Single-Loop Stochastic Algorithms for Difference of Max-Structured Weakly Convex Functions

Quanqi Hu [1] **Qi Qi** [2] **Zhaosong Lu** [3] **Tianbao Yang** [1]

[1] Department of Computer Science & Engineering, Texas A&M University
[2] Department of Computer Science, The University of Iowa
[3] Department of Industrial and Systems Engineering, University of Minnesota
{quanqi-hu, tianbao-yang}@tamu.edu qi-qi@uiowa.edu zhaosong@umn.edu

Abstract

In this paper, we study a class of non-smooth non-convex problems in the form of $\min_x[\max_{y\in\mathcal{Y}}\phi(x,y)-\max_{z\in\mathcal{Z}}\psi(x,z)]$, where both $\Phi(x)=\max_{y\in\mathcal{Y}}\phi(x,y)$ and $\Psi(x)=\max_{z\in\mathcal{Z}}\psi(x,z)$ are weakly convex functions, and $\phi(x,y),\psi(x,z)$ are strongly concave functions in terms of y and z, respectively. It covers two families of problems that have been studied but are missing single-loop stochastic algorithms, i.e., difference of weakly convex functions and weakly convex strongly-concave min-max problems. We propose a stochastic Moreau envelope approximate gradient method dubbed SMAG, the first single-loop algorithm for solving these problems, and provide a state-of-the-art non-asymptotic convergence rate. The key idea of the design is to compute an approximate gradient of the Moreau envelopes of Φ,Ψ using only one step of stochastic gradient update of the primal and dual variables. Empirically, we conduct experiments on positive-unlabeled (PU) learning and partial area under ROC curve (pAUC) optimization with an adversarial fairness regularizer to validate the effectiveness of our proposed algorithms.

1 Introduction

In this paper, we consider a class of non-convex, non-smooth problems in the following form

$$\min_{x\in\mathbb{R}^{d_x}}\big\{F(x):=\max_{y\in\mathcal{Y}}\phi(x,y)-\max_{z\in\mathcal{Z}}\psi(x,z)\big\},\tag{1}$$

where the sets $\mathcal{Y}\subset\mathbb{R}^{d_y}$, $\mathcal{Z}\subset\mathbb{R}^{d_z}$ are convex and compact, and the two component functions $\phi(x,y)$ and $\psi(x,z)$ are weakly-convex in terms of x and strongly-concave in the terms of y and z, respectively. Both component functions are in expectation forms, i.e., $\phi(x,y)=\mathbb{E}_{\xi\sim\mathcal{D}_\phi}[\phi(x,y;\xi)]$ and $\psi(x,z)=\mathbb{E}_{\zeta\sim\mathcal{D}_\psi}[\psi(x,z;\zeta)]$. We refer to this class of problems as the Difference of Max-Structured Weakly Convex Functions (DMax) Optimization. DMax optimization unifies two emerging families of problems in optimization field, difference-of-weakly-convex (DWC) optimization

$$\min_{x\in\mathbb{R}^{d_x}}\{F(x):=\phi(x)-\psi(x)\},\tag{2}$$

and weakly-convex-strongly-concave (WCSC) min-max optimization

$$\min_{x\in\mathbb{R}^{d_x}}\big\{F(x):=\max_{y\in\mathcal{Y}}\phi(x,y)\big\}.\tag{3}$$

Thus, DMax optimization has a wide range of applications in machine learning and AI, including applications of DWC optimization (e.g., positive-unlabeled (PU) Learning [39], non-convex

Correspondence to: Tianbao Yang <tianbao-yang@tamu.edu>.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1807

Table 1: Comparison with existing stochastic methods for solving DWC problems with non-asymptotic convergence guarantee. * The method SBCD is designed to solve a problem in the form of $\min_x\{\min_y \phi(x,y) - \min_z \psi(x,z)\}$ with a specific formulation of ϕ and ψ. However, the method and analysis can be generalized to solving non-smooth DWC problems.

Method	Smoothness of ϕ, ψ	Complexity	Loops
SDCA [26]	ϕ: Smooth	$\mathcal{O}(\epsilon^{-4})$	Double
SSDC [39]	ϕ or ψ: ν-Hölder continuous gradient	$\mathcal{O}(\epsilon^{-4/\nu})$	Double
SBCD* [45]	Non-smooth	$\mathcal{O}(\epsilon^{-6})$	Double
SMAG (ours)	Non-smooth	$\mathcal{O}(\epsilon^{-4})$	Single

sparsity-promoting regularizers [39], Boltzmann machines [26]) and applications of min-max optimization (e.g., adversarial learning [31, 22], distributional robust learning [8, 28], learning with non-decomposable loss [28]). In recent years, the scale of data and models significantly increased, leading to the demand of more efficient optimization methods. However, all existing stochastic methods for DWC optimization and non-smooth WCSC min-max optimization with state-of-the-art non-asymptotic convergence rate $\mathcal{O}(\epsilon^{-4})$ are double-loop. As a result, these methods are complex regarding the implementation and require extensive hyperparameter tuning. To close this gap, we propose a single-loop stochastic algorithm for DMax optimization and provide non-asymptotic convergence analysis to match the state-of-the-art non-asymptotic convergence rate.

The main challenges of designing a single-loop method for DMax optimization are threefold. 1) given the weakly-convex nature of the component functions, their difference $F(x)$ is not necessarily weakly-convex, resulting in a non-smooth non-convex optimization problem. 2) the component functions $\max_{y\in\mathcal{Y}} \phi(x,y)$ and $\max_{z\in\mathcal{Z}} \psi(x,z)$ require solving maximization subproblems, making unbiased estimations of their subgradients inaccessible. 3) existing work on non-smooth problems with DC or/and min-max structures heavily rely on inner loops to solve subproblems to a certain accuracy.

To address the first challenge, we apply Moreau envelope smoothing technique [24, 3] to the component functions individually and take their difference as a smooth approximation of the original objective. Inspired by existing work [32, 45], we show that solving the original DMax problem can be achieved by solving this smooth approximation. Consequently, the problem is transformed into a smooth problem with two layer of nested optimization structure, the Moreau envelope and the maximization from the min-max structure. In order to avoid inner-loop, we perform only one step of update for each of the nested optimization problems. Our analysis leverages the fast convergence of strongly convex/concave problems, proving that single-step updates are sufficient to achieve a state-of-the-art convergence rate. Although the Moreau envelope smoothing is not new for solving DC and min-max optimization [32, 45, 47, 43], the existing results either require double loops [32, 45] or require smoothness of the objective function [47, 43].

Contributions. We summarize the main contribution of this work as following.

- We construct a new framework DMax optimization that unifies the DWC optimization and WCSC min-max optimization. Based on a Moreau envelope smoothing technique, we propose a single-loop stochastic algorithm, namely SMAG, for DMax optimization in non-smooth setting, which achieves $\mathcal{O}(\epsilon^{-4})$ convergence rate.
- We show that the proposed method leads to the first single-loop stochastic algorithms for DWC optimization and non-smooth WCSC min-max optimization achieving $\mathcal{O}(\epsilon^{-4})$ convergence rate.
- Finally, we present experimental results on applications including Positive-Unlabeled (PU) Learning and partial AUC optimization with an adversarial fairness regularizer to validate the effectiveness of our proposed algorithms.

2 Related Work

Stochastic DC Optimization. DWC can be converted into Difference-of-convex (DC) programming. DC programming was initially introduced in [33] and has been extensively studied since then. A comprehensive review on the developments of DC programming can be found in [18]. Despite the

Table 2: Comparison with existing stochastic methods for solving non-convex non-smooth min-max problems. The objective function is in the form of $\phi(x,y) = f(x,y) - g(y) + h(x)$. NS and S stand for non-smooth and smooth respectively, and NSP means non-smooth and its proximal mapping is easily solved. WC, C stand for weakly-convex and convex respectively. WCSC stands for weakly-convex-strongly-concave, SSC stands for smooth and strongly concave and WCC means weakly-convex-concave. Note that Epoch-GDA and SMAG studies the general formulation $\phi(x,y) = f(x,y)$.

Method	$f(x,y)$	$g(y)$	$h(x)$	Complexity	Loops
PG-SMD [29]	NS, WCC	NSP, SC	NSP, C	$\mathcal{O}(\epsilon^{-4})$	Double
SAPD+ [48]	SSC	NSP, C	NSP,C	$\mathcal{O}(\epsilon^{-4})$	Double
Epoch-GDA [40]	NS, WCSC	-	-	$\mathcal{O}(\epsilon^{-4})$	Double
StocAGDA [1]	SSC	NSP, C	NSP, C	$\mathcal{O}(\epsilon^{-4})$	Single
SMAG (ours)	NS,WCSC	-	-	$\mathcal{O}(\epsilon^{-4})$	Single

rich literature on DC programming, DC in stochastic setting has rarely been mentioned until recently. Most of the existing studies on stochastic DC optimization are based on the classical method, DC Algorithm (DCA) in deterministic DC optimization. The main idea of DCA is to approximate the DC problem by a convex problem by taking the linear approximation of the second component. In other words, DCA solves $\min_x \{\phi(x) - \langle \nabla \phi(x_k), x\rangle\}$ to update x_k and thus forms a double-loop algorithm. [34] first proposed stochastic DCA (SDCA) for solving large sum problems of non-convex smooth functions, which was further generalized to solving large sum non-smooth problems in [16]. [15] is the first work that allows both components in DC problems to be non-smooth. The authors proposed a SDCA scheme in the aggregated update style, where all past information needs to be stored for constructing future subproblems. [17] improved the efficiency of the SDCA scheme by removing the need of storing historical information. So far, none of the above work provides non-asymptotic convergence guarantee. The first non-asymptotic convergence analysis was established in [26]. The authors proposed a stochastic proximal DC algorithm (SPD), which modifies SDCA by adding an extra quadratic term after linearizing the second component function, and proved that SPD has a convergence rate of $\mathcal{O}(\epsilon^{-4})$. The main drawback of their analysis is that they need the smoothness assumption of the first component function. With very similar algorithm design, [39] managed to partially relax the smoothness assumption. Given at least one of the two component functions having ν-Hölder continuous gradient, i.e., $\|\nabla f(x) - \nabla f(x')\| \leq \|x - x'\|^{\nu}$ for all x, x', they proved a convergence rate of $\mathcal{O}(\epsilon^{-4/\nu})$. In fact, the Hölder continuous gradient assumption is still fairly strong as some of the common non-smooth functions do not satisfy, for example the hinge loss function.

Recently, another approach to tackling the non-smoothness in DC problems has been considered. Following the smoothing technique in non-smooth weakly-convex optimization literature [3], [32, 25] constructed Moreau envelope smoothing approximations for both of the component functions respectively and established non-asymptotic convergence analysis under deterministic setting and the assumption that either one component function is smooth or the proximal-point subproblems can be solved exactly. Following a similar idea, [45] studied a problem in the form of $\min_x F(x) := \min_y \phi(x,y) - \min_z \psi(x,z)$, where ϕ and ψ are in some specific formulations, and proposed a double-loop algorithm with $\mathcal{O}(\epsilon^{-6})$ convergence rate. Although the ϕ and ψ are non-smooth, their analysis heavily relies on the properties in the given formulation, especially the structures in the dual variables y, z, thus is not trivial to generalize.

Note that none of the aforementioned work is able to solve the DMax problem, as they require unbiased stochastic gradient estimations of the two component functions, which are not accessible in DMax due to the presence of the maximization structure.

