point. To this end we first establish a bound of the duality gap for the regularized problem $\hat{f}_k(x,y)$ for the k-th epoch (Lemma 4). Then we connect it to $\gamma\|\hat{x}_k^* - x_0^k\|^2$ (Lemma 5). Finally, we bound $\gamma\|\hat{x}_k^* - x_0^k\|^2$ by a telescoping sum of $\text{E}[\widehat{\text{Gap}}_k(x_0^k, y_0^k)] - \text{E}[\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})]$ and $\text{E}[P(x_0^k) - P(x_0^{k+1})]$.

Lemma 4. *Suppose Assumption 3 holds and $\gamma = 2\rho$. For $k \ge 1$, Lines 4 to 8 of Algorithm 2 guarantee*

$$\text{E}[\widehat{Gap}_k(\bar{x}_k, \bar{y}_k)] = \text{E}[\max_{y\in Y} \hat{f}_k(\bar{x}_k, y) - \min_{x\in X} \hat{f}_k(x, \bar{y}_k)] = \text{E}[\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_k(\bar{y}_k), \bar{y}_k)]$$
$$\le \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{1}{T_k}\Big\{(\frac{1}{\eta_x^k} + \frac{\rho}{2})\text{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{1}{\eta_y^k}\text{E}[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\Big\}. \quad (5)$$

For RHS of (5), particularly, due to $k \ge 1$, $T_k = \frac{106(k+1)}{3}$, $\eta_x^k = \frac{4}{\rho(k+1)}$ and $\eta_y^k = \frac{2}{\lambda(k+1)}$ in Algorithm 2, we have $\frac{1}{T_k}(\frac{1}{\eta_x^k} + \frac{\rho}{2}) \le \frac{3\rho}{212}$ and $\frac{1}{T_k\eta_y^k} = \frac{3\lambda}{212}$. Then for the last two terms in the RHS of (5), we could have the following upper bound by the key lemma (Lemma 1)

$$\frac{3}{53}\Big(\frac{\rho}{4}\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2 + \frac{\lambda}{4}\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2\Big) \le \frac{3}{53}\Big(\widehat{\text{Gap}}_k(x_0^k, y_0^k) + \widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k)\Big). \quad (6)$$

On the other hand, the following lemma lower bounds LHS of (5) to construct telescoping sums.

Lemma 5. *We could derive the following lower bound for $\widehat{Gap}_k(\bar{x}_k, \bar{y}_k)$*

$$\widehat{Gap}_k(\bar{x}_k, \bar{y}_k) \ge \frac{3}{50}\widehat{Gap}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) + \frac{\gamma}{80}\|x_0^k - \hat{x}_k^*\|^2. \quad (7)$$

Lemma 5 lower bounds $\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k)$ in LHS of (5) by three parts. The first part constructs telescoping sum of $\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \widehat{\text{Gap}}_k(x_0^k, y_0^k)$ together with (6). The second part itself is an element of telescoping sums over the primal gap. The third part $\|x_0^k - \hat{x}_k^*\|^2$ can be used as the measure of nearly ϵ-stationary point, which is further explored in Theorem 2.

Proof. (of Theorem 2) Consider the k-th stage. Let us start from (5) in Lemma 4 as follows

$$\begin{aligned}
&\text{E}[\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k)] \\
\le& \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{1}{T_k}\Big\{(\frac{1}{\eta_x} + \frac{\rho}{2})\text{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{1}{\eta_y}\text{E}[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\Big\} \\
\overset{(a)}{\le}& \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{3}{53}\Big(\frac{\rho}{4}\text{E}[\|\hat{x}_k(\bar{y}_k) - x_0^k\|^2] + \frac{\lambda}{4}\text{E}[\|\hat{y}_k(\bar{x}_k) - y_0^k\|^2]\Big) \\
\overset{(6)}{\le}& \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{3}{53}\text{E}[\widehat{\text{Gap}}_k(x_0^k, y_0^k)] + \frac{3}{53}\text{E}[\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k)],
\end{aligned}$$

where (a) is due to settings $T_k = \frac{106(k+1)}{3}$, $\eta_x^k = \frac{4}{\rho(k+1)}$, and $\eta_y^k = \frac{2}{\lambda(k+1)}$. Re-organizing the above inequality, we have

$$\frac{50}{53}\mathrm{E}[\widehat{\mathrm{Gap}}_k(\bar{x}_k, \bar{y}_k)] \leq \frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{3}{53}\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^k, y_0^k)]. \tag{8}$$

Then for the LHS of (8), we apply (7) of Lemma 5 as follows

$$\begin{aligned}
&\frac{50}{53}\Big(\frac{3}{50}\widehat{\mathrm{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) + \frac{\gamma}{80}\|x_0^k - \hat{x}_k^*\|^2\Big) \\
\leq &\frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{3}{53}\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^k, y_0^k)].
\end{aligned} \tag{9}$$

Next we have

$$\begin{aligned}
\frac{5\gamma}{424}\mathrm{E}[\|x_0^k - \hat{x}_k^*\|^2] \leq &\frac{5\eta_x^k M_1^2}{2} + \frac{5\eta_y^k M_2^2}{2} + \frac{40}{53}\mathrm{E}[P(x_0^k) - P(x_0^{k+1})] \\
&+ \frac{3}{53}\Big(\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^k, y_0^k)] - \mathrm{E}[\widehat{\mathrm{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})]\Big)
\end{aligned} \tag{10}$$

Summing from $k = 1$ to $k = K$, we have

$$\begin{aligned}
\frac{5\gamma}{424}\sum_{k=1}^K \mathrm{E}[\|x_0^k - \hat{x}_k^*\|^2] \leq &\underbrace{\sum_{k=1}^K \frac{5\eta_x^k M_1^2}{2} + \sum_{k=1}^K \frac{5\eta_y^k M_2^2}{2}}_{:=A} + \underbrace{\frac{40}{53}\sum_{k=1}^K \mathrm{E}[P(x_0^k) - P(x_0^{k+1})]}_{:=B} \\
&+ \frac{3}{53}\underbrace{\sum_{k=1}^K \Big(\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^k, y_0^k)] - \mathrm{E}[\widehat{\mathrm{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})]\Big)}_{:=C} \qquad (11) \\
\leq &5\Big(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda}\Big)\ln(K+1) + \frac{43}{53}\mathrm{E}[\mathrm{Gap}(x_0^1, y_0^1)], \qquad (12)
\end{aligned}$$

where the last inequality is due to the upper bounds the three terms A, B and C as follows.

For the term A, we have

$$\begin{aligned}
A = &\sum_{k=1}^K \frac{5\eta_x^k M_1^2}{2} + \sum_{k=1}^K \frac{5\eta_y^k M_2^2}{2} = \frac{10M_1^2}{\rho}\sum_{k=1}^K \frac{1}{k+1} + \frac{5M_2^2}{\lambda}\sum_{k=1}^K \frac{1}{k+1} \\
\leq &5\Big(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda}\Big)\ln(K+1),
\end{aligned}$$

where the second equality is due to the setting of $\eta_x^k = \frac{4}{\rho(k+1)}$ and $\eta_y^k = \frac{2}{\lambda(k+1)}$. The last inequality is due to $\sum_{k=1}^{K+1} \frac{1}{k} \leq \ln(K+1) + 1$.

For the term B, we have

$$\begin{aligned}
B = &\sum_{k=1}^K \mathrm{E}[P(x_0^k) - P(x_0^{k+1})] = \mathrm{E}[P(x_0^1) - P(x_0^{K+1})] = \mathrm{E}[f(x_0^1, \hat{y}(x_0^1)) - f(x_0^{K+1}, \hat{y}(x_0^{K+1}))] \\
\leq &\mathrm{E}[f(x_0^1, \hat{y}(x_0^1)) - f(x_0^{K+1}, y_0^1)] \leq \mathrm{E}[f(x_0^1, \hat{y}(x_0^1)) - f(\hat{x}(y_0^1), y_0^1)] = \mathrm{E}[\mathrm{Gap}(x_0^1, y_0^1)],
\end{aligned}$$

where the two inequalities are due to $f(x_0^{K+1}, \hat{y}(x_0^{K+1})) \geq f(x_0^{K+1}, y_0^1) \geq f(\hat{x}(y_0^1), y_0^1)$.

For the term C, we have

$$\begin{aligned}
C = &\sum_{k=1}^K \Big(\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^k, y_0^k) - \widehat{\mathrm{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})]\Big) \\
= &\mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^1, y_0^1) - \widehat{\mathrm{Gap}}_{K+1}(x_0^{K+1}, y_0^{K+1})] \leq \mathrm{E}[\widehat{\mathrm{Gap}}_k(x_0^1, y_0^1)] \\
= &\mathrm{E}[f(x_0^1, \hat{y}(x_0^1)) + \frac{\gamma}{2}\|x_0^1 - x_0^1\|^2 - f(\hat{x}_1(y_0^1), y_0^1) - \frac{\gamma}{2}\|\hat{x}_1(y_0^1) - x_0^1\|^2] \\
\leq &\mathrm{E}[f(x_0^1, \hat{y}(x_0^1)) - f(\hat{x}(y_0^1), y_0^1)] = \mathrm{E}[\mathrm{Gap}(x_0^1, y_0^1)],
\end{aligned}$$

where the first inequality is due to $\widehat{\text{Gap}}_{K+1}(x_0^{K+1}, y_0^{K+1}) \geq 0$. By plugging the above upper bounds of the three terms A, B and C into (11), we have (12).

