# Large-scale Optimization of Partial AUC in a Range of False Positive Rates

**Yao Yao**
Department of Mathematics
The University of Iowa
yao-yao-2@uiowa.edu

**Qihang Lin**
Tipple College of Business
The University of Iowa
qihang-lin@uiowa.edu

**Tianbao Yang**
Department of Computer Science & Engineering
Texas A&M University
tianbao-yang@tamu.edu

## Abstract

The area under the ROC curve (AUC) is one of the most widely used performance measures for classification models in machine learning. However, it summarizes the true positive rates (TPRs) over all false positive rates (FPRs) in the ROC space, which may include the FPRs with no practical relevance in some applications. The partial AUC, as a generalization of the AUC, summarizes only the TPRs over a specific range of the FPRs and is thus a more suitable performance measure in many real-world situations. Although partial AUC optimization in a range of FPRs had been studied, existing algorithms are not scalable to big data and not applicable to deep learning. To address this challenge, we cast the problem into a non-smooth difference-of-convex (DC) program for any smooth predictive functions (e.g., deep neural networks), which allowed us to develop an efficient approximated gradient descent method based on the Moreau envelope smoothing technique, inspired by recent advances in non-smooth DC optimization. To increase the efficiency of large data processing, we used an efficient stochastic block coordinate update in our algorithm. Our proposed algorithm can also be used to minimize the sum of ranked range loss, which also lacks efficient solvers. We established a complexity of $\tilde{O}(1/\epsilon^6)$ for finding a nearly $\epsilon$-critical solution. Finally, we numerically demonstrated the effectiveness of our proposed algorithms in training both linear models and deep neural networks for partial AUC maximization and sum of ranked range loss minimization.

## 1 Introduction

The area under the receiver operating characteristic (ROC) curve (AUC) is one of the most widely used performance measures for classifiers in machine learning, especially when the data is imbalanced between the classes [7, 19]. Typically, the classifier produces a score for each data point. Then a data point is classified as positive if its score is above a chosen threshold; otherwise, it is classified as negative. Varying the threshold will change the true positive rate (TPR) and the false positive rate (FPR) of the classifier. The ROC curve shows the TPR as a function of the FPR that corresponds to the same threshold. Hence, maximizing the AUC of a classifier is essentially maximizing the classifier's average TPR over all FPRs from zero to one. However, for some applications, some FPR regions have no practical relevance. So does the TPR over those regions. For example, in clinical practice, a high FPR in diagnostic tests often results in a high monetary cost, so people may only

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

need to maximize the TPR when the FPR is low [13, 34, 64]. Moreover, since two models with the same AUC can still have different ROCs, the AUC does not always reflect the true performance of a model that is needed in a particular production environment [8].

As a generalization of the AUC, the partial AUC (pAUC) only measures the area under the ROC curve that is restricted between two FPRs. A probabilistic interpretation of the pAUC can be found in [13]. In contrast to the AUC, the pAUC represents the average TPR only over a relevant range of FPRs and provides a performance measure that is more aligned with the practical needs in some applications.

In literature, the existing algorithms for training a classifier by maximizing the pAUC include the boosting method [29] and the cutting plane algorithm [41, 42, 43]. However, the former has no theoretical guarantee, and the latter applies only to linear models. More importantly, both methods require processing all the data in each iteration and thus, become computationally inefficient for large datasets.

In this paper, we proposed an approximate gradient method for maximizing the pAUC that works for nonlinear models (e.g., deep neural networks) and only needs to process randomly sampled positive and negative data points of any size in each iteration. In particular, we formulated the maximization of the pAUC as a non-smooth difference-of-convex (DC) program [30, 54]. Due to non-smoothness, most existing DC optimization algorithms cannot be applied to our formulation. Motivated by [52], we approximate the two non-smooth convex components in the DC program by their Moreau envelopes and obtain a smooth approximation of the problem, which will be solved using the gradient descent method. Since the gradient of the smooth problem cannot be calculated explicitly, we approximated the gradient by solving the two proximal-point subproblems defined by each convex component using the stochastic block coordinate descent (SBCD) method. Our method, besides its low per-iteration cost, has a rigorous theoretical guarantee, unlike the existing methods. *In fact, we show that our method finds a nearly $\epsilon$-critical point of the pAUC optimization problem in $\tilde{O}(\epsilon^{-6})$ iterations with only small samples of positive and negative data points processed per iteration.*[1] *This is the main contribution of this paper.*

Note that, for non-convex non-smooth optimization, the existing stochastic methods [10, 11] find an nearly $\epsilon$-critical point in $O(\epsilon^{-4})$ iterations under a weak convexity assumption. Our method needs $O(\epsilon^{-6})$ iterations because our problem is a DC problem with both convex components non-smooth which is much more challenging than a weakly non-convex minimization problem. In addition, our iteration number matches the known best iteration complexity for non-smooth non-convex min-max optimization [33, 46] and non-smooth non-convex constrained optimization [35].

In addition to pAUC optimization, our method can be also used to minimize the sum of ranked range (SoRR) loss, which can be viewed as a special case of pAUC optimization. Many machine learning models are trained by minimizing an objective function, which is defined as the sum of losses over all training samples [60]. Since the sum of losses weights all samples equally, it is insensitive to samples from minority groups. Hence, the sum of top-$k$ losses [17, 49] is often used as an alternative objective function because it provides the model with robustness to non-typical instances. However, the sum of top-$k$ losses can be very sensitive to outliers, especially when $k$ is small. To address this issue, [24] proposed the SoRR loss as a new learning objective, which is defined as the sum of a consecutive sequence of losses from any range after the losses are sorted. Compared to the sum of all losses and the sum of top-$k$ losses, the SoRR loss maintains a model's robustness to a minority group but also reduces the model's sensitivity to outliers. See Fig.1 in [24] for an illustration of the benefit of using the SoRR loss over other ways of aggregating individual losses.

To minimize the SoRR loss, [24] applied a difference-of-convex algorithm (DCA) [3, 54], which linearizes the second convex component and solves the resulting subproblem using the stochastic subgradient method. DCA has been well studied in literature; but when the both components are non-smooth, as in our problem, only asymptotic convergence results are available. To establish the total number of iterations needed to find an $\epsilon$-critical point in a non-asymptotic sense, most existing studies had to assume that at least one of the components is differentiable, which is not the case in this paper. Using the approximate gradient method presented in this paper, one can find a nearly $\epsilon$-critical point of the SoRR loss optimization problem in $\tilde{O}(\epsilon^{-6})$ iterations.

[1] Throughout the paper, $\tilde{O}(\cdot)$ suppresses all logarithmic factors.

## 2 Related Works

The pAUC has been studied for decades [26, 37, 57]. However, most studies focused on its estimation [13] and application as a performance measure, while only a few studies were devoted to numerical algorithms for optimizing the pAUC. Efficient optimization methods have been developed for maximizing AUC and multiclass AUC by [68] and [65], but they cannot be applied to pAUC. Besides the boosting method [29] and the cutting plane algorithm [41, 42, 43] mentioned in the previous section, [59, 66, 67, 72] developed surrogate optimization techniques that directly maximize a smooth approximation of the pAUC or the two-way pAUC [64]. However, their approaches can only be applied when the FPR starts from exactly zero. On the contrary, our method allows the FPR to start from any value between zero and one. [61] and [47] developed algorithms that use the pAUC as a criterion for creating a linear combination of multiple existing classifiers while we consider directly train a classifier using the pAUC.

DC optimization has been studied since the 1950s [2, 20]. We refer interested readers to [30, 45, 54, 55, 58], and the references therein. The actively studied numerical methods for solving a DC program include DCA [3, 50, 54, 55], which is also known as the concave-convex procedure [32, 51, 71], the proximal DCA [5, 38, 40, 53], and the direct gradient methods [28]. However, when the two convex components are both non-smooth, the existing methods have only asymptotic convergence results except the method by [1], who considered a stopping criterion different from ours. When at least one component is smooth, non-asymptotic convergence rates have been established with and without the Kurdyka-Łojasiewicz (KL) condition [5, 6, 28, 50, 62].

The algorithms mentioned above are deterministic and require processing the entire dataset per iteration. Stochastic algorithms that process only a small data sample per iteration have been studied [12, 21, 31, 36, 44]. However, they all assumed smoothness on at least one of the two convex components in the DC program. The stochastic methods of [4, 56, 63] can be applied when both components are non-smooth but their methods require an unbiased stochastic estimation of the gradient and/or value of the two components, which is not available in the DC formulation of the pAUC maximization problem in this paper.

The technique most related to our work is the smoothing method based on the Moreau envelope [16, 18, 22, 23, 39, 52]. Our work is motivated by [39, 52], but the important difference is that they studied deterministic methods and assumed either that one function is smooth or that the proximal-point subproblems can be solved exactly, which we do not assume. However, [39, 52] consider a more general problem and study the fundamental properties of the smoothed function such as its Lipschitz smoothness and how its stationary points correspond to those of the original problems. We mainly focus on partial AUC optimization which has a special structure we can utilize when solving the proximal-point subproblems. Additionally, [52] developed an algorithm when there were linear equality constraints, which we do not consider in this paper.

## 3 Preliminary

We consider a classical binary classification problem, where the goal is to build a predictive model that predicts a binary label $y \in \{1, -1\}$ based on a feature vector $\mathbf{x} \in \mathbb{R}^p$. Let $h_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}$ be the predictive model parameterized by a vector $\mathbf{w} \in \mathbb{R}^d$, which produces a score $h_{\mathbf{w}}(\mathbf{x})$ for $\mathbf{x}$. Then $\mathbf{x}$ is classified as positive ($y = 1$) if $h_{\mathbf{w}}(\mathbf{x})$ is above a chosen threshold and classified as negative ($y = -1$), otherwise.

Let $\mathcal{X}_+ = \{\mathbf{x}_i^+\}_i^{N_+}$ and $\mathcal{X}_- = \{\mathbf{x}_i^-\}_i^{N_-}$ be the sets of feature vectors of positive and negative training data, respectively. The problem of learning $h_{\mathbf{w}}$ through maximizing its empirical AUC on the training data can be formulated as

$$\max_{\mathbf{w}} \frac{1}{N_+ N_-} \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} \mathbf{1}(h_{\mathbf{w}}(\mathbf{x}_i^+) > h_{\mathbf{w}}(\mathbf{x}_j^-)), \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function which equals one if the inequality inside the parentheses holds and equals zero, otherwise. According to the introduction, pAUC can be a better performance measure of $h_{\mathbf{w}}$ than AUC. Consider two FPRs $\alpha$ and $\beta$ with $0 \leq \alpha < \beta \leq 1$. For simplicity of exposition, we assume $N_-\alpha$ and $N_-\beta$ are both integers. Let $m = N_-\alpha$ and $n = N_-\beta$. The problem of maximizing

the empirical pAUC with FPR between $\alpha$ and $\beta$ can be formulated as

$$\max_{\mathbf{w}} \frac{1}{N_+(n-m)} \sum_{i=1}^{N_+} \sum_{j=m+1}^{n} \mathbf{1}(h_{\mathbf{w}}(\mathbf{x}_i^+) > h_{\mathbf{w}}(\mathbf{x}_{[j]}^-)), \tag{2}$$

where $[j]$ denotes the index of the $j$th largest coordinate in vector $(h_{\mathbf{w}}(\mathbf{x}_j^-))_{j=1}^{N_-}$ with ties broken arbitrarily. Note that $N_+(n-m)$ in (2) is a normalizer that makes the objective value between zero and one. Solving (2) is challenging due to discontinuity. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a differential non-increasing loss function. Problem (2) can be approximated by the loss minimization problem

$$\min_{\mathbf{w}} \frac{1}{N_+(n-m)} \sum_{i=1}^{N_+} \sum_{j=m+1}^{n} \ell(h_{\mathbf{w}}(\mathbf{x}_i^+) - h_{\mathbf{w}}(\mathbf{x}_{[j]}^-)). \tag{3}$$

To facilitate the discussion, we first introduce a few notations. Given a vector $S = (s_i)_{i=1}^N \in \mathbb{R}^N$ and an integer $l$ with $0 \le l \le N$, the sum of the top-$l$ values in $S$ is

$$\phi_l(S) := \textstyle\sum_{j=1}^{l} s_{[j]}, \tag{4}$$

where $[j]$ denotes the index of the $j$th largest coordinate in $S$ with ties broken arbitrarily. For integers $l_1$ and $l_2$ with $0 \le l_1 < l_2 \le N$, $\phi_{l_2}(S) - \phi_{l_1}(S)$ is the sum from the $(l_1+1)$th to the $l_2$th (inclusive) largest coordinates of $S$, also called a *sum of ranked range* (SoRR). In addition, we define vectors

$$S_i(\mathbf{w}) := (s_{ij}(\mathbf{w}))_{j=1}^{N_-}$$

for $i = 1, \ldots, N_+$, where $s_{ij}(\mathbf{w}) := \ell(h_{\mathbf{w}}(\mathbf{x}_i^+) - h_{\mathbf{w}}(\mathbf{x}_j^-))$ for $i = 1, \ldots, N_+$ and $j = 1, \ldots, N_-$. Since $\ell$ is non-increasing, the $j$th largest coordinate of $S_i(\mathbf{w})$ is $\ell(h_{\mathbf{w}}(\mathbf{x}_i^+) - h_{\mathbf{w}}(\mathbf{x}_{[j]}^-))$. As a result, we have, for $i = 1, \ldots, N_+$,

$$\textstyle\sum_{j=m+1}^{n} \ell(h_{\mathbf{w}}(\mathbf{x}_i^+) - h_{\mathbf{w}}(\mathbf{x}_{[j]}^-)) = \phi_n(S_i(\mathbf{w})) - \phi_m(S_i(\mathbf{w})).$$

Hence, after dropping the normalizer, (3) can be equivalently written as

$$F^* = \min_{\mathbf{w}} \left\{ F(\mathbf{w}) := f^n(\mathbf{w}) - f^m(\mathbf{w}) \right\}, \tag{5}$$

where

$$f^l(\mathbf{w}) = \textstyle\sum_{i=1}^{N_+} \phi_l(S_i(\mathbf{w})) \quad \text{for } l = m, n. \tag{6}$$

Next, we introduce an interesting special case of (5), namely, the problem of minimizing SoRR loss. We still consider a supervised learning problem but the target $y \in \mathbb{R}$ does not need to be binary. We want to predict $y$ based on a feature vector $\mathbf{x} \in \mathbb{R}^p$ using $h_{\mathbf{w}}(\mathbf{x})$. With a little abuse of notation, we measure the discrepancy between $h_{\mathbf{w}}(\mathbf{x})$ and $y$ by $\ell(h_{\mathbf{w}}(\mathbf{x}), y)$, where $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ is a loss function. We consider learning the model's parameter $\mathbf{w}$ from a training set $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^N$, where $\mathbf{x}_j \in \mathbb{R}^p$ and $y_j \in \mathbb{R}$ for $j = 1, \ldots, N$, by minimizing the SoRR loss. More specifically, we define vector

$$S(\mathbf{w}) = (s_j(\mathbf{w}))_{j=1}^N,$$

where $s_j(\mathbf{w}) := \ell(h_{\mathbf{w}}(\mathbf{x}_j), y_j),\ j = 1, \ldots, N$. Recall (4). For any integers $m$ and $n$ with $0 \le m < n \le N$, the problem of minimizing the SoRR loss with a range from $m+1$ to $n$ is formulated as $\min_{\mathbf{w}} \{\phi_n(S(\mathbf{w})) - \phi_m(S(\mathbf{w}))\}$, which is an instance of (5) with

$$f^l = \phi_l(S(\mathbf{w})) \text{ for } l = m, n. \tag{7}$$

If we view $S_i(\mathbf{w})$ and $S(\mathbf{w})$ only as vector-value functions of $\mathbf{w}$ but ignore how they are formulated using data, (7) is a special case of (6) with $N_+ = 1$ and $N_- = N$.

## 4 Nearly Critical Point and Moreau Envelope Smoothing

We first develop a stochastic algorithm for (5) with $f^l$ defined in (6). To do so, we make the following assumptions, which are satisfied by many smooth $h_{\mathbf{w}}$'s and $\ell$'s.

