When AUC meets DRO: Optimizing Partial AUC for Deep Learning with Non-Convex Convergence Guarantee

Dixian Zhu [*1] **Gang Li** [*1] **Bokun Wang** [1] **Xiaodong Wu** [2] **Tianbao Yang** [1]

Abstract

In this paper, we propose systematic and efficient gradient-based methods for both one-way and two-way partial AUC (pAUC) maximization that are applicable to deep learning. We propose new formulations of pAUC surrogate objectives by using the distributionally robust optimization (DRO) to define the loss for each individual positive data. We consider two formulations of DRO, one of which is based on conditional-value-at-risk (CVaR) that yields a non-smooth but exact estimator for pAUC, and another one is based on a KL divergence regularized DRO that yields an inexact but smooth (soft) estimator for pAUC. For both one-way and two-way pAUC maximization, we propose two algorithms and prove their convergence for optimizing their two formulations, respectively. Experiments demonstrate the effectiveness of the proposed algorithms for pAUC maximization for deep learning on various datasets. The proposed methods are implemented with tutorials in our open-sourced library LibAUC (`www.libauc.org`).

1. Introduction

AUC, short for the area under the ROC curve, is a performance measure of a model, where the ROC curve is a curve of true positive rate (TPR) vs false positive rate (FPR) for all possible thresholds. AUC maximization in machine learning has a long history dating back to early 2000s (Herbrich et al., 1999). It has four ages in the twenty-years history, full-batch based methods in the first age, online methods in the second age, stochastic methods in the third age, and deep learning methods in the recent age. The first three ages focus on learning linear models or kernelized models. In each age, there have been seminal works in rigorous optimization algorithms that play important roles in the evolution of AUC maximization methods. Recent advances in non-convex optimization (in particular non-convex min-max optimization) (Liu et al., 2020) has driven large-scale deep AUC maximization to succeed in real-world tasks, e.g., medical image classification (Yuan et al., 2020) and molecular properties prediction (Wang et al., 2021).

Nevertheless, the research on efficient optimization algorithms for partial AUC (pAUC) lag behind. In many applications, there are large monetary costs due to high false positive rates (FPR) and low true positive rates (TPR), e.g., in medical diagnosis. Hence, a measure of primary interest is the region of the curve corresponding to low FPR and/or high TPR, i.e., pAUC. There are two commonly used versions of pAUC, namely one-way pAUC (OPAUC) (Dodd & Pepe, 2003) and two-way pAUC (TPAUC) (Yang et al., 2019), where OPAUC puts a restriction on the range of FPR, i.e., FPR$\in [\alpha, \beta]$ (Figure 1 middle) and TPAUC puts a restriction on the lower bound of TPR and the upper bound of FPR, i.e., TPR$\geq \alpha$, FPR$\leq \beta$ (Figure 1 right). Compared with standard AUC maximization, pAUC maximization is more challenging since its estimator based on training examples involves selection of examples whose prediction scores are in certain ranks.

To the best of our knowledge, there are few rigorous and efficient algorithms developed for pAUC maximization for deep learning. Some earlier works have focused on pAUC maximization for learning linear models. For example, Narasimhan & Agarwal (2017) have proposed a structured SVM approach for one-way pAUC maximization, which is guaranteed to converge for optimizing the surrogate objective of pAUC. However, their approach is not efficient for big data and is not applicable to deep learning, which needs to evaluate the prediction scores of all examples and sort them at each iteration. There are some heuristic approaches, e.g., updating the model parameters according to the gradient of surrogate pAUC computed based on a mini-batch data (Kar et al., 2014) or using an ad-hoc weighting function for each example for computing the stochastic gradient (Yang et al., 2021). However, such approaches are

[*]Equal contribution [1]Department of Computer Science, University of Iowa, Iowa City, Iowa, USA [2]Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa, USA. Correspondence to: Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

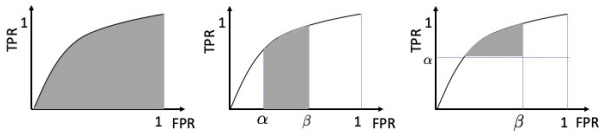

Figure 1. From left to right: AUC, one-way pAUC, two-way pAUC

either not guaranteed to converge or could suffer a large approximation error.