Then by randomly sampling τ from $\{1, ..., K\}$, we have

$$\mathrm{E}[\|x_0^\tau - \hat{x}_\tau^*\|^2] \leq \frac{424}{\gamma K}\Big(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda}\Big)\ln(K+1) + \frac{344}{5\gamma K}\mathrm{E}[\text{Gap}(x_0^1, y_0^1)].$$

Since $\mathrm{E}[\text{Dist}(0, \partial P(\hat{x}_\tau^*))^2] \leq \gamma^2 \mathrm{E}[\|x_\tau^* - x_0^\tau\|^2]$ and $\gamma = 2\rho$, we could set

$$K = \max\left\{\frac{1696\rho(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}\ln(\frac{1696\rho(\frac{2M_1^2}{\rho} + \frac{M_2^2}{\lambda})}{\epsilon^2}), \frac{1376\rho\text{Gap}(x_0^1, y_0^1)}{5\epsilon^2}\right\},$$

which leads to $\gamma^2\mathrm{E}[\|x_\tau^* - x_0^\tau\|^2] \leq \epsilon^2$. Recall $T_k = \frac{106(k+1)}{3}$. To compute the total number of iterations, we have

$$T_{tot} = \sum_{k=1}^{K} T_k = \frac{106}{3}\sum_{k=1}^{K}(k+1) = O(K^2) = \tilde{O}\left(\frac{1}{\epsilon^4}\right).$$

We would highlight that we prove the expectation result for WCSC in Theorem 2 for consistency with previous results [36]. Theorem 2 can also be extended to the high probability statement, as Theorem 1. In particular, we can prove a high-probability result of Lemma 4 similar to Lemma 2. Then by appropriately setting the radius R_k according to η_k and T_k we can prove a similar result as in Lemma 3, which leads to a high-probability upper bound for the duality gap of $\hat{f}_k(x, y)$. From this point, we can prove the high-prob convergence for the WCSC similar to the existing proof of Theorem 2 except replacing expectation result with high-probability result. We will leave the detailed proof to the longer version. □

B Proof of Lemma 1

Proof. Let us first consider the first term in LHS of (2) as follows,

$$\begin{aligned}
&\frac{\mu}{4}\|\hat{x}(y_1) - x_0\|^2 \\
\leq &\frac{\mu}{2}\|\hat{x}(y_1) - x^*\|^2 + \frac{\mu}{2}\|x^* - x_0^k\|^2 \\
\overset{(a)}{\leq} &f(x^*, y_1) - f(\hat{x}(y_1), y_1) + f(x_0, y^*) - f(x^*, y^*) \\
\overset{(b)}{\leq} &f(x^*, y^*) - f(\hat{x}(y_1), y_1) + f(x_0, y^*) - f(x^*, y^*) \\
\overset{(c)}{\leq} &f(x_0, \hat{y}(x_0)) - f(\hat{x}(y_1), y_1),
\end{aligned} \tag{13}$$

where inequality (a) is due to μ-strong convexity of $f(x, y_1)$ in x with fixed y_1 (with optimality at $\hat{x}(y_1)$) and μ-strong convexity of $f(x, y^*)$ in x with fixed y^* (with optimality at x^*). Inequality (b) is due to $f(x^*, y_1) \leq f(x^*, y^*)$. Inequality (c) is due to $f(x_0, y^*) \leq f(x_0, \hat{y}(x_0))$.

In a similar way, for the second term, we have

$$\begin{aligned}
&\frac{\lambda}{4}\|\hat{y}(x_1) - y_0\|^2 \\
\leq &\frac{\lambda}{2}\|\hat{y}(x_1) - y^*\|^2 + \frac{\lambda}{2}\|y^* - y_0\|^2 \\
\overset{(a)}{\leq} &f(x_1, \hat{y}(x_1)) - f(x_1, y^*) + f(x^*, y^*) - f(x^*, y_0) \\
\overset{(b)}{\leq} &f(x_1, \hat{y}(x_1)) - f(x^*, y^*) + f(x^*, y^*) - f(x^*, y_0) \\
\overset{(c)}{\leq} &f(x_1, \hat{y}(x_1)) - f(\hat{x}(y_0), y_0),
\end{aligned} \tag{14}$$

where inequality (a) is due to λ-strong concavity of $f(x_1, y)$ in y with fixed x_1 (optimality at $\hat{y}(x_1)$) and $f(x^*, y)$ in y with fixed x^* (optimality at $\hat{y}^*$). Inequality (b) is due to $f(x_1, y^*) \geq f(x^*, y^*)$. Inequality (c) is due to $f(x^*, y_0) \geq f(\hat{x}(y_0), y_0)$.

Then, combining inequalities (13) and (14), we have

$$\begin{aligned}
&\frac{\mu}{4}\|\hat{x}(y_1)-x_0\|^2+\frac{\lambda}{4}\|\hat{y}(x_1)-y_0\|^2\\
\leq& f(x_0,\hat{y}(x_0))-f(\hat{x}(y_1),y_1)+f(x_1,\hat{y}(x_1))-f(\hat{x}(y_0),y_0)\\
=&\Big(\max_{y'\in\Omega_2} f(x_0,y')-\min_{x'\in\Omega_1} f(x',y_0)\Big)+\Big(\max_{y'\in\Omega_2} f(x_1,y')-\min_{x'\in\Omega_1} f(x',y_1)\Big).
\end{aligned}$$

□

C Proof of Lemma 2

Proof. Before the proof, we first present the following two lemmas as follows.

Lemma 6. *Let $X_1, X_2, ..., X_T$ be independent random variables and $\mathrm{E}_t[\exp(\frac{X_t^2}{B^2})] \leq \exp(1)$ for any $t \in \{1, ..., T\}$. Then we have with probability at least $1-\tilde{\delta}$*

$$\sum_{t=1}^{T} X_t \leq B^2(T+\log(1/\tilde{\delta})).$$

Lemma 7. *(Lemma 2 of [24]) Let $X_1, ..., X_T$ be a martingale difference sequence, i.e., $E_t[X_t]=0$ for all t. Suppose that for some values σ_t, for $t=1,2,...,T$, we have $E_t[\exp(\frac{X_t^2}{\sigma_t^2})] \leq \exp(1)$. Then with probability at least $1-\delta$, we have*

$$\sum_{t=1}^{T} X_t \leq \sqrt{3\log(1/\delta)\sum_{t=1}^{T}\sigma_t^2}.$$

For simplicity of presentation, we use the notations $\Delta_x^t = \partial_x f(x_t, y_t; \xi_t)$, $\Delta_y^t = \partial y f(x_t, y_t, ; \xi_t)$, $\partial_x^t = \partial_x f(x_t, y_t)$ and $\partial_y^t = \partial_y f(x_t, y_t)$. To prove Lemma 2, we would leverage the following two update approaches:

$$\begin{aligned}
&\begin{cases} x_{t+1} = \arg\min_{x\in X\cap\mathcal{B}(x_0,R)} & x^\top\Delta_x^t+\frac{1}{2\eta_x}||x-x_t||^2 \\ y_{t+1} = \arg\min_{y\in Y\cap\mathcal{B}(y_0,R)} & -y^\top\Delta_y^t+\frac{1}{2\eta_y}||y-y_t||^2 \end{cases}\\
&\begin{cases} \tilde{x}_{t+1} = \arg\min_{x\in X\cap\mathcal{B}(x_0,R)} & x^\top(\partial_x^t-\Delta_x^t)+\frac{1}{2\eta_x}||x-\tilde{x}_t||^2 \\ \tilde{y}_{t+1} = \arg\min_{y\in Y\cap\mathcal{B}(y_0,R)} & -y^\top(\partial_y^t-\Delta_y^t)+\frac{1}{2\eta_y}||y-\tilde{y}_t||^2, \end{cases}
\end{aligned} \tag{15}$$

where $x_0=\tilde{x}_0$ and $y_0=\tilde{y}_0$. The first two updates are identical to Line 4 and Line 5 in Algorithm 1. This can be verified easily. Take the first one as example:

$$\begin{aligned}
x_{t+1} =&\Pi_X(x_t-\eta_x\Delta_x^t) = \arg\min_{x\in X\cap\mathcal{B}(x_0,R)}||x-(x_t-\eta_x\Delta_x^t)||^2\\
=&\arg\min_{x\in X\cap\mathcal{B}(x_0,R)}\frac{1}{2\eta_x}||x-x_t||^2+x^\top\Delta_x^t.
\end{aligned}$$