**Assumption 1** *(a) $s_{ij}(\mathbf{w})$ is smooth and there exists $L \geq 0$ such that*[2] $\|\nabla s_{ij}(\mathbf{w}) - \nabla s_{ij}(\mathbf{v})\| \leq L\|\mathbf{w} - \mathbf{v}\|$ *for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, $i = 1, \ldots, N_+$ and $j = 1, \ldots, N_-$. (b) There exists $B \geq 0$ such that $\|\nabla s_{ij}(\mathbf{w})\| \leq B$ for any $\mathbf{w} \in \mathbb{R}^d$, $i = 1, \ldots, N_+$ and $j = 1, \ldots, N_-$. (c) $F^* > -\infty$.*

Given $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the subdifferential of $f$ is

$$\partial f(\mathbf{w}) = \left\{ \boldsymbol{\xi} \in \mathbb{R}^d \,\middle|\, f(\mathbf{v}) \geq h(\mathbf{w}) + \boldsymbol{\xi}^\top(\mathbf{v} - \mathbf{w}) + o(\|\mathbf{v} - \mathbf{w}\|_2),\ \mathbf{v} \to \mathbf{w} \right\},$$

where each element in $\partial f(\mathbf{w})$ is called a subgradient of $f$ at $\mathbf{w}$. We say $f$ is **$\rho$-weakly convex** for some $\rho \geq 0$ if $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \boldsymbol{\xi}, \mathbf{v} - \mathbf{w} \rangle - \frac{\rho}{2}\|\mathbf{v} - \mathbf{w}\|^2$ for any $\mathbf{v}$ and $\mathbf{w}$ and $\boldsymbol{\xi} \in \partial f(\mathbf{w})$ and say $f$ is **$\rho$-strongly convex** for some $\rho \geq 0$ if $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \boldsymbol{\xi}, \mathbf{v} - \mathbf{w} \rangle + \frac{\rho}{2}\|\mathbf{v} - \mathbf{w}\|^2$ for any $\mathbf{v}$ and $\mathbf{w}$ and $\boldsymbol{\xi} \in \partial f(\mathbf{w})$. It is known that, if $f$ is $\rho$-weakly convex, then $f(\mathbf{w}) + \frac{1}{2\mu}\|\mathbf{w}\|^2$ is a $(\mu^{-1} - \rho)$-strongly convex function when $\mu^{-1} > \rho$.

Under Assumption 1, $\phi_l(S_i(\mathbf{w}))$ is a composite of the closed convex function $\phi_l$ and the smooth map $S_i(\mathbf{w})$. According to Lemma 4.2 in [14], we have the following lemma.

**Lemma 1** *Under Assumption 1, $f^m(\mathbf{w})$ and $f^n(\mathbf{w})$ in (6) are $\rho$-weakly convex with $\rho := N_+N_-L$.*

To solve (5) numerically, we need to overcome the following challenges. (i) $F(\mathbf{w})$ is non-convex even if each $s_{ij}(\mathbf{w})$ is convex. In fact, $F(\mathbf{w})$ is a DC function because, by Lemma 1, we can represent $F(\mathbf{w})$ as the difference of the convex functions $f^n(\mathbf{w}) + \frac{1}{2\mu}\|\mathbf{w}\|^2$ and $f^m(\mathbf{w}) + \frac{1}{2\mu}\|\mathbf{w}\|^2$ with $\mu^{-1} > \rho$. (ii) $F(\mathbf{w})$ is non-smooth due to $\phi_l$ so that finding an approximate critical point (defined below) of $F(\mathbf{w})$ is difficult. (iii) Computing the exact subgradient of $f^l(\mathbf{w})$ for $l = m, n$ requires processing $N_+N_-$ data pairs, which is computationally expensive for a large data set.

Because of challenges (i) and (ii), we have to consider a reasonable goal when solving (5). We say $\mathbf{w} \in \mathbb{R}^d$ is a *critical point* of (5) if $\mathbf{0} \in \partial f^n(\mathbf{w}) - \partial f^m(\mathbf{w})$. Given $\epsilon > 0$, we say $\mathbf{w} \in \mathbb{R}^d$ is an *$\epsilon$-critical point* of (5) if there exists $\boldsymbol{\xi} \in \partial f^n(\mathbf{w}) - \partial f^m(\mathbf{w})$ such that $\|\boldsymbol{\xi}\| \leq \epsilon$. A critical point can only be achieved asymptotically in general.[3] Within finitely many iterations, there also exists no algorithm that can find an $\epsilon$-critical point unless at least one of $f^m$ and $f^n$ is smooth, e.g., [63]. Since $f^m$ and $f^n$ are both non-smooth, we have to consider a weaker but achievable target, which is a nearly $\epsilon$-critical point defined below.

**Definition 1** *Given $\epsilon > 0$, we say $\mathbf{w} \in \mathbb{R}^d$ is a nearly $\epsilon$-critical point of (5) if there exist $\boldsymbol{\xi}$, $\mathbf{w}'$, and $\mathbf{w}'' \in \mathbb{R}^d$ such that $\boldsymbol{\xi} \in \partial f^n(\mathbf{w}') - \partial f^m(\mathbf{w}'')$ and $\max\{\|\boldsymbol{\xi}\|, \|\mathbf{w} - \mathbf{w}'\|, \|\mathbf{w} - \mathbf{w}''\|\} \leq \epsilon$.*

Definition 1 is reduced to the $\epsilon$-stationary point defined by [39, 52] when $\mathbf{w}$ equals $\mathbf{w}'$ or $\mathbf{w}''$. However, obtaining their $\epsilon$-stationary point requires exactly solving the proximal mapping of $f^m$ or $f^n$ while finding a nearly $\epsilon$-critical point requires only solving the proximal mapping inexactly. When $\mathbf{w}$ is generated by a stochastic algorithm, we also call $\mathbf{w}$ a nearly $\epsilon$-critical point if it satisfies Definition 1 with each $\|\cdot\|$ replaced by $\mathbb{E}\|\cdot\|$.

Motivated by [52] and [39], we approximate non-smooth $F(\mathbf{w})$ by a smooth function using the Moreau envelopes. Given a proper, $\rho$-weakly convex and closed function $f$ on $\mathbb{R}^d$, the *Moreau envelope* of $f$ with the smoothing parameter $\mu \in (0, \rho^{-1})$ is defined as

$$f_\mu(\mathbf{w}) := \min_{\mathbf{v}} \left\{ f(\mathbf{v}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2 \right\} \tag{8}$$

and the *proximal mapping* of $f$ is defined as

$$\mathbf{v}_{\mu f}(\mathbf{w}) := \arg\min_{\mathbf{v}} \left\{ f(\mathbf{v}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2 \right\}. \tag{9}$$

Note that the $\mathbf{v}_{\mu f}(\mathbf{w})$ is unique because the minimization above is strongly convex. Standard results show that $f_\mu(\mathbf{w})$ is smooth with $\nabla f_\mu(\mathbf{w}) = \mu^{-1}(\mathbf{w} - \mathbf{v}_{\mu f}(\mathbf{w}))$ and $\mathbf{v}_{\mu f}(\mathbf{w})$ is $(1 - \mu\rho)^{-1}$-Lipschitz continuous. See Proposition 13.37 in [48] and Proposition 1 in [52]. Hence, using the Moreau envelope, we can construct a smooth approximation of (5) as follows

$$\min_{\mathbf{w}} \left\{ F^\mu := f^n_\mu(\mathbf{w}) - f^m_\mu(\mathbf{w}) \right\}. \tag{10}$$

[2] In this paper, $\|\cdot\|$ represents Euclidean norm.

[3] A stronger notion than criticality is (directional) stationarity, which can also be achieved asymptotically [45].

Function $F^\mu$ has the following properties. The first property is shown in [52]. We give the proof for the second in Appendix B.

**Lemma 2** *Suppose Assumption 1 holds and $\mu > \rho^{-1}$ with $\rho$ defined in Lemma 1. The following claims hold*

1. $\nabla F^\mu(\mathbf{w}) = \mu^{-1}(\mathbf{v}_{\mu f^m}(\mathbf{w}) - \mathbf{v}_{\mu f^n}(\mathbf{w}))$ *and it is $L_\mu$-Lipschitz continuous with* $L_\mu = \frac{2}{\mu - \mu^2\rho}$.
2. *If $\bar{\mathbf{v}}$ and $\mathbf{w}$ are two random vectors such that $\mathbb{E}\|\nabla F^\mu(\mathbf{w})\|^2 \leq \min\{1, \mu^{-2}\}\epsilon^2/4$ and $\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^l}(\mathbf{w})\|^2 \leq \epsilon^2/4$ for either $l = m$ or $l = n$, then $\bar{\mathbf{v}}$ is a nearly $\epsilon$-critical points of (5).*

Since $F^\mu$ is smooth, we can directly apply a first-order method for smooth non-convex optimization to (10). To do so, we need to evaluate $\nabla F^\mu(\mathbf{w})$, which requires computing $\mathbf{v}_{\mu f^m}(\mathbf{w})$ and $\mathbf{v}_{\mu f^n}(\mathbf{w})$, i.e., exactly solving (9) with $f = f^m$ and $f = f^n$, respectively. Computing the subgradients of $f^m$ and $f^n$ require processing $N_+N_-$ data pairs which is costly. Unfortunately, the standard approach of sampling over data pairs does not produce unbiased stochastic subgradients of $f^m$ and $f^n$ due to the composite structure $\phi_l(S_i(\mathbf{w}))$. In the next section, we will discuss a solution to overcome this challenge and approximate $\mathbf{v}_{\mu f^m}(\mathbf{w})$ and $\mathbf{v}_{\mu f^n}(\mathbf{w})$, which leads to an efficient algorithm for (10).

## 5 Algorithm for pAUC Optimization

Consider (10) with $f^l$ defined in (6) for $l = m$ and $n$. To avoid of processing $N_+N_-$ data points, one method is to introduce dual variables $\mathbf{p}_i = (p_{ij})_{j=1}^{N_-}$ for $i = 1, \ldots, N_+$ and formulate $f^l$ as

$$f^l(\mathbf{w}) = \max_{\mathbf{p}_i \in \mathcal{P}^l, i=1,\ldots,N_+} \left\{ \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} p_{ij} s_{ij}(\mathbf{w}) \right\}, \tag{11}$$

where $\mathcal{P}^l = \{\mathbf{p} \in \mathbb{R}^{N_-} | \sum_{j=1}^{N_-} p_j = l,\ p_j \in [0,1]\}$. Then (10) can be reformulated as a min-max problem and solved by a primal-dual stochastic gradient method (e.g. [46]). However, the maximization in (11) involves $N_+N_-$ decision variables and equality constraints, so the per-iteration cost is still $O(N_+N_-)$ even after using stochastic gradients.

To further reduce the per-iteration cost, we take the dual form of the maximization in (11) (see Lemma 4 in Appendix B) and formulate $f^l$ as

$$f^l(\mathbf{w}) = \min_{\boldsymbol{\lambda}} \left\{ g^l(\mathbf{w}, \boldsymbol{\lambda}) := l\mathbf{1}^\top \boldsymbol{\lambda} + \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} [s_{ij}(\mathbf{w}) - \lambda_i]_+ \right\}, \tag{12}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{N_+})$. Hence, (9) with $f = f^l$ for $l = m$ and $n$ can be reformulated as

$$\min_{\mathbf{v}, \boldsymbol{\lambda}} \left\{ g^l(\mathbf{v}, \boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2 \right\}. \tag{13}$$

Note that $g^l(\mathbf{v}, \boldsymbol{\lambda})$ is jointly convex in $\mathbf{v}$ and $\boldsymbol{\lambda}$ when $\mu^{-1} > \rho = N_+N_-L$ (see Lemma 3 in Appendix B). Thanks to formulation (13), we can construct stochastic subgradient of $g^l$ and apply coordinate update to $\boldsymbol{\lambda}$ by sampling indexes $i$'s and $j$'s, which significantly reduce the computational cost when $N_+$ and $N_-$ are both large. We present this standard stochastic block coordinate descent (SBCD) method for solving (13) in Algorithm 1 and present its convergence property as follows.

**Proposition 1** *Suppose Assumption 1 holds and $\mu^{-1} > \rho = N_+N_-L$, $\theta_t = \frac{\text{dist}(\boldsymbol{\lambda}^{(0)}, \Lambda^*)}{\sqrt{IT}N_-}$ and $\eta_t = \frac{\|\mathbf{v}_{\mu f^l}(\mathbf{w}) - \mathbf{w}\|}{N_+N_-B\sqrt{T}}$ for any $t$ in Algorithm 1. It holds that*

$$\left(\frac{1}{2\mu} - \frac{\rho}{2}\right)\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^l}(\mathbf{w}))\|^2 \leq \frac{N_+N_-}{\sqrt{IT}}\text{dist}(\boldsymbol{\lambda}^{(0)}, \Lambda^*) + \frac{N_+N_-B}{2\sqrt{T}}\|\mathbf{v}_{\mu f^l}(\mathbf{w}) - \mathbf{w}\| + \frac{\|\mathbf{v}_{\mu f^l}(\mathbf{w}) - \mathbf{w}\|^2}{2\mu T},$$

*where* $\Lambda^* = \arg\min_{\boldsymbol{\lambda}} g^l(\mathbf{v}_{\mu f^l}(\mathbf{w}), \boldsymbol{\lambda})$.

Using Algorithm 1 to compute an approximation of $\mathbf{v}_{\mu f^l}(\mathbf{w})$ for $l = m$ and $n$ and thus, an approximation of $\nabla F^\mu(\mathbf{w})$, we can apply an approximate gradient descent (AGD) method to (10) and find a nearly $\epsilon$-critical point of (5) according to Lemma 2. We present the AGD method in Algorithm 2 and its convergence property as follows.

**Algorithm 1** Stochastic Block Coordinate Descent for (13): $(\bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}})$ =SBCD$(\mathbf{w}, \boldsymbol{\lambda}, T, \mu, l)$

1: **Input:** Initial solution $(\mathbf{w}, \boldsymbol{\lambda})$, the number of iterations $T$, $\mu > 0$, an integer $l > 0$ and sample sizes $I$ and $J$.
2: Set $(\mathbf{v}^{(0)}, \boldsymbol{\lambda}^{(0)}) = (\mathbf{w}, \boldsymbol{\lambda})$ and choose $(\eta_t, \theta_t)_{t=0}^{T-1}$.
3: **for** $t = 0$ **to** $T-1$ **do**
4: Sample $\mathcal{I}_t \subset \{1, \dots, N_+\}$ with $|\mathcal{I}_t| = I$ and sample $\mathcal{J}_t \subset \{1, \dots, N_-\}$ with $|\mathcal{J}_t| = J$.
5: Compute stochastic subgradient w.r.t. $\mathbf{v}$:

$$G_{\mathbf{v}}^{(t)} = \frac{N_+ N_-}{IJ} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t} \nabla s_{ij}(\mathbf{v}^{(t)}) \mathbf{1}\left(s_{ij}(\mathbf{v}^{(t)}) > \lambda_i^{(t)}\right)$$

6: Proximal stochastic subgradient update on $\mathbf{v}$:

$$\mathbf{v}^{(t+1)} = \arg\min_{\mathbf{v}} (G_{\mathbf{v}}^{(t)})^\top \mathbf{v} + \frac{\|\mathbf{v} - \mathbf{w}\|^2}{2\mu} + \frac{\|\mathbf{v} - \mathbf{v}^{(t)}\|^2}{2\eta_t} \tag{14}$$

7: Compute stochastic subgradient w.r.t. $\lambda_i$ for $i \in \mathcal{I}_t$:

$$G_{\lambda_i}^{(t)} = l - \frac{N_-}{J} \sum_{j \in \mathcal{J}_t} \mathbf{1}\left(s_{ij}(\mathbf{v}^{(t)}) > \lambda_i^{(t)}\right) \quad \text{for } i \in \mathcal{I}_t$$

8: Stochastic block subgradient update on $\lambda_i$ for $i \in \mathcal{I}_t$:

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} - \theta_t G_{\lambda_i}^{(t)} \quad \text{for } i \in \mathcal{I}_t \quad \text{and} \quad \lambda_i^{(t+1)} = \lambda_i^{(t)} \quad \text{for } i \notin \mathcal{I}_t. \tag{15}$$

9: **end for**
10: **Output:** $(\bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}}) = \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)})$.