Stochastic Non-smooth Weakly-Convex-Strongly-Concave Min-Max Optimization. Stochastic WCSC min-max optimization has been an emerging topic in recent years. Most of the existing works focuses on the smooth setting, i.e., the objective is smooth [12, 19, 49, 43, 47, 43] or the stochastic gradient oracles are Lipschitz continuous [23, 42, 11, 21, 38]. To the best of our knowledge, [29] is the first work that considers non-smooth WCSC min-max problems. They considered a special structure where the maximization over y given x can be simply solved and it is solved with $O(1/\epsilon^2)$ times. They proposed a nested method Proximally Guided Stochastic Mirror Descent Method (PG-SMD) that achieves a convergence rate of $\mathcal{O}(\epsilon^{-4})$. Later, [40] further relaxed the assumption by removing the requirement of the special structure, and proved that their nested method Epoch-GDA

https://doi.org/10.52202/079017-1807

has a similar convergence rate of $\mathcal{O}(\epsilon^{-4})$. Another line of work studies a special case of the general non-smooth non-convex min-max optimization, where the objective is assumed to be composite, i.e., $\phi(x,y) = f(x,y) - g(y) + h(x)$, so that f is smooth while g, h are potentially non-smooth [1, 48]. Both works established $\mathcal{O}(\epsilon^{-4})$ convergence rate, and assume f is smooth and strongly concave, g and h are convex but potentially non-smooth and their proximal mappings can be easily solved. However, none of them is applicable to the general non-smooth WCSC min-max optimization.

3 Preliminaries

Notations. For simplicity, we denote $\Phi(x) := \max_{y\in\mathcal{Y}} \phi(x,y)$, $\Psi(x) := \max_{z\in\mathcal{Z}} \psi(x,z)$, $y^*(\cdot) := \arg\max_{y\in\mathcal{Y}} \phi(\cdot,y)$, and $z^*(\cdot) := \arg\max_{z\in\mathcal{Z}} \psi(\cdot,z)$. We use $\|\cdot\|$ to denote the Euclidean norm of a vector and $P_{\mathcal{C}}(\cdot)$ to denote the Euclidean projection onto a closed set $\mathcal{C}$. We use the following definitions of general subgradient and subdifferential [3, 30].

Definition 3.1 (subgradient and subdifferential)**.** Consider a function $f : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ and a point x with finite $f(x)$. A vector $v \in \mathbb{R}^d$ is a general subgradient of f at x if

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \quad \text{as } y \to x.$$

The subdifferential $\partial f(x)$ is the set of subgradients of f at point x.

For simplicity, we abuse the notation $\partial f(x)$ to denote one subgradient from the corresponding subdifferential when no confusion could be caused. We use $\tilde{\partial} f(x)$ and $\tilde{\nabla} f(x)$ to represent an unbiased stochastic estimator of the subgradient $\partial f(x)$ and the gradient $\nabla f(x)$ respectively. A function $f : \mathcal{D} \to \mathbb{R}$ is said to be L-smooth if $\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|$ for all $x, x' \in \mathcal{D}$. A function $f : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is δ-weakly convex if $f(\cdot) + \frac{\delta}{2}\|\cdot\|^2$ is convex. A mapping $\mathcal{M} : \mathcal{D} \to \mathbb{R}^l$ is said to be C-Lipschitz continuous if $\|\mathcal{M}(x) - \mathcal{M}(x')\| \leq C\|x - x'\|$ for all $x, x' \in \mathcal{D}$.

Consider solving a non-smooth problem $\min_x f(x)$. One of the main challenges is that the ϵ-stationary point, i.e., a point x such that $\text{dist}(0, \partial f(x)) \leq \epsilon$, which is the typical goal for smooth problems, may not exist in the neighborhood of its optimal solution. A classical counter example would be $f(x) = |x|$, where for $\epsilon \in [0, 1)$ the only ϵ-stationary point is the optimal solution $x = 0$. A standard solution to this issue in weakly-convex setting is to use a relaxed convergence criteria, that is to find a point no more than ϵ away from an ϵ-stationary point. This is called a nearly ϵ-stationary point, and is widely used in non-smooth weakly-convex optimization literature [4, 29, 41, 50, 51, 19]. In fact, finding a nearly ϵ-stationary point for $f(x)$ can be achieved by finding an ϵ-stationary point of $f_\gamma(x)$, the Moreau envelope of $f(x)$. Assume function f is δ-weakly-convex, then its Moreau envelope and proximal map are given by

$$f_\gamma(x) := \min_{x'} \Big\{ f(x') + \frac{1}{2\gamma}\|x' - x\|^2 \Big\}, \quad \text{prox}_{\gamma f}(x) := \arg\min_{x'} \Big\{ f(x') + \frac{1}{2\gamma}\|x' - x\|^2 \Big\}.$$

Existing work [3] has shown that with $\gamma \in (0, \delta^{-1})$ and $\hat{x} = \text{prox}_{\gamma f}(x)$, we have

$$\nabla f_\gamma(x) = \gamma^{-1}(x - \hat{x}), \quad f(\hat{x}) \leq f(x), \quad \text{dist}(0, \partial f(\hat{x})) \leq \|\nabla f_\gamma(x)\|.$$

Moreover, $\text{prox}_{\gamma f}(x)$ is $\frac{1}{1-\gamma\delta}$ - Lipschitz continuous [32].

Now we consider the DMax problem (1). By Danskin's Theorem, the weak convexity assumption of $\phi(\cdot, y)$ and $\psi(\cdot, z)$ naturally leads to the weak convexity of $\Phi(\cdot)$ and $\Psi(\cdot)$. Since the weak convexity assumption of component functions does not guarantee the weak convexity of their difference function $F(x)$, one may neither 1) use nearly ϵ-stationary point of $F(x)$ as the convergence metric, nor 2) directly apply Moreau envelope smoothing technique to $F(x)$. To tackle the first issue, we follow the existing work [45] to use the following convergence metric for non-smooth DWC problems.

Definition 3.2 (Definition 2 in [45])**.** Given $\epsilon > 0$, we say x is a **nearly ϵ-critical point** of $\min_x\{F(x) := \Phi(x) - \Psi(x)\}$ if there exist v, x', x'' such that $v \in \partial\Phi(x') - \partial\Psi(x'')$ and $\max\{\mathbb{E}\|v\|, \mathbb{E}\|x - x'\|, \mathbb{E}\|x - x''\|\} \leq \epsilon$.

To tackle the second issue, we take the Moreau envelope of $\Phi(\cdot)$ and $\Psi(\cdot)$ individually and define the smooth approximation of $F(x)$ as

$$F_\gamma(x) = \Phi_\gamma(x) - \Psi_\gamma(x). \tag{4}$$

The recent work [32] has proven that $F_\gamma(x)$ is indeed smooth.

Algorithm 1 Stochastic Moreau Envelope Approximate Gradient Method (SMAG)

Input Initial points: x_ϕ^0, x_ψ^0, x_0, y_0, z_0. Hyper-parameters: γ, η_0, η_1.
Stochastic (sub)gradients: $\tilde{\partial}_x\phi$, $\tilde{\nabla}_y\phi$, $\tilde{\partial}_x\psi$, $\tilde{\nabla}_z\psi$.
for $t = 0, \ldots, T-1$ **do**
  $x_\phi^{t+1} = x_\phi^t - \eta_1(\tilde{\partial}_x\phi(x_\phi^t, y_t) + \frac{1}{\gamma}(x_\phi^t - x_t))$
  $y_{t+1} = P_{\mathcal{Y}}\big(y_t + \eta_1\tilde{\nabla}_y\phi(x_\phi^t, y_t)\big)$
  $x_\psi^{t+1} = x_\psi^t - \eta_1(\tilde{\partial}_x\psi(x_\psi^t, z_t) + \frac{1}{\gamma}(x_\psi^t - x_t))$
  $z_{t+1} = P_{\mathcal{Z}}\big(z_t + \eta_1\tilde{\nabla}_z\psi(x_\psi^t, z_t)\big)$
  $G_{t+1} = \frac{1}{\gamma}(x_t - x_\phi^{t+1}) - \frac{1}{\gamma}(x_t - x_\psi^{t+1})$
  $x_{t+1} = x_t - \eta_0 G_{t+1}$
end for
return $x_\phi^{\bar{t}}$ or $x_\psi^{\bar{t}}$ with $\bar{t}$ uniformly sampled from $\{1, \ldots, T\}$

Proposition 3.3 (Proposition EC.1.2 in [32]). *Assume $\Phi(\cdot)$ and $\Psi(\cdot)$ are δ_ϕ, δ_ψ-weakly convex respectively. Then $F_\gamma(x) = \Phi_\gamma(x) - \Psi_\gamma(x)$ is L_F-smooth, where $L_F = \frac{2}{\gamma - \gamma^2\min\{\delta_\psi, \delta_\phi\}}$.*

Moreover, one can show that a good approximate stationary point x of $F_\gamma(\cdot)$ and a good approximation point x' to the proximal points $\text{prox}_{\gamma\Phi}(x)$ and $\text{prox}_{\gamma\Psi}(x)$ can guarantee that x' is a nearly ϵ-critical point of $\min_{\hat{x}} F(\hat{x})$.

Lemma 3.4 (Lemma 3 in [45]). *Assume $\Phi(\cdot)$ and $\Psi(\cdot)$ are δ_ϕ, δ_ψ-weakly convex respectively, and $0 < \gamma < \min\{\delta_\phi^{-1}, \delta_\psi^{-1}\}$. If x is a vector such that $\mathbb{E}[\|\nabla F_\gamma(x)\|^2] \le \min\{1, \gamma^{-2}\}\epsilon^2/4$, and x' is a vector such that $\mathbb{E}[\|x' - prox_{\gamma\Phi}(x)\|^2] \le \epsilon^2/4$ or $\mathbb{E}[\|x' - prox_{\gamma\Psi}(x)\|^2] \le \epsilon^2/4$, then x' is a nearly ϵ-critical point of $\min_{\hat{x}} F(\hat{x})$.*

4 Algorithms and Convergence

Since we aim to minimize the smooth function $F_\gamma(x)$, the natural strategy is to perform gradient descent to update the variable x. Following from the properties of Moreau envelope, the gradient of $F_\gamma(x)$ is given by

$$\nabla F_\gamma(x) = \frac{1}{\gamma}(x - \text{prox}_{\gamma\Phi}(x)) - \frac{1}{\gamma}(x - \text{prox}_{\gamma\Psi}(x)), \tag{5}$$

where the blue component is the gradient of $\Phi_\gamma(x)$ and the green component is the gradient of $\Psi_\gamma(x)$. However, the proximal points $\text{prox}_{\gamma\Psi}(x)$ and $\text{prox}_{\gamma\Phi}(x)$ are not accessible in general. Indeed, these proximal points are the optimal solutions to $\min_{x'}\{\Phi(x') + \frac{1}{2\gamma}\|x - x'\|^2\}$ and $\min_{x'}\{\Psi(x') + \frac{1}{2\gamma}\|x - x'\|^2\}$ respectively, and $\Phi(\cdot)$ and $\Psi(\cdot)$ are typically not accessible because they are the value functions of possibly sophisticated maximization problems. Thus, we maintain two variables x_ϕ^t and x_ψ^t as the estimators of $\text{prox}_{\gamma\Phi}(x_t)$ and $\text{prox}_{\gamma\Psi}(x_t)$ respectively, and maintain another two variables y_t and z_t as the estimators of $\arg\max_{y\in\mathcal{Y}} \phi(\text{prox}_{\gamma\Phi}(x_t), y)$ and $\arg\max_{z\in\mathcal{Z}} \psi(\text{prox}_{\gamma\Psi}(x_t), z)$ respectively. At each iteration, we update x_ϕ^t and x_ψ^t by one step of stochastic gradient descent, and update y_t and z_t by one step of stochastic gradient ascent. Finally, we compute the gradient estimator $G_{t+1} = \frac{1}{\gamma}(x_t - x_\phi^{t+1}) - \frac{1}{\gamma}(x_t - x_\psi^{t+1})$ of $\nabla F_\gamma(x_t)$ and update x_t by one step of gradient descent. The resulting algorithm is presented in Algorithm 1.