Let $\psi(x)=x^\top u+\frac{1}{2\gamma}||x-v||^2$ with $x'=\arg\min_{x\in X'}\psi(x)$, which includes the four update approaches in (15) as special cases. By using the strong convexity of $\psi(x)$ and the first order optimality condition ($\partial\psi(x')^\top(x-x')\geq 0$), for any $x\in X'$, we have

$$\psi(x)-\psi(x') \geq \partial\psi(x')^T(x-x')+\frac{1}{2\gamma}||x-x'||^2 \geq \frac{1}{2\gamma}||x-x'||^2,$$

which implies

$$
\begin{aligned}
0 \leq & (x-x')^\top u + \frac{1}{2\gamma}||x-v||^2 - \frac{1}{2\gamma}||x'-v||^2 - \frac{1}{2\gamma}||x-x'||^2 \\
= & (v-x')^\top u - (v-x)^\top u + \frac{1}{2\gamma}||x-v||^2 - \frac{1}{2\gamma}||x'-v||^2 - \frac{1}{2\gamma}||x-x'||^2 \\
= & -\frac{1}{2\gamma}||x'-v||^2 + (v-x')^\top u + \frac{1}{2\gamma}||x-v||^2 - \frac{1}{2\gamma}||x-x'||^2 - (v-x)^\top u \\
\leq & \frac{\gamma}{2}||u||^2 + \frac{1}{2\gamma}||x-v||^2 - \frac{1}{2\gamma}||x-x'||^2 - (v-x)^\top u.
\end{aligned}
$$

Then

$$
(v-x)^\top u \leq \frac{\gamma}{2}||u||^2 + \frac{1}{2\gamma}||x-v||^2 - \frac{1}{2\gamma}||x-x'||^2. \tag{16}
$$

Applying the above result to the updates in (15), we have for any $x \in X \cap \mathcal{B}(x_0, R)$ and $y \in Y \cap \mathcal{B}(y_0, R)$,

$$
\begin{aligned}
& (x_t - x)^\top \Delta_x^t \leq \frac{1}{2\eta_x}||x-x_t||^2 - \frac{1}{2\eta_x}||x-x_{t+1}||^2 + \frac{\eta_x}{2}||\Delta_x^t||^2 \\
& (y - y_t)^\top \Delta_y^t \leq \frac{1}{2\eta_y}||y-y_t||^2 - \frac{1}{2\eta_y}||y-y_{t+1}||^2 + \frac{\eta_y}{2}||\Delta_y^t||^2 \\
& (\tilde{x}_t - x)^\top (\partial_x^t - \Delta_x^t) \leq \frac{1}{2\eta_x}||x-\tilde{x}_t||^2 - \frac{1}{2\eta_x}||x-\tilde{x}_{t+1}||^2 + \frac{\eta_x}{2}||\partial_x^t - \Delta_x^t||^2 \\
& (y - \tilde{y}_t)^\top (\partial_y^t - \Delta_y^t) \leq \frac{1}{2\eta_y}||y-\tilde{y}_t||^2 - \frac{1}{2\eta_y}||y-\tilde{y}_{t+1}||^2 + \frac{\eta_y}{2}||\partial_y^t - \Delta_y^t||^2.
\end{aligned} \tag{17}
$$

Adding the above four inequalities together, we have

$$
\begin{aligned}
\text{LHS} = & (x_t - x)^\top \Delta_x^t + (y - y_t)^\top \Delta_y^t + (\tilde{x}_t - x)^\top (\partial_x^t - \Delta_x^t) + (y - \tilde{y}_t)^\top (\partial_y^t - \Delta_y^t) \\
= & (x_t - x)^\top \partial_x^t + (x_t - x)^\top (\Delta_x^t - \partial_x^t) + (y - y_t)^\top \partial_y^t + (y - y_t)^\top (\Delta_y^t - \partial_y^t) \\
& + (\tilde{x}_t - x)^\top (\partial_x^t - \Delta_x^t) + (y - \tilde{y}_t)^\top (\partial_y^t - \Delta_y^t) \\
= & -(x - x_t)^\top \partial_x^t + (y - y_t)^\top \partial_y^t - (x_t - \tilde{x}_t)^\top (\partial_x^t - \Delta_x^t) - (\tilde{y}_t - y_t)^\top (\partial_y^t - \Delta_y^t) \\
\overset{(a)}{\geq} & -(f(x, y_t) - f(x_t, y_t)) + (f(x_t, y) - f(x_t, y_t)) - (x_t - \tilde{x}_t)^\top (\partial_x^t - \Delta_x^t) - (\tilde{y}_t - y_t)^\top (\partial_y^t - \Delta_y^t) \\
= & f(x_t, y) - f(x, y_t) - (x_t - \tilde{x}_t)^\top (\partial_x^t - \Delta_x^t) - (\tilde{y}_t - y_t)^\top (\partial_y^t - \Delta_y^t) \\
\text{RHS} = & \frac{1}{2\eta_x}\Big\{||x-x_t||^2 - ||x-x_{t+1}||^2 + ||x-\tilde{x}_t||^2 - ||x-\tilde{x}_{t+1}||^2\Big\} + \frac{\eta_x}{2}\Big\{||\Delta_x^t||^2 + ||\partial_x^t - \Delta_x^t||^2\Big\} \\
& + \frac{1}{2\eta_y}\Big\{||y-y_t||^2 - ||y-y_{t+1}||^2 + ||y-\tilde{y}_t||^2 - ||y-\tilde{y}_{t+1}||^2\Big\} + \frac{\eta_y}{2}\Big\{||\Delta_y^t||^2 + ||\partial_y^t - \Delta_y^t||^2\Big\} \\
\overset{(b)}{\leq} & \frac{1}{2\eta_x}\Big\{||x-x_t||^2 - ||x-x_{t+1}||^2 + ||x-\tilde{x}_t||^2 - ||x-\tilde{x}_{t+1}||^2\Big\} + \frac{\eta_x}{2}\Big\{3||\Delta_x^t||^2 + 2||\partial_x^t||^2\Big\} \\
& + \frac{1}{2\eta_y}\Big\{||y-y_t||^2 - ||y-y_{t+1}||^2 + ||y-\tilde{y}_t||^2 - ||y-\tilde{y}_{t+1}||^2\Big\} + \frac{\eta_y}{2}\Big\{3||\Delta_y^t||^2 + 2||\partial_y^t||^2\Big\}
\end{aligned} \tag{18}
$$

where inequality (a) above is due to the convexity of $f(x, y_t)$ in x and concavity of $f(x_t, y)$ in y. Inequality (b) is due to $(a+b)^2 \leq 2a^2 + 2b^2$.

Then we combine the LHS and RHS by summing up $t = 0, ..., T-1$:

$$\begin{aligned}\sum_{t=0}^{T-1}(f(x_t, y) - f(x, y_t)) \leq & \frac{1}{2\eta_x}\Big\{||x - x_0||^2 - ||x - x_T||^2 + ||x - \tilde{x}_0||^2 - ||x - \tilde{x}_T||^2\Big\} \\ & \frac{1}{2\eta_y}\Big\{||y - y_0||^2 - ||y - y_T||^2 + ||y - \tilde{y}_0||^2 - ||y - \tilde{y}_T||^2\Big\} \\ & + \frac{3\eta_x}{2}\underbrace{\sum_{t=1}^{T}||\Delta_x^t||^2}_{:=A} + \eta_x \underbrace{\sum_{t=1}^{T}||\partial_x^t||^2}_{:=B} \\ & + \frac{3\eta_y}{2}\underbrace{\sum_{t=1}^{T}||\Delta_y^t||^2}_{:=C} + \eta_y \underbrace{\sum_{t=1}^{T}||\partial_y^t||^2}_{:=D} \\ & + \underbrace{\sum_{t=0}^{T-1}\Big((x_t - \tilde{x}_t)^\top(\partial_x^t - \Delta_x^t) + (y_t - \tilde{y}_t)^\top(\partial_y^t - \Delta_y^t)\Big)}_{:=E}. \end{aligned} \tag{19}$$

In the following, we show how to bound the above A to E terms. To bound the above term A in (19), we apply Lemma 6 as follows, which holds with probability $1 - \tilde{\delta}$,

$$\sum_{t=1}^{T}\|\Delta_x^t\|^2 \leq B_1^2(T + \log(1/\tilde{\delta})). \tag{20}$$

Similarly, term C in (19) can be bounded with probability $1 - \tilde{\delta}$ as follows

$$\sum_{t=1}^{T}\|\Delta_y^t\|^2 \leq B_2^2(T + \log(1/\tilde{\delta})). \tag{21}$$

To bound term B of (19), which contains only the full subgradients ∂_x^t, we have

$$\|\partial_x^t\|^2 = \|\mathrm{E}[\Delta_x^t]\|^2 \leq \mathrm{E}[\|\Delta_x^t\|^2] \leq B_1^2,$$

where the first inequality is due to Jensen's inequality and the second inequality is due to

$$\exp(\mathrm{E}[\frac{\|\Delta_x^t\|^2}{B_1^2}]) \leq \mathrm{E}[\exp(\frac{\|\Delta_x^t\|^2}{B_1^2})] \leq \exp(1) \;\Rightarrow\; \mathrm{E}[\frac{\|\Delta_x^t\|^2}{B_1^2}] \leq 1 \;\Rightarrow\; \mathrm{E}[\|\Delta_x^t\|^2] \leq B_1^2.$$