**Algorithm 2** Approximate Gradient Descent (AGD) for (10)

1: **Input:** Initial solutions $(\mathbf{w}^{(0)}, \bar{\boldsymbol{\lambda}}_m^{(0)}, \bar{\boldsymbol{\lambda}}_n^{(0)})$, the number of iterations $K$, $\mu > \rho^{-1}$, $\gamma > 0$, $m = \alpha N_-$ and $n = \beta N_-$.
2: **for** $k = 0$ **to** $K-1$ **do**
3: $(\bar{\mathbf{v}}_m^{(k)}, \bar{\boldsymbol{\lambda}}_m^{(k+1)})$ =SBMD$(\mathbf{w}^{(k)}, \bar{\boldsymbol{\lambda}}_m^{(k)}, T_k, \mu, m)$
4: $(\bar{\mathbf{v}}_n^{(k)}, \bar{\boldsymbol{\lambda}}_n^{(k+1)})$ =SBMD$(\mathbf{w}^{(k)}, \bar{\boldsymbol{\lambda}}_n^{(k)}, T_k, \mu, n)$
5: $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \gamma\mu^{-1}(\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)})$
6: **end for**
7: **Output:** $\bar{\mathbf{v}}_n^{(\bar{k})}$ with $\bar{k}$ sampled from $\{0, \dots, K-1\}$.

**Theorem 1** *Suppose Assumption 1 holds and Algorithm 1 is called in iteration $k$ of Algorithm 2 with parameters $\mu^{-1} > \rho = N_+ N_- L$, $\theta_t = \frac{dist(\bar{\boldsymbol{\lambda}}^{(k)}, \Lambda_k^*)}{\sqrt{T T_k} N_-}$, $\eta_t = \frac{\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|}{N_+ N_- B \sqrt{T_k}}$ for any $t$, and*

$$T_k = \max\left\{\frac{144 N_+^2 N_-^2 D_l^2 (k+1)^2}{I(\mu^{-1} - \rho)^2}, \frac{4 N_+^2 N_-^2 \mu^2 l^2 B^2 (k+1)^2}{(\mu^{-1} - \rho)^2}, \frac{6\mu l^2 B^2 (k+1)}{2(\mu^{-1} - \rho)^2}\right\}$$

*where $\Lambda_k^* = \arg\min_{\boldsymbol{\lambda}} g^l(\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}), \boldsymbol{\lambda})$ and*

$$D_l := \max\left\{dist(\bar{\boldsymbol{\lambda}}_l^{(0)}, \Lambda_0^*), \quad \frac{1}{2}\left(\frac{1}{\mu} - \rho\right) + \frac{\mu l^2 B^2}{2} + N_+ B + \frac{N_+ B}{1 - \mu\rho}\left(\frac{2\gamma}{\mu} + \gamma n B + \gamma m B\right)\right\}. \tag{16}$$

*Then $\bar{\mathbf{v}}_n^{(\bar{k})}$ is a nearly $\epsilon$-critical point of (5) with $f^l$ defined in (6) with $K$ no more than*

$$K = \max\left\{\frac{16\mu^2}{\gamma \min\{1, \mu^2\}\epsilon^2}\left(F(\mathbf{v}_{\mu f^n}(\mathbf{w}^{(0)})) - F^*\right), \frac{96}{\min\{1, \mu^2\}\epsilon^2} \log\left(\frac{96}{\min\{1, \mu^2\}\epsilon^2}\right)\right\}. \tag{17}$$

According to Theorem 1, to find a nearly $\epsilon$-critical point of (5), we need $K = \tilde{O}(\epsilon^{-2})$ iterations in Algorithm 2 and $\sum_{k=0}^{K-1} T_k = O(K^3) = \tilde{O}(\epsilon^{-6})$ iterations of Algorithm 1 in total across all calls.

**Remark 1 (Challenges in proving Theorem 1)** *Suppose we can set $T_k$ in lines 3 and 4 of Algorithm 2 appropriately such that the approximation errors $\mathbb{E}\|\bar{\mathbf{v}}_m^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)})\|^2$ and $\mathbb{E}\|\bar{\mathbf{v}}_n^{(k)} -$*

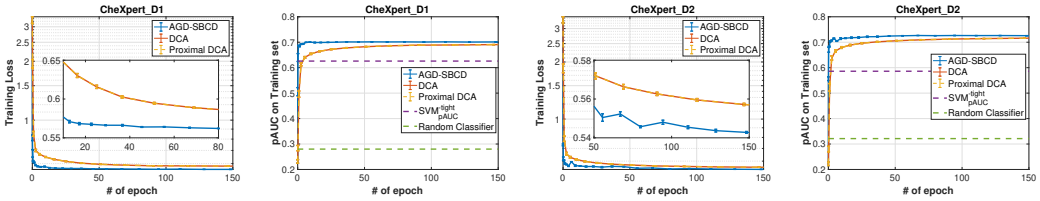


Figure 1: Results for Patial AUC Maximization of D1 and D2. (Results of D3, D4 and D5 are shown in Appendix E.3 Figure 3)

$\mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2$ *are both* $O(1/k)$. *We can then prove that Algorithm 2 finds a nearly* $\epsilon$*-critical point within* $K = \tilde{O}(\epsilon^{-2})$ *iterations and the total complexity is* $\sum_{k=0}^{K-1} T_k$. *This is just a standard idea. However, by Proposition 1, such a* $T_k$ *must be* $\Theta(k^2(dist^2(\bar{\boldsymbol{\lambda}}^{(k)}, \Lambda_k^*) + \|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|^2))$ *where* $dist^2(\bar{\boldsymbol{\lambda}}^{(k)}, \Lambda_k^*)$ *and* $\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|^2$ *also change with* $k$. *Then it is not clear what the order of* $T_k$ *is. By a novel proving technique based on the (linear) error-bound condition of* $g^l(\mathbf{w}, \boldsymbol{\lambda})$ *with respect to* $\boldsymbol{\lambda}$, *we prove that both* $dist^2(\bar{\boldsymbol{\lambda}}^{(k)}, \Lambda_k^*)$ *and* $\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|^2$ *are* $O(1)$ *(see (27) and (30) in Appendix D) which ensures that* $T_k = \Theta(k^2)$ *and thus the total complexity is* $\sum_{k=0}^{K-1} T_k = O(K^3) = \tilde{O}(\epsilon^{-6})$.

**Remark 2 (Analysis of sensitivity of the algorithm to $\mu$)** *For the interesting case where* $\rho \geq 1$, *we have* $\mu < 1/\rho < 1$. *In this case, we can derive that the order of dependency on* $\mu$ *is* $O(\frac{1}{\epsilon^6 \mu^6})$ *and the optimal choice of* $\mu$ *is thus* $\Theta(\rho^{-1})$, *e.g.,* $\mu = \frac{1}{2\rho}$, *which leads to a complexity of* $O(\rho^6/\epsilon^6)$. *We present the convergence curves and the test performance of our method when applied to training a linear model with* $\mu$ *of different values in Appendix E.7.*

The technique in the previous sections can be directly applied to minimize the SoRR loss, which is formulated as (5) but with $f^l$ defined in (7). Due to the limit of space, we present the algorithm for minimizing the SoRR loss and its convergence result in Appendix A.

## 6 Numerical Experiments

In this section, we demonstrate the effectiveness of our algorithm AGD-SBCD for pAUC maximization and SoRR loss minimization problems (see Appendix E.1 for details). All experiments are conducted in Python and Matlab on a computer with the CPU 2GHz Quad-Core Intel Core i5 and the GPU NVIDIA GeForce RTX 2080 Ti. All datasets we used are publicly available and contain no personally identifiable information and offensive contents.

### 6.1 Partial AUC Maximization

For maximizing pAUC, we focus on large-scale imbalanced medical dataset CheXpert [25], which is licensed under CC-BY-SA and has 224,316 images. We construct five binary classification tasks with the logistic loss $\ell(z) = \log(1 + \exp(-z))$ for predicting five popular diseases, Cardiomegaly (D1), Edema (D2), Consolidation (D3), Atelectasis (D4), and P. Effusion (D5).

For comparison of training convergence, we consider different methods for optimizing the partial AUC. We compare with three baselines, DCA [24] (see Appendix E.4 for details), proximal DCA [62] (see Appendix E.5 for details) and SVM$_{pAUC}$-tight [42]. Since DCA, proximal DCA and SVM$_{pAUC}$-tight cannot be applied to deep neural networks, we focus on linear model and use a pre-trained deep neural network to extract a fixed dimensional feature vectors of 1024. The deep neural network was trained by optimizing the cross-entropy loss following the same setting as in [69].

For three baselines and our algorithm, the process to tune the hyper-parameters is explained in Appendix E.2. In Figure 1 and Figure 3 in Appendix E.3, we show how the training loss (the objective value of (3)) and normalized partial AUC on the training data change with the number of epochs.

Table 1: Comparison on the CheXpert training data. From left to right, the columns are the tasks, the pAUCs returned by SVM$_{pAUC}$-tight, the CPU time (in seconds) SVM$_{pAUC}$-tight takes, the CPU and GPU time AGD-SBCD uses to exceed SVM$_{pAUC}$-tight's pAUCs, the final pAUCs returned by AGD-SBCD, and the CPU and GPU time (in seconds) AGD-SBCD takes to return the final pAUCs.

| Methods | SVM$_{pAUC}$-tight | | AGD-SBCD | | | | |
|---|---|---|---|---|---|---|---|
| Tasks | pAUC | CPU time | CPU time (epoch) to outperform | GPU time to outperform | pAUC | CPU time | GPU time |
| D1 | 0.6259 | 95.14 | 2.91 (0.23) | 1.85 | 0.7005±0.0003 | 118.32 | 82.13 |
| D2 | 0.5860 | 90.83 | 3.36 (0.23) | 1.93 | 0.7214±0.0024 | 415.66 | 247.29 |
| D3 | 0.3745 | 90.56 | 3.26 (0.23) | 1.84 | 0.4910±0.0006 | 181.70 | 104.55 |
| D4 | 0.3895 | 89.64 | 10.09 (0.63) | 8.38 | 0.4616±0.0006 | 187.36 | 158.14 |
| D5 | 0.7267 | 90.86 | 3.97 (0.23) | 1.89 | 0.8272±0.0001 | 238.10 | 142.91 |

We observe that for all of these five diseases, our algorithm converges much faster than DCA and proximal DCA and we get a better partial AUC than DCA and proximal DCA.

The comparison between our AGD-SBCD and SVM$_{pAUC}$-tight on training data are shown in Table 1. As shown from the second to the fifth column of Table 1, our algorithm needs only a few seconds to exceed the pAUCs that SVM$_{pAUC}$-tight takes more than one minute to return. As shown from sixth to eighth column, our algorithm eventually improves the pAUC by at least $12\%$ compared with SVM$_{pAUC}$-tight. DCA and proximal DCA are not included in the tables because it computes deterministic subgradients, which leads to a runtime significantly longer than the other two methods. We plot the convergence curves of training pAUC over GPU time for DCA and our algorithm in Figure 7 in Appendix E.8.

To compare the testing performances, we consider the deep neural networks besides the linear model. For linear model, we still compare with DCA and SVM$_{pAUC}$-tight. For deep neural networks, we compare with the naive mini-batch based method (MB) [27] and methods based on different optimization objectives, including the cross-entropy loss (CE) and the AUC min-max margin loss (AUC-M) [70]. We learn the model DenseNet121 from scratch with the CheXpert training data split in train/val=9:1 and the CheXpert validation dataset as the testing set, which has 234 samples. The range of FPRs in pAUC is [0.05, 0.5]. For optimizing CE, we use the standard Adam optimizer. For optimizing AUC-M, we use the PESG optimizer in [70]. We run each method 10 epochs and the learning rate ($c$ in AGD-SBCD) of all methods is tuned from $\{10^{-5} \sim 10^{0}\}$. The mini-batch size is 32. For AGD-SBCD, $T_k$ is set to $50(k+1)^2$, $\mu$ is set to $\frac{10^3}{N_+N_-}$ and $\gamma$ is tuned from $\{0.1, 1, 2\} \times 10^3/(N_+N_-)$. For MB, the learning rate decays in the same way as in [27]. For CE and AUC-M, the learning rate decays 10-fold after every 5 epochs. For AUC-M, we tune the hyperparameter $\gamma$ in {100, 500, 1000}. For each method, the validation set is used to tune the hyperparameters and select the best model across all iterations. The results of the pAUCs on the testing set are reported in Table 2, which shows that our algorithm performs the best for all diseases. The complete ROC curves on the testing set are shown in Appendix E.3.

Table 2: The pAUCs with FPRs between $0.05$ and $0.5$ on the testing sets from the CheXpert data.

| | **Method** | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|---|
| Linear Model | SVM$_{pAUC}$-tight | 0.6538±0.0042 | 0.6038±0.0009 | 0.6946±0.0020 | **0.6521±0.0006** | 0.7994±0.0004 |
| | DCA | 0.6636±0.0093 | 0.8078±0.0030 | 0.7427±0.0257 | 0.6169±0.0208 | 0.8371±0.0022 |
| | Proximal DCA | 0.6615±0.0103 | 0.8041±0.0033 | 0.7064±0.0253 | 0.5945±0.0266 | 0.8352±0.0023 |
| | AGD-SBCD | **0.6721±0.0081** | **0.8257±0.0025** | **0.8016±0.0075** | 0.6340±0.0165 | **0.8500±0.0017** |
| Deep Model | MB | 0.7510±0.0248 | 0.8197±0.0127 | 0.6339±0.0328 | 0.5698±0.0343 | 0.8461±0.0188 |
| | CE | 0.6994±0.0453 | 0.8075±0.0244 | 0.7673±0.0266 | 0.6499±0.0184 | 0.7884±0.0080 |
| | AUC-M | 0.7403±0.0339 | 0.8002±0.0274 | 0.8533±0.0469 | 0.7420±0.0277 | 0.8504±0.0065 |
| | AGD-SBCD | **0.7535±0.0255** | **0.8345±0.0130** | **0.8689±0.0184** | **0.7520±0.0079** | **0.8513±0.0107** |

For deep neural networks, we also learn the model ResNet-20 from scratch with the CIFAR-10-LT and the Tiny-ImageNet-200-LT datasets, which are constructed similarly as in [66]. Details about these two datasets are summarized in Appendix E.6. The range of FPRs in pAUC is [0.05, 0.5]. The process of tuning hyperparameters is the same as that for CheXpert. The results of the pAUCs on the

testing set are reported in Table 3, which shows that our algorithm performs the best for these two long-tailed datasets.

Table 3: The pAUCs with FPRs between $0.05$ and $0.5$ on the testing sets from the CIFAR-10-LT and the Tiny-ImageNet-200-LT Datasets.

| | **Dataset** | MB | CE | AUC-M | AGD-SBCD |
|---|---|---|---|---|---|
| Deep Model | CIFAR-10-LT | 0.9337±0.0043 | 0.9016±0.0137 | 0.9323±0.0055 | **0.9408±0.0084** |
| | Tiny-ImageNet-200-LT | 0.6445±0.0214 | 0.6549±0.008 | 0.6497±0.009 | **0.6594±0.0192** |

## 7 Conclusion

Most existing methods for optimizing pAUC are deterministic and only have an asymptotic convergence property. We formulate pAUC optimization as a non-smooth DC program and develop a stochastic subgradient method based on the Moreau envelope smoothing technique. We show that our method finds a nearly $\epsilon$-critical point in $\tilde{O}(\epsilon^{-6})$ iterations and demonstrate its performance numerically. A limitation of this paper is the smoothness assumption on $s_{ij}(\mathbf{w})$, which does not hold for some models, e.g., neural networks using ReLU activation functions. It is a future work to extend our results for non-smooth models.