In this paper, we propose more systematic and rigorous optimization algorithms for pAUC maximization with convergence guarantee, which are applicable to deep learning. We consider both OPAUC maximization and TPAUC maximization, where for OPAUC we focus on maximizing pAUC in the region where FPR $\in [0, \beta]$ and for TPAUC we focus on maximizing pAUC in the region where FPR $\leq \beta$ and TPR $\geq \alpha$ for some $\alpha, \beta \in (0, 1)$. In order to tackle the challenge of computing unbiased stochastic gradients of the surrogate objective of pAUC, we propose new formulations based on distributionally robust optimization (DRO), which allows us to formulate the problem into weakly convex optimization, and novel compositional optimization problems, and to develop efficient stochastic algorithms with convergence guarantee. We summarize our contributions below.

- For OPAUC maximization, for each positive example, we define a loss over all negative examples based on DRO. We consider two special formulations of DRO, with one based on the conditional-value-at-risk (CVaR) function that yields an exact estimator of the surrogate objective of OPAUC, and another one based on Kullback–Leibler (KL) divergence regularized DRO that yields a soft estimator of the surrogate objective.
- We propose efficient stochastic algorithms for optimizing both formulations of OPAUC and establish their convergence guarantee and complexities for finding a (nearly) stationary solution. We also demonstrate that the algorithm for optimizing the soft estimator based on the KL divergence regularized DRO can enjoy parallel speed-up.
- For TPAUC maximization, we apply another level of DRO with respect to the positive examples on top of OPAUC formulations, yielding both exact and soft estimators for TPAUC. We also provide two rigorous stochastic algorithms with provable convergence for optimizing both the exact and soft estimator of TPAUC, with the latter problem formulated as a novel three-level compositional stochastic optimization problem.
- We conduct extensive experiments for deep learning on image classification and graph classification tasks with imbalanced data. We compare with heuristic and ad-hoc approaches for pAUC maximization and multiple baseline methods, and observe superior performance of the proposed algorithms.

To the best our knowledge, this work is the first one that provides rigorous stochastic algorithms and convergence guarantee for pAUC maximization that are efficient and applicable to deep learning. We expect the proposed novel formulations for OPAUC and TPAUC will allow researchers to develop even faster algorithms than the proposed algorithms in this paper.

2. Related Work

In this section, we provide a brief overview of related work for pAUC maximization.

Earlier works have considered indirect methods for pAUC maximization (Rudin, 2009; Agarwal, 2011; Rakotomamonjy, 2012; Li et al., 2014; Wu et al., 2008). They did not directly optimize the surrogate objective of pAUC but instead some objectives that have some relationship to the right corner of ROC curve, e.g., p-norm push (Rudin, 2009), infinite-push (Agarwal, 2011; Rakotomamonjy, 2012; Li et al., 2014), and asymmetric SVM objective (Wu et al., 2008). Nevertheless, none of these studies propose algorithms that are scalable and applicable for deep learning.