DWC Optimization. For DWC problem (2), the associated functions $\Phi(\cdot) = \phi(\cdot)$ and $\Psi(\cdot) = \psi(\cdot)$ are directly accessible. Thus the variables y_t and z_t in SMAG are no longer needed. The simplified SMAG algorithm for DWC optimization is presented in Algorithm 2.

WCSC Min-Max Optimization. For WCSC Min-Max problem (3), the second component function $\Psi = 0$ can be ignored, and thus variables x_ψ^t and z_t are no longer needed. However, this brings a

https://doi.org/10.52202/079017-1807

change to the gradient of $F_\gamma(x_t)$ as it now becomes

$$\nabla F_\gamma(x_t) = \gamma^{-1}(x_t - \text{prox}_{\gamma\Phi}(x_t)).$$

The simplified SMAG algorithm for WCSC Min-Max optimization is presented in Algorithm 3.

Algorithm 2 SMAG for DWC Optimization

for $t = 0, \ldots, T-1$ **do**
  $x_\phi^{t+1} = x_\phi^t - \eta_1(\tilde{\partial}_x \phi(x_\phi^t) + \frac{1}{\gamma}(x_\phi^t - x_t))$
  $x_\psi^{t+1} = x_\psi^t - \eta_1(\tilde{\partial}_x \psi(x_\psi^t) + \frac{1}{\gamma}(x_\psi^t - x_t))$
  $G_{t+1} = \frac{1}{\gamma}(x_\psi^{t+1} - x_\phi^{t+1})$
  $x_{t+1} = x_t - \eta_0 G_{t+1}$
end for
return $x_\phi^{\bar{t}}$ or $x_\psi^{\bar{t}}$ with $\bar{t} \sim \{1, \ldots, T\}$

Algorithm 3 SMAG for WCSC Min-Max Optimization

for $t = 0, \ldots, T-1$ **do**
  $x_\phi^{t+1} = x_\phi^t - \eta_1(\tilde{\partial}_x \phi(x_\phi^t, y_t) + \frac{1}{\gamma}(x_\phi^t - x_t))$
  $y_{t+1} = P_{\mathcal{Y}}\big(y_t + \eta_1 \tilde{\nabla}_y \phi(x_\phi^t, y_t)\big)$
  $G_{t+1} = \frac{1}{\gamma}(x_t - x_\phi^{t+1})$
  $x_{t+1} = x_t - \eta_0 G_{t+1}$
end for
return $x_{\bar{t}}$ with $\bar{t} \sim \{0, \ldots, T-1\}$

4.1 Convergence Analysis

In this section, we present convergence results for Algorithms 1-3. To proceed, we make the following assumption for DMax problem (1).

Assumption 4.1. Considering DMax problem (1), we assume that

(i) $\phi(\cdot, y)$ is δ_ϕ-weakly convex, and $\psi(\cdot, z)$ is δ_ψ-weakly convex.

(ii) $\phi(x, \cdot)$ is μ_ϕ-strongly concave, and $\psi(x, \cdot)$ is μ_ψ-strongly concave.

(iii) $\phi(x, y)$ and $\psi(x, z)$ are differentiable in terms of y and z respectively, $\nabla_y \phi(\cdot, y)$ is $L_{\phi,yx}$-Lipschitz continuous, and $\nabla_z \psi(\cdot, z)$ is $L_{\psi,zx}$-Lipschitz continuous.

(iv) There exists a constant $F_\gamma^* > -\infty$ such that $F_\gamma^* \leq F_\gamma(x)$ for all x.

(v) There exists a finite constant M such that $\mathbb{E}\|\tilde{\partial}_x \phi(x, y)\|^2 \leq M^2$, $\mathbb{E}\|\tilde{\nabla}_y \phi(x, y)\|^2 \leq M^2$, $\mathbb{E}\|\tilde{\partial}_x \psi(x, z)\|^2 \leq M^2$, $\mathbb{E}\|\tilde{\nabla}_z \psi(x, z)\|^2 \leq M^2$ for all $x \in \mathbb{R}^{d_x}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$.

It shall be noted that Assumption 4.1(iii) only requires partial smoothness of ϕ and ψ, and is to ensure the Lipschitz continuity of $y^*(\cdot) := \arg\max_{y \in \mathcal{Y}} \phi(\cdot, y)$ and $z^*(\cdot) := \arg\max_{z \in \mathcal{Z}} \psi(\cdot, z)$. This follows from existing results.

Lemma 4.2 (Lemma 4.3 in [19])**.** *Consider problem* $\max_{y \in \hat{\mathcal{Y}}} f(x, y)$ *for any* $x \in \mathbb{R}^{d_x}$*, where* $\hat{\mathcal{Y}} \subset \mathbb{R}^{d_y}$ *is a closed convex set. Assume that* $f(x, y)$ *is* μ*-strongly concave in* y *for each* $x \in \mathbb{R}^{d_x}$*, and* $\nabla_y f(\cdot, y)$ *is* L_{yx}*-Lipschitz for each* $y \in \hat{\mathcal{Y}}$*. Then* $\arg\max_y f(\cdot, y)$ *is* $\frac{L_{yx}}{\mu}$*-Lipschitz continuous.*

A Lipschitz smooth function $f(x, y)$ is guaranteed to have Lipschitz continuous partial gradient $\nabla_y f(\cdot, y)$, while the reverse statement is not necessarily true. For example, consider a function $f(x, y) = y^\top h(x) - g(y)$ with non-smooth C-Lipschitz continuous $h(\cdot)$ and strongly convex g. Then $f(x, y)$ is non-smooth but the partial subgradient $\nabla_y f(\cdot, y) = h(\cdot) - \nabla g(y)$ is Lipschitz continuous with respect to the first argument. Another example is given by $f(x, y) = f_1(x) + f_2(x, y)$, where f_1 is weakly convex and f_2 is smooth and strongly concave in terms of y. The latter is indeed seen in our considered application for pAUC maximization with adversarial fairness. In fact, one may replace Assumption 4.1(iii) by directly assuming that $y^*(\cdot)$ and $z^*(\cdot)$ are Lipschitz continuous. In addition, Assumption 4.1(v) is standard in non-smooth optimization literature [3, 39, 10].

Here we give a brief outline of the convergence analysis. First of all, we present a standard result [7].

Lemma 4.3. *Suppose that* $F_\gamma(\cdot)$ *is* L_F*-smooth and* $x_{t+1} = x_t - \eta_0 G_{t+1}$ *with* $0 < \eta_0 \leq \frac{1}{2L_F}$*. Then we have*

$$F_\gamma(x_{t+1}) \leq F_\gamma(x_t) + \frac{\eta_0}{2}\|\nabla F_\gamma(x_t) - G_{t+1}\|^2 - \frac{\eta_0}{2}\|\nabla F_\gamma(x_t)\|^2 - \frac{\eta_0}{4}\|G_{t+1}\|^2.$$

This implies that the key to bounding the gradient $\|\nabla F_\gamma(x_t)\|^2$ is to obtain a recursive bound for the gradient estimation error $\|\nabla F_\gamma(x_t) - G_{t+1}\|^2$. Following from the true gradient formulation 5, we have

$$\|\nabla F_\gamma(x_t) - G_{t+1}\|^2 \leq \frac{2}{\gamma^2}\left(\|x_\phi^{t+1} - \text{prox}_{\gamma\Phi}(x_t)\|^2 + \|x_\psi^{t+1} - \text{prox}_{\gamma\Psi}(x_t)\|^2\right). \tag{6}$$

In other words, the error of the gradient estimation G_{t+1} can be bounded by the estimation errors of x_ϕ^{t+1} and x_ψ^{t+1}. Thus, we construct recursive bound for the proximal point estimation errors $\|x_\phi^{t+1} - \text{prox}_{\gamma\Phi}(x_t)\|^2$ and $\|x_\psi^{t+1} - \text{prox}_{\gamma\Psi}(x_t)\|^2$ individually. In fact, these two errors share almost identical analysis due to similar assumptions and updates. Here we only present the result for function ϕ, as the result for ψ directly follows.

Lemma 4.4. *Suppose that Assumption 4.1 holds,* $0 < \gamma < 1/\delta_\phi$*, and* $\eta_1 \leq \frac{\gamma^2(1/\gamma - \delta_\phi)}{2}$*. Then the sequences* $\{x_t\}$*,* $\{y_t\}$*,* $\{x_\phi^t\}$ *and* $\{G_t\}$ *generated by Algorithm 1 satisfy*

$$\begin{aligned}
&\mathbb{E}\|x_\phi^{t+1} - \textit{prox}_{\gamma\Phi}(x_t)\|^2 + \mathbb{E}_t\|y_{t+1} - y^*(\textit{prox}_{\gamma\Phi}(x_t))\|^2 \\
&\leq (1 - \frac{\eta_1(1/\gamma - \delta_\phi)}{2})\mathbb{E}\|x_\phi^t - \textit{prox}_{\gamma\Phi}(x_{t-1})\|^2 + (1 - \eta_1\mu_\phi)\mathbb{E}\|y_t - y^*(\textit{prox}_{\gamma\Phi}(x_{t-1}))\|^2 \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma - \delta_\phi)^3} + \frac{L_{\phi,yx}^2\eta_0^2}{\eta_1\mu_\phi^3\gamma^2(1/\gamma - \delta_\phi)^2}\right)\mathbb{E}\|G_t\|^2 + 12M^2\eta_1^2.
\end{aligned}$$

Finally, combining Lemma 4.3, inequality (6) and Lemma 4.4 yields the following convergence result for Algorithm 1.

Theorem 4.5. *Suppose that Assumption 4.1 holds,* $0 < \gamma < \min\{\delta_\phi^{-1}, \delta_\psi^{-1}\}$*,* $\eta_1 = \mathcal{O}(\epsilon^2)$*, and* $\eta_0 = \tau\eta_1$*. Then after* $T \geq \mathcal{O}(\epsilon^{-4})$ *iterations, the sequences* $\{x_t\}$*,* $\{x_\phi^t\}$ *and* $\{x_\psi^t\}$ *generated by Algorithm 1 satisfy* $\mathbb{E}[\|x_\phi^{\bar{t}} - \textit{prox}_{\gamma\Phi}(x_{\bar{t}-1})\|^2 + \|x_\psi^{\bar{t}} - \textit{prox}_{\gamma\Psi}(x_{\bar{t}-1})\|^2 + \|\nabla F_\gamma(x_{\bar{t}-1})\|^2] \leq \min\{1, \gamma^{-2}\}\epsilon^2/4$*, and the outputs* $x_\phi^{\bar{t}}$ *and* $x_\psi^{\bar{t}}$ *are both nearly* ϵ*-critical points of problem (1).*

Since DMax optimization is a unified framework covering DWC optimization and WCSC min-max optimization, the convergence results of Algorithms 2 and 3 directly follow from Theorem 4.5. To present them, we first provide a reduced version of Assumption 4.1 for DWC problem (2).

Assumption 4.6. Considering DWC problem (2), we assume that

(i) $\phi(\cdot)$ is δ_ϕ-weakly convex, and $\psi(\cdot)$ is δ_ψ-weakly convex.

(ii) There exists a constant $F_\gamma^* > -\infty$ such that $F_\gamma^* \leq F_\gamma(x)$ for all x.

(iii) There exists a finite constant M such that $\mathbb{E}\|\tilde{\partial}\phi(x)\|^2 \leq M^2$ and $\mathbb{E}\|\tilde{\partial}\psi(x)\|^2 \leq M^2$ for all $x \in \mathbb{R}^{d_x}$.