## Acknowledgements

This work was jointly supported by the University of Iowa Jumpstarting Tomorrow Program and NSF award 2147253. T. Yang was also supported by NSF awards 2110545 and 1844403, and Amazon research award. We thank Zhishuai Guo, Zhuoning Yuan and Qi Qi for discussing about processing the image dataset.

## References

[1] Hadi Abbaszadehpeivasti, Etienne de Klerk, and Moslem Zamani. On the rate of convergence of the difference-of-convex algorithm (dca). *arXiv preprint arXiv:2109.13566*, 2021. 3

[2] AD Alexandroff. Surfaces represented by the difference of convex functions. In *Doklady Akademii Nauk SSSR (NS)*, volume 72, pages 613–616, 1950. 3

[3] Le Thi Hoai An and Pham Dinh Tao. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of operations research*, 133(1):23–46, 2005. 2, 3

[4] Le Thi Hoai An, Huynh Van Ngai, Pham Dinh Tao, and Luu Hoang Phuc Hau. Stochastic difference-of-convex algorithms for solving nonconvex optimization problems. *arXiv preprint arXiv:1911.04334*, 2019. 3

[5] Nguyen Thai An and Nguyen Mau Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2017. 3

[6] Francisco J Aragón Artacho, Ronan MT Fleming, and Phan T Vuong. Accelerating the dc algorithm for smooth functions. *Mathematical Programming*, 169(1):95–118, 2018. 3

[7] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. 1

[8] Andrew P Bradley. Half-auc for the evaluation of sensitive or specific classifiers. *Pattern Recognition Letters*, 38:93–98, 2014. 2

[9] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011. 14, 22

[10] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018. 2

[11] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019. 2

[12] Qi Deng and Chenghao Lan. Efficiency of coordinate descent methods for structured nonconvex optimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 74–89. Springer, 2020. 3

[13] Lori E Dodd and Margaret S Pepe. Partial auc estimation and regression. *Biometrics*, 59(3):614–623, 2003. 2, 3

[14] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019. 5, 16

[15] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 14, 22

[16] Rachid Ellaia. *Contribution à l'analyse et l'optimisation de différence de fonctions convexes*. PhD thesis, Université Paul Sabatier, 1984. 3

[17] Yanbo Fan, Siwei Lyu, Yiming Ying, and Bao-Gang Hu. Learning with average top-k loss. *arXiv preprint arXiv:1705.08826*, 2017. 2

[18] D. Gabay. Minimizing the difference of two convex functions. I. Algorithms based on exact regularization. 1982. 3

[19] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 1

[20] Philip Hartman. On functions representable as a difference of convex functions. *Pacific Journal of Mathematics*, 9(3):707–713, 1959. 3

[21] Lulu He, Jimin Ye, et al. Accelerated proximal stochastic variance reduction for dc optimization. *Neural Computing and Applications*, 33(20):13163–13181, 2021. 3

[22] J-B Hiriart-Urruty. Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. In *Convexity and duality in optimization*, pages 37–70. Springer, 1985. 3

[23] J-B Hiriart-Urruty. How to regularize a difference of convex functions. *Journal of mathematical analysis and applications*, 162(1):196–209, 1991. 3

[24] Shu Hu, Yiming Ying, Siwei Lyu, et al. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 8, 23

[25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. 8, 14

[26] Yulei Jiang, Charles E Metz, and Robert M Nishikawa. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750, 1996. 3

[27] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Online and stochastic gradient methods for non-decomposable loss functions. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 9

[28] Koulik Khamaru and Martin Wainwright. Convergence guarantees for a class of non-convex and non-smooth optimization problems. In *International Conference on Machine Learning*, pages 2601–2610. PMLR, 2018. 3

[29] Osamu Komori and Shinto Eguchi. A boosting method for maximizing the partial area under the roc curve. *BMC bioinformatics*, 11(1):1–17, 2010. 2, 3

[30] Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018. 2, 3

[31] 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 *International Conference on Machine Learning*, pages 3394–3403. PMLR, 2017. 3

[32] Thomas Lipp and Stephen Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016. 3

[33] Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021. 2

[34] Hua Ma, Andriy I Bandos, Howard E Rockette, and David Gur. On use of partial area under the roc curve for evaluation of diagnostic performance. *Statistics in medicine*, 32(20):3449–3458, 2013. 2

[35] Runchao Ma, Qihang Lin, and Tianbao Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564. PMLR, 2020. 2

[36] Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. *arXiv preprint arXiv:1306.4650*, 2013. 3

[37] Donna Katzman McClish. Analyzing a portion of the roc curve. *Medical decision making*, 9(3):190–195, 1989. 3

[38] Abdellatif Moudafi. On the difference of two maximal monotone operators: Regularization and algorithmic approaches. *Applied mathematics and computation*, 202(2):446–452, 2008. 3

[39] Abdellatif Moudafi. A complete smooth regularization of dc optimization problems. 2021. 3, 5

[40] Abdellatif Moudafi and Paul-Emile Maingé. On the convergence of an approximate proximal method for dc functions. *Journal of computational Mathematics*, pages 475–480, 2006. 3

[41] Harikrishna Narasimhan and Shivani Agarwal. A structural svm based approach for optimizing partial auc. In *International Conference on Machine Learning*, pages 516–524. PMLR, 2013. 2, 3

[42] Harikrishna Narasimhan and Shivani Agarwal. $\text{SVM}^{\text{tight}}_{\text{pAUC}}$: a new support vector method for optimizing partial auc based on a tight convex upper bound. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 167–175, 2013. 2, 3, 8, 23

[43] Harikrishna Narasimhan and Shivani Agarwal. Support vector algorithms for optimizing the partial area under the roc curve. *Neural computation*, 29(7):1919–1963, 2017. 2, 3

[44] Atsushi Nitanda and Taiji Suzuki. Stochastic difference of convex algorithm and its application to training deep boltzmann machines. In *Artificial intelligence and statistics*, pages 470–478. PMLR, 2017. 3

[45] Jong-Shi Pang, Meisam Razaviyayn, and Alberth Alvarado. Computing b-stationary points of nonsmooth dc programs. *Mathematics of Operations Research*, 42(1):95–118, 2017. 3, 5

[46] 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*, pages 1–35, 2021. 2, 6

[47] Maria Teresa Ricamato and Francesco Tortorella. Partial auc maximization in a linear combination of dichotomizers. *Pattern Recognition*, 44(10-11):2669–2677, 2011. 3

[48] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. 5

[49] Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: How and why. In *International Conference on Machine Learning*, pages 793–801. PMLR, 2016. 2

[50] João Carlos O Souza, Paulo Roberto Oliveira, and Antoine Soubeyran. Global convergence of a proximal linearized algorithm for difference of convex functions. *Optimization Letters*, 10(7):1529–1539, 2016. 3

[51] Bharath K Sriperumbudur and Gert RG Lanckriet. On the convergence of the concave-convex procedure. In *Nips*, volume 9, pages 1759–1767. Citeseer, 2009. 3

[52] Kaizhao Sun and Xu Andy Sun. Algorithms for difference-of-convex (dc) programs based on difference-of-moreau-envelopes smoothing. *arXiv preprint arXiv:2104.01470*, 2021. 2, 3, 5, 6, 15, 22

[53] Wen-yu Sun, Raimundo JB Sampaio, and MAB Candido. Proximal point algorithm for minimization of dc function. *Journal of computational Mathematics*, pages 451–462, 2003. 3

[54] Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997. 2, 3

[55] Pham Dinh Tao and Le Thi Hoai An. A dc optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998. 3

[56] Hoai An Le Thi, Hoang Phuc Hau Luu, and Tao Pham Dinh. Online stochastic dca with applications to principal component analysis. *arXiv preprint arXiv:2108.02300*, 2021. 3

[57] Mary Lou Thompson and Walter Zucchini. On the statistical analysis of roc curves. *Statistics in medicine*, 8(10):1277–1290, 1989. 3

[58] Hoang Tuy. Dc optimization: theory, methods and algorithms. In *Handbook of global optimization*, pages 149–216. Springer, 1995. 3

[59] Naonori Ueda and Akinori Fujino. Partial auc maximization via nonlinear scoring functions. *arXiv preprint arXiv:1806.04838*, 2018. 3

[60] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992. 2

[61] Zhanfeng Wang and Yuan-Chin Ivan Chang. Marker selection via maximizing the partial area under the roc curve of linear risk scores. *Biostatistics*, 12(2):369–385, 2011. 3

[62] Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational optimization and applications*, 69(2):297–324, 2018. 3, 8

[63] 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 *International Conference on Machine Learning*, pages 6942–6951. PMLR, 2019. 3, 5

[64] Hanfang Yang, Kun Lu, Xiang Lyu, and Feifang Hu. Two-way partial auc and its properties. *Statistical methods in medical research*, 28(1):184–195, 2019. 2, 3

[65] Zhiyong Yang, Qianqian Xu, Shilong Bao, Xiaochun Cao, and Qingming Huang. Learning with multiclass auc: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[66] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. When all we need is a piece of the pie: A generic framework for optimizing two-way partial auc. In *International Conference on Machine Learning*, pages 11820–11829. PMLR, 2021. 3, 9

[67] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. Optimizing two-way partial auc with an end-to-end framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[68] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016. 3

[69] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 2020. 8

[70] 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. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3020–3029. IEEE, 2021. 9

[71] Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003. 3

[72] 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. *arXiv preprint arXiv:2203.00176*, 2022. 3

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Algorithm 1 and 2 and Section 6.
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We are not aware of any negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 6.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite them as [9, 15, 25]
   (b) Did you mention the license of the assets? [Yes] See Section 6.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We only use publicly available data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 6.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendix

## A Algorithm for Sum of Range Optimization

The technique in the previous sections can be directly applied to minimize the SoRR loss, which is formulated as (5) but with $f^l$ defined in (7). Since (7) is a special case of (6) with $N_+ = 1$ and $N_- = N$, we can again formulate subproblem (9) with $f = f^l$ as (13) with $\boldsymbol{\lambda} = \lambda$ being a scalar. Since $\lambda$ is a scalar, when solving (13), we no longer use block coordinate update but only need to sample over indexes $j = 1, \ldots, N$ to construct stochastic subgradients. We present the stochastic subgradient (SGD) method for (13) in Algorithm 3. Next, we apply Algorithm 2 with SBCD in lines 3 and 4 replaced by SGD. The convergence result in this case is directly from Theorem 1.

**Algorithm 3** Stochastic Subgradient Descent for SoRR: $(\bar{\mathbf{v}}, \bar{\lambda})$ =SGD$(\mathbf{w}, \lambda, T, \mu, l)$

1: **Input:** Initial solution $(\mathbf{w}, \lambda)$, the number of iterations $T$, $\mu > \rho^{-1}$, an integer $l > 0$ and sample size $J$.
2: Set $(\mathbf{v}^{(0)}, \lambda^{(0)}) = (\mathbf{w}, \lambda)$ and choose $(\eta_t, \theta_t)_{t=0}^{T-1}$.
3: **for** $t = 0$ **to** $T - 1$ **do**
4: Sample $\mathcal{J}_t \subset \{1, \ldots, N\}$ with $|\mathcal{J}_t| = J$.
5: Compute stochastic subgradient w.r.t. $\mathbf{v}$:

$$G_{\mathbf{v}}^{(t)} = \frac{N}{J} \sum_{j \in \mathcal{J}_t} \nabla s_j(\mathbf{v}^{(t)}) \mathbf{1}\left(s_j(\mathbf{v}^{(t)}) > \lambda_i^{(t)}\right)$$

6: Proximal stochastic subgradient update on $\mathbf{v}$:

$$\mathbf{v}^{(t+1)} = \arg\min_{\mathbf{v}} (G_{\mathbf{v}}^{(t)})^\top \mathbf{v} + \frac{\|\mathbf{v} - \mathbf{w}\|^2}{2\mu} + \frac{\|\mathbf{v} - \mathbf{v}^{(t)}\|^2}{2\eta_t}$$

7: Compute stochastic subgradient w.r.t. $\lambda$:

$$G_{\lambda}^{(t)} = l - \frac{N}{J} \sum_{j \in \mathcal{J}_t} \mathbf{1}\left(s_j(\mathbf{v}^{(t)}) > \lambda^{(t)}\right)$$

8: Stochastic subgradient update on $\lambda$:

$$\lambda^{(t+1)} = \lambda^{(t)} - \eta_t G_{\lambda}^{(t)}$$

9: **end for**
10: **Output:** $(\bar{\mathbf{v}}, \bar{\lambda}) = \frac{1}{T} \sum_{t=0}^{T-1} (\mathbf{v}^{(t)}, \lambda^{(t)})$.

**Corollary 1** *Suppose Assumption 1 holds with $N_+ = 1$, $N_- = N$ and $s_{ij}(\mathbf{w}) = s_j(\mathbf{w})$ and SBCD in Algorithm 2 are replaced by SGD (Algorithm 3). Suppose $\theta_t$, $\eta_t$, and $T_k$ are set the same as in Theorem 1 when Algorithm 3 is called in iteration $k$ of Algorithm 2. Then $\bar{\mathbf{v}}_n^{(k)}$ is an nearly $\epsilon$-critical point of (5) with $f^l$ defined in (7) with $K$ no more than (17).*

## B Proofs of Lemmas

*Proof.*[of Lemma 2] We will only prove the second conclusion in Lemma 2 since the first conclusion has been shown in Proposition 1 in [52].

Suppose $\mathbb{E}\|\nabla F^\mu(\mathbf{w})\|^2 \leq \min\{1, \mu^{-2}\}\epsilon^2/4$. By the first conclusion in Lemma 2, we must have $\mu^{-2}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}) - \mathbf{v}_{\mu f^n}(\mathbf{w})\|^2 \leq \min\{1, \mu^{-2}\}\epsilon^2/4$. By the optimality conditions satisfied by $\mathbf{v}_{\mu f^m}(\mathbf{w})$ and $\mathbf{v}_{\mu f^n}(\mathbf{w})$, there exist $\boldsymbol{\xi}_m \in \partial f^m(\mathbf{v}_{\mu f^m}(\mathbf{w}))$ and $\boldsymbol{\xi}_n \in \partial f^n(\mathbf{v}_{\mu f^n}(\mathbf{w}))$ such that

$$\boldsymbol{\xi}_m + \mu^{-1}(\mathbf{v}_{\mu f^m}(\mathbf{w}) - \mathbf{w}) = \mathbf{0} = \boldsymbol{\xi}_n + \mu^{-1}(\mathbf{v}_{\mu f^n}(\mathbf{w}) - \mathbf{w}),$$

which implies $\boldsymbol{\xi} = \boldsymbol{\xi}_n - \boldsymbol{\xi}_m = \mu^{-1}(\mathbf{v}_{\mu f^m}(\mathbf{w}) - \mathbf{v}_{\mu f^n}(\mathbf{w})) \in \partial f^n(\mathbf{v}_{\mu f^n}(\mathbf{w})) - \partial f^m(\mathbf{v}_{\mu f^m}(\mathbf{w}))$ and $\mathbb{E}\|\boldsymbol{\xi}\| \leq \sqrt{\mathbb{E}\|\boldsymbol{\xi}\|^2} \leq \epsilon/2$. Suppose $\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^n}(\mathbf{w})\|^2 \leq \epsilon^2/4$. We have $\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^n}(\mathbf{w})\| \leq \epsilon/2$ and $\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^m}(\mathbf{w})\| \leq \mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^n}(\mathbf{w})\| + \mathbb{E}\|\mathbf{v}_{\mu f^n}(\mathbf{w}) - \mathbf{v}_{\mu f^m}(\mathbf{w})\| \leq \epsilon$. Hence, $\bar{\mathbf{v}}$ satisfies Definition 1 with $\mathbf{w}' = \mathbf{v}_{\mu f^n}(\mathbf{w})$ and $\mathbf{w}'' = \mathbf{v}_{\mu f^m}(\mathbf{w})$. Suppose $\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}_{\mu f^m}(\mathbf{w})\|^2 \leq \epsilon^2/4$. The conclusion can be also proved similarly. □

We first present the following lemma which is similar to Lemma 4.2 in [14].