Therefore, we have

$$\sum_{t=1}^{T}\|\partial_x^t\|^2 \leq TB_1^2. \tag{22}$$

Similarly, for term D in (19), we have

$$\sum_{t=1}^{T}\|\partial_y^t\|^2 \leq TB_2^2. \tag{23}$$

To bound term E of (19), let $U_t = (x_t - \tilde{x}_t)^\top(\partial_x^t - \Delta_x^t)$ and $V_t = (y_t - \tilde{y}_t)^\top(\partial_y^t - \Delta_y^t)$ for $t \in \{0, ..., T-1\}$, which are Martingale difference sequences. We thus would like to use Lemma 7 to handle these terms. To this end, we can first upper bound $|U_t|$ and $|V_t|$ as follows

$$\begin{aligned}|U_t| =& |(x_t - \tilde{x}_t)^\top(\partial_x^t - \Delta_x^t)| \leq \|x_t - x_0 + x_0 - \tilde{x}_t\| \cdot \|\partial_x^t - \Delta_x^t\| \\ \leq& 2R(\|\partial_x^t\| + \|\Delta_x^t\|) \leq 2R(B_1 + \|\Delta_x^t\|), \\ |V_t| =& |(y_t - \tilde{y}_t)^\top(\partial_y^t - \Delta_y^t)| \leq \|(y_t - y_0 + y_0 - \tilde{y}_t\| \cdot \|\partial_y^t - \Delta_y^t)\| \\ \leq& 2R(\|\partial_y^t\| + \|\Delta_y^t)\|) \leq 2R(B_2 + \|\Delta_y^t)\|).\end{aligned}$$

Then the above two inequalities implies that

$$\begin{aligned}\mathrm{E}_t[\exp(\frac{U_t^2}{16B_1^2R^2})] \leq & \mathrm{E}_t[\exp(\frac{(2R(B_1+\|\Delta_x^t\|))^2}{16B_1^2R^2})] \overset{(a)}{\leq} \mathrm{E}_t[\exp(\frac{4R^2(2B_1^2+2\|\Delta_x^t\|^2)}{16B_1^2R^2})] \\ = & \mathrm{E}_t[\exp(\frac{B_1^2+\|\Delta_x^t\|^2}{2B_1^2})] = \mathrm{E}_t[\exp(\frac{1}{2}+\frac{\|\Delta_x^t\|^2}{2B_1^2})] \\ = & \exp(\frac{1}{2})\cdot \mathrm{E}_t\Big[\sqrt{\exp(\frac{\|\Delta_x^t\|^2}{B_1^2})}\Big] \overset{(b)}{\leq} \exp(\frac{1}{2})\cdot\sqrt{\mathrm{E}_t[\exp(\frac{\|\Delta_x^t\|^2}{B_1^2})]} \\ \overset{(c)}{\leq} & \exp(\frac{1}{2})\sqrt{\exp(1)} = \exp(1), \end{aligned} \tag{24}$$

where inequality (a) is due to $(a+b)^2 \leq 2a^2+2b^2$, inequality (b) is due to the concavity of $\sqrt{\cdot}$ and Jensen's inequality. Inequality (c) is due to the assumption. In a similar way, we have

$$\mathrm{E}_t[\exp(\frac{V_t^2}{16B_2^2R^2})] \leq \exp(1). \tag{25}$$

Next, applying Lemma 7 with (24) and (25), we have with probability at least $1-\tilde{\delta}$

$$\begin{aligned}\sum_{t=0}^{T-1} U_t &\leq 4B_1R\sqrt{3\log(1/\tilde{\delta})T}, \\ \sum_{t=0}^{T-1} V_t &\leq 4B_2R\sqrt{3\log(1/\tilde{\delta})T}. \end{aligned} \tag{26}$$

For LHS of (19), by Jensen's inequality, we have

$$\sum_{t=0}^{T-1}(f(x_t,y)-f(x,y_t)) \geq T(f(\bar{x},y)-f(x,\bar{y})), \tag{27}$$

where $\bar{x}=\frac{1}{T}\sum_{t=0}^{T-1}x_t$ and $\bar{y}=\frac{1}{T}\sum_{t=0}^{T-1}$.

Suppose $T \geq 1$. By plugging (27), (20), (21), (22), (23) and (26) back into (19), with probability at least $1-\tilde{\delta}$, we have

$$\begin{aligned} f(\bar{x},y)-f(x,\bar{y}) \leq & \frac{\|x-x_0\|^2}{\eta_x T}+\frac{\|y-y_0\|^2}{\eta_y T}+\frac{\eta_x B_1^2}{2}(5+3\log(1/\tilde{\delta}))+\frac{\eta_y B_2^2}{2}(5+3\log(1/\tilde{\delta})) \\ & +\frac{4(B_1+B_2)R\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{T}} \end{aligned} \tag{28}$$

□

D Proof of Lemma 6

Proof. First, we start from

$$\begin{aligned}\mathrm{E}[\exp(\frac{\sum_{t=1}^{T}X_t}{B^2})] = & \mathrm{E}[\mathrm{E}_T[\exp(\frac{\sum_{t=1}^{T-1}X_t+X_T}{B^2})]] \\ = & \mathrm{E}[\exp(\frac{\sum_{t=1}^{T-1}X_t}{B^2})\cdot\mathrm{E}_T[\exp(\frac{X_T}{B^2})]] \\ \leq & \mathrm{E}[\exp(\frac{\sum_{t=1}^{T-1}X_t}{B^2})\cdot\exp(1)] \\ \leq & \mathrm{E}[\exp(\frac{\sum_{t=1}^{T-2}X_t}{B^2})\cdot\exp(2)] \\ \leq & \exp(T), \end{aligned}$$

where the first inequality is due to the assumption.

Markov inequality indicates that $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$ for a random variable X, which, by additionally introducing $\tilde{\delta}$, leads to

$$P\Big(\exp(\frac{\sum_{t=1}^{T} X_t}{B^2}) \geq \frac{\mathrm{E}[\exp(\frac{\sum_{t=1}^{T} X_t}{B^2})]}{\tilde{\delta}}\Big) \leq \tilde{\delta}.$$

Therefore, with probability at least $1-\tilde{\delta}$, we have

$$\begin{aligned}
&\exp(\frac{\sum_{t=1}^{T} X_t}{B^2}) \leq \frac{\mathrm{E}[\exp(\frac{\sum_{t=1}^{T} X_t}{B^2})]}{\tilde{\delta}} \leq \frac{\exp(T)}{\tilde{\delta}} \\
\Rightarrow &\frac{\sum_{t=1}^{T} X_t}{B^2} \leq \log(\frac{\exp(T)}{\tilde{\delta}}) = \log(\exp(T)) + \log(1/\tilde{\delta}) = T + \log(1/\tilde{\delta}) \\
\Rightarrow &\sum_{t=1}^{T} X_t \leq B^2(T + \log(1/\tilde{\delta})).
\end{aligned}$$

□

E Proof of Lemma 3

Proof. Here we consider the following problem

$$\min_{x \in X \cap \mathcal{B}(x_0, R)} \max_{y \in Y \cap \mathcal{B}(y_0, R)} f(x, y)$$

with two solutions (x_0, y_0) and $(\bar{x}, \bar{y})$.

By (1) of Lemma 1, we have

$$\begin{aligned}
\frac{\mu}{4}\|\hat{x}_R(\bar{y}) - x_0\|^2 + \frac{\lambda}{4}\|\hat{y}_R(\bar{x}) - y_0\|^2 \leq &\underbrace{\max_{y' \in Y \cap \mathcal{B}(y_0, R)} f(x_0, y') - \min_{x \in X \cap \mathcal{B}(x_0, R)} f(x', y_0)}_{:=A} \\
&+ \underbrace{\max_{y' \in Y \cap \mathcal{B}(y_0, R)} f(\bar{x}, y') - \min_{x \in X \cap \mathcal{B}(x_0, R)} f(x', \bar{y})}_{:=B}. \quad (29)
\end{aligned}$$

We can bound the above term A as follows

$$\begin{aligned}
&\max_{y' \in Y \cap \mathcal{B}(y_0, R)} f(x_0, y') - \min_{x \in X \cap \mathcal{B}(x_0, R)} f(x', y_0) \\
\leq &\max_{y' \in Y} f(x_0, y') - \min_{x \in X} f(x', y_0) \leq \frac{\min\{\mu, \lambda\} R^2}{8}, \quad (30)
\end{aligned}$$

where the last inequality is due to the setting of R.