In (Kar et al., 2014), the authors proposed mini-batch based stochastic methods for pAUC maximization. At each iteration, a gradient estimator is simply computed based on the pAUC surrogate function of the mini-batch data. However, this heuristic approach is not guaranteed to converge for minimizing the pAUC objective and its error scales as $O(1/\sqrt{B})$, where B is the mini-batch size. Narasimhan & Agarwal (2013b;a; 2017) developed rigorous algorithms for optimizing pAUC with FPR restricted in a range (α, β) based on the structured SVM formulation. However, their algorithms are only applicable to learning linear models and are not efficient for big data due to per-iteration costs proportional to the size of training data. Recently, Yang et al. (2021) considered optimizing two-way partial AUC with FPR less than β and TPR larger than α. Their paper focuses on simplifying the optimization problem that involves selection of top ranked negative examples and bottom ranked positive examples. They use ad-hoc weight functions for each positive and negative examples to relax the objective function into decomposable over pairs. The weight function is designed such that the larger the scores of negative examples the higher are their weights, the smaller the scores of positive examples the higher are their weights. Nevertheless, their objective function might have a large approximation error for the pAUC estimator.

There are also some studies about partial AUC maximization without providing rigorous convergence guarantee on their methods, including greedy methods (Wang & Chang, 2011; Ricamato & Tortorella, 2011) and boosting methods (Komori & Eguchi, 2010; Takenouchi et al., 2012). Some works also use pAUC maximization for learning non-linear neural networks (Ueda & Fujino, 2018; Iwata et al., 2020). However, it is unclear how the optimization algorithms were designed as there were no discussion on the algorithm design and convergence analysis. Finally, it was brought to our attention that a recent work (Yao et al., 2022) also considered partial AUC maximization with a non-convex objective. The difference between this work and (Yao et al., 2022) is that: (i) they focus on optimizing one-way pAUC with FPR in a certain range (α, β) where $\alpha > 0$; in contrast we consider optimizing both one-way pAUC and two-way pAUC, but for one-way pAUC we only consider FPR in a range of $(0, \beta)$; (ii) the second difference is that the proposed algorithms in this paper for one-way pAUC maximization has a better complexity than that established in (Yao et al., 2022).

3. Preliminaries

In this section, we present some notations and preliminaries. Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ denote a set of training data, where $\mathbf{x}_i$ represents an input training example (e.g., an image), and $\mathbf{y}_i \in \{1, -1\}$ denotes its corresponding label (e.g., the indicator of a certain disease). Let $h_{\mathbf{w}}(\mathbf{x}) = h(\mathbf{w}, \mathbf{x})$ denote the score function of the neural network on an input data $\mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^d$ denotes the parameters of the network. Denote by $\mathbb{I}(\cdot)$ an indicator function of a predicate, and by $[s]_+ = \max(s, 0)$. For a set of given training examples $\mathcal{S}$, let $\mathcal{S}_+$ and $\mathcal{S}_-$ be the subsets of $\mathcal{S}$ with only positive and negative examples, respectively, with $n_+ = |\mathcal{S}_+|$ and $n_- = |\mathcal{S}_-|$. Let $\mathcal{S}^{\downarrow}[k_1, k_2] \subset \mathcal{S}$ be the subset of examples whose rank in terms of their prediction scores in the descending order are in the range of $[k_1, k_2]$, where $k_1 \leq k_2$. Similarly, let $\mathcal{S}^{\uparrow}[k_1, k_2] \subset \mathcal{S}$ denote the subset of examples whose rank in terms of their prediction scores in the ascending order are in the range of $[k_1, k_2]$, where $k_1 \leq k_2$. We denote by $\mathrm{E}_{\mathbf{x}\sim\mathcal{S}}$ the average over $\mathbf{x} \in \mathcal{S}$. Let $\mathbf{x}_+ \sim \mathbb{P}_+$ denote a random positive example and $\mathbf{x}_- \sim \mathbb{P}_-$ denote a random negative example. We use Δ to denote a simplex of a proper dimension.