**Lemma 3** *Suppose Assumption 1 holds. For any* $\mathbf{v}$, $\boldsymbol{\lambda}$, $\mathbf{v}'$, $\boldsymbol{\lambda}'$, *and* $(\boldsymbol{\xi}_{\mathbf{v}}, \boldsymbol{\xi}_{\boldsymbol{\lambda}}) \in \partial g^l(\mathbf{v}', \boldsymbol{\lambda}')$, *we have*

$$g^l(\mathbf{v}, \boldsymbol{\lambda}) \geq g^l(\mathbf{v}', \boldsymbol{\lambda}') + \boldsymbol{\xi}_{\mathbf{v}}^\top(\mathbf{v} - \mathbf{v}') + \boldsymbol{\xi}_{\boldsymbol{\lambda}}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \frac{\rho}{2}\|\mathbf{v} - \mathbf{v}'\|^2,$$

*where* $\rho = N_+ N_- L$. *Moreover,* $g^l(\mathbf{v}, \boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2$ *is jointly convex in* $\boldsymbol{\lambda}$ *and* $\mathbf{v}$ *for any* $\mathbf{w}$ *and any* $\mu^{-1} > \rho$.

*Proof.* By Assumption 1, we have

$$s_{ij}(\mathbf{v}) - s_{ij}(\mathbf{v}') \geq \nabla s_{ij}(\mathbf{v}')^\top(\mathbf{v} - \mathbf{v}') - \frac{L}{2}\|\mathbf{v} - \mathbf{v}'\|^2. \tag{18}$$

Let $\xi_{ij} \in \partial[s_{ij}(\mathbf{v}') - \lambda_i']_+$. We have

$$\begin{aligned}
& g^l(\mathbf{v}, \boldsymbol{\lambda}) \\
= \;& l\mathbf{1}^\top\boldsymbol{\lambda} + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}[s_{ij}(\mathbf{v}) - \lambda_i]_+ \\
\geq \;& l\mathbf{1}^\top\boldsymbol{\lambda}' + l\mathbf{1}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}') + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}[s_{ij}(\mathbf{v}') - \lambda_i']_+ + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}\xi_{ij}(s_{ij}(\mathbf{v}) - s_{ij}(\mathbf{v}') - \lambda_i + \lambda_i') \\
\geq \;& g^l(\mathbf{v}', \boldsymbol{\lambda}') + \boldsymbol{\xi}_{\boldsymbol{\lambda}}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}') + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}\xi_{ij}\nabla s_{ij}(\mathbf{v}')^\top(\mathbf{v} - \mathbf{v}') - \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}\xi_{ij}\frac{L}{2}\|\mathbf{v} - \mathbf{v}'\|^2 \\
\geq \;& g^l(\mathbf{v}', \boldsymbol{\lambda}') + \boldsymbol{\xi}_{\mathbf{v}}^\top(\mathbf{v} - \mathbf{v}') + \boldsymbol{\xi}_{\boldsymbol{\lambda}}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}') - \frac{N_+N_-L}{2}\|\mathbf{v} - \mathbf{v}'\|^2,
\end{aligned}$$

where the first inequality is by the convexity of $[\cdot]_+$, the second inequality is from (18) and the last inequality is by the definitions of $(\boldsymbol{\xi}_{\mathbf{v}}, \boldsymbol{\xi}_{\boldsymbol{\lambda}})$ and the fact that $\xi_{ij} \in [0, 1]$.

Combining the inequality from the first conclusion with the equality $\frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2 = \frac{1}{2\mu}\|\mathbf{v} - \mathbf{v}'\|^2 + \frac{1}{\mu}(\mathbf{v} - \mathbf{v}')^\top(\mathbf{v}' - \mathbf{w}) + \frac{1}{2\mu}\|\mathbf{v}' - \mathbf{w}\|^2$ and using the fact that $\mu^{-1} > \rho$, we can obtain

$$g^l(\mathbf{v}, \boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2 \geq g^l(\mathbf{v}', \boldsymbol{\lambda}') + \frac{1}{2\mu}\|\mathbf{v}' - \mathbf{w}\|^2 + (\mu^{-1}(\mathbf{v}' - \mathbf{w}) + \boldsymbol{\xi}_{\mathbf{v}})^\top(\mathbf{v} - \mathbf{v}') + \boldsymbol{\xi}_{\boldsymbol{\lambda}}^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}'),$$

which proves the second conclusion. □

**Lemma 4** *The dual problem of the maximization problem in* (11) *is the minimization problem in* (12).

*Proof.* For $i = 1, \ldots, N_+$, we introduce a Lagrangian multiplier $\lambda_i$ for the constraint $\sum_{j=1}^{N_-} p_j = l$ in (11). Let $[z]_- = \min\{z, 0\}$ which equals $-[-z]_+$. Then, for each $i$, we have

$$\begin{aligned}
\max_{\mathbf{p}_i \in \mathcal{P}^l}\left\{\sum_{j=1}^{N_-} p_{ij}s_{ij}(\mathbf{w})\right\} &= -\min_{p_{ij}\in[0,1]\ \forall j}\max_{\lambda_i}\left\{-\sum_{j=1}^{N_-} p_{ij}s_{ij}(\mathbf{w}) + \lambda_i(\sum_{j=1}^{N_-} p_j - l)\right\} \\
&= -\max_{\lambda_i}\min_{p_{ij}\in[0,1]\ \forall j}\left\{-l\lambda_i + \sum_{j=1}^{N_-} p_{ij}[\lambda_i - s_{ij}(\mathbf{w})]\right\} \\
&= -\max_{\lambda_i}\left\{-l\lambda_i + \sum_{j=1}^{N_-}[\lambda_i - s_{ij}(\mathbf{w})]_-\right\} \\
&= \min_{\lambda_i}\left\{l\lambda_i + \sum_{j=1}^{N_-}[s_{ij}(\mathbf{w}) - \lambda_i]_+\right\}.
\end{aligned}$$

The conclusion is thus proved by summing up the equality above for $i = 1, \ldots, N_+$. □

## C Proof of Proposition 1

Let $\mathbf{v}^* = \mathbf{v}_{\mu f^l}(\mathbf{w})$ be the unique optimal solution of (8) and let $s_{i[j]}(\mathbf{v}^*)$ be the $j$th largest coordinate of $S_i(\mathbf{v}^*)$. It is easy to show that the set of optimal solutions of (13) is $\{\mathbf{v}^*\} \times \Lambda^*$ where

$$\Lambda^* = \arg\min_{\boldsymbol{\lambda}} g^l(\mathbf{v}^*, \boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 = \prod_{i=1}^{N_+} \left(\Lambda_i^* := \left[s_{i[l]}(\mathbf{v}^*), s_{i[l+1]}(\mathbf{v}^*)\right]\right). \tag{19}$$

Given any $\boldsymbol{\lambda} \in \mathbb{R}^{N_+}$, we denote its projection onto $\Lambda^*$ as $\text{Proj}_{\Lambda^*}(\boldsymbol{\lambda})$. By the structure of $\Lambda^*$, the $i$th coordinate of $\text{Proj}_{\Lambda^*}(\boldsymbol{\lambda})$ is just the projection of $\lambda_i$ onto $\Lambda_i^*$, which we denote by $\text{Proj}_{\Lambda_i^*}(\lambda_i)$. Moreover, we denote the distance from $\boldsymbol{\lambda}$ to $\Lambda^*$ as $\text{dist}(\boldsymbol{\lambda}, \Lambda^*)$ and it satisfies

$$\text{dist}^2(\boldsymbol{\lambda}, \Lambda^*) = \sum_{i=1}^{N_+} \text{dist}^2(\lambda_i, \Lambda_i^*) = \sum_{i=1}^{N_+} (\lambda_i - \text{Proj}_{\Lambda_i^*}(\lambda_i))^2 \tag{20}$$

With these definitions, we can present the following lemma.

**Lemma 5** *Suppose Assumption 1 holds and $\mu^{-1} > \rho$. For any $\mathbf{w}$, $\mathbf{v}$, $\boldsymbol{\lambda}$ and $\mathbf{v}^* = \mathbf{v}_{\mu f^l}(\mathbf{w})$, we have*

$$N_+ B\|\mathbf{v} - \mathbf{v}^*\| + g^l(\mathbf{v}, \boldsymbol{\lambda}) - g^l(\mathbf{v}^*, \text{Proj}_{\Lambda^*}(\boldsymbol{\lambda})) \geq \sum_{i=1}^{N_+} \textit{dist}(\lambda_i, \Lambda_i^*) \geq \textit{dist}(\boldsymbol{\lambda}, \Lambda^*).$$

*Proof.* It is easy to observe that $g^l(\mathbf{v}, \boldsymbol{\lambda}) := \sum_{i=1}^{N_+} g_i^l(\mathbf{v}, \lambda_i)$, where

$$g_i^l(\mathbf{v}, \lambda_i) := l\lambda_i + \sum_{j=1}^{N_-} [s_{ij}(\mathbf{v}) - \lambda_i]_+,$$

and $\Lambda_i^* = \arg\min_{\lambda_i} g_i^l(\mathbf{v}^*, \lambda_i)$. Since $g_i^l(\mathbf{v}^*, \lambda_i)$ is a piecewise linear in $\lambda_i$ with an outward slope of at least one at either end of the interval $\Lambda_i^*$, we must have

$$g_i^l(\mathbf{v}^*, \lambda_i) - g_i^l(\mathbf{v}^*, \text{Proj}_{\Lambda_i^*}(\lambda_i)) \geq |\lambda_i - \text{Proj}_{\Lambda_i^*}(\lambda_i)|,$$

which implies

$$g^l(\mathbf{v}^*, \boldsymbol{\lambda}) - g^l(\mathbf{v}^*, \text{Proj}_{\Lambda^*}(\boldsymbol{\lambda})) \geq \sum_{i=1}^{N_+} |\lambda_i - \text{Proj}_{\Lambda_i^*}(\lambda_i)| = \sum_{i=1}^{N_+} \text{dist}(\lambda_i, \Lambda_i^*) \geq \text{dist}(\boldsymbol{\lambda}, \Lambda^*).$$

Moreover, by Assumption 1, $g_i^l(\mathbf{v}, \lambda_i)$ is $B$-Lipschitz continuous in $\mathbf{v}$ so $g^l(\mathbf{v}, \boldsymbol{\lambda})$ is $N_+B$-Lipschitz continuous in $\mathbf{v}$. We then have $N_+B\|\mathbf{v} - \mathbf{v}^*\| + g^l(\mathbf{v}, \boldsymbol{\lambda}) \geq g^l(\mathbf{v}^*, \boldsymbol{\lambda})$, which implies the conclusion together with the previous inequality. □

We present the proof of Proposition 1 below.

*Proof.*[of **Proposition** 1] Let us denote $\mathcal{I}_{[t]} = \{\mathcal{I}_0, \ldots, \mathcal{I}_t\}$ and $\mathcal{J}_{[t]} = \{\mathcal{J}_0, \ldots, \mathcal{J}_t\}$. Let $\mathbb{E}_t$ be the expectation conditioning on $\mathcal{I}_{[t-1]}$ and $\mathcal{J}_{[t-1]}$.

By Assumption 1 and the definitions of $G_{\mathbf{v}}^{(t)}$ and $G_{\lambda_i}^{(t)}$ in Algorithm 1, we have

$$\|G_{\mathbf{v}}^{(t)}\| \leq N_+ N_- B \;\text{ and }\; |G_{\lambda_i}^{(t)}| \leq N_- \text{ for } t = 0, \ldots, T-1 \text{ and } i = 1, \ldots, N_+. \tag{21}$$

By the optimality condition satisfied by $\mathbf{v}^{(t+1)}$ and $\left(\frac{1}{\mu} + \frac{1}{\eta_t}\right)$-strong convexity of the objective function in (14), we have

$$\begin{aligned}
& \frac{1}{2\mu}\|\mathbf{v}^{(t+1)} - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}\|^2 + \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\eta_t}\right)\|\mathbf{v}^* - \mathbf{v}^{(t+1)}\|^2 \\
\leq\; & (G_{\mathbf{v}}^{(t)})^\top \left(\mathbf{v}^* - \mathbf{v}^{(t+1)}\right) + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2 \\
=\; & (G_{\mathbf{v}}^{(t)})^\top \left(\mathbf{v}^* - \mathbf{v}^{(t)}\right) + (G_{\mathbf{v}}^{(t)})^\top \left(\mathbf{v}^{(t)} - \mathbf{v}^{(t+1)}\right) + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2 \\
\leq\; & (G_{\mathbf{v}}^{(t)})^\top \left(\mathbf{v}^* - \mathbf{v}^{(t)}\right) + \frac{\eta_t (G_{\mathbf{v}}^{(t)})^2}{2} + \frac{1}{2\eta_t}\|\mathbf{v}^{(t)} - \mathbf{v}^{(t+1)}\|^2 + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2,
\end{aligned}$$

where the last inequality is by Young's inequality. Since $\mathbf{v}^* - \mathbf{v}^{(t)}$ is deterministic conditioning on $\mathcal{I}_{[t-1]}$ and $\mathcal{J}_{[t-1]}$, applying (21) and taking expectation $\mathbb{E}_t$ on the both sides of the inequality above yield

$$\begin{aligned}
&\frac{1}{2\mu}\mathbb{E}_t\|\mathbf{v}^{(t+1)} - \mathbf{w}\|^2 + \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\eta_t}\right)\mathbb{E}_t\|\mathbf{v}^* - \mathbf{v}^{(t+1)}\|^2 \\
\leq\ & \left(\mathbb{E}_t G_{\mathbf{v}}^{(t)}\right)^\top \left(\mathbf{v}^* - \mathbf{v}^{(t)}\right) + \frac{\eta_t N_+^2 N_-^2 B^2}{2} + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2. \quad (22)
\end{aligned}$$

By the updating equation (15) for $\boldsymbol{\lambda}^{(t+1)}$, we have

$$\begin{aligned}
\mathrm{dist}^2(\boldsymbol{\lambda}^{(t+1)}, \Lambda^*) &= \sum_{i\in\mathcal{I}_t^c}(\lambda_i^{(t+1)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t+1)}))^2 + \sum_{i\in\mathcal{I}_t}(\lambda_i^{(t+1)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t+1)}))^2 \\
&\leq \sum_{i\in\mathcal{I}_t^c}(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 + \sum_{i\in\mathcal{I}_t}(\lambda_i^{(t+1)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 \\
&= \sum_{i\in\mathcal{I}_t^c}(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 + \sum_{i\in\mathcal{I}_t}(\lambda_i^{(t)} - \theta_t G_{\lambda_i}^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 \\
&= \sum_{i\in\mathcal{I}_t^c}(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 + \sum_{i\in\mathcal{I}_t}(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}))^2 \\
&\quad -2\theta_t \sum_{i\in\mathcal{I}_t}(G_{\lambda_i}^{(t)})^\top(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)})) + \theta_t^2 \sum_{i\in\mathcal{I}_t}(G_{\lambda_i}^{(t)})^2.
\end{aligned}$$

Since $\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)})$ is deterministic conditioning on $\mathcal{I}_{[t-1]}$ and $\mathcal{J}_{[t-1]}$, applying (21) and taking expectation $\mathbb{E}_t$ on the both sides of the inequality above yield

$$\mathbb{E}_t\mathrm{dist}^2(\boldsymbol{\lambda}^{(t+1)}, \Lambda^*) \leq \mathrm{dist}^2(\boldsymbol{\lambda}^{(t)}, \Lambda^*) - \frac{2\theta_t I}{N_+}\sum_{i=1}^{N_+}\mathbb{E}_t G_{\lambda_i}^{(t)}(\lambda_i^{(t)} - \mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)})) + \theta_t^2 I N_-^2 \quad (23)$$

Multiplying both sides of (23) by $\frac{N_+}{2I\theta_t}$ and adding it with (22), we have

$$\begin{aligned}
&\frac{N_+}{2I\theta_t}\mathbb{E}_t\mathrm{dist}^2(\boldsymbol{\lambda}^{(t+1)}, \Lambda^*) + \frac{1}{2\mu}\mathbb{E}_t\|\mathbf{v}^{(t+1)} - \mathbf{w}\|^2 + \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\eta_t}\right)\mathbb{E}_t\|\mathbf{v}^* - \mathbf{v}^{(t+1)}\|^2 \\
\leq\ & \frac{N_+}{2I\theta_t}\mathrm{dist}^2(\boldsymbol{\lambda}^{(t)}, \Lambda^*) + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\eta_t}\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2 + \frac{\eta_t N_+^2 N_-^2 B^2}{2} + \frac{\theta_t N_+ N_-^2}{2} \\
&+ \left[\mathbb{E}_t G_{\mathbf{v}}^{(t)}\right]^\top \left(\mathbf{v}^* - \mathbf{v}^{(t)}\right) + \sum_{i=1}^{N_+}\mathbb{E}_t G_{\lambda_i}^{(t)}(\mathrm{Proj}_{\Lambda_i^*}(\lambda_i^{(t)}) - \lambda_i^{(t)}) \\
\leq\ & \frac{N_+}{2I\theta_t}\mathrm{dist}^2(\boldsymbol{\lambda}^{(t)}, \Lambda^*) + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2}\left(\rho + \frac{1}{\eta_t}\right)\|\mathbf{v}^* - \mathbf{v}^{(t)}\|^2 + \frac{\eta_t N_+^2 N_-^2 B^2}{2} + \frac{\theta_t N_+ N_-^2}{2} \\
&+ g^l(\mathbf{v}^*, \mathrm{Proj}_{\Lambda^*}(\boldsymbol{\lambda}^{(t)})) - g^l(\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)}), \quad (24)
\end{aligned}$$

where the second inequality is because $(\mathbb{E}_t G_{\mathbf{v}}^{(t)}, \mathbb{E}_t G_{\lambda_1}^{(t)}, \ldots, \mathbb{E}_t G_{\lambda_{N_+}}^{(t)}) \in \partial g^l(\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)})$, which allows us to apply Lemma 3 with $(\mathbf{v}, \boldsymbol{\lambda}) = (\mathbf{v}^*, \mathrm{Proj}_{\Lambda^*}(\boldsymbol{\lambda}^{(t)}))$ and $(\mathbf{v}', \boldsymbol{\lambda}') = (\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)})$.