Recall the definitions

$$\hat{x}_R(\bar{y}) = \arg \min_{x' \in x \cap \mathcal{B}(x_0, R)} f(x', \bar{y}), \qquad \hat{y}_R(\bar{x}) = \arg \max_{y' \in Y \cap \mathcal{B}(y_0, R)} f(\bar{x}, y').$$

To Bound term B in (29), we apply Lemma 2 as follows

$$\begin{aligned}
&\max_{y' \in Y \cap \mathcal{B}(y_0, R)} f(\bar{x}, y') - \min_{x' \in \cap \mathcal{B}(x_0, R)} f(x, \bar{y}) \\
\leq &\frac{\|\hat{x}_R(\bar{y}) - x_0\|^2}{\eta_x T} + \frac{\|\hat{y}_R(\bar{x}) - y_0\|^2}{\eta_y T} + \frac{\eta_x B_1^2}{2}(5 + 3\log(1/\tilde{\delta})) + \frac{\eta_y B_2^2}{2}(5 + 3\log(1/\tilde{\delta})) \\
&+ \frac{4(B_1 + B_2)R\sqrt{2\log(1/\tilde{\delta})}}{\sqrt{T}} \\
\leq &\frac{R^2}{\eta_x T} + \frac{R^2}{\eta_y T} + \frac{\eta_x B_1^2}{2}(5 + 3\log(1/\tilde{\delta})) + \frac{\eta_y B_2^2}{2}(5 + 3\log(1/\tilde{\delta})) \\
&+ \frac{4(B_1 + B_2)R\sqrt{2\log(1/\tilde{\delta})}}{\sqrt{T}} \leq \frac{\min\{\mu, \lambda\} R^2}{16}, \quad (31)
\end{aligned}$$

where the last inequality holds with probability at least $1-\tilde{\delta}$ with the setting of η_x, η_y and T as follows

$$\begin{aligned}\eta_x =& \frac{\min\{\mu,\lambda\}R^2}{40(5+3\log(1/\tilde{\delta}))B_1^2}, \eta_y = \frac{\min\{\mu,\lambda\}R^2}{40(5+3\log(1/\tilde{\delta}))B_2^2}\\ T \geq& \frac{\max\left\{320^2(B_1+B_1)^2 3\log(1/\tilde{\delta}), 3200(5+3\log(1/\tilde{\delta}))\max\{B_1^2,B_2^2\}\right\}}{\min\{\mu,\lambda\}^2R^2}. \end{aligned} \tag{32}$$

Finally, we use (30) and (31) to bound term A and term B in (29) as follows

$$\begin{aligned}\frac{\mu}{4}\|\hat{x}_R(\bar{y})-x_0\|^2+\frac{\lambda}{4}\|\hat{y}_R(\bar{x})-y_0\|^2 \leq& \frac{\min\{\mu,\lambda\}R^2}{8}+\frac{\min\{\mu,\lambda\}R^2}{16}=\frac{3\min\{\mu,\lambda\}R^2}{16}\\ <& \frac{\min\{\mu,\lambda\}R^2}{4}.\end{aligned}$$

It implies

$$\begin{aligned}\|\hat{x}_R(\bar{y})-x_0\|<R,\\ \|\hat{y}_R(\bar{x})-y_0\|<R,\end{aligned}$$

which shows $\hat{x}_R(\bar{y})$ and $\hat{y}_R(\bar{y})$ are interior points of $\mathcal{B}(x_0,R)$ and $\mathcal{B}(y_0,R)$, respectively, so that $\hat{x}_R(\bar{y})=\hat{x}(\bar{y})$ and $\hat{y}_R(\bar{x})=\hat{y}(\bar{x})$.

□

F Proof of Theorem 1

Proof. Let

$$T_1 = \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(3\log(1/\tilde{\delta})+2)\max\{B_1^2,B_2^2\}\right\}}{\min\{\mu,\lambda\}^2R_1^2},$$

where $\text{Gap}(x_0,y_0)=\max_{y\in Y}f(x_0,y)-\min_{x\in X}f(x,y_0)\leq\epsilon_0$ and $R_1\geq 2\sqrt{\frac{2\epsilon_0}{\min\{\mu,\lambda\}}}$.

Given $T_{k+1}=2T_k$ in Algorithm 1 and $K=\lceil\log(\frac{\epsilon_0}{\epsilon})\rceil$, the total number of iterations can be computed by

$$\begin{aligned}T_{tot} =& \sum_{k=1}^K T_k = T_1\sum_{k=1}^K 2^{k-1} = T_1(2^K-1)\leq T_1 2^{\lceil\log(\frac{\epsilon_0}{\epsilon})\rceil}\leq T_1\frac{2\epsilon_0}{\epsilon}\\ =& \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(3\log(1/\tilde{\delta})+2)\max\{B_1^2,B_2^2\}\right\}}{\min\{\mu,\lambda\}^2R_1^2}\cdot\frac{2\epsilon_0}{\epsilon}\\ \leq& \frac{\max\left\{320^2(B_1+B_2)^2 3\log(1/\tilde{\delta}), 3200(3\log(1/\tilde{\delta})+2)\max\{B_1^2,B_2^2\}\right\}}{8\min\{\mu,\lambda\}\epsilon_0}\cdot\frac{2\epsilon_0}{\epsilon}\\ =& \frac{\max\left\{320^2(B_1+B_2)^2 3\log(\frac{1}{\tilde{\delta}}), 3200(3\log(1/\tilde{\delta})+2)\max\{B_1^2,B_2^2\}\right\}}{4\min\{\mu,\lambda\}\epsilon}\end{aligned}$$

□

G Proof of Lemma 4

Proof. In this proof, we focus on the analysis of one inner loop and thus omit the index of k for simpler presentation. Let $\Delta_x^t=\partial_x f(x_t,y_t;\xi_t)$, $\Delta_y^t=\partial_y f(x_t,y_t;\xi^t)$, $\partial_x^t=\partial_x f(x_t,y_t)$ and $\partial_y^t=\partial_y f(x_t,y_t)$. Denote $\hat{f}(x,y)=f(x,y)+\frac{\gamma}{2}\|x-x_0\|^2$.

Let $\psi_x^t(x)=x^\top\Delta_x^t+\frac{1}{2\eta_x}\|x-x_t\|^2+\frac{\gamma}{2}\|x-x_0\|^2$ and $\psi_y^t(y)=-y^\top\Delta_y^t+\frac{1}{2\eta_y}\|y-y_t\|^2$. According to the update of x_{t+1} and y_{t+1}, we have $x_{t+1}=\arg\min_{x\in X}\psi_x^t(x)$ and $y_{t+1}=\arg\max_{y\in Y}\psi_y^t(y)$. It is easy to verify that ψ_x^t and ψ_y^t are strongly convex in x and y, respectively.

By $\Big(\frac{1}{\eta_x}+\gamma\Big)$-strong convexity of $\psi_x^t(x)$ and the optimality condition at x_{t+1}, we have

$$
\begin{aligned}
&\Big(\frac{1}{2\eta_x}+\frac{\gamma}{2}\Big)\|x-x_{t+1}\|^2 \le \psi_x^t(x)-\psi_x^t(x_{t+1})\\
=&x^\top\Delta_x^t+\frac{1}{2\eta_x}\|x-x_t\|^2+\frac{\gamma}{2}\|x-x_0\|^2-\Big(x_{t+1}^\top\Delta_x^t+\frac{1}{2\eta_x}\|x_{t+1}-x_t\|^2+\frac{\gamma}{2}\|x_{t+1}-x_0\|^2\Big)\\
=&(x-x_t)^\top\partial_x^t+(x_t-x_{t+1})^\top\partial_x^t+(x-x_{t+1})^\top(\Delta_x^t-\partial_x^t)\\
&+\frac{1}{2\eta_x}\|x-x_t\|^2+\frac{\gamma}{2}\|x-x_0\|^2-\frac{1}{2\eta_x}\|x_{t+1}-x_t\|^2-\frac{\gamma}{2}\|x_{t+1}-x_0\|^2\\
\overset{(a)}{\le}&f(x,y_t)-f(x_t,y_t)+\frac{\gamma}{2}\|x-x_0\|^2-\frac{\gamma}{2}\|x_t-x_0\|^2+\frac{\gamma}{2}\Big(\|x_t-x_0\|^2-\|x_{t+1}-x_0\|^2\Big)\\
&+\Big(\frac{1}{2\eta_x}+\frac{\rho}{2}\Big)\|x-x_t\|^2+(x-x_t)^\top(\Delta_x^t-\partial^t)+(x_t-x_{t+1})^\top\Delta_x^t-\frac{1}{2\eta_x}\|x_{t+1}-x_t\|^2\\
\overset{(b)}{\le}&\hat f(x,y_t)-\hat f(x_t,y_t)+\frac{\gamma}{2}\Big(\|x_t-x_0\|^2-\|x_{t+1}-x_0\|^2\Big)\\
&+\Big(\frac{1}{2\eta_x}+\frac{\rho}{2}\Big)\|x-x_t\|^2+(x-x_t)^\top(\Delta_x^t-\partial^t)+\frac{\eta_x}{2}\|\Delta_x^t\|^2, \qquad (33)
\end{aligned}
$$

where inequality (a) is due to ρ-weakly convexity of f in x. Inequality (b) is due to Young's inequality, i.e., $(x_t-x_{t+1})^\top\Delta_x^t-\frac{1}{2\eta_x}\|x_{t+1}-x_t\|^2\le\frac{\eta_x}{2}\|\Delta_x^t\|^2$.