A function $F(\mathbf{w})$ is weakly convex if there exists $C > 0$ such that $F(\mathbf{w}) + \frac{C}{2}\|\mathbf{w}\|^2$ is a convex function. A function $F(\mathbf{w})$ is L-smooth if its gradient is Lipchitz continuous, i.e., $\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$.

pAUC and its non-parametric estimator. For a given threshold t and a score function $h(\cdot)$, the TPR can be written as $\mathrm{TPR}(t) = \Pr(h(\mathbf{x}) \geq t | y = 1)$, and the FPR can be written as $\mathrm{FPR}(t) = \Pr(h(\mathbf{x}) > t | y = -1)$. For a given $u \in [0, 1]$, let $\mathrm{FPR}^{-1}(u) = \inf\{t \in \mathbb{R} : \mathrm{FPR}(t) \leq u\}$ and $\mathrm{TPR}^{-1}(u) = \inf\{t \in \mathbb{R} : \mathrm{TPR}(t) \leq u\}$. The ROC curve defined as $\{u, \mathrm{ROC}(u)\}$, where $u \in [0, 1]$ and $\mathrm{ROC}(u) = \mathrm{TPR}(\mathrm{FPR}^{-1}(u))$. OPAUC (non-normalized) with FRP restricted in the range (α_0, α_1) is equal to (Dodd & Pepe, 2003)

$$\mathrm{OPAUC}(h, \alpha_0, \alpha_1) = \int_{\alpha_0}^{\alpha_1} \mathrm{ROC}(u)du = \Pr(h(\mathbf{x}_+) > h(\mathbf{x}_-), h(\mathbf{x}_-) \in [\mathrm{FPR}^{-1}(\alpha_1), \mathrm{FPR}^{-1}(\alpha_0)]), \tag{1}$$

where $h(\mathbf{x}_-) \in [\mathrm{FPR}^{-1}(\alpha_1), \mathrm{FPR}^{-1}(\alpha_0)]$ means that only negative examples whose prediction scores are in certain quantiles are considered. As a result, we have the following non-parametric estimator of OPAUC:

$$\widehat{\mathrm{OPAUC}}(h, \alpha_0, \alpha_1) = \frac{1}{n_+}\frac{1}{n_-}\sum_{\mathbf{x}_i \in \mathcal{S}_+}\sum_{\mathbf{x}_j \in \mathcal{S}_-^{\downarrow}[k_1+1, k_2]} \mathbb{I}(h(\mathbf{x}_i) > h(\mathbf{x}_j)), \tag{2}$$

where $k_1 = \lceil n_-\alpha_0 \rceil, k_2 = \lfloor n_-\alpha_1 \rfloor$. In this work, we will focus on optimizing $\widehat{\mathrm{OPAUC}}(h, 0, \beta)$ for some $\beta \in (0, 1)$.

Similarly, a non-parametric estimator of TPAUC with $\mathrm{FPR} \leq \beta, \mathrm{TPR} \geq \alpha$ is given by

$$\widehat{\mathrm{TPAUC}}(h, \alpha, \beta) = \frac{1}{n_+}\frac{1}{n_-}\sum_{\mathbf{x}_i \in \mathcal{S}_+^{\uparrow}[1, k_1]}\sum_{\mathbf{x}_j \in \mathcal{S}_-^{\downarrow}[1, k_2]} \mathbb{I}(h(\mathbf{x}_i) > h(\mathbf{x}_j)), \tag{3}$$

where $k_1 = \lfloor n_+\alpha \rfloor, k_2 = \lfloor n_-\beta \rfloor$.

Distributionally Robust Optimization (DRO). For a set of random loss functions $\ell_1(\cdot), \ldots, \ell_n(\cdot)$, a DRO loss can be written as

$$\hat{L}_\phi(\cdot) = \max_{\mathbf{p} \in \Delta} \sum_j p_j \ell_j(\cdot) - \lambda D_\phi(\mathbf{p}, 1/n), \tag{4}$$