Notice that $\theta_t$ and $\eta_t$ do not change with $t$ that $\rho < \mu^{-1}$. Summing up (24) for $t = 0, \ldots, T-1$, taking full expectation, and organizing terms give us

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left(g^l(\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)}) + \frac{1}{2\mu}\|\mathbf{v}^{(t)} - \mathbf{w}\|^2 - g^l(\mathbf{v}^*, \mathrm{Proj}_{\Lambda^*}(\boldsymbol{\lambda}^{(t)})) - \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2\right) \\
&+ \frac{N_+}{2I\theta_{T-1}}\mathbb{E}\mathrm{dist}^2(\boldsymbol{\lambda}^{(T)}, \Lambda^*) + \frac{1}{2\mu}\mathbb{E}\|\mathbf{v}^{(T)} - \mathbf{w}\|^2 + \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\eta_{T-1}}\right)\mathbb{E}\|\mathbf{v}^* - \mathbf{v}^{(T)}\|^2 \\
\leq\ & \frac{N_+}{2I\theta_0}\mathrm{dist}^2(\boldsymbol{\lambda}^{(0)}, \Lambda^*) + \frac{1}{2\mu}\|\mathbf{v}^{(0)} - \mathbf{w}\|^2 + \frac{1}{2}\left(\frac{1}{\mu} + \frac{1}{\eta_0}\right)\|\mathbf{v}^* - \mathbf{v}^{(0)}\|^2 \\
&+ \sum_{t=0}^{T-1}\left(\frac{\eta_t N_+^2 N_-^2 B^2}{2} + \frac{\theta_t N_+ N_-^2}{2}\right). \quad (25)
\end{aligned}$$

Because $f^l(\bar{\mathbf{v}}) + \frac{1}{2\mu}\|\bar{\mathbf{v}} - \mathbf{w}\|^2$ is $(\mu^{-1} - \rho)$-strongly convex and the facts that $f^l(\bar{\mathbf{v}}) \leq g^l(\bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}})$ and that $f^l(\mathbf{v}^*) = g^l(\mathbf{v}^*, \boldsymbol{\lambda}^*)$ for any optimal $\boldsymbol{\lambda}^*$, we have that

$$\begin{aligned}
& \frac{1}{2}\left(\frac{1}{\mu} - \rho\right)\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}^*\|^2 \\
\leq\ & \mathbb{E}\left(f^l(\bar{\mathbf{v}}) + \frac{1}{2\mu}\|\bar{\mathbf{v}} - \mathbf{w}\|^2 - f^l(\mathbf{v}^*) - \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2\right) \\
\leq\ & \mathbb{E}\left(g^l(\bar{\mathbf{v}}, \bar{\boldsymbol{\lambda}}) + \frac{1}{2\mu}\|\bar{\mathbf{v}} - \mathbf{w}\|^2 - g^l(\mathbf{v}^*, \boldsymbol{\lambda}^*) - \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2\right) \\
\leq\ & \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left(g^l(\mathbf{v}^{(t)}, \boldsymbol{\lambda}^{(t)}) + \frac{1}{2\mu}\|\mathbf{v}^{(t)} - \mathbf{w}\|^2 - g^l(\mathbf{v}^*, \mathrm{Proj}_{\Lambda^*}(\boldsymbol{\lambda}^{(t)})) - \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2\right), \quad (26)
\end{aligned}$$

where the last inequality is because $g^l(\mathbf{v}, \boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v} - \mathbf{w}\|^2$ is jointly convex in $\mathbf{v}$ and $\boldsymbol{\lambda}$. Recall that $\mathbf{v}^{(0)} = \mathbf{w}$. Applying (25) to the left-hand side of the inequality above, we obtain

$$\begin{aligned}
& \frac{1}{2}\left(\frac{1}{\mu} - \rho\right)\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}^*\|^2 \\
\leq\ & \frac{N_+}{2I\theta_0 T}\mathrm{dist}^2(\boldsymbol{\lambda}^{(0)}, \Lambda^*) + \frac{1}{2\mu T}\|\mathbf{v}^{(0)} - \mathbf{w}\|^2 + \frac{1}{2T}\left(\frac{1}{\mu} + \frac{1}{\eta_0}\right)\|\mathbf{v}^* - \mathbf{v}^{(0)}\|^2 \\
& + \frac{\eta_0 N_+^2 N_-^2 B^2}{2} + \frac{\theta_0 N_+ N_-^2}{2} \\
\leq\ & \frac{N_+ N_-}{\sqrt{IT}}\mathrm{dist}(\boldsymbol{\lambda}^{(0)}, \Lambda^*) + \frac{N_+ N_- B}{2\sqrt{T}}\|\mathbf{v}^* - \mathbf{w}\| + \frac{1}{2\mu T}\|\mathbf{v}^* - \mathbf{w}\|^2.
\end{aligned}$$

The conclusion is then proved given that $\mathbf{v}_{\mu f^l}(\mathbf{w}) = \mathbf{v}^*$. □

## D Proof of Theorem 1

We first present a technical lemma.

**Lemma 6** *Given two intervals $[a, b]$ and $[a', b']$ with $\max\{|a - a'|, |b - b'|\} \leq \delta$. We have*

$$\mathit{dist}(z, [a, b]) \leq \mathit{dist}(z, [a', b']) + \delta, \quad \forall z \in \mathbb{R}.$$

*Proof.* We will prove the result in three cases, $z < a'$, $z > b'$ and $a' \leq z \leq b'$. Suppose $z < a'$ so that $\mathrm{dist}(z, [a', b']) = |z - a'|$. We have $\mathrm{dist}(z, [a, b]) \leq |z - a| \leq |z - a'| + |a - a'| \leq \mathrm{dist}(z, [a', b']) + \delta$. The result when $z > b'$ can be proved similarly. Suppose $a' \leq z \leq b'$ so that $\mathrm{dist}(z, [a', b']) = 0$. Note that $z = \frac{z - a'}{b' - a'} \cdot b' + \frac{b' - z}{b' - a'} \cdot a'$. We define $z' = \frac{z - a'}{b' - a'} \cdot b + \frac{b' - z}{b' - a'} \cdot a \in [a, b]$. Then $\mathrm{dist}(z, [a, b]) \leq |z - z'| \leq \frac{z - a'}{b' - a'} \cdot |b - b'| + \frac{b' - z}{b' - a'} \cdot |a - a'| \leq \mathrm{dist}(z, [a', b']) + \delta$. □

Under Assumption 1, we have $\|\nabla s_{ij}(\mathbf{v})\| \leq B$ for any $\mathbf{v}$ so that $\|\boldsymbol{\xi}\| \leq lB$ for any $\boldsymbol{\xi} \in \partial f^l(\mathbf{v})$ and any $\mathbf{v}$. By the definition of $\mathbf{v}_{\mu f^l}(\mathbf{w})$, we have $\mu^{-1}(\mathbf{w} - \mathbf{v}_{\mu f^l}(\mathbf{w})) \in \partial f^l(\mathbf{v}_{\mu f^l}(\mathbf{w}))$, which implies

$$\|\mathbf{w} - \mathbf{v}_{\mu f^l}(\mathbf{w})\| \leq \mu l B \text{ for any } \mathbf{w}. \quad (27)$$

We then provide the following proposition as an additional conclusion from the proof of Proposition 1.

**Proposition 2** *Suppose Assumption 1 holds. Algorithm 1 guarantees that*

$$\begin{aligned}
& \mathbb{E}\mathit{dist}(\bar{\boldsymbol{\lambda}}, \Lambda^*) \leq \sum_{i=1}^{N_+}\mathbb{E}\mathit{dist}(\bar{\lambda}_i, \Lambda_i^*) \\
\leq\ & \frac{N_+ N_-}{\sqrt{IT}}\mathit{dist}(\boldsymbol{\lambda}^{(0)}, \Lambda^*) + \frac{N_+ N_- B}{2\sqrt{T}}\|\mathbf{v}^* - \mathbf{w}\| + \frac{1}{2\mu T}\|\mathbf{v}^* - \mathbf{w}\|^2 + \frac{1}{2\mu}\|\mathbf{v}^* - \mathbf{w}\|^2 \\
& + N_+ B\mathbb{E}\|\bar{\mathbf{v}} - \mathbf{v}^*\|.
\end{aligned}$$

*Proof.* By (25), the last inequality in (26), and Lemma 5, we have

$$\begin{aligned}
& \mathbb{E}\left(\sum_{i=1}^{N_+}\text{dist}(\bar{\lambda}_i,\Lambda_i^*) - N_+B\|\bar{\mathbf{v}}-\mathbf{v}^*\| - \frac{1}{2\mu}\|\mathbf{v}^*-\mathbf{w}\|^2\right) \\
\leq\ & \mathbb{E}\left(g^l(\bar{\mathbf{v}},\bar{\boldsymbol{\lambda}}) - g^l(\mathbf{v}^*,\boldsymbol{\lambda}^*) + \frac{1}{2\mu}\|\bar{\mathbf{v}}-\mathbf{w}\|^2 - \frac{1}{2\mu}\|\mathbf{v}^*-\mathbf{w}\|^2\right) \\
\leq\ & \frac{N_+N_-}{\sqrt{IT}}\text{dist}(\boldsymbol{\lambda}^{(0)},\Lambda^*) + \frac{N_+N_-B}{2\sqrt{T}}\|\mathbf{v}^*-\mathbf{w}\| + \frac{1}{2\mu T}\|\mathbf{v}^*-\mathbf{w}\|^2,
\end{aligned}$$

which implies the conclusion by reorganizing terms. □

Next we are ready to present the proof of of Theorem 1.

*Proof.*[**of Theorem 1**] Let $\epsilon_k = \frac{1}{\sqrt{k+1}}$ for $k = 1, \ldots$. In the $k$th iteration of Algorithm 2, Algorithm 1 is applied to the subproblem

$$\min_{\mathbf{v},\boldsymbol{\lambda}}\left\{g^l(\mathbf{v},\boldsymbol{\lambda}) + \frac{1}{2\mu}\|\mathbf{v}-\mathbf{w}^{(k)}\|^2\right\}. \tag{28}$$

with initial solution $(\mathbf{w}^{(k)}, \bar{\boldsymbol{\lambda}}_l^{(k)})$ and runs for $T_k$ iterations. Let $\Lambda_k^*$ be the set of optimal $\boldsymbol{\lambda}$ for (28). We will prove by induction the following two inequalities for $k = 0, 1, \ldots$.

$$\mathbb{E}\|\bar{\mathbf{v}}_l^{(k)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\|^2 \leq \epsilon_k^2/4 \tag{29}$$

$$\mathbb{E}\text{dist}(\bar{\boldsymbol{\lambda}}_l^{(k)},\Lambda_k^*) \leq D_l, \tag{30}$$

where $D_l$ is defined in (16) for $l = m$ and $n$.

Applying Proposition 1 to (28) and using (27), we have, for $k = 0, 1, \ldots,$

$$\begin{aligned}
& \frac{1}{2}\left(\frac{1}{\mu}-\rho\right)\mathbb{E}\|\bar{\mathbf{v}}_l^{(k)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\|^2 \\
\leq\ & \frac{N_+N_-}{\sqrt{IT_k}}\mathbb{E}\text{dist}(\bar{\boldsymbol{\lambda}}_l^{(k)},\Lambda_k^*) + \frac{N_+N_-B}{2\sqrt{T_k}}\mathbb{E}\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\| + \frac{1}{2\mu T_k}\mathbb{E}\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|^2 \\
\leq\ & \frac{N_+N_-}{\sqrt{IT_k}}\mathbb{E}\text{dist}(\bar{\boldsymbol{\lambda}}_l^{(k)},\Lambda_k^*) + \frac{N_+N_-\mu l B^2}{2\sqrt{T_k}} + \frac{\mu^2 l^2 B^2}{2\mu T_k}.
\end{aligned} \tag{31}$$

Moreover, applying Proposition 2 to (28) and using (27), we have

$$\begin{aligned}
\sum_{i=1}^{N_+}\mathbb{E}\text{dist}(\bar{\lambda}_{l,i}^{(k+1)},\Lambda_i^*) \leq\ & \frac{N_+N_-}{\sqrt{IT_k}}\mathbb{E}\text{dist}(\bar{\boldsymbol{\lambda}}_l^{(k)},\Lambda_{k,i}^*) + \frac{N_+N_-\mu l B^2}{2\sqrt{T_k}} + \frac{\mu^2 l^2 B^2}{2\mu T_k} + \frac{\mu l^2 B^2}{2} \\
& + N_+B\mathbb{E}\|\bar{\mathbf{v}}_l^{(k)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\|.
\end{aligned} \tag{32}$$

Suppose $k = 0$. By (31) and the choice of $T_0$, we have $\mathbb{E}\|\bar{\mathbf{v}}_l^{(0)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(0)})\|^2 \leq \epsilon_0^2/4$, so (29) holds for $k = 0$. In addition, (30) holds trivially for $k = 0$. Next, we assume (29) and (30) both hold for $k$ and prove they also hold for $k + 1$.