By setting $\phi(x, y) = \phi(x)$ and $\psi(x, z) = \psi(x)$, namely independent of y and z, in DMax problem (1), we obtain the following convergence result for Algorithm 2, which is an immediate consequence of Theorem 4.5.

Corollary 4.7. *Suppose that Assumption 4.6 holds,* $0 < \gamma < \min\{\delta_\phi^{-1}, \delta_\psi^{-1}\}$*,* $\eta_1 = \mathcal{O}(\epsilon^2)$*, and* $\eta_0 = \tau\eta_1$*. Then after* $T \geq \mathcal{O}(\epsilon^{-4})$ *iterations, the outputs* $x_\phi^{\bar{t}}$ *and* $x_\psi^{\bar{t}}$ *of Algorithm 2 are both nearly* ϵ*-critical points of problem (2).*

For WCSC min-max problem (3), we reduce Assumption 4.1 to the following.

Assumption 4.8. Considering WCSC min-max problem (3), we assume that

(i) $\phi(\cdot, y)$ is δ_ϕ-weakly convex, and $\phi(x, \cdot)$ is μ_ϕ-strongly concave.

(ii) $\phi(x, y)$ is differentiable in terms of y, and $\nabla_y\phi(\cdot, y)$ is $L_{\phi,yx}$-Lipschitz continuous.

(iii) There exists a constant $F_\gamma^* > -\infty$ such that $F_\gamma^* \leq F_\gamma(x)$ for all x.

(iv) There exists a finite constant M such that $\mathbb{E}\|\tilde{\partial}_x\phi(x, y)\|^2 \leq M^2$ and $\mathbb{E}\|\tilde{\nabla}_y\phi(x, y)\|^2 \leq M^2$ for all $x \in \mathbb{R}^{d_x}$ and $y \in \mathcal{Y}$.

https://doi.org/10.52202/079017-1807

By setting $\psi(x, z) = 0$ in DMax problem (1), we obtain the following convergence result for Algorithm 3, which is an immediate consequence of Theorem 4.5.

Corollary 4.9. *Suppose that Assumption 4.8 holds,* $0 < \gamma < 1/\delta_\phi$, $\eta_1 = \mathcal{O}(\epsilon^2)$, *and* $\eta_0 = \tau\eta_1$, *Then after* $T \geq \mathcal{O}(\epsilon^{-4})$ *iterations, the output* $x_{\bar{t}}$ *of Algorithm 3 is a nearly* ϵ*-stationary point of problem (3).*

It shall be mentioned that for WCSC min-max problem (3, we use nearly ϵ-stationary point as the convergence metric. This is standard in weakly-convex optimization literature [3].

5 Applications

In this section, we introduce two applications of DMax optimization, PU learning for DWC optimization and partial AUC optimization with adversarial fairness regularization for WCSC min-max optimization. We also show experimental results on both applications.

5.1 Positive-Unlabeled Learning

In binary classification task, the optimization problem is commonly formulated as the minimization of empirical risk, i.e., $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}_i\in\mathcal{S}} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ where $\ell(\mathbf{w}; \mathbf{x}_i, y_i)$ is the loss given the model parameter $\mathbf{w}$ on a data point $\mathbf{x}_i$ and its ground truth label y_i. Given the scenario where only positive data $\mathcal{S}_+$ are observed, then the standard approach becomes problematic. One way to address this issue is to utilize unlabeled data $\mathcal{S}_u$ to construct unbiased risk estimators. To be specific, [13] formulated the PU learning problem as following

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\pi_p}{n_+} \sum_{\mathbf{x}_i\in\mathcal{S}_+} [\ell(\mathbf{w}; \mathbf{x}_i, +1) - \ell(\mathbf{w}; \mathbf{x}_i, -1)] + \frac{1}{n_u} \sum_{\mathbf{x}_j^u\in\mathcal{S}_u} \ell(\mathbf{w}; x_j^u, -1) \tag{7}$$

where $n_+ = |\mathcal{S}_+|$, $n_u = |\mathcal{S}_u|$, $\pi_p = Pr(y = 1)$ is the prior probability of the positive class. If $\ell(\mathbf{w}; \mathbf{x}, y)$ is weakly convex in terms of $\mathbf{w}$, then Problem (7) is a DWC problems. In particular, in our experiments we consider linear classification model and hinge loss.

Baselines. We implemented five baselines and compared them with our proposed method SMAG for DWC optimization. The first baseline, stochastic gradient descent (SGD), does not have theoretical convergence guarantee for DWC problems. However, since it is the fundamental method for convex optimization, we include it to show its performance. We also implemented existing stochastic methods for solving DC or DWC problems with non-smooth components, including SDCA [26], SSDC-SPG [39], SSDC-Adagrad [39] and SBCD [45].

Datasets. We use four multi-class classification datasets, Fashion-MNIST [36], MNIST [5] CIFAR10 [14] and FER2013 [6]. To fit them in binary classification task, we consider the first five classes as negative for Fashion-MNIST, MNIST and CIFAR10, and the first four classes as negative for FER2013. For Fashion-MNIST, MNIST, CIFAR10, we follow the standard train-test split. For FER2013, we take the first 25709 samples as the training data, and the rest as for testing.

Setup. For all datasets, we use a batch size of 64 and set $\pi_p = 0.5$. We train 40 epochs and decay the learning rate by 10 at epoch 12 and 24. The learning rates of SGD, SDCA, SSDC-SPG and SSDC-Adagrad, the learning rate of the inner loop of SBCD (i.e., $\mu\eta_t/(\mu + \eta_t)$), and η_1 in SMAG are all tuned from $\{10, 1, 0.2, 0.1, 0.01, 0.001\}$. The learning rate of the outer loop in SDCA and η_0 in SMAG are tuned from $\{0.1, 0.5, 0.9\}$. The numbers of inner loops for all double-loop methods are tuned from $\{2, 5, 10\}$. The μ in SBCD, $1/\gamma$ in SSDC-SPG and SSDC-Adagrad, γ in SMAG are tuned in $\{0.05, 0.1, 0.2, 0.5, 1, 2\}$. We run 4 trails for each setting and plot the average curves.

Results. We plot the curves of training losses in Figure 1. For all tested datasets, the performance of SMAG surpasses the baselines. Among the baselines, SBDC is the generally the next best choice. However, since SBDC is a double-loop method, it has one more hyperparameter compared to SMAG. We also present the ablation study of SMAG regarding the parameter γ in Figure 2 included in the Appendix.

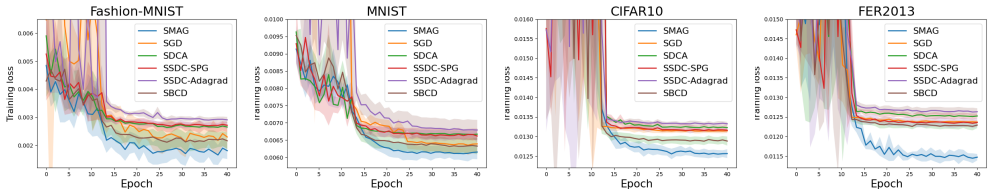

Figure 1: Training Curves of PU Learning

5.2 Partial AUC Maximization with Fairness Regularization

AUC Maximization aims to maximize the area under the curve of true positive rate (TPR) vs false positive rate (FPR). It has been studied extensively [44, 46, 20, 9] and has shown great success in large-scale real-world tasks, e.g., medical image classification [46] and molecular properties prediction [35]. One-way partial AUC (OPAUC) is an extension of AUC that has a primary interest in the curve corresponding to low FPR. To be specific, OPAUC restrict the FPR to the region $[0, \rho]$ where $\rho \in (0, 1)$. A recent work [52] proposed to formulate OPAUC problem into a non-smooth weakly convex optimization problem using conditional-value-at-risk (CVaR) based distributionally robust optimization (DRO). The formulation is given by

$$\min_{\mathbf{w},\mathbf{s}\in\mathbb{R}^{n_+}} F_{\text{pauc}}(\mathbf{w}, \mathbf{s}) = \frac{1}{n_+} \sum_{\mathbf{x}_i \in \mathcal{S}_+} \left(s_i + \frac{1}{\rho n_-} \sum_{\mathbf{x}_j \in \mathcal{S}_-} (L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - s_i)_+ \right), \tag{8}$$

where $\mathcal{S}_+, \mathcal{S}_-$ are the sets of positive and negative samples respectively, $n_+ = |\mathcal{S}_+|$, $n_- = |\mathcal{S}_-|$, and $\mathbf{w}$ denotes the weights of encoder network and classification layer. The pairwise surrogate loss is defined by $L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) = \ell(h(\mathbf{w}, \mathbf{x}_i) - h(\mathbf{w}, \mathbf{x}_j))$ and we use squared hinge loss as the surrogate loss, i.e., $\ell(\cdot) = (c - \cdot)^2$, where $c > 0$ is a parameter.

However, directly solving the above problem may end up with a model that is unfair with respect to some protected groups (e.g., female patients). Hence, we consider a formulation that incorporates an adversarial fairness regularization:

$$\max_{\mathbf{w}_a} F_{\text{fair}}(\mathbf{w}, \mathbf{w}_a) := \mathbb{E}_{(\mathbf{x},a)\sim\mathcal{D}_a} \left\{ \mathbb{I}(a = 1) \log(\sigma(\mathbf{w}, \mathbf{w}_a, \mathbf{x})) + \mathbb{I}(a = -1) \log(1 - \sigma(\mathbf{w}, \mathbf{w}_a, \mathbf{x})) \right\},$$

where $\sigma(\mathbf{w}, \mathbf{w}_a, \mathbf{x})$ denotes a predicted probability that the data has a sensitive attribute $a = 1$ by using a classification head $\mathbf{w}_a$ on top of the encoded representation of $\mathbf{x}$. This adversarial fairness regularization has been demonstrated effective for promoting fairness [37]. As a result, we consider OPAUC problem with a fairness regularization:

$$\min_{\mathbf{w},\mathbf{s}\in\mathbb{R}^{n_+}} \max_{\mathbf{w}_a} F_{\text{pauc}}(\mathbf{w}, \mathbf{s}) + \alpha F_{\text{fair}}(\mathbf{w}, \mathbf{w}_a) + \frac{\lambda_0}{2} \|\mathbf{w}_a\|_2^2 \tag{9}$$

It is clear that the problem is WCSC.

Baseline. We implement our proposed method SMAG for solving OPAUC problem (8) and OPAUC problem with adversarial fairness regularization (9). We refer the former as SMAG* and the latter as SMAG. The baseline on OPAUC problem (8) is SOPA, proposed in [52]. The baselines on OPAUC problem with adversarial fairness regularization (9) are SGDA [19] and Epoch-GDA [40].

Dataset. CelebA contains 200k celebrity face images with 40 binary attributes each, including the gender-sensitive attribute denoted as *Male*. In our experiments, we conduct experiments on three independent attribute prediction tasks: *Attractive, Big Nose, and Bags Under Eyes*, which have high Pearson correlations [2, 27] with the sensitive attribute *Male*. We divide the dataset into training, validation, and test data with an 80%/10%/10% split.

Setup. For all experiments, we adopt ResNet-18 as our backbone model architecture and initialize it with ImageNet pre-trained weights. The batch size is 128. We set the FPR upper bound to be $\rho = 0.3$. We train the model for 3 epochs with cosine decay learning rates for all baselines. The regularizer parameter α is tuned in $0.1, 0.2, 0.5$ for SGDA, Epoch-GDA, and SMAG, and the adversarial learning rates are tuned in $0.001, 0.01, 0.1$. $\alpha = 0$ for SOPA and SMAG*. The initial learning rates for optimizing $\mathbf{w}$ are tuned in $0.1, 0.01, 0.001$ for all methods, while the weight interpolation parameters,

https://doi.org/10.52202/079017-1807

i.e., γ in Epoch-GDA and SMAG, are also tuned in $0.1, 0.01, 0.001$. The inner loop step is tuned in $\{5, 10, 15\}$ for Epoch-GDA. η_1 in SMAG are tuned from $\{10, 1, 0.2, 0.1, 0.01, 0.001\}$.