Similarly, due to the $\frac{1}{\eta_y}$-strong convexity of ψ_y^t in y and the optimality condition of y_{t+1}, we have

$$
\begin{aligned}
&\frac{1}{2\eta_y}\|y-y_{t+1}\|^2 \le \psi_y^t(y)-\psi_y^t(y_{t+1})\\
=&-y^\top\Delta_y^t+\frac{1}{2\eta_y}\|y-y_t\|^2-\Big(-y_{t+1}^\top\Delta_y^t+\frac{1}{2\eta_y}\|y_{t+1}-y_t\|^2\Big)\\
=&(y_t-y)^\top\partial_y^t+(y_{t+1}-y_t)^\top\partial_y^t+(y_{t+1}-y)^\top(\Delta_y^t-\partial_y^t)\\
&+\frac{1}{2\eta_y}\|y-y_t\|^2-\frac{1}{2\eta_y}\|y_{t+1}-y_t\|^2\\
\overset{(a)}{\le}&f(x_t,y_t)-f(x_t,y)+(y_{t+1}-y_t)^\top\Delta_y^t+(y_t-y)^\top(\Delta_y^t-\partial_y^t)\\
&+\frac{1}{2\eta_y}\|y-y_t\|^2-\frac{1}{2\eta_y}\|y_{t+1}-y_t\|^2\\
\overset{(b)}{\le}&\hat f(x_t,y_t)-\hat f(x_t,y)+(y_t-y)^\top(\Delta_y^t-\partial_y^t)+\frac{1}{2\eta_y}\|y-y_t\|^2+\frac{\eta_y}{2}\|\Delta_y^t\|^2, \qquad (34)
\end{aligned}
$$

where inequality (a) is due to concavity of f in y. Inequality (b) is due to Young's inequality, i.e., $(y_{t+1}-y_t)^\top\Delta_y^t-\frac{1}{2\eta_y}\|y_{t+1}-y_t\|^2\le\frac{\eta_y}{2}\|\Delta_y^t\|^2$.

Combining (33) and (34), we have

$$
\begin{aligned}
&\hat f(x_t,y)-\hat f(x,y_t)\le\frac{\eta_x}{2}\|\Delta_x^t\|^2+\frac{\eta_y}{2}\|\Delta_y^t\|^2\\
&+(x-x_t)^\top(\Delta_x^t-\partial^t)+(y_t-y)^\top(\Delta_y^t-\partial^t)+\frac{\gamma}{2}\Big(\|x_t-x_0\|^2-\|x_{t+1}-x_0\|^2\Big)\\
&+\Big(\frac{1}{2\eta_x}+\frac{\rho}{2}\Big)\|x-x_t\|^2-\Big(\frac{1}{2\eta_x}+\frac{\gamma}{2}\Big)\|x-x_{t+1}\|^2+\frac{1}{2\eta_y}\Big(\|y-y_t\|^2-\|y-y_{t+1}\|^2\Big). \qquad (35)
\end{aligned}
$$

Now we do not take expectation, since we aim to eliminate the randomness of x and y in $(x-x_t)$ and (y_t-y), respectively. To achieve this, we use the following updates

$$
\begin{aligned}
\tilde x_{t+1}&=\arg\min_{x\in X}x^\top(\partial_x^t-\Delta_x^t)+\frac{1}{2\eta_x}\|x-\tilde x_t\|^2\\
\tilde y_{t+1}&=\arg\min_{y\in Y}-y^\top(\partial_y^t-\Delta_y^t)+\frac{1}{2\eta_y}\|y-\tilde y_t\|^2,
\end{aligned}
$$

where $\tilde{x}_0 = x_0$ and $\tilde{y}_0 = y_0$.

Using similar analysis as the beginning, we have

$$\begin{aligned}
\frac{1}{2\eta_x}\|x-\tilde{x}_{t+1}\|^2 \leq & x^\top(\partial_x^t - \Delta_x^t) + \frac{1}{2\eta_x}\|x-\tilde{x}_t\|^2 - \Big(\tilde{x}_{t+1}^\top(\partial_x^t - \Delta_x^t) + \frac{1}{2\eta_x}\|\tilde{x}_{t+1}-\tilde{x}_t\|^2\Big) \\
= & (\tilde{x}_t - x)^\top(\Delta_x^t - \partial_x^t) + \frac{1}{2\eta_x}\|x-\tilde{x}_t\|^2 + (\tilde{x}_t - \tilde{x}_{t+1})^\top(\partial_x^t - \Delta_x^t) - \frac{1}{2\eta_x}\|\tilde{x}_{t+1}-\tilde{x}_t\|^2 \\
\leq & (\tilde{x}_t - x)^\top(\Delta_x^t - \partial_x^t) + \frac{1}{2\eta_x}\|x-\tilde{x}_t\|^2 + \frac{\eta_x}{2}\|\partial_x^t - \Delta_x^t\|^2 \\
\leq & (\tilde{x}_t - x)^\top(\Delta_x^t - \partial_x^t) + \frac{1}{2\eta_x}\|x-\tilde{x}_t\|^2 + \eta_x\|\partial_x^t\|^2 + \eta_x\|\Delta_x^t\|^2.
\end{aligned}$$

We could also derive the similar result for y as follows

$$\begin{aligned}
\frac{1}{2\eta_y}\|y-\tilde{y}_{t+1}\|^2 \leq & -y^\top(\partial_y^t - \Delta_y^t) + \frac{1}{2\eta_y}\|y-\tilde{y}_t\|^2 - \Big(-\tilde{y}_{t+1}^\top(\partial_y^t - \Delta_y^t) + \frac{1}{2\eta_y}\|\tilde{y}_{t+1}-\tilde{y}_t\|^2\Big) \\
= & (y - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \frac{1}{2\eta_y}\|y-\tilde{y}_t\|^2 + (\tilde{y}_{t+1} - \tilde{y}_t)^\top(\partial_y^t - \Delta_y^t) - \frac{1}{2\eta_y}\|\tilde{y}_{t+1}-\tilde{y}_t\|^2 \\
\leq & (y - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \frac{1}{2\eta_y}\|y-\tilde{y}_t\|^2 + \frac{\eta_y}{2}\|\partial_y^t - \Delta_y^t\|^2 \\
\leq & (y - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \frac{1}{2\eta_y}\|y-\tilde{y}_t\|^2 + \eta_y\|\partial_y^t\|^2 + \eta_y\|\Delta_y^t\|^2.
\end{aligned}$$

Summing the above two inequalities, we have

$$\begin{aligned}
0 \leq & \frac{1}{2\eta_x}\Big(\|x-\tilde{x}_t\|^2 - \|x-\tilde{x}_{t+1}\|^2\Big) + (\tilde{x}_t - x)^\top(\Delta_x^t - \partial_x^t) + \eta_x\|\partial_x^t\|^2 + \eta_x\|\Delta_x^t\|^2 \\
& + \frac{1}{2\eta_y}\Big(\|y-\tilde{y}_t\|^2 - \|y-\tilde{y}_{t+1}\|^2\Big) + (y - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \eta_y\|\partial_y^t\|^2 + \eta_y\|\Delta_y^t\|^2 \qquad (36)
\end{aligned}$$

Combining (35) and (36), we have

$$\begin{aligned}
& \hat{f}(x_t, y) - \hat{f}(x, y_t) \leq \frac{\eta_x}{2}\|\Delta_x^t\|^2 + \frac{\eta_y}{2}\|\Delta_y^t\|^2 \\
& + (x - x_t)^\top(\Delta_x^t - \partial_x^t) + (y_t - y)^\top(\Delta_y^t - \partial_y^t) + \frac{\gamma}{2}\Big(\|x_t - x_0\|^2 - \|x_{t+1} - x_0\|^2\Big) \\
& + \Big(\frac{1}{2\eta_x} + \frac{\rho}{2}\Big)\|x - x_t\|^2 - \Big(\frac{1}{2\eta_x} + \frac{\gamma}{2}\Big)\|x - x_{t+1}\|^2 + \frac{1}{2\eta_y}\Big(\|y-y_t\|^2 - \|y-y_{t+1}\|^2\Big) \\
& + \frac{1}{2\eta_x}\Big(\|x-\tilde{x}_t\|^2 - \|x-\tilde{x}_{t+1}\|^2\Big) + (\tilde{x}_t - x)^\top(\Delta_x^t - \partial_x^t) + \eta_x\|\partial_x^t\|^2 + \eta_x\|\Delta_x^t\|^2 \\
& + \frac{1}{2\eta_y}\Big(\|y-\tilde{y}_t\|^2 - \|y-\tilde{y}_{t+1}\|^2\Big) + (y - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \eta_y\|\partial_y^t\|^2 + \eta_y\|\Delta_y^t\|^2 \\
= & \frac{3\eta_x}{2}\|\Delta_x^t\|^2 + \eta_x\|\partial_x^t\|^2 + \frac{3\eta_y}{2}\|\Delta_y^t\|^2 + \eta_y\|\partial_y^t\|^2 \\
& + (\tilde{x}_t - x_t)^\top(\Delta_x^t - \partial_x^t) + (y_t - \tilde{y}_t)^\top(\Delta_y^t - \partial_y^t) + \frac{\gamma}{2}\Big(\|x_t - x_0\|^2 - \|x_{t+1} - x_0\|^2\Big) \\
& + \Big(\frac{1}{2\eta_x} + \frac{\rho}{2}\Big)\|x - x_t\|^2 - \Big(\frac{1}{2\eta_x} + \frac{\gamma}{2}\Big)\|x - x_{t+1}\|^2 + \frac{1}{2\eta_y}\Big(\|y-y_t\|^2 - \|y-y_{t+1}\|^2\Big) \\
& + \frac{1}{2\eta_x}\Big(\|x-\tilde{x}_t\|^2 - \|x-\tilde{x}_{t+1}\|^2\Big) + \frac{1}{2\eta_y}\Big(\|y-\tilde{y}_t\|^2 - \|y-\tilde{y}_{t+1}\|^2\Big)
\end{aligned}$$