where $D_\phi(\mathbf{p}, 1/n) = \frac{1}{n}\sum_i \phi(np_i)$ is a divergence measure, and $\lambda > 0$ is a parameter. The idea of the DRO loss is to assign an importance weight p_i to each individual loss and take the uncertainty into account by maximization over $\mathbf{p} \in \Delta$ with a proper constraint/regularization on $\mathbf{p}$. In the literature, several divergence measures have been considered (Levy et al., 2020). In this paper, we will consider two special divergence measures that are of most interest for our purpose, i.e., the KL divergence $\phi_{kl}(t) = t\log t - t + 1$, which gives $D_\phi(\mathbf{p}, 1/n) = \sum_i p_i \log(np_i)$, and the CVaR divergence $\phi_c(t) = \mathbb{I}(0 < t \leq 1/\gamma)$ with a parameter $\gamma \in (0, 1)$, which gives $D_\phi(\mathbf{p}, 1/n) = 0$ if $p_i \leq 1/(n\gamma)$ and infinity otherwise. The following lemma gives the closed form of $\hat{L}_\phi$ for ϕ_c and ϕ_{kl}.

Lemma 1. *By using KL divergence measure, we have*

$$\hat{L}_{kl}(\cdot;\lambda) = \lambda \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{\ell_i(\cdot)}{\lambda}\right)\right). \quad (5)$$

By using the CVaR divergence $\phi_c(t)$ for some γ such that $n\gamma$ is an integer, we have,

$$\hat{L}_{cvar}(\cdot;\gamma) = \frac{1}{n\gamma}\sum_{i=1}^{n\gamma}\ell_{[i]}(\cdot), \quad (6)$$

where $\ell_{[i]}(\cdot)$ denotes the i-th largest value in $\{\ell_1,\ldots,\ell_n\}$.

The estimator in (6) is also known as the estimator of conditional-value-at-risk (Rockafellar et al., 2000).

4. AUC meets DRO for OPAUC Maximization

Since the non-parametric estimator of OPAUC in (2) is non-continuous and non-differentiable, a continuous surrogate objective for OPAUC($h_{\mathbf{w}}, 0, \beta$) is usually defined by using a continuous pairwise surrogate loss $L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) = \ell(h_{\mathbf{w}}(\mathbf{x}_i) - h_{\mathbf{w}}(\mathbf{x}_j))$, resulting in the following problem:

$$\min_{\mathbf{w}} \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\frac{1}{n_-\beta}\sum_{\mathbf{x}_j\in\mathcal{S}_-^{\downarrow}[1,n_-\beta]} L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j), \quad (7)$$

where we assume $n_-\beta$ is an positive integer for simplicity of presentation. For the surrogate loss $\ell(\cdot)$, we assume it satisfies the following properties.

Assumption 1. *We assume $\ell(\cdot)$ is a convex, differentiable and monotonically decreasing function when $\ell(\cdot) > 0$, and $\ell'(0) < 0$.*

It is notable that the above condition is a sufficient condition to ensure that the surrogate $\ell(\cdot)$ is consistent for AUC maximization (Gao & Zhou, 2015). There are many surrogate loss functions that have the above properties, e.g., squared hinge loss $\ell(s) = (c-s)_+^2$, logistic loss $\ell(s) = \log(1+\exp(-s/c))$ where $c > 0$ is a parameter.

The challenge of optimizing a surrogate objective of pAUC in (7) lies at tackling the selection of top ranked negative examples from $\mathcal{S}_-$, i.e., $\mathcal{S}_-^{\downarrow}[1,k]$ for some fixed k. It is impossible to compute an unbiased stochastic gradient of the objective in (7) based on a mini-batch of examples that include only a part of negative examples.