Results. We report the experimental results on three fairness metrics [27], equalized odds difference (EOD), equalized opportunity (EOP), and demographic disparity (DP) in Table 3. We observe that SMAG consistently achieves the highest pAUC score and lowest disparities metrics across all tasks compared to all other baseline min-max methods.

Table 3: Mean $\pm$ std of fairness results on CelebA test dataset with *Attractive and Big Nose* task labels, and *Male* sensitive attribute. Results are reported on 3 independent runs. We use bold font to denote the best result and use underline to denote the second best. Results on *Bags Under Eyes* are included in the appendix due to limited space.

	Attractive, Male				Big Nose, Male			
Methods	**pAUC**↑	**EOD**↓	**EOP**↓	**DP**↓	**pAUC**↑	**EOD**↓	**EOP**↓	**DP**↓
SOPA	0.8485 ± 0.012	0.2638 ± 0.035	0.2438 ± 0.032	0.4753 ± 0.023	0.8039 ± 0.005	0.2829 ± 0.024	0.2269 ± 0.019	0.4424 ± 0.034
SMAG*	**0.8606** ± 0.003	$\underline{0.2192} \pm 0.020$	0.2333 ± 0.068	0.4510 ± 0.027	**0.8078** ± 0.002	$\underline{0.2735} \pm 0.012$	$\underline{0.2205} \pm 0.030$	$\underline{0.4364} \pm 0.019$
SGDA	0.8509 ± 0.001	0.2701 ± 0.020	0.2549 ± 0.025	0.4860 ± 0.015	0.8038 ± 0.002	0.2846 ± 0.023	0.2398 ± 0.029	0.4390 ± 0.028
EGDA	0.8546 ± 0.004	0.2290 ± 0.006	$\underline{0.1735} \pm 0.059$	$\underline{0.4305} \pm 0.032$	0.8023 ± 0.005	0.3293 ± 0.027	0.3076 ± 0.012	0.4620 ± 0.031
SMAG	$\underline{0.8605} \pm 0.002$	**0.1900** ± 0.023	**0.1648** ± 0.064	**0.4116** ± 0.031	$\underline{0.8058} \pm 0.001$	**0.2708** ± 0.021	**0.2148** ± 0.021	**0.4333** ± 0.013

6 Conclusion

In this study, we have introduced a new framework namely DMax optimization, that unifies DWC optimization and non-smooth WCSC min-max optimization. We proposed a single-loop stochastic method for solving DMax optimization and presented a novel convergence analysis showing that the proposed method achieves a non-asymptotic convergence rate of $\mathcal{O}(\epsilon^{-4})$. Experimental results on two applications, PU learning and OPAUC optimization with adversarial fairness regularization demonstrate strong performance of our method. One limitation of this work is the strong convexity assumption on the $\phi(x, \cdot)$ and $\psi(x, \cdot)$. This strong assumption may limit the applicability of our method. Future work will focus on exploring DMax optimization with weaker assumptions.

Acknowledgment

We thank anonymous reviewers for constructive comments. Q. Hu and T. Yang were partially supported by the National Science Foundation Career Award 2246753, the National Science Foundation Award 2246757, 2246756 and 2306572. Z. Lu was partially supported by the National Science Foundation Award IIS-2211491, the Office of Naval Research Award N00014-24-1-2702, and the Air Force Office of Scientific Research Award FA9550-24-1-0343.

References

[1] Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM J. Optim.*, 33:1884–1913, 2020.

[2] Luigi Celona, Simone Bianco, and Raimondo Schettini. Fine-grained face annotation using deep multi-task cnn. *Sensors*, 18(8):2666, 2018.

[3] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions, 2018.

[4] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[5] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[6] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests, 2013.

[7] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family and beyond, 2022.

[8] Mert Gürbüzbalaban, A. Ruszczynski, and Landi Zhu. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 194:1014 – 1041, 2022.

[9] Quanqi Hu, Yongjian Zhong, and Tianbao Yang. Multi-block min-max bilevel optimization with applications in multi-task deep auc maximization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29552–29565. Curran Associates, Inc., 2022.

[10] Quanqi Hu, Dixian Zhu, and Tianbao Yang. Non-smooth weakly-convex finite-sum coupled compositional optimization. *ArXiv*, abs/2310.03234, 2023.

[11] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *ArXiv*, abs/2008.08170, 2020.

[12] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 13–18 Jul 2020.

[13] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *ArXiv*, abs/1703.00593, 2017.

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

[15] Hoai An Le Thi, Van Ngai Huynh, Tao Pham Dinh, and Hoang Phuc Hau Luu. Stochastic difference-of-convex-functions algorithms for nonconvex programming. *SIAM Journal on Optimization*, 32(3):2263–2293, 2022.

[16] Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran. Stochastic dca for minimizing a large sum of dc functions with application to multi-class logistic regression. *Neural Networks*, 132:220–231, 2020.

[17] Hoai An Le Thi, Hoang Phuc Hau Luu, and Tao Pham Dinh. Online stochastic dca with applications to principal component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):7035–7047, 2024.

 https://doi.org/10.52202/079017-1807

[18] Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169, 01 2018.

[19] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6083–6093. PMLR, 13–18 Jul 2020.

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

[21] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20566–20577. Curran Associates, Inc., 2020.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[23] Gabriel Mancino-Ball and Yangyang Xu. Variance-reduced accelerated methods for decentralized stochastic double-regularized nonconvex strongly-concave minimax problems. *ArXiv*, abs/2307.07113, 2023.

[24] J.J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[25] Abdellatif Moudafi. A Regularization of DC Optimization. *Pure and Applied Functional Analysis*, 2022.

[26] Atsushi Nitanda and Taiji Suzuki. Stochastic Difference of Convex Algorithm and its Application to Training Deep Boltzmann Machines. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 470–478. PMLR, 20–22 Apr 2017.

[27] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022.

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

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

[30] R.T. Rockafellar, M. Wets, and R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009.

[31] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training, 2020.

[32] Kaizhao Sun and Xu Andy Sun. Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing. *INFORMS J. Optim.*, 5:321–339, 2022.

[33] Pham Dinh Tao and El Bernoussi Souad. Algorithms for solving a class of nonconvex optimization problems. methods of subgradients. *North-holland Mathematics Studies*, 129:249–271, 1986.

[34] Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran. Stochastic DCA for the large-sum of non-convex functions problem and its application to group variable selection in classification. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3394–3403. PMLR, 06–11 Aug 2017.

[35] Zhengyang Wang, Meng Liu, Youzhi Luo, Zhao Xu, Yaochen Xie, Limei Wang, Lei Cai, Qi Qi, Zhuoning Yuan, Tianbao Yang, and Shuiwang Ji. Advanced graph and sequence neural networks for molecular property prediction and drug discovery. *Bioinformatics*, 38(9):2579–2586, 02 2022.

[36] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[37] Qizhe Xie, Zihang Dai, Yulun Du, Eduard H. Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Neural Information Processing Systems*, 2017.

[38] Tengyu Xu, Zhe Wang, Yingbin Liang, and H. Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *ArXiv*, abs/2006.09361, 2020.

[39] Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6942–6951. PMLR, 2019.

[40] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5789–5800. Curran Associates, Inc., 2020.

[41] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization. *arXiv preprint arXiv:2002.05309*, 2020.

[42] Junchi Yang, Xiang Li, and Niao He. Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. *ArXiv*, abs/2206.00743, 2022.

[43] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5485–5517. PMLR, 28–30 Mar 2022.

[44] Tianbao Yang and Yiming Ying. AUC maximization in the era of big data and AI: A survey. *ACM Comput. Surv.*, 55(8):172:1–172:37, 2023.

[45] Yao Yao, Qihang Lin, and Tianbao Yang. Large-scale optimization of partial auc in a range of false positive rates, 2022.

[46] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification, 2021.

[47] Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7377–7389. Curran Associates, Inc., 2020.

[48] Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21668–21681. Curran Associates, Inc., 2022.

[49] Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. In *Neural Information Processing Systems*, 2022.

[50] Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems, 2023.

 https://doi.org/10.52202/079017-1807

[51] Renbo Zhao. A primal-dual smoothing framework for max-structured non-convex optimization, 2022.

[52] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27548–27573. PMLR, 17–23 Jul 2022.

A Convergence Analysis

Recall that $\Phi(x) := \max_{y\in\mathcal{Y}} \phi(x,y)$, $\Psi(x) := \max_{z\in\mathcal{Z}} \psi(x,z)$, $y^*(\cdot) := \arg\max_{y\in\mathcal{Y}} \phi(\cdot,y)$, and $z^*(\cdot) := \arg\max_{z\in\mathcal{Z}} \psi(\cdot,z)$. Before presenting the proof of Theorem 4.5, we first give the proof of the proximal point estimation error bounds. As we have stated the bound for $\|x_\phi^{t+1} - \text{prox}_{\gamma\Phi}(x_t)\|^2$ in Lemma 4.4, here we present the corresponding lemma for $\|x_\psi^{t+1} - \text{prox}_{\gamma\Psi}(x_t)\|^2$.

Lemma A.1. *Suppose that Assumption 4.1 holds, $0 < \gamma < 1/\delta_\psi$, and $\eta_1 \le \frac{\gamma^2(1/\gamma-\delta_\psi)}{2}$. Then the sequences $\{x_t\}$, $\{z_t\}$, $\{x_\psi^t\}$ and $\{G_t\}$ generated by Algorithm 1 satisfy*

$$\begin{aligned}
&\mathbb{E}\|x_\psi^{t+1} - \textit{prox}_{\gamma\Psi}(x_t)\|^2 + \mathbb{E}\|z_{t+1} - z^*(\textit{prox}_{\gamma\Psi}(x_t))\|^2 \\
&\le (1 - \frac{\eta_1(1/\gamma-\delta_\psi)}{2})\mathbb{E}\|x_\psi^t - \textit{prox}_{\gamma\Psi}(x_{t-1})\|^2 + (1-\eta_1\mu_\psi)\mathbb{E}\|z_t - z^*(\textit{prox}_{\gamma\Psi}(x_{t-1}))\|^2 \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma-\delta_\psi)^3} + \frac{L_{\psi,zx}^2\eta_0^2}{\eta_1\mu_\psi^3\gamma^2(1/\gamma-\delta_\psi)^2}\right)\mathbb{E}\|G_t\|^2 + 12M^2\eta_1^2
\end{aligned}$$

Since Lemma 4.4 and Lemma A.1 share the same proof strategy, we only present the proof of Lemma 4.4.

A.1 Proof of Lemma 4.4

Proof. Recall that $\Phi(x) = \max_{y\in\mathcal{Y}} \phi(x,y)$ and $y^*(\cdot) = \arg\max_{y\in\mathcal{Y}} \phi(\cdot,y)$. Observe from Assumption 4.1(i) that Φ is δ_ϕ-weakly convex. It then follows that $\text{prox}_{\gamma\Phi}(\cdot)$ is $1/(1-\gamma\delta_\phi)$-Lipschitz continuous. By this, Assumption 4.1(iii) and Lemma 4.2, it is not hard to see that $y^*(\text{prox}_{\gamma\Phi}(\cdot))$ is $L_{\phi,yx}/(\mu_\phi(1-\gamma\delta_\phi))$-Lipschitz continuous.