Summing the above inequality from $t=0$ to $T-1$ and using Jensen's inequality, we have

$$
\begin{aligned}
&T\Big(\hat{f}(\bar{x},y)-\hat{f}(x,\bar{y})\Big)\leq\sum_{t=0}^{T-1}\Big(\hat{f}(x_t,y)-\hat{f}(x,y_t)\Big)\\
\leq&\frac{\eta_x}{2}\sum_{t=0}^{T-1}(3\|\Delta_x^t\|^2+2\|\partial_x^t\|^2)+\frac{\eta_y}{2}\sum_{t=0}^{T-1}(3\|\Delta_y^t\|^2+2\|\partial_y^t\|^2)\\
&+\sum_{t=0}^{T-1}(\tilde{x}_t-x_t)^\top(\Delta_x^t-\partial_x^t)+\sum_{t=0}^{T-1}(y_t-\tilde{y}_t)^\top(\Delta_y^t-\partial_y^t)+\frac{\gamma}{2}\Big(\|x_0-x_0\|^2-\|x_T-x_0\|^2\Big)\\
&+\Big(\frac{1}{2\eta_x}+\frac{\rho}{2}\Big)\|x-x_0\|^2-\Big(\frac{1}{2\eta_x}+\frac{\gamma}{2}\Big)\|x-x_T\|^2+\frac{1}{2\eta_y}\Big(\|y-y_0\|^2-\|y-y_T\|^2\Big)\\
&+\frac{1}{2\eta_x}\Big(\|x-x_0\|^2-\|x-\tilde{x}_T\|^2\Big)+\frac{1}{2\eta_y}\Big(\|y-y_0\|^2-\|y-\tilde{y}_T\|^2\Big)
\end{aligned}
$$

where $\bar{x}=\frac{1}{T}\sum_{t=0}^{T-1}x_t$ and $\bar{y}=\frac{1}{T}\sum_{t=0}^{T-1}y_t$.

Plugging in $x=\hat{x}(\bar{y})$ and $y=\hat{y}(\bar{x})$, we have

$$
\begin{aligned}
&\widehat{\text{Gap}}(\bar{x},\bar{y})=\hat{f}(\bar{x},\hat{y}(\bar{x}))-\hat{f}(\hat{x}(\bar{y}),\bar{y})\leq\frac{1}{T}\sum_{t=0}^{T-1}\Big(\hat{f}(x_t,\hat{y}(\bar{x}))-\hat{f}(\hat{x}(\bar{y}),y_t)\Big)\\
\leq&\frac{\eta_x}{2T}\sum_{t=0}^{T-1}(3\|\Delta_x^t\|^2+2\|\partial_x^t\|^2)+\frac{\eta_y}{2T}\sum_{t=0}^{T-1}(3\|\Delta_y^t\|^2+2\|\partial_y^t\|^2)\\
&+\frac{1}{T}\sum_{t=0}^{T}(\tilde{x}_t-x_t)^\top(\Delta_x^t-\partial_x^t)+\frac{1}{T}\sum_{t=0}^{T}(y_t-\tilde{y}_t)^\top(\Delta_y^t-\partial_y^t)\\
&+\frac{1}{T}\Big(\frac{1}{\eta_x}+\frac{\rho}{2}\Big)\|\hat{x}(\bar{y})-x_0\|^2+\frac{1}{\eta_y T}\|\hat{y}(\bar{x})-y_0\|^2
\end{aligned}
$$

Taking expectation over both sides and recalling that $\mathrm{E}[\|\partial_x f(x,y;\xi)\|^2]\leq M_1^2$ and $\mathrm{E}[\|\partial_y f(x,y;\xi)\|^2]\leq M_2^2$, we have

$$
\begin{aligned}
&\mathrm{E}[\widehat{\text{Gap}}(\bar{x},\bar{y})]=\mathrm{E}[\hat{f}(\bar{x},\hat{y}(\bar{x}))-\hat{f}(\hat{x}(\bar{y}),\bar{y})]\leq\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{E}[\hat{f}(x_t,\hat{y}(\bar{x}))-\hat{f}(\hat{x}(\bar{y}),y_t)]\\
\leq&\frac{5\eta_x M_1^2}{2}+\frac{5\eta_y M_2^2}{2}+\frac{1}{T}\Big(\frac{1}{\eta_x}+\frac{\rho}{2}\Big)\mathrm{E}[\|\hat{x}(\bar{y})-x_0\|^2]+\frac{1}{\eta_y T}\mathrm{E}[\|\hat{y}(\bar{x})-y_0\|^2].
\end{aligned}
$$

□

H Proof of Lemma 5

Before proving Lemma 5, we first state the following lemma, whose proof is in the next section.

Lemma 8. *$\widehat{Gap}_k(\bar{x}_k,\bar{y}_k)$ could be lower bounded by the following inequalities*

$$
\begin{aligned}
&1)\ \widehat{Gap}_k(\bar{x}_k,\bar{y}_k)\geq(1-\frac{\gamma}{\rho}(\frac{1}{\alpha}-1))\widehat{Gap}_{k+1}(x_0^{k+1},y_0^{k+1})+\frac{\gamma}{2}(1-\frac{1}{1-\alpha})\|x_0^{k+1}-x_0^k\|^2,\\
&2)\ \widehat{Gap}_k(\bar{x}_k,\bar{y}_k)\geq P(x_0^{k+1})-P(x_0^k)+\frac{\gamma}{2}\|\bar{x}_k-x_0^k\|^2,\ \textit{where}\ P(x)=\max_{y\in Y}f(x,y),\\
&3)\ \widehat{Gap}_k(\bar{x}_k,\bar{y}_k)\geq\frac{\rho(1-\beta)}{2(\frac{1}{\beta}-1)}\|x_0^k-\hat{x}_k^*\|^2-\frac{\rho}{2(\frac{1}{\beta}-1)}\|\bar{x}_k-x_0^k\|^2,
\end{aligned}
\tag{37}
$$

where $0<\alpha\leq 1$ and $0<\beta\leq 1$.

Proof. (of Lemma 5)

$$
\begin{aligned}
&\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) \\
=&\frac{1}{10}\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) + \frac{4}{5}\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) + \frac{1}{10}\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) \\
\overset{(a)}{\geq}&\frac{1}{10}\Big\{\Big(1-\frac{\gamma}{\rho}(\frac{1}{\alpha}-1)\Big)\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{\gamma}{2}(1-\frac{1}{1-\alpha})\|x_0^{k+1}-x_0^k\|^2)\Big\} \\
&+\frac{4}{5}\Big\{P(x_0^{k+1}) + \frac{\gamma}{2}\|\bar{x}_k - x_0^k\|^2 - P(x_0^k)\Big\} \\
&+\frac{1}{10}\Big\{\frac{\rho(1-\beta)}{2(\frac{1}{\beta}-1)}\|x_0^k - \hat{x}_k^*\|^2 - \frac{\rho}{2(\frac{1}{\beta}-1)}\|\bar{x}_k - x_0^k\|^2\Big\} \\
=&\frac{1}{10}\Big(1-\frac{\gamma}{\rho}(\frac{1}{\alpha}-1)\Big)\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) \\
&+\Big(\frac{1}{10}\cdot\frac{\gamma}{2}(1-\frac{1}{1-\alpha}) + \frac{4}{5}\cdot\frac{\gamma}{2} - \frac{1}{10}\cdot\frac{\rho}{2(\frac{1}{\beta}-1)}\Big)\|\bar{x}_k - x_0^k\|^2 \\
&+\frac{1}{10}(\frac{\rho(1-\beta)}{2(\frac{1}{\beta}-1)})\|x_0^k - \hat{x}_k^*\|^2] \\
\overset{(b)}{=}&\frac{1}{10}(1-2(\frac{1}{\frac{5}{6}}-1))\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) \\
&+\Big(\frac{1}{10}\cdot\frac{\gamma}{2}(1-\frac{1}{1-\frac{5}{6}}) + \frac{4}{5}\cdot\frac{\gamma}{2} - \frac{1}{10}\cdot\frac{\gamma}{4(\frac{1}{\frac{1}{2}}-1)}\Big)\|\bar{x}_k - x_0^k\|^2 \\
&+\frac{1}{10}(\frac{\rho(1-\frac{1}{2})}{2(\frac{1}{\frac{1}{2}}-1)})\|x_0^k - \hat{x}_k^*\|^2 \\
=&\frac{3}{50}\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1})] + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) \\
&+\frac{\gamma}{8}\|\bar{x}_k - x_0^k\|^2 + \frac{\gamma}{80}\|x_0^k - \hat{x}_k^*\|^2 \\
\overset{(c)}{\geq}&\frac{3}{50}\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{4}{5}(P(x_0^{k+1}) - P(x_0^k)) \\
&+\frac{\gamma}{80}\|x_0^k - \hat{x}_k^*\|^2, \qquad (38)
\end{aligned}
$$

where inequality (a) is due to Lemma 8, inequality (b) is due to the setting of $\gamma = 2\rho$, $\alpha = \frac{5}{6}$, $\beta = \frac{1}{2}$. Inequality (c) is due to $\|\bar{x}_k - x_0^k\|^2 \geq 0$. □