4.1. AUC meets DRO for OPAUC

To address the above challenge, we define new formulations for OPAUC maximization by leveraging the DRO. In particular, we define a robust loss for each positive data by

$$\hat{L}_{\phi}(\mathbf{w};\mathbf{x}_i) = \max_{\mathbf{p}\in\Delta}\sum_{\mathbf{x}_j\in\mathcal{S}_-} p_j L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - \lambda D_{\phi}(\mathbf{p}, 1/n_-).$$

Then we define the following objective for OPAUC maximization:

$$\min_{\mathbf{w}} \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\hat{L}_{\phi}(\mathbf{w};\mathbf{x}_i). \quad (8)$$

When $\phi(\cdot) = \phi_c(\cdot)$, we refer to the above estimator (i.e., the objective function) as CVaR-based OPAUC estimator; and when $\phi(\cdot) = \phi_{kl}(\cdot)$, we refer to the above estimator as KLDRO-based OPAUC estimator. Below, we present two theorems to state the equivalent form of the objective, and the relationship between the two estimators and the surrogate objective in (7) of OPAUC.

Theorem 1. *By choosing $\phi(\cdot) = \phi_c(\cdot) = \mathbb{I}(\cdot \in (0, 1/\beta])$, then the problem (8) is equivalent to*

$$\min_{\mathbf{w}}\min_{\mathbf{s}\in\mathbb{R}^{n_+}} F(\mathbf{w},\mathbf{s}) = \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\left(s_i + \frac{1}{\beta}\psi_i(\mathbf{w},s_i)\right), \quad (9)$$

where $\psi_i(\mathbf{w},s_i) = \frac{1}{n_-}\sum_{\mathbf{x}_j\in\mathcal{S}_-}(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - s_i)_+$. If ℓ is a monotonically decreasing function for $\ell(\cdot) > 0$, then the objective in (8) is equivalent to (7) of OPAUC.

Remark: The above theorem indicates that CVaR-based OPAUC estimator is an exact estimator of OPAUC, which is consistent for OPAUC maximization. The variable s_i can be considered as the threshold variable to select the top-ranked negative examples for each positive data.

Theorem 2. *By choosing $\phi(\cdot) = \phi_{kl}(\cdot)$, then the problem (8) becomes*

$$\min_{\mathbf{w}} \frac{1}{n_+}\sum_{\mathbf{x}_i\sim\mathcal{S}_+}\lambda\log \mathrm{E}_{\mathbf{x}_j\in\mathcal{S}_-}\exp\left(\frac{L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)}{\lambda}\right). \quad (10)$$

If $\ell(\cdot)$ is a monotonically decreasing function for $\ell(\cdot) > 0$, when $\lambda = 0$, the above objective is a surrogate of $\widehat{OPAUC}(h_{\mathbf{w}}, 0, \frac{1}{n_-})$; and when $\lambda = +\infty$, the above objective is a surrogate of $\widehat{OPAUC}(h_{\mathbf{w}}, 0, 1)$, i.e., the AUC.

Remark: Theorem 2 indicates that KLDRO-based OPAUC estimator is a soft estimator, which interpolates between OPAUC($h_{\mathbf{w}}, 0, 1/n_-$) and OPAUC($h_{\mathbf{w}}, 0, 1$) by varying λ.

It is also notable that when $\beta = 1/n_-$ in CVaR-based estimator, the objective in (8) becomes the infinite-push (or top-push) objective considered in the literature (Agarwal, 2011; Rakotomamonjy, 2012; Li et al., 2014), and hence our algorithm for solving (9) can be also used for solving the infinite-push objective for deep learning. In contrast, the previous works for the infinite-push objective focus on learning linear models. Similarly, when $\lambda = 0$ in KLDRO-based estimator, the objective in (10) becomes the infinite-push objective. Nevertheless, our algorithm for optimizing KLDRO-based estimator is not exactly applicable to optimizing the infinite-push objective as we focus on the cases $\lambda > 0$, which yields a smooth objective function under proper conditions of $\ell(\cdot)$ and $h(\cdot;\mathbf{x})$. As a result, we could

have stronger convergence by optimizing the KLDRO-based estimator as indicated by our convergence results in next subsection.