For notational convenience, we let

$$\begin{aligned}
&\Phi_t(x,y) = \phi(x,y) + \frac{1}{2\gamma}\|x - x_t\|^2, \\
&x_{\Phi,t}^* = \text{prox}_{\gamma\Phi}(x_t), \quad y_t^* = y^*(\text{prox}_{\gamma\Phi}(x_t)).
\end{aligned} \tag{10}$$

In view of (10) and the update rule of x_ϕ^{t+1}, one has

$$\begin{aligned}
&\mathbb{E}_t\|x_\phi^{t+1} - x_{\Phi,t}^*\|^2 = \mathbb{E}_t\|x_\phi^t - \eta_1\tilde{\partial}_x\Phi_t(x_\phi^t, y_t) - x_{\Phi,t}^*\|^2 \\
&= \|x_\phi^t - x_{\Phi,t}^*\|^2 - 2\mathbb{E}_t\langle\eta_1\tilde{\partial}_x\Phi_t(x_\phi^t,y_t), x_\phi^t - x_{\Phi,t}^*\rangle + \mathbb{E}_t\|\eta_1\tilde{\partial}_x\Phi_t(x_\phi^t,y_t)\|^2 \\
&\le \|x_\phi^t - x_{\Phi,t}^*\|^2 + 2\eta_1\underbrace{\langle\partial_x\Phi_t(x_\phi^t,y_t), x_{\Phi,t}^* - x_\phi^t\rangle}_{(A)} + 8M^2\eta_1^2 + \frac{2\eta_1^2}{\gamma^2}\|x_\phi^t - x_{\Phi,t}^*\|^2,
\end{aligned} \tag{11}$$

where we use the inequality

$$\begin{aligned}
&\mathbb{E}_t\|\tilde{\partial}_x\Phi_t(x_\phi^t,y_t)\|^2 = \mathbb{E}_t\|\tilde{\partial}_x\phi(x_\psi^t,y_t) + \frac{1}{\gamma}(x_\phi^t - x_t)\|^2 \\
&= \mathbb{E}_t\|\tilde{\partial}_x\phi(x_\phi^t,y_t) + \frac{1}{\gamma}(x_\phi^t - x_t) - \partial_x\phi(x_{\Phi,t}^*, y_t^*) - \frac{1}{\gamma}(x_{\Phi,t}^* - x_t)\|^2 \\
&\le 4\mathbb{E}_t\|\tilde{\partial}\phi(x_\phi^t,y_t)\|^2 + 4\|\partial_x\phi(x_{\Phi,t}^*, y_t^*)\|^2 + \frac{2}{\gamma^2}\|x_\phi^t - x_{\Phi,t}^*\|^2 \\
&\le 8M^2 + \frac{2}{\gamma^2}\|x_\phi^t - x_{\Phi,t}^*\|^2.
\end{aligned}$$

By $(\gamma^{-1}-\delta_\phi)$-strong convexity of $\Phi_t(\cdot,y)$ and the definition of $x_{\Phi,t}^*$ in (10), one has

$$\begin{aligned}
&\langle\partial_x\Phi_t(x_\phi^t,y_t), x_{\Phi,t}^* - x_\phi^t\rangle \le \Phi_t(x_{\Phi,t}^*,y_t) - \Phi_t(x_\phi^t,y_t) - \frac{(1/\gamma-\delta_\phi)}{2}\|x_{\Phi,t}^* - x_\phi^t\|^2, \\
&0 \le \Phi_t(x_\phi^t,y_t^*) - \Phi_t(x_{\Phi,t}^*,y_t^*) - \frac{(1/\gamma-\delta_\phi)}{2}\|x_{\Phi,t}^* - x_\phi^t\|^2.
\end{aligned}$$

https://doi.org/10.52202/079017-1807

Summing up these two inequalities gives

$$(A) \le \Phi_t(x^*_{\Phi,t}, y_t) - \Phi_t(x^t_\phi, y_t) + \Phi_t(x^t_\phi, y^*_t) - \Phi_t(x^*_{\Phi,t}, y^*_t) - (1/\gamma - \delta_\phi)\|x^*_{\Phi,t} - x^t_\phi\|^2. \quad (12)$$

Notice from the definition of y^*_t in (10) that there exists a particular subgradient $\nabla_y \phi(x^*_{\Phi,t}, y^*_t)$ such that

$$y^*_t = P_{\mathcal{Y}}\big(y^*_t + \eta_1 \nabla_y \phi(x^*_{\Phi,t}, y^*_t)\big).$$

Using this and the update rule of y_{t+1}, we have

$$\begin{aligned}
&\mathbb{E}_t\|y_{t+1} - y^*_t\|^2 = \mathbb{E}_t\|P_{\mathcal{Y}}(y_t + \eta_1 \tilde{\nabla}_y \Phi(x^t_\phi, y_t)) - y^*_t\|^2 \\
&= \mathbb{E}_t\|P_{\mathcal{Y}}(y_t + \eta_1 \tilde{\nabla}_y \Phi_t(x^t_\phi, y_t)) - P_{\mathcal{Y}}(y^*_t + \eta_1 \nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t))\|^2 \\
&\le \mathbb{E}_t\|y_t + \eta_1 \tilde{\nabla}_y \Phi_t(x^t_\phi, y_t) - (y^*_t + \eta_1 \nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t))\|^2 \\
&\le \|y_t - y^*_t\|^2 + 2\eta_1 \langle \nabla_y \Phi_t(x^t_\phi, y_t) - \nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t), y_t - y^*_t\rangle \\
&\quad + \eta_1^2 \mathbb{E}_t\|\tilde{\nabla}_y \Phi_t(x^t_\phi, y_t) - \nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t)\|^2 \\
&\le \|y_t - y^*_t\|^2 + 2\eta_1 \underbrace{\langle \nabla_y \Phi_t(x^t_\phi, y_t) - \nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t), y_t - y^*_t\rangle}_{(B)} + 4\eta_1^2 M^2.
\end{aligned} \quad (13)$$

By μ_ϕ-strong concavity of $\Phi_t(x, \cdot)$, we have

$$\begin{aligned}
(B) &= \langle -\nabla_y \Phi_t(x^t_\phi, y_t), y^*_t - y_t\rangle + \langle -\nabla_y \Phi_t(x^*_{\Phi,t}, y^*_t), y_t - y^*_t\rangle \\
&\le -\Phi_t(x^t_\phi, y^*_t) + \Phi_t(x^t_\phi, y_t) - \frac{\mu_\phi}{2}\|y^*_t - y_t\|^2 \\
&\quad - \Phi_t(x^*_{\Phi,t}, y_t) + \Phi_t(x^*_{\Phi,t}, y^*_t) - \frac{\mu_\phi}{2}\|y^*_t - y_t\|^2 \\
&= -\Phi_t(x^t_\phi, y^*_t) + \Phi_t(x^t_\phi, y_t) - \Phi_t(x^*_{\Phi,t}, y_t) + \Phi_t(x^*_{\Phi,t}, y^*_t) - \mu_\phi\|y^*_t - y_t\|^2.
\end{aligned} \quad (14)$$

Combining (12) and (14) yields

$$(A) + (B) \le -(1/\gamma - \delta_\phi)\|x^*_{\Phi,t} - x^t_\phi\|^2 - \mu_\phi\|y^*_t - y_t\|^2.$$

Using this inequality, (11) and (13), we have

$$\begin{aligned}
&\mathbb{E}_t\|x^{t+1}_\phi - x^*_{\Phi,t}\|^2 + \mathbb{E}_t\|y_{t+1} - y^*_t\|^2 \\
&\le (1 - 2\eta_1(1/\gamma - \delta_\phi) + 2\eta_1^2/\gamma^2)\|x^*_{\Phi,t} - x^t_\phi\|^2 + (1 - 2\eta_1\mu_\phi)\|y^*_t - y_t\|^2 + 12M^2\eta_1^2 \\
&\overset{(a)}{\le} (1 - \eta_1(1/\gamma - \delta_\phi))\|x^*_{\Phi,t} - x^t_\phi\|^2 + (1 - 2\eta_1\mu_\phi)\|y^*_t - y_t\|^2 + 12M^2\eta_1^2 \\
&\overset{(b)}{\le} (1 - \eta_1(1/\gamma - \delta_\phi)) \left(\Big(1 + \frac{\eta_1(1/\gamma - \delta_\phi)}{2}\Big)\|x^t_\phi - x^*_{\Phi,t-1}\|^2 + \Big(1 + \frac{2}{\eta_1(1/\gamma - \delta_\phi)}\Big)\|x^*_{\Phi,t-1} - x^*_{\Phi,t}\|^2 \right) \\
&\quad + (1 - 2\eta_1\mu_\phi)\left((1 + \eta_1\mu_\phi)\|y_t - y^*_{t-1}\|^2 + \big(1 + (\eta_1\mu_\phi)^{-1}\big)\|y^*_{t-1} - y^*_t\|^2\right) + 12M^2\eta_1^2 \\
&\overset{(c)}{\le} \Big(1 - \frac{\eta_1(1/\gamma - \delta_\phi)}{2}\Big)\|x^t_\phi - x^*_{\Phi,t-1}\|^2 + \frac{2}{\eta_1(1/\gamma - \delta_\phi)}\|x^*_{\Phi,t-1} - x^*_{\Phi,t}\|^2 \\
&\quad + (1 - \eta_1\mu_\phi)\|y_t - y^*_{t-1}\|^2 + (\eta_1\mu_\phi)^{-1}\|y^*_{t-1} - y^*_t\|^2 + 12M^2\eta_1^2 \\
&\overset{(d)}{\le} \left(1 - \frac{\eta_1(1/\gamma - \delta_\phi)}{2}\right)\|x^t_\phi - x^*_{\Phi,t-1}\|^2 + (1 - \eta_1\mu_\phi)\|y_t - y^*_{t-1}\|^2 \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma - \delta_\phi)^3} + \frac{L^2_{\phi,yx}\eta_0^2}{\eta_1\mu_\phi^3\gamma^2(1/\gamma - \delta_\phi)^2}\right)\|G_t\|^2 + 12M^2\eta_1^2,
\end{aligned}$$

where (a) follows from the assumption $\eta_1 \le \frac{\gamma^2(1/\gamma - \delta_\phi)}{2}$, (b) uses the fact that $\|a + b\|^2 \le (1 + \alpha)\|a\|^2 + (1 + \frac{1}{\alpha})\|b\|^2$ for any $\alpha > 0$, (c) follows from bounding the coefficient of each term from above, and (d) uses $1/(1 - \gamma\delta_\phi)$-Lipschitz continuity of $\text{prox}_{\gamma\Phi}(\cdot)$, $L_{\phi,yx}/(\mu_\phi(1 - \gamma\delta_\phi))$-Lipschitz continuity of $y^*(\text{prox}_{\gamma\Phi}(\cdot))$ and the update rule of x_t.

Finally taking expectation over all randomness yields the desired result. □

A.2 Proof of Theorem 4.5

We first present a detailed version of Theorem 4.5.