I Proof of Lemma 8

Proof. Before we prove the three results, we first state two results of Young's inequality as follows

$$
\begin{aligned}
&\|x-y\|^2 = \|x-z+z-y\|^2 = \|x-z\|^2 + \|z-y\|^2 - 2\langle x-z, y-z\rangle \\
\geq&\|x-z\|^2 + \|z-y\|^2 - \alpha\|x-z\|^2 - \frac{1}{\alpha}\|y-z\|^2 \\
=&(1-\alpha)\|x-z\|^2 + (1-\frac{1}{\alpha})\|z-y\|^2 \\
\Rightarrow&\|x-z\|^2 \leq \frac{1}{1-\alpha}\|x-y\|^2 + \frac{1}{\alpha}\|y-z\|^2 \qquad (39) \\
\Rightarrow&-\|y-z\|^2 \leq -\alpha\|x-z\|^2 + \frac{\alpha}{1-\alpha}\|x-y\|^2 \qquad (40) \\
\Rightarrow&\|x-y\|^2 \geq (1-\alpha)\|x-z\|^2 + (1-\frac{1}{\alpha})\|y-z\|^2, \qquad (41)
\end{aligned}
$$

where $0 < \alpha \leq 1$.

We first consider the result 1).

$$
\begin{aligned}
&\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) \\
=&\hat{f}_{k+1}(x_0^{k+1}, \hat{y}_{k+1}(x_0^{k+1})) - \hat{f}_{k+1}(\hat{x}_{k+1}(y_0^{k+1}), y_0^{k+1}) \\
=&f(x_0^{k+1}, \hat{y}_{k+1}(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 \\
=&f(x_0^{k+1}, \hat{y}_{k+1}(x_0^{k+1})) + \frac{\gamma}{2}\|x_0^{k+1} - x_0^{k}\|^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}\|^2 \\
&+ \frac{\gamma}{2}\|\hat{x}_{k+1}(y_0^{k+1}) - x_0^{k}\|^2 - \frac{\gamma}{2}\|\hat{x}_{k+1}(y_0^{k+1}) - x_0^{k+1}\|^2 - \frac{\gamma}{2}\|x_0^{k+1} - x_0^{k}\|^2 \\
\overset{(a)}{\leq}&f(\bar{x}_k, \hat{y}_{k+1}(\bar{x}_k)) + \frac{\gamma}{2}\|\bar{x}_k - x_0^{k}\|^2 - f(\hat{x}_{k+1}(\bar{y}_k), \bar{y}_k) - \frac{\gamma}{2}\|\hat{x}_{k+1}(\bar{y}_k) - x_0^{k}\|^2 \\
&+ \frac{\gamma}{2}\Big\{\frac{1}{\alpha}\|\hat{x}_{k+1}(y_0^{k+1}) - x_0^{k+1}\|^2 + \frac{1}{1-\alpha}\|x_0^{k+1} - x_0^{k}\|^2\Big\} \\
&- \frac{\gamma}{2}\|\hat{x}_{k+1}(y_0^{k+1}) - x_0^{k+1}\|^2 - \frac{\gamma}{2}\|x_0^{k+1} - x_0^{k}\|^2 \\
=&\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_{k+1}(\bar{y}_k), \bar{y}_k) \\
&+ \frac{\gamma}{2}(\frac{1}{\alpha} - 1)\|\hat{x}_{k+1}(y_0^{k+1}) - x_0^{k+1}\|^2 + \frac{\gamma}{2}(\frac{1}{1-\alpha} - 1)\|x_0^{k+1} - x_0^{k}\|^2 \\
\overset{(b)}{\leq}&\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_{k+1}(\bar{y}_k), \bar{y}_k) \\
&+ \frac{\gamma}{2}(\frac{1}{\alpha} - 1)\frac{2}{\rho}(\hat{f}_{k+1}(x_0^{k+1}, y_0^{k+1}) - \hat{f}_{k+1}(\hat{x}_{k+1}(y_0^{k+1}), y_0^{k+1})) + \frac{\gamma}{2}(\frac{1}{1-\alpha} - 1)\|x_0^{k+1} - x_0^{k}\|^2 \\
\overset{(c)}{\leq}&\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) + \frac{\gamma}{\rho}(\frac{1}{\alpha} - 1)\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{\gamma}{2}(\frac{1}{1-\alpha} - 1)\|x_0^{k+1} - x_0^{k}\|^2,
\end{aligned}
$$

where inequality (a) is due to (39) ($0 < \alpha \leq 1$). Inequality (b) is due to ρ-strong convexity of $\hat{f}_{k+1}(x, y_0^{k+1})$ in x and optimality at $\hat{x}_{k+1}(y_0^{k+1})$. Inequality (c) is due to $\hat{f}_k(\hat{x}_{k+1}(\bar{y}_k), \bar{y}_k) \geq \hat{f}_k(\hat{x}_k(\bar{y}_k), \bar{y}_k)$ and $\hat{f}_{k+1}(x_0^{k+1}, y_0^{k+1}) \leq \hat{f}_{k+1}(x_0^{k+1}, \hat{y}_{k+1}(x_0^{k+1}))$.

Re-organizing the above inequality, we have

$$
\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) \geq (1 - \frac{\gamma}{\rho}(\frac{1}{\alpha} - 1))\widehat{\text{Gap}}_{k+1}(x_0^{k+1}, y_0^{k+1}) + \frac{\gamma}{2}(1 - \frac{1}{1-\alpha})\|x_0^{k+1} - x_0^{k}\|^2,
$$

which proves result 1).

Then we turn to result 2) as follows.

$$
\begin{aligned}
\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) =&\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_k(\bar{y}_k), \bar{y}_k) \\
\geq&\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(x_0^k, \bar{y}_k) \\
\geq&\hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(x_0^k, \hat{y}_k(x_0^k)) \\
=&f(\bar{x}_k, \hat{y}_k(\bar{x}_k)) + \frac{\gamma}{2}\|\bar{x}_k - x_0^k\|^2 - f(x_0^k, \hat{y}_k(x_0^k)) - 0 \\
=&f(x_0^{k+1}, \hat{y}_k(x_0^{k+1})) + \frac{\gamma}{2}\|\bar{x}_k - x_0^k\|^2 - f(x_0^k, \hat{y}_k(x_0^k)) \\
=&P(x_0^{k+1}) - P(x_0^k) + \frac{\gamma}{2}\|\bar{x}_k - x_0^k\|^2,
\end{aligned}
$$

which proves result 2).

Result 3) can be proved as follows

$$\begin{aligned}
\|\bar{x}_k - x_0^k\|^2 \overset{(a)}{\geq} & (1-\beta)\|x_0^k - \hat{x}_k^*\|^2 + (1-\frac{1}{\beta})\|\hat{x}_k^* - \bar{x}_k\|^2 \\
\overset{(b)}{\geq} & (1-\beta)\|x_0^k - \hat{x}_k^*\|^2 + (1-\frac{1}{\beta})\frac{2}{\rho}(\hat{f}_k(\bar{x}_k, \hat{y}_k^*) - \hat{f}_k(\hat{x}_k^*, \hat{y}_k^*)) \\
\overset{(c)}{\geq} & (1-\beta)\|x_0^k - \hat{x}_k^*\|^2 + (1-\frac{1}{\beta})\frac{2}{\rho}\widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k)) \\
\Rightarrow \widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k) \geq & \frac{\rho(1-\beta)}{2(\frac{1}{\beta}-1)}\|x_0^k - \hat{x}_k^*\|^2 - \frac{\rho}{2(\frac{1}{\beta}-1)}\|\bar{x}_k - x_0^k\|^2,
\end{aligned}$$

where inequality (a) is due to (41) and $0 < \beta \leq 1$. Inequality (b) is due to ρ-storng convexity of $\hat{f}_k$ in x. Inequality (c) is due to $0 < \beta \leq 1$ and

$$\begin{aligned}
\hat{f}_k(\bar{x}_k, \hat{y}_k^*) - \hat{f}_k(\hat{x}_k^*, \hat{y}_k^*) \leq & \hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_k^*, \bar{y}_k) \\
\leq & \hat{f}_k(\bar{x}_k, \hat{y}_k(\bar{x}_k)) - \hat{f}_k(\hat{x}_k(\bar{y}_k), \bar{y}_k) \\
= & \widehat{\text{Gap}}_k(\bar{x}_k, \bar{y}_k).
\end{aligned}$$

□