4.2. Optimization Algorithms and Convergence Results

In this subsection, we present the optimization algorithms for solving both (9) and (10), and then present their convergence results for finding a nearly stationary solution. The key to our development is to formulate the two optimization problems into known non-convex optimization problems that have been studied in the literature, and then to develop stochastic algorithms by borrowing the existing techniques.

Optimizing CVaR-based estimator. We first consider optimizing the CVaR-based estimator, which is equivalent to (9). A benefit for solving (9) is that an unbiased stochastic subgradient can be computed in terms of $(\mathbf{w}, \mathbf{s})$. However, this problem is still challenging because the objective function $F(\mathbf{w}, \mathbf{s})$ is non-smooth non-convex. In order to develop a stochastic algorithm with convergence guarantee, we prove that $F(\mathbf{w}, \mathbf{s})$ is weakly convex in terms of $(\mathbf{w}, \mathbf{s})$, which allows us to borrow the techniques of optimizing weakly convex function (Davis & Drusvyatskiy, 2018) for solving our problem and to establish the convergence. We first establish the weak convexity of $F(\mathbf{w}, \mathbf{s})$.

Lemma 2. *If $L(\cdot; \mathbf{x}_i, \mathbf{x}_j)$ is a L_s-smooth function for any $\mathbf{x}_i, \mathbf{x}_j$, then $F(\mathbf{w}, \mathbf{s})$ is ρ-weakly convex with $\rho = L_s/\beta$.*

Another challenge for optimizing $F(\mathbf{w}, \mathbf{s})$ is that $\mathbf{s}$ is of high dimensionality and computing the gradient for all entries in $\mathbf{s}$ at each iteration is expensive. Therefore, we develop a tailored stochastic algorithm for solving (9), which is shown in Algorithm 1. This algorithm uses stochastic gradient descent (SGD) updates for updating $\mathbf{w}$ and stochastic coordinate gradient descent (SCGD) updates for updating $\mathbf{s}$. We refer to the algorithm as SOPA. A key feature of SOPA is that the stochastic gradient estimator for $\mathbf{w}$ is a weighted average gradient of the pairwise losses for all pairs in the mini-batch, i.e., step 6. The hard weights p_{ij} (either 0 or 1) are dynamically computed by step 4, which compares the pairwise loss $(\ell(h(\mathbf{w}_t, \mathbf{x}_i) - h(\mathbf{w}_t, \mathbf{x}_j))$ with the threshold variable s_i^t, which is also updated by a SGD step.

Optimizing KLDRO-based estimator of OPAUC. Next, we consider optimizing the KLDRO-based estimator, which is equivalent to (10). A nice property of the objective function is that it is smooth under a proper condition as stated in Assumption 2. However, the challenge for solving (10) is that an unbiased stochastic gradient is not readily computed. To highlight the issue, the problem (10) can be written as

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{n_+} \sum\nolimits_{\mathbf{x}_i \in \mathcal{S}_+} f(g_i(\mathbf{w})), \qquad (11)$$

where $g_i(\mathbf{w}) = \mathrm{E}_{\mathbf{x}_j \sim \mathcal{S}_-} \exp(\frac{L(\mathbf{w}; x_i, \mathbf{x}_j)}{\lambda})$ and $f(\cdot) = \lambda \log(\cdot)$. A similar optimization problem has been studied in (Qi et al., 2021) for maximizing average precision, which is referred to as finite-sum coupled compositional stochastic optimization, where $f(g_i(\mathbf{w}))$ is a compositional function and $g_i(\mathbf{w})$ depends on the index i for the outer summation. A full gradient of $f(g_i(\mathbf{w}))$ is given by $f'(g_i(\mathbf{w}))\nabla g_i(\mathbf{w})$. With a mini-batch of samples, $g_i(\mathbf{w})$ can be estimated by an unbiased estimator $\hat{g}_i(\mathbf{w})$. However, $f'(\hat{g}_i(\mathbf{w}))\nabla \hat{g}_i(\mathbf{w})$ is a biased estimator due to the compositional form. To address this challenge, Qi et al. (2021) proposed a novel stochastic algorithm that maintains a moving average estimator for $g_i(\mathbf{w})$ denoted by u_i. Recently, Wang & Yang (2022) has also studied the finite-sum coupled compositional optimization problem comprehensively and proposed a similar algorithm (SOX) and derived better convergence results than that in (Qi et al., 2021). Hence, we employ the same algorithm in (Wang & Yang, 2022) for solving (10), which is in shown in Algorithm 2 and is referred to as SOPA-s.