Since (29) and (30) hold for $k$, by (32) and the choice of $T_k$, we have

$$\begin{aligned}
\sum_{i=1}^{N_+}\mathbb{E}\text{dist}(\bar{\lambda}_{l,i}^{(k+1)},\Lambda_{k,i}^*) \leq\ & \frac{1}{2}\left(\frac{1}{\mu}-\rho\right)\epsilon_k^2/4 + \frac{\mu l^2 B^2}{2} + N_+B\mathbb{E}\|\bar{\mathbf{v}}_l^{(k)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\| \\
\leq\ & \frac{1}{2}\left(\frac{1}{\mu}-\rho\right)\epsilon_k^2/4 + \frac{\mu l^2 B^2}{2} + N_+B\sqrt{\mathbb{E}\|\bar{\mathbf{v}}_l^{(k)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\|^2} \\
\leq\ & \frac{1}{2}\left(\frac{1}{\mu}-\rho\right) + \frac{\mu l^2 B^2}{2} + N_+B.
\end{aligned} \tag{33}$$

From the updating equation $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \gamma\mu^{-1}(\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)})$ in Algorithm 2, we know that

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| &\leq \gamma\mu^{-1}\mathbb{E}\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)}\| \\
&\leq \gamma\mu^{-1}\Big(\mathbb{E}\|\bar{\mathbf{v}}_m^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)})\| + \mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\| \\
&\qquad + \mathbb{E}\|\bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\| + \mathbb{E}\|\mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)}) - \mathbf{w}^{(k)}\|\Big) \\
&\leq \frac{2\gamma\epsilon_k}{\mu} + \gamma nB + \gamma mB, \qquad (34)
\end{aligned}$$

where the second inequality is by triangle inequality and the last inequality is by (27) and the fact that (29) holds for $k$.

Let $\bar{\lambda}_{l,i}^{(k+1)}$ denote the $i$th coordinate of $\bar{\boldsymbol{\lambda}}_l^{(k+1)}$ for $i = 1, \ldots, N_+$ and $l = m$ and $n$. Recall that $s_{i[j]}(\mathbf{v})$ be the $j$th largest coordinate of $S_i(\mathbf{v})$. By Assumption 1 and elementary argument, we can show that $s_{i[j]}(\mathbf{v})$ is $B$-Lipschitz continuous for any $i$ and $j$. Since $\mathbf{v}_{\mu f^l}(\mathbf{w})$ is $\frac{1}{1-\mu\rho}$-Lipschitz continuous, we have

$$\begin{aligned}
\mathbb{E}\left|s_{i[j]}(\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k+1)})) - s_{i[j]}(\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)}))\right| &\leq B\mathbb{E}\|\mathbf{v}_{\mu f^l}(\mathbf{w}^{(k+1)}) - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k)})\| \\
&\leq \frac{B}{1-\mu\rho}\mathbb{E}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\| \\
&\leq \frac{B}{1-\mu\rho}\left(\frac{2\gamma\epsilon_k}{\mu} + \gamma nB + \gamma mB\right) \qquad (35)
\end{aligned}$$

for $j = l$ and $j = l+1$, where the last inequality is by (34). According to (19), (20), (35), and Lemma 6, we can prove that, for $i = 1, \ldots, N_+$,

$$\mathbb{E}\text{dist}\left(\bar{\lambda}_{l,i}^{(k+1)}, \Lambda_{k+1,i}^*\right) \leq \mathbb{E}\text{dist}\left(\bar{\lambda}_{l,i}^{(k+1)}, \Lambda_{k,i}^*\right) + \frac{B}{1-\mu\rho}\left(\frac{2\gamma}{\mu} + \gamma nB + \gamma mB\right).$$

Combining this inequality with (33) yields (30) for case $k+1$.

Stating (31) for case $k+1$ gives

$$\frac{1}{2}\left(\frac{1}{\mu} - \rho\right)\mathbb{E}\|\bar{\mathbf{v}}_l^{(k+1)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k+1)})\|^2 \leq \frac{N_+N_-}{\sqrt{IT_{k+1}}}\mathbb{E}\text{dist}(\bar{\boldsymbol{\lambda}}_l^{(k+1)}, \Lambda_{k+1}^*) + \frac{N_+N_-\mu lB^2}{2\sqrt{T_{k+1}}} + \frac{\mu^2 l^2 B^2}{2\mu T_{k+1}}. \quad (36)$$

By (36), (30) for case $k+1$, and the choice of $T_{k+1}$, we have $\mathbb{E}\|\bar{\mathbf{v}}_l^{(k+1)} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(k+1)})\|^2 \leq \epsilon_{k+1}^2/4$, which means (29) holds for case $k+1$. By induction, we have proved that (29) and (30) holds for $k = 0, 1, \ldots$.

Because $\nabla F^\mu(\mathbf{w}) = \mu^{-1}(\mathbf{v}_{\mu f^m}(\mathbf{w}) - \mathbf{v}_{\mu f^n}(\mathbf{w}))$ is $L_\mu$-Lipschitz continuous, we have

$$\begin{aligned}
&F^\mu(\mathbf{w}^{(k)}) - F^\mu(\mathbf{w}^{(k+1)}) \\
\geq &\langle -\nabla F^\mu(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rangle - \frac{L_\mu}{2}\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\|^2 \\
= &\left\langle \mu^{-1}(\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})), \gamma\mu^{-1}(\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)})\right\rangle - \frac{\gamma^2 L_\mu}{2\mu^2}\left\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)}\right\|^2 \\
= &\frac{\gamma}{\mu^2}\left\langle \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)}), \bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\right\rangle \\
&- \frac{\gamma^2 L_\mu}{2\mu^2}\left\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\right\|^2 \\
\geq &\frac{\gamma}{\mu^2}\left\langle \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)}), \bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\right\rangle \\
&+ \left(\frac{\gamma}{\mu^2} - \frac{\gamma^2 L_\mu}{\mu^2}\right)\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 - \frac{\gamma^2 L_\mu}{\mu^2}\left\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\right\|^2.
\end{aligned}$$

Applying Young's inequality to the first term on the right-hand side of the last inequality above gives

$$
\begin{aligned}
&\mathbb{E}\left[F^{\mu}(\mathbf{w}^{(k)}) - F^{\mu}(\mathbf{w}^{(k+1)})\right] \\
\geq& -\frac{\gamma}{\mu^2}\left(\frac{1}{2}\mathbb{E}\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 + \frac{1}{2}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2\right) \\
&+ \left(\frac{\gamma}{\mu^2} - \frac{\gamma^2 L_\mu}{\mu^2}\right)\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 \\
&- \frac{\gamma^2 L_\mu}{\mu^2}\mathbb{E}\left\|\bar{\mathbf{v}}_m^{(k)} - \bar{\mathbf{v}}_n^{(k)} - \mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) + \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\right\|^2 \\
\geq& \frac{\gamma - 2\gamma^2 L_\mu}{2\mu^2}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 - \left(\frac{\gamma}{\mu^2} + \frac{2\gamma^2 L_\mu}{\mu^2}\right)\epsilon_k^2 \\
\geq& \frac{\gamma}{4\mu^2}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 - \frac{3\gamma}{2\mu^2}\epsilon_k^2,
\end{aligned}
$$

where the second inequality is because of (30) for $l = m$ and $n$ and the last because of $\gamma \leq \frac{1}{L_\mu}$.

Summing the above inequality over $k = 0, \cdots, K-1$, we have

$$
\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(k)}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(k)})\|^2 &\leq \frac{4\mu^2}{\gamma K}\mathbb{E}(F^{\mu}(\mathbf{w}^{(0)}) - F^{\mu}(\mathbf{w}^{(K)})) + \frac{6}{K}\sum_{k=0}^{K-1}\epsilon_k^2 \\
&\leq \frac{4\mu^2}{\gamma K}\left(F(\mathbf{v}_{\mu f^n}(\mathbf{w}^{(0)})) - F^*\right) + \frac{6\log K}{K},
\end{aligned}
$$

where the second inequality is because $F^* \leq F(\mathbf{v}_{\mu f^n}(\mathbf{w}^{(K)})) \leq F^{\mu}(\mathbf{w}^{(K)})$ and $F^{\mu}(\mathbf{w}^{(0)}) \leq F(\mathbf{v}_{\mu f^m}(\mathbf{w}^{(0)}))$ (see Lemma 1 in [52]). Let $\bar{k}$ be randomly sampled from $\{0, \ldots, K-1\}$, then it holds

$$
\mathbb{E}\|\nabla F^{\mu}(\mathbf{w}^{(\bar{k})})\|^2 = \mu^{-2}\mathbb{E}\|\mathbf{v}_{\mu f^m}(\mathbf{w}^{(\bar{k})}) - \mathbf{v}_{\mu f^n}(\mathbf{w}^{(\bar{k})})\|^2 \leq \mu^{-2}\min\{1, \mu^2\}\epsilon^2 = \min\{1, \mu^{-2}\}\epsilon^2
$$

by the definition of $K$. On the other hands, by (29), we have for $l = m$ and $n$ that

$$
\mathbb{E}\|\bar{\mathbf{v}}_l^{(\bar{k})} - \mathbf{v}_{\mu f^l}(\mathbf{w}^{(\bar{k})})\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\epsilon_k^2 \leq \frac{\log K}{K} \leq \frac{\min\{1, \mu^2\}\epsilon^2}{48} \leq \frac{\epsilon^2}{4},
$$

where the last inequality is because of the value of $K$. Hence, by the second claim in Lemma 2 (with $\mathbf{w} = \mathbf{w}^{(\bar{k})}$ and $\bar{\mathbf{v}} = \bar{\mathbf{v}}_n^{(\bar{k})}$), $\bar{\mathbf{v}}_n^{(\bar{k})}$ is an nearly $\epsilon$-critical point of (5) with $f^l$ defined in (6). This completes the proof. □

# E Additional Materials for Numerical Experiments

In this section, we present some additional details of our numerical experiments in Section 6 which we are not able to show due to the limit of space.

## E.1 Details of SoRR Loss Minimization Experiment

For SoRR loss minimization, we compare with the baseline DCA. We focus on learning a linear model $h_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with four benchmark datasets from the UCI [15] data repository preprocessed by Libsvm [9]: a9a, w8a, ijcnn1 and covtype. The statistics of these datasets are summarized in Table 4.We use logistic loss $l(h_{\mathbf{w}}(\mathbf{x}), y) = -y\log(h_{\mathbf{w}}(\mathbf{x})) - (1-y)\log(1 - h_{\mathbf{w}}(\mathbf{x}))$ where label $y \in \{0, 1\}$. Due to the limit space, we present the process of tuning hyperparameters and the convergence curves (Figure 2). From the curves, we notice that our algorithm reduce SoRR loss faster than DCAs for all of these four datasets. In this section, we first present some summary statistics on the datasets.

For AGD-SBCD, we fix the $J = 100$, $m = 10^3$ and $n = 2 \times 10^4$. In the spirit of Corollary 1, within Algorithm 3 called in the $k$th main iteration, $\eta_t$ and $\theta_t$ are both set to $\frac{c}{(k+1)N}$ for any $t$ with $c$ tuned in

Table 4: Statistics of the UCI datasets.

| DATASETS | # SAMPLES | # FEATURES |
|---|---|---|
| A9A | 32,561 | 123 |
| W8A | 49,749 | 300 |
| IJCNN1 | 49,990 | 22 |
| COVTYPE | 581,012 | 54 |

the range of $\{0.1, 2, 5, 10, 1\}$ and $T_k$ is set to $C(k+1)^2$ with $C$ selected from $\{30, 50, 100\}$. Parameter $\mu$ is chosen from $\{2\times10^2, 10^3\}/N$ and $\gamma$ is tuned from $\{2\times10^2, 5\times10^2, 8\times10^2, 4\times10^3, 2\times10^4\}/N$.

According to the experiments in [24], when solving (37), we first use the same step size and the same number of iterations for all $k$'s. We choose the step size from $\{0.01, 0.1, 0.5, 1\}$ and the iteration number from $\{2000, 3000\}$. However, we find that the performance of DCA improves if we let the step size and the number of iterations vary with $k$. Hence, we apply the same process as in AGD-SBCD to select $\theta_t$, $\eta_t$, and $T$ in Algorithm 4 in the $k$th main iteration of DCA. We report the performances of DCA under both settings (named DCA.Constant.lr and DCA.Changing.lr, respectively). The changes of the SoRR loss with the number of epochs are shown in Figure 2. From the curves, we notice that our algorithm reduce SoRR loss faster than DCAs for all of these four datasets.

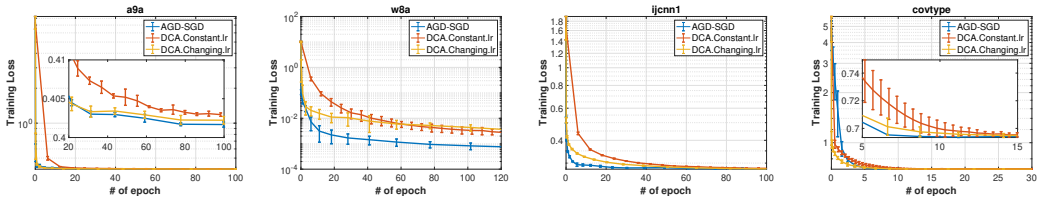


Figure 2: Results for SoRR Loss Minimization

### E.2 Process of Tuning Hyperparameters

For AGD-SBCD, we fix the $I = J = 100$, $m = 10^4$ and $n = 10^5$. In the spirit of Theorem 1, within Algorithm 3 called in the $k$th main iteration, $\eta_t$ is set to $\frac{c}{(k+1)N^+N^-}$, $\theta_t$ is set to $\frac{c}{(k+1)N^-}$ for any $t$ with $c$ tuned in $\{0.1, 1, 5, 10, 20, 30\}$ and $T_k$ is set to $50(k+1)^2$. Parameters $\mu$ is chosen from $\{10^2, 10^3\}/(N_+N_-)$ and $\gamma$ is tuned from $\{2\times10^3, 5\times10^3\}/(N_+N_-)$. For DCA, we only implement the setting of DCA.Changing.lr (see Appendix E.1 for descriptions). In Algorithm 4 called in the $k$th main iteration, $\eta_t$ and $\theta_t$ are selected following the same process as in AGD-SBCD and $T$ is set to $C(k+1)^2$ with $C$ chosen from $\{100, 200, 500\}$. For the SVM$_{pAUC}$-tight method, we use their MATLAB code in [42] and tune hyper-parameter $C$ from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$.

### E.3 Additional Plots of Partial AUC Maximization Experiment

The results for partial AUC maximization of diseases D3∼D5 are shown in Figure 3.

We plot the ROC curves of three algorithms with linear model on the CheXpert testing set in Figure 4. The range of false positive rate for the pAUC is set as $[0.05, 0.5]$.

We plot the convergence curves of patial AUC on the CheXpert testing set of our algorithm AGD-SBCD and baseline MB in Figure 5.

### E.4 DCA and Algorithm for its Subproblem

Although DCA is originally only studied for SoRR loss minimization in [24], it can also be applied to pAUC maximization. Hence, we only describe DCA for pAUC maximization which cover SoRR loss minimization as a special case. At the $k$th iteration of DCA, it computes a deterministic subgradient of $f^m(\mathbf{w}) = \sum_{i=1}^{N_+} \phi_m(S_i(\mathbf{w}^{(k)}))$ at iterate $\mathbf{w}^{(k)}$, denoted by $\boldsymbol{\xi}^{(k)}$. Then DCA updates $\mathbf{w}^{(k)}$ by

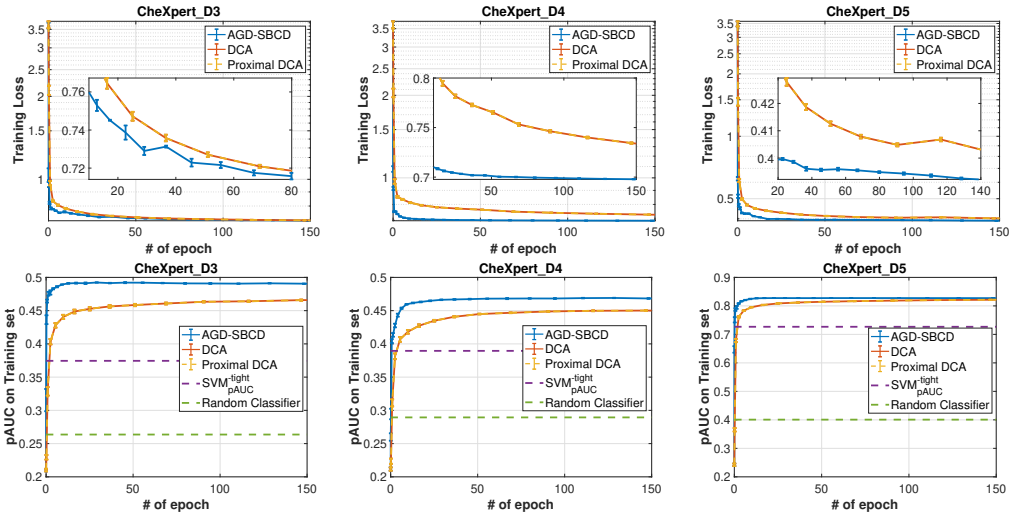


Figure 3: Results for Patial AUC Maximization of D3, D4 and D5.