Theorem A.2. *Consider Problem 1 and assume Assumption 4.1 holds. Suppose that the parameters γ, η_0 and η_1 in Algorithm 1 are chosen as follows:*

$$0<\gamma<\min\{\delta_\phi^{-1},\delta_\psi^{-1}\},\quad \alpha=\min\left\{\frac{1/\gamma-\delta_\phi}{4},\frac{1/\gamma-\delta_\psi}{4},\mu_\phi,\mu_\psi\right\},$$

$$\tau=\min\left\{\frac{\gamma^2\alpha^2}{4},\frac{\mu_\phi^{1.5}\gamma^2\alpha^{1.5}}{4L_{\phi,yx}},\frac{\mu_\psi^{1.5}\gamma^2\alpha^{1.5}}{4L_{\psi,zx}}\right\},\quad \nu=\min\left\{1,\frac{2\tau}{\gamma^2\alpha}\right\},\ L_F=\frac{2}{\gamma-\gamma^2\min\{\delta_\psi,\delta_\phi\}},$$

$$\eta_1=\min\left\{\frac{\gamma^2(1/\gamma-\delta_\phi)}{2},\frac{\gamma^2(1/\gamma-\delta_\psi)}{2},\frac{1}{2L_F\tau},\frac{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha}{768\tau M^2}\epsilon^2\right\},\quad \eta_0=\tau\eta_1.$$

Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1}-\mathit{prox}_{\gamma\Phi}(x_t)\|^2+\mathbb{E}\|x_\psi^{t+1}-\mathit{prox}_{\gamma\Psi}(x_t)\|^2+\mathbb{E}\|\nabla F_\gamma(x_t)\|^2)\le\min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4},$$

and consequently $x_\phi^{\bar{t}}$ and $x_\psi^{\bar{t}}$ are both nearly ϵ-critical points of problem (1)*, whenever*

$$T\ge\frac{16(F_\gamma(x_0)-F_\gamma^*+P_0)}{\min\{1,\gamma^{-2}\}\min\{\alpha,\tau\}\nu\epsilon^2}\max\left\{\frac{2}{\gamma^2(1/\gamma-\delta_\phi)},\frac{2}{\gamma^2(1/\gamma-\delta_\psi)},2L_F\tau,\frac{768\tau M^2}{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha\epsilon^2}\right\}\tag{15}$$

with

$$P_0=\frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(\mathbb{E}\|x_\phi^1-\mathit{prox}_{\gamma\Phi}(x_0)\|^2+\mathbb{E}\|y_1-y_0^*\|^2+\mathbb{E}\|x_\psi^1-\mathit{prox}_{\gamma\Psi}(x_0)\|^2+\mathbb{E}\|z_1-z_0^*\|^2\right).$$

Proof. For notational convenience, let

$$x_{\Psi,t}^*=\mathrm{prox}_{\gamma\Psi}(x_t),\quad z_t^*=\arg\max_{z\in\mathcal{Z}}\psi(x_{\Psi,t}^*,z).\tag{16}$$

From Proposition 3.3, we know that $F_\gamma(\cdot)$ is L_F-smooth. By this, $0<\eta_0\le\frac{1}{2L_F}$, and Lemma 4.3, one has

$$F_\gamma(x_{t+1})\le F_\gamma(x_t)+\frac{\eta_0}{2}\|\nabla F_\gamma(x_t)-G_{t+1}\|^2-\frac{\eta_0}{2}\|\nabla F_\gamma(x_t)\|^2-\frac{\eta_0}{4}\|G_{t+1}\|^2.\tag{17}$$

Notice that

$$\nabla F_\gamma(x_t)=\gamma^{-1}(\mathrm{prox}_{\gamma\Psi}(x_t)-x_t+x_t-\mathrm{prox}_{\gamma\Phi}(x_t))=\gamma^{-1}(\mathrm{prox}_{\gamma\Psi}(x_t)-\mathrm{prox}_{\gamma\Phi}(x_t)),$$
$$G_{t+1}=\gamma^{-1}(x_\psi^{t+1}-x_\phi^{t+1}).$$

Using these, (10) and (16), we have

$$\begin{aligned}\|\nabla F_\gamma(x_t)-G_{t+1}\|^2&=\|\gamma^{-1}(\mathrm{prox}_{\gamma\Psi}(x_t)-\mathrm{prox}_{\gamma\Phi}(x_t))-\gamma^{-1}(x_\psi^{t+1}-x_\phi^{t+1})\|^2\\&=\|\gamma^{-1}(x_{\Psi,t}^*-x_{\Phi,t}^*)-\gamma^{-1}(x_\psi^{t+1}-x_\phi^{t+1})\|^2\\&\le 2\gamma^{-2}\left(\|x_{\Psi,t}^*-x_\psi^{t+1}\|^2+\|x_{\Phi,t}^*-x_\phi^{t+1}\|^2\right).\end{aligned}\tag{18}$$

It follows from this and (17) that

$$\begin{aligned}\mathbb{E}[F_\gamma(x_{t+1})]\le\mathbb{E}[F_\gamma(x_t)]&+\frac{\eta_0}{\gamma^2}\mathbb{E}\|x_\psi^{t+1}-x_{\Psi,t}^*\|^2+\frac{\eta_0}{\gamma^2}\mathbb{E}\|x_\phi^{t+1}-x_{\Phi,t}^*\|^2\\&-\frac{\eta_0}{2}\mathbb{E}\|\nabla F_\gamma(x_t)\|^2-\frac{\eta_0}{4}\mathbb{E}\|G_{t+1}\|^2.\end{aligned}\tag{19}$$

Let $x_{\Phi,t}^*$ and y_t^* be defined in (10). Invoking Lemma 4.4, we have

$$\begin{aligned}&\mathbb{E}\|x_\phi^{t+2}-x_{\Phi,t+1}^*\|^2+\mathbb{E}\|y_{t+2}-y_{t+1}^*\|^2\\&\le\left(1-\frac{\eta_1(1/\gamma-\delta_\phi)}{2}\right)\mathbb{E}\|x_\phi^{t+1}-x_{\Phi,t}^*\|^2+(1-\eta_1\mu_\phi)\mathbb{E}\|y_{t+1}-y_t^*\|^2\\&\quad+\left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma-\delta_\phi)^3}+\frac{L_{\phi,yx}^2\eta_0^2}{\eta_1\mu_\phi^3\gamma^2(1/\gamma-\delta_\phi)^2}\right)\mathbb{E}\|G_{t+1}\|^2+12M^2\eta_1^2.\end{aligned}$$

https://doi.org/10.52202/079017-1807

Recall that $x^*_{\Psi,t}$ and z^*_t are defined in (16). By Lemma A.1, one has

$$\begin{aligned}
&\mathbb{E}\|x_\psi^{t+2} - x^*_{\Psi,t+1}\|^2 + \mathbb{E}\|z_{t+2} - z^*_{t+1}\|^2 \\
&\le \left(1 - \frac{\eta_1(1/\gamma - \delta_\psi)}{2}\right)\mathbb{E}\|x_\psi^{t+1} - x^*_{\Psi,t}\|^2 + (1-\eta_1\mu_\psi)\mathbb{E}\|z_{t+1} - z^*_t\|^2 \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma-\delta_\psi)^3} + \frac{L^2_{\psi,zx}\eta_0^2}{\eta_1\mu_\psi^3\gamma^2(1/\gamma-\delta_\psi)^2}\right)\mathbb{E}\|G_{t+1}\|^2 + 12M^2\eta_1^2.
\end{aligned}$$

Let α be given in the statement of this theorem. Using this and the last two inequalities above, we have

$$\begin{aligned}
&\mathbb{E}\|x_\phi^{t+2} - x^*_{\Phi,t+1}\|^2 + \mathbb{E}\|y_{t+2} - y^*_{t+1}\|^2 \\
&\le (1-\alpha\eta_1)\big(\mathbb{E}\|x_\phi^{t+1} - x^*_{\Phi,t}\|^2 + \mathbb{E}\|y_{t+1} - y^*_t\|^2\big) \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma-\delta_\phi)^3} + \frac{L^2_{\phi,yx}\eta_0^2}{\eta_1\mu_\phi^3\gamma^2(1/\gamma-\delta_\phi)^2}\right)\mathbb{E}\|G_{t+1}\|^2 + 12M^2\eta_1^2,
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\mathbb{E}\|x_\psi^{t+2} - x^*_{\Psi,t+1}\|^2 + \mathbb{E}\|z_{t+2} - z^*_{t+1}\|^2 \\
&\le (1-\alpha\eta_1)\big(\mathbb{E}\|x_\psi^{t+1} - x^*_{\Psi,t}\|^2 + \mathbb{E}\|z_{t+1} - z^*_t\|^2\big) \\
&\quad + \left(\frac{2\eta_0^2}{\eta_1\gamma^2(1/\gamma-\delta_\psi)^3} + \frac{L^2_{\psi,zx}\eta_0^2}{\eta_1\mu_\psi^3\gamma^2(1/\gamma-\delta_\psi)^2}\right)\mathbb{E}\|G_{t+1}\|^2 + 12M^2\eta_1^2.
\end{aligned} \tag{21}$$

Summing up inequalities (19), (20)$\times\frac{2\eta_0}{\eta_1\gamma^2\alpha}$ and (21)$\times\frac{2\eta_0}{\eta_1\gamma^2\alpha}$ yields

$$\begin{aligned}
&\mathbb{E}[F_\gamma(x_{t+1})] + \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(\mathbb{E}\|x_\phi^{t+2} - x^*_{\Phi,t+1}\|^2 + \mathbb{E}\|y_{t+2} - y^*_{t+1}\|^2\right) \\
&\quad + \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(\mathbb{E}\|x_\psi^{t+2} - x^*_{\Psi,t+1}\|^2 + \mathbb{E}\|z_{t+2} - z^*_{t+1}\|^2\right) \\
&\le \mathbb{E}[F_\gamma(x_t)] + \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(1-\frac{\eta_1\alpha}{2}\right)\left(\mathbb{E}\|x_\phi^{t+1} - x^*_{\Phi,t}\|^2 + \mathbb{E}\|y_{t+1} - y^*_t\|^2\right) \\
&\quad + \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(1-\frac{\eta_1\alpha}{2}\right)\left(\mathbb{E}\|x_\psi^{t+1} - x^*_{\Psi,t}\|^2 + \mathbb{E}\|y_{t+1} - y^*_t\|^2\right) \\
&\quad + \left(\frac{4\eta_0^3}{\eta_1^2\gamma^4\alpha(1/\gamma-\delta_\phi)^3} + \frac{4\eta_0^3}{\eta_1^2\gamma^4\alpha(1/\gamma-\delta_\psi)^3} + \frac{2L^2_{\phi,yx}\eta_0^3}{\eta_1^2\mu_\phi^3\gamma^4\alpha(1/\gamma-\delta_\phi)^2}\right. \\
&\quad \left. + \frac{2L^2_{\psi,zx}\eta_0^3}{\eta_1^2\mu_\psi^3\gamma^4\alpha(1/\gamma-\delta_\psi)^2} - \frac{\eta_0}{4}\right)\mathbb{E}\|G_{t+1}\|^2 \\
&\quad - \frac{\eta_0}{2}\mathbb{E}\|\nabla F_\gamma(x_t)\|^2 + \frac{24\eta_0\eta_1M^2}{\gamma^2\alpha} + \frac{24\eta_0\eta_1M^2}{\gamma^2\alpha}.
\end{aligned} \tag{22}$$

We now introduce a potential function

$$P_t = \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(\mathbb{E}\|x_\phi^{t+1} - x^*_{\Phi,t}\|^2 + \mathbb{E}\|y_{t+1} - y^*_t\|^2 + \mathbb{E}\|x_\psi^{t+1} - x^*_{\Psi,t}\|^2 + \mathbb{E}\|z_{t+1} - z^*_t\|^2\right), \tag{23}$$

and rewrite inequality (22) as

$$\begin{aligned}
&\mathbb{E}[F_\gamma(x_{t+1})] + P_{t+1} \\
&\le \mathbb{E}[F_\gamma(x_t)] + (1-\beta)P_t - \beta\mathbb{E}\|\nabla F_\gamma(x_t)\|^2 + \frac{48\eta_0\eta_1M^2}{\gamma^2\alpha} \\
&\quad + \left(\frac{\eta_0^3}{\eta_1^2\gamma^4\alpha^4} + \frac{L^2_{\phi,yx}\eta_0^3}{\eta_1^2\mu_\phi^3\gamma^4\alpha^3} + \frac{L^2_{\psi,zx}\eta_0^3}{\eta_1^2\mu_\psi^3\gamma^4\alpha^3} - \frac{\eta_0}{4}\right)\mathbb{E}\|G_{t+1}\|^2,
\end{aligned}$$

where

$$\beta = \min\left\{\frac{\eta_1\alpha}{2}, \frac{\eta_0}{2}\right\}. \tag{24}$$

This inequality, together with the choice of η_0 and τ specified in this theorm, yields

$$E[F_\gamma(x_{t+1})] + P_{t+1} \le \mathbb{E}[F_\gamma(x_t)] + (1-\beta)P_t - \beta\mathbb{E}\|\nabla F_\gamma(x_t)\|^2 + \frac{48\eta_0\eta_1 M^2}{\gamma^2\alpha}.$$

Taking average of these inequalities over $t = 0, \ldots, T-1$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(P_t + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \frac{F_\gamma(x_0) - F_\gamma^* + P_0}{\beta T} + \frac{48\eta_0\eta_1 M^2}{\beta\gamma^2\alpha}, \tag{25}$$

where we use $F_\gamma^* \le F_\gamma(x_T)$ due to Assumption 4.1(iii). Recall that $\eta_0 = \tau\eta_1$ and $\nu = \min\{1, \frac{2\tau}{\gamma^2\alpha}\}$. Using these, (23) and (25), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1} - x_{\Phi,t}^*\|^2 + \mathbb{E}\|x_\psi^{t+1} - x_{\Psi,t}^*\|^2 + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \\
&\le \frac{1}{\nu T}\sum_{t=0}^{T-1}(P_t + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \frac{F_\gamma(x_0) - F_\gamma^* + P_0}{\nu\beta T} + \frac{48\eta_0\eta_1 M^2}{\nu\beta\gamma^2\alpha}.
\end{aligned}$$

By (24) and the choice of α, η_0 and η_1 specified in this theorem, one has

$$\frac{\min\{1,\gamma^{-2}\}\nu\beta\gamma^2\alpha\epsilon^2}{384\eta_0 M^2} = \frac{\min\{1,\gamma^2\}\min\left\{\frac{\eta_1\alpha}{2}, \frac{\eta_1\tau}{2}\right\}\nu\alpha\epsilon^2}{384\eta_1\tau M^2} = \frac{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha}{768\tau M^2}\epsilon^2 \ge \eta_1,$$

which implies that

$$\frac{48\eta_0\eta_1 M^2}{\nu\beta\gamma^2\alpha} \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{8}.$$

Suppose that T satisfies (15). It then follows from (24), $\eta_0 = \tau\eta_1$, and the expression of η_1 that

$$\begin{aligned}
T &\ge \frac{16(F_\gamma(x_0) - F_\gamma^* + P_0)}{\min\{1,\gamma^{-2}\}\min\{\alpha,\tau\}\nu\epsilon^2}\max\left\{\frac{2}{\gamma^2(1/\gamma - \delta_\phi)}, \frac{2}{\gamma^2(1/\gamma - \delta_\psi)}, 2L_F\tau, \frac{768\tau M^2}{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha\epsilon^2}\right\} \\
&= \frac{8(F_\gamma(x_0) - F_\gamma^* + P_0)}{\min\{1,\gamma^{-2}\}\nu\beta\epsilon^2},
\end{aligned}$$

which implies that

$$\frac{F_\gamma(x_0) - F_\gamma^* + P_0}{\nu\beta T} \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{8}.$$

Hence, for any T satisfying (15), one has

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1} - x_{\Phi,t}^*\|^2 + \mathbb{E}\|x_\psi^{t+1} - x_{\Psi,t}^*\|^2 + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4},$$

which together with $x_{\Phi,t}^* = \text{prox}_{\gamma\Phi}(x_t)$ and $x_{\Psi,t}^* = \text{prox}_{\gamma\Psi}(x_t)$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1} - \text{prox}_{\gamma\Phi}(x_t)\|^2 + \mathbb{E}\|x_\psi^{t+1} - \text{prox}_{\gamma\Psi}(x_t)\|^2 + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4}.$$

Since $\bar{t}$ is uniformly sampled from $\{1, \ldots, T\}$, we have

$$\mathbb{E}[\|x_\phi^{\bar{t}} - \text{prox}_{\gamma\Phi}(x_{\bar{t}-1})\|^2 + \|x_\psi^{\bar{t}} - \text{prox}_{\gamma\Psi}(x_{\bar{t}-1})\|^2 + \|\nabla F_\gamma(x_{\bar{t}-1})\|^2] \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4}.$$

It then follows from Lemma 3.4 that $x_\phi^{\bar{t}}$ and $x_\psi^{\bar{t}}$ are both nearly ϵ-critical points of problem (1).

□

https://doi.org/10.52202/079017-1807

A.3 Proof of Corollary 4.7

We present a detailed version of Corollary 4.7

Corollary A.3. *Consider Problem 2 and assume Assumption 4.6 holds. Suppose that the parameters γ, η_0 and η_1 in Algorithm 2 are chosen as follows:*

$$0 < \gamma < \min\{\delta_\phi^{-1}, \delta_\psi^{-1}\}, \quad \alpha = \min\left\{\frac{1/\gamma - \delta_\phi}{2}, \frac{1/\gamma - \delta_\psi}{2}\right\}, \quad \tau = \frac{\gamma^2\alpha^2}{4}, \quad \nu = \min\left\{1, \frac{2\tau}{\gamma^2\alpha}\right\},$$

$$\eta_1 = \min\left\{\frac{\gamma^2(1/\gamma - \delta_\phi)}{2}, \frac{\gamma^2(1/\gamma - \delta_\psi)}{2}, \frac{1}{2L_F\tau}, \frac{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha}{768\tau M^2}\epsilon^2\right\}, \quad \eta_0 = \tau\eta_1.$$

Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1} - \mathit{prox}_{\gamma\phi}(x_t)\|^2 + \mathbb{E}\|x_\psi^{t+1} - \mathit{prox}_{\gamma\psi}(x_t)\|^2 + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4},$$

and consequently $x_\phi^{\bar{t}}$ and $x_\psi^{\bar{t}}$ are both nearly ϵ-critical points of problem (2)*, whenever*

$$T \ge \frac{16(F_\gamma(x_0) - F_\gamma^* + P_0)}{\min\{1,\gamma^{-2}\}\min\{\alpha,\tau\}\nu\epsilon^2}\max\left\{\frac{2}{\gamma^2(1/\gamma-\delta_\phi)}, \frac{2}{\gamma^2(1/\gamma-\delta_\psi)}, 2L_F\tau, \frac{768\tau M^2}{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha\epsilon^2}\right\}.$$

with

$$P_0 = \frac{2\tau}{\gamma^2\alpha}\left(\mathbb{E}\|x_\phi^1 - \mathit{prox}_{\gamma\phi}(x_0)\|^2 + \mathbb{E}\|x_\psi^1 - \mathit{prox}_{\gamma\psi}(x_0)\|^2\right).$$

Since problem (2) and Algorithm 2 are special cases of problem (1) and Algorithm 1 respectively, Corollary A.3 directly follows from Theorem A.2.

A.4 Proof of Corollary 4.9

We present a detailed version of Corollary 4.9

Corollary A.4. *Consider Problem 3 and assume Assumption 4.8 holds. Suppose that the parameters γ, η_0 and η_1 in Algorithm 3 are chosen as follows:*

$$0 < \gamma < \delta_\phi^{-1}, \quad \alpha = \min\left\{\frac{1/\gamma - \delta_\phi}{2}, \mu_\phi\right\}, \quad \tau = \min\left\{\frac{\gamma^2\alpha^2}{4}, \frac{\mu_\phi^{1.5}\gamma^2\alpha^{1.5}}{4L_{\phi,yx}}\right\}, \quad \nu = \min\left\{1, \frac{2\tau}{\gamma^2\alpha}\right\},$$

$$\eta_1 = \min\left\{\frac{\gamma^2(1/\gamma - \delta_\phi)}{2}, \frac{1}{2L_F\tau}, \frac{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha}{384\tau M^2}\epsilon^2\right\}, \quad \eta_0 = \tau\eta_1.$$

Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\|x_\phi^{t+1} - \mathit{prox}_{\gamma F}(x_t)\|^2 + \mathbb{E}\|\nabla F_\gamma(x_t)\|^2) \le \min\{1,\gamma^{-2}\}\frac{\epsilon^2}{4},$$

and consequently $x_{\bar{t}}$ is a nearly ϵ-critical point of problem (3)*, whenever*

$$T \ge \frac{16(F_\gamma(x_0) - F_\gamma^* + P_0)}{\min\{1,\gamma^{-2}\}\min\{\alpha,\tau\}\nu\epsilon^2}\max\left\{\frac{2}{\gamma^2(1/\gamma-\delta_\phi)}, 2L_F\tau, \frac{384\tau M^2}{\min\{1,\gamma^2\}\min\{\alpha,\tau\}\nu\alpha\epsilon^2}\right\},$$

with

$$P_0 = \frac{2\eta_0}{\eta_1\gamma^2\alpha}\left(\mathbb{E}\|x_\phi^1 - \mathit{prox}_{\gamma F}(x_0)\|^2 + \mathbb{E}\|y_1 - y_0^*\|^2\right).$$

Proof. This proof is similar to that of Theorem A.2 except that the inequality (18) is replaced by

$$\begin{aligned}\|\nabla F_\gamma(x_t) - G_{t+1}\|^2 &= \left\|\frac{1}{\gamma}(x_t - \mathrm{prox}_{\gamma F}(x_t)) - \frac{1}{\gamma}(x_t - x_\phi^{t+1})\right\|^2 \\ &= \frac{1}{\gamma^2}\|\mathrm{prox}_{\gamma F}(x_t) - x_\phi^{t+1}\|^2.\end{aligned}$$

□

B More Experimental Results

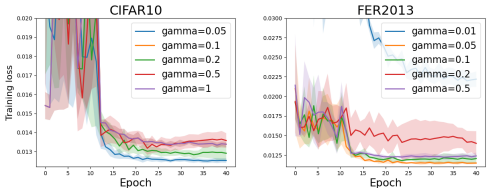

Figure 2: Ablation Study of SMAG for PU Learning

Table 4: Mean ± std of fairness results on CelebA test dataset with *Bags Under Eyes* task labels, and *Male* sensitive attribute. Results are reported on 3 independent runs. We use bold font to denote the best result and use underline to denote the second best.

	Bags Under Eyes, Male			
Methods	**pAUC**↑	**EOD**↓	**EOP**↓	**DP** ↓
SOPA	<u>0.8293</u> ± 0.006	0.2015 ± 0.041	<u>0.1000</u>± 0.043	0.4055 ± 0.027
SMAG*	0.8261 ± 0.004	<u>0.1848</u> ± 0.023	0.1065 ± 0.046	<u>0.3754</u> ± 0.033
SGDA	**0.8307** ± 0.003	0.2026 ± 0.028	0.1096 ± 0.031	0.4028 ± 0.039
EGDA	0.8262 ± 0.004	0.2223 ± 0.032	0.1287 ± 0.038	0.4200 ± 0.024
SMAG	0.8278 ± 0.002	**0.1642** ± 0.025	**0.0982**± 0.034	**0.3690** ± 0.029

https://doi.org/10.52202/079017-1807

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have detailed explanation on all items described in the abstract presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discussed the limitation of this in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are presented in the main paper. The detailed theorem statements and proofs are presented in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information needed to reproduce the experimental results in the application section.

   Guidelines:

https://doi.org/10.52202/079017-1807

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is included in the supplemental material. The data we used are public datasets.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided all the implementation details in the application section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We ran multiple trails for each experiment setting and present the error bars in the plot or table.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. Compute resources information is provided in the application section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-1807

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer:[NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For all assets we used in this paper, we cited their original source.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

https://doi.org/10.52202/079017-1807

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.