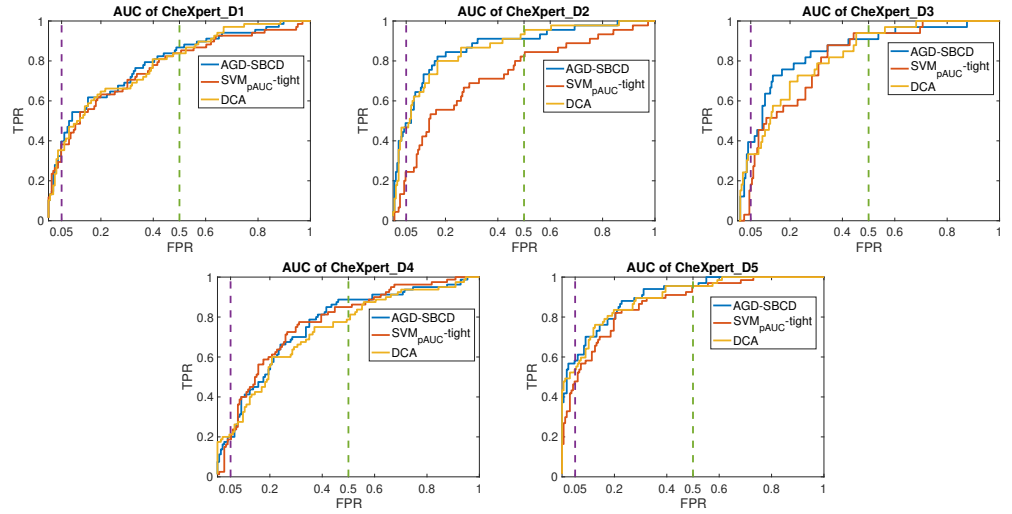


Figure 4: ROC Curves of CheXpert Testing Set.

approximately solving the following subproblem using a SBCD method similar to Algorithm 1 by sampling indexes $i$ and $j$ (see Algorithm 4 for details).

$$(\mathbf{w}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}) \approx \arg\min_{\mathbf{w},\boldsymbol{\lambda}} n\mathbf{1}^\top\boldsymbol{\lambda} + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}[s_{ij}(\mathbf{w}) - \lambda_i]_+ - \mathbf{w}^\top\boldsymbol{\xi}^{(k)} \tag{37}$$

Algorithm 4 is used to solve the subproblem (37) in the $k$th main iteration of DCA. It is similar to the SBCD method in Algorithm 1.

### E.5 Details about Proximal DCA

At the $k$th iteration of proximal DCA, it computes a deterministic subgradient of $f^m(\mathbf{w}) = \sum_{i=1}^{N_+}\phi_m(S_i(\mathbf{w}^{(k)}))$ at iterate $\mathbf{w}^{(k)}$, denoted by $\boldsymbol{\xi}^{(k)}$. Then proximal DCA updates $\mathbf{w}^{(k)}$ by approximately solving the subproblem

$$(\mathbf{w}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}) \approx \arg\min_{\mathbf{w},\boldsymbol{\lambda}} n\mathbf{1}^\top\boldsymbol{\lambda} + \sum_{i=1}^{N_+}\sum_{j=1}^{N_-}[s_{ij}(\mathbf{w}) - \lambda_i]_+ + \frac{L}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 - \mathbf{w}^\top\boldsymbol{\xi}^{(k)} \tag{38}$$

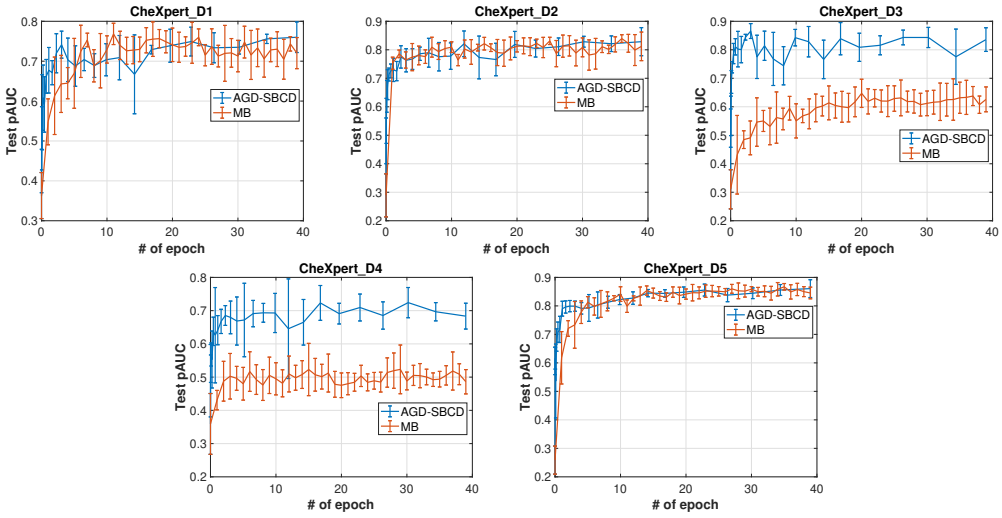


Figure 5: Convergence Curves of Partial AUC on the CheXpert Testing Set.

**Algorithm 4** Stochastic Block Coordinate Descent for (37): $(\bar{\mathbf{w}}, \bar{\boldsymbol{\lambda}})$ =SBCD$(\mathbf{w}, \boldsymbol{\lambda}, T, \boldsymbol{\xi}^{(k)})$

1: **Input:** Initial solution $(\mathbf{w}, \boldsymbol{\lambda})$, the number of iterations $T$, sample sizes $I$ and $J$ and a deterministic subgradient $\boldsymbol{\xi}^{(k)}$.
2: Set $(\mathbf{w}^{(0)}, \boldsymbol{\lambda}^{(0)}) = (\mathbf{w}, \boldsymbol{\lambda})$ and choose $(\eta_t, \theta_t)_{t=0}^{T-1}$.
3: **for** $t = 0$ **to** $T-1$ **do**
4: Sample $\mathcal{I}_t \subset \{1, \ldots, N_+\}$ with $|\mathcal{I}_t| = I$.
5: Sample $\mathcal{J}_t \subset \{1, \ldots, N_-\}$ with $|\mathcal{J}_t| = J$.
6: Compute stochastic subgradient w.r.t. $\mathbf{w}$:

$$G_{\mathbf{w}}^{(t)} = \frac{N_+ N_-}{IJ} \sum_{i \in \mathcal{I}_t} \sum_{j \in \mathcal{J}_t} \nabla s_{ij}(\mathbf{w}^{(t)}) \mathbf{1}\left(s_{ij}(\mathbf{w}^{(t)}) > \lambda_i^{(t)}\right) - \boldsymbol{\xi}^{(k)}$$

7: Stochastic subgradient update on $\mathbf{w}$:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t G_{\mathbf{w}}^{(t)}$$

8: Compute stochastic subgradient w.r.t. $\lambda_i$ for $i \in \mathcal{I}_t$:

$$G_{\lambda_i}^{(t)} = n - \frac{N_-}{J} \sum_{j \in \mathcal{J}_t} \mathbf{1}\left(s_{ij}(\mathbf{w}^{(t)}) > \lambda_i^{(t)}\right) \quad \text{for } i \in \mathcal{I}_t$$

9: Stochastic block subgradient update on $\lambda_i$ for $i \in \mathcal{I}_t$:

$$\lambda_i^{(t+1)} = \begin{cases} \lambda_i^{(t)} - \theta_t G_{\lambda_i}^{(t)} & i \in \mathcal{I}_t, \\ \lambda_i^{(t)} & i \notin \mathcal{I}_t. \end{cases}$$

10: **end for**
11: **Output:** $(\bar{\mathbf{w}}, \bar{\boldsymbol{\lambda}}) = (\mathbf{w}^{(T)}, \boldsymbol{\lambda}^{(T)})$.

using a SBCD method similar to Algorithm 1 by sampling indexes $i$ and $j$. In proximal DCA, $L$ is tuned from $\{10^{-5} \sim 10^{0}\}$ and other hyper-parameters are tuned from the same range as DCA.

### E.6 Details about CIFAR-10-LT and Tiny-Imagenet-200-LT Datasets

**Binary CIFAR-10-LT Dataset**. To construct a binary classification, we set labels of category 'cats' to 1 and labels of other categories to 0. We split the training data in train/val = 9:1 and use the validation dataset as the testing set. More details are provided in Table 5.

**Binary Tiny-Imagenet-200-LT Dataset**. To construct a binary classification, we set labels of category 'birds' to 1 and labels of other categories to 0. We split the training data in train/val = 9:1 and use the validation dataset as the testing set. More details are provided in Table 5.

Table 5: Statistics of the Long-Tailed Datesets.

| Dataset | Pos. Class ID | Pos. Class Name | # Pos. Samples | # Neg. Samples |
|---|---|---|---|---|
| CIFAR-10-LT | 3 | cats | 1077 | 11329 |
| Tiny-Imagenet-200-LT | 11,20,21,22 | birds | 1308 | 20241 |

### E.7 Convergence Curves and Testing Performance of AGD-SBCD on Different $\mu$

Table 6: The pAUCs with FPRs between $0.05$ and $0.5$ on the testing sets from the CheXpert data of AGD-SBCD on different $\mu$.

| | $\mu$ | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|---|
| AGD-SBCD | $10^3/N_+N_-$ | 0.6721±0.0081 | 0.8257±0.0025 | 0.8016±0.0075 | 0.6340±0.0165 | 0.8500±0.0017 |
| | $10^8/N_+N_-$ | 0.6617±0.0073 | 0.8242±0.0057 | 0.8272±0.0070 | 0.6323±0.0028 | 0.8463±0.0003 |
| | $10^{10}/N_+N_-$ | 0.6636±0.0056 | 0.8242±0.0057 | 0.8237±0.0077 | 0.6332±0.0072 | 0.8463±0.0002 |

### E.8 Convergence Curves of Training pAUC over GPU Time

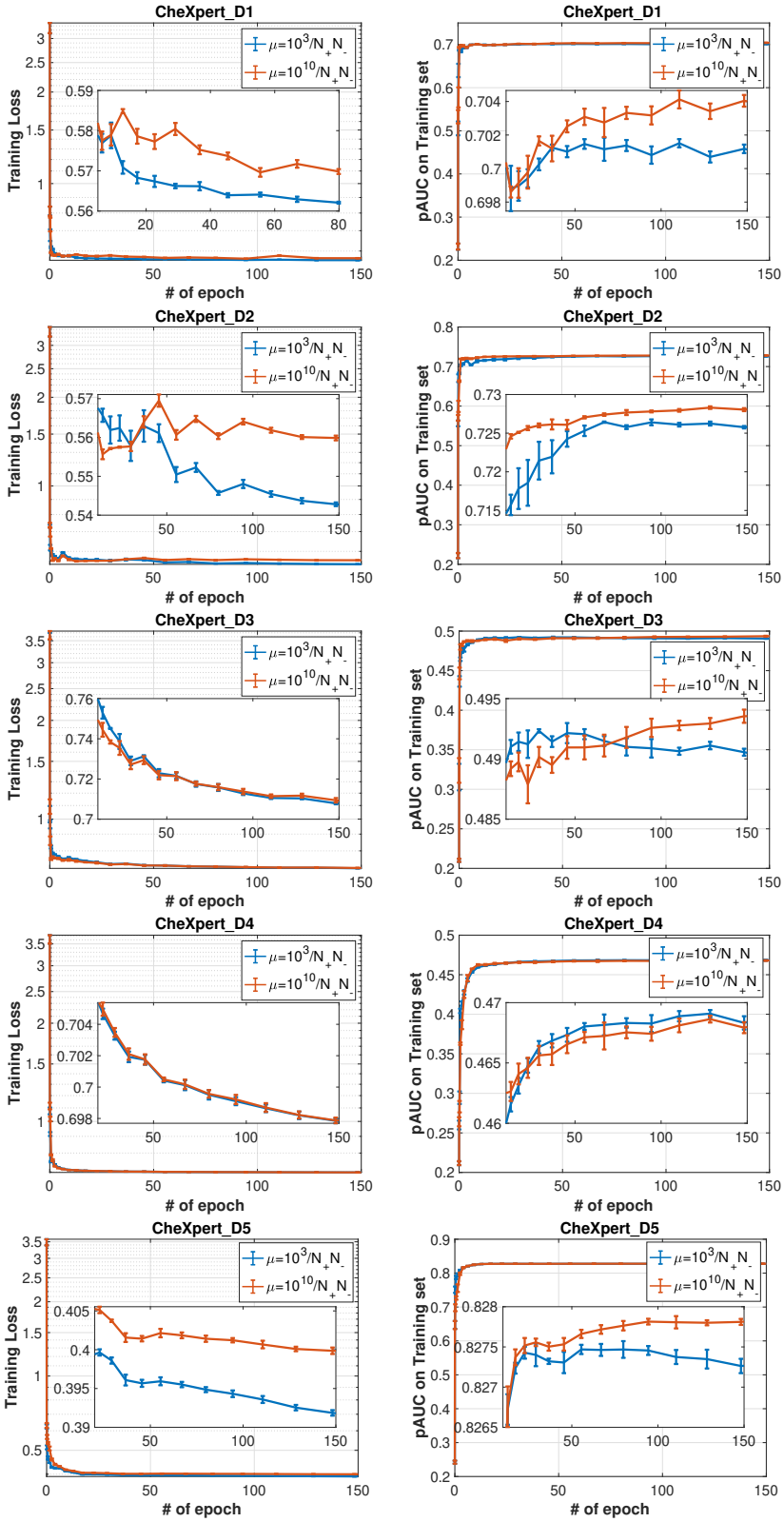


Figure 6: Convergence curves of training loss and training pAUC of AGD-SBCD on different $\mu$.

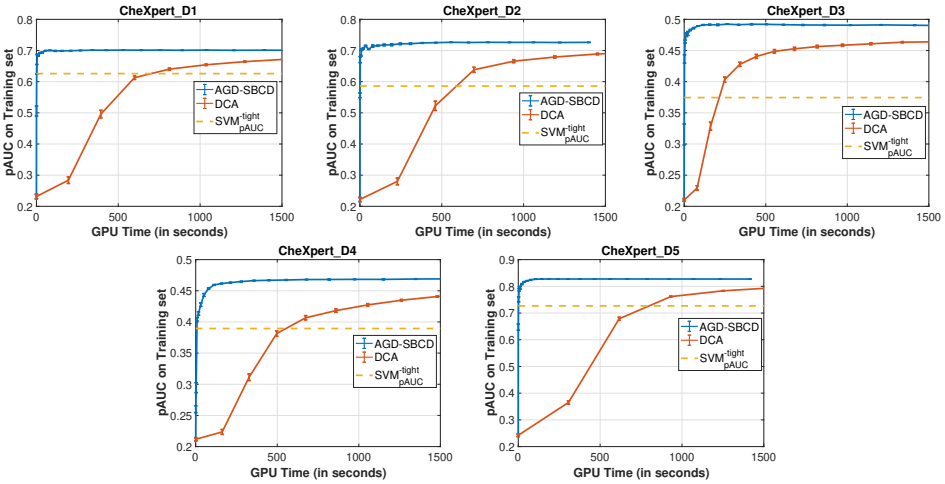


Figure 7: Convergence curves of training pAUC over GPU time. (The dashed line of SVM$_{pAUC}$-tight does not reflect its convergence with GPU time. It is only reported for reference since we use the authors' MATLAB implementation which does not support GPU.)