Algorithm 1 SOPA

1: Set $\mathbf{s}^1 = 0$ and initialize $\mathbf{w}$
2: **for** $t = 1, \ldots, T$ **do**
3: Sample two mini-batches $\mathcal{B}_+ \subset \mathcal{S}_+, \mathcal{B}_- \subset \mathcal{S}_-$
4: Let $p_{ij} = \mathbb{I}(\ell(h(\mathbf{w}_t, \mathbf{x}_i) - h(\mathbf{w}_t, \mathbf{x}_j)) - s_i^t > 0)$
5: Update $s_i^{t+1} = s_i^t - \frac{\eta_2}{n_+}(1 - \frac{\sum_j p_{ij}}{\beta|\mathcal{B}_-|})$ for $\mathbf{x}_i \in \mathcal{B}_+$
6: Compute a gradient estimator ∇_t by

$$\nabla_t = \frac{1}{\beta|\mathcal{B}_+||\mathcal{B}_-|} \sum_{\mathbf{x}_i \in \mathcal{B}_+} \sum_{\mathbf{x}_j \in \mathcal{B}_-} p_{ij} \nabla_{\mathbf{w}} L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)$$

7: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_1 \nabla_t$
8: **end for**

There are two key differences between SOPA-s and SOPA. First, the pairwise weights p_{ij} in SOPA-s (step 5) are soft weights between 0 and 1, in contrast to the hard weights $p_{ij} \in \{0, 1\}$ in SOPA. Second, the update for $\mathbf{w}_{t+1}$ is a momentum-based update where $\gamma_1 \in (0, 1)$. We can also use an Adam-style update, which shares similar convergence as the momentum-based update (Guo et al., 2021).

4.3. Convergence Analysis

For convergence analysis, we make the following assumption about h and $\ell(\cdot)$.

Assumption 2. *Assume $h(\cdot; \mathbf{x})$ is Lipschitz continuous, smooth and bounded, $\ell(\cdot)$ is a smooth function and has a bounded gradient for a bounded argument.*

A bounded smooth score function $h(\cdot; \mathbf{x})$ is ensured if the activation function of the neural network is smooth and the output layer uses a bounded and smooth activation function. For example, let $\hat{h}(\mathbf{w}; \mathbf{x})$ denote the plain output of the neural network, then the score function $h(\mathbf{w}; \mathbf{x}) =$

Algorithm 2 SOPA-s

1: Set $\mathbf{u}^1 = 0$ and initialize $\mathbf{w}$
2: **for** $t = 1, \ldots, T$ **do**
3: Sample two mini-batches $\mathcal{B}_+ \subset \mathcal{S}_+, \mathcal{B}_- \subset \mathcal{S}_-$
4: For each $\mathbf{x}_i \in \mathcal{B}_+$, update $u_i^{t+1} = (1-\gamma_0)u_i^t + \gamma_0 \frac{1}{|\mathcal{B}_-|}\sum_{\mathbf{x}_j \in \mathcal{B}_-} \exp\left(\frac{L(\mathbf{w}_t;\mathbf{x}_i,\mathbf{x}_j)}{\lambda}\right)$
5: Let $p_{ij} = \exp(L(\mathbf{w}_t;\mathbf{x}_i,\mathbf{x}_j)/\lambda)/u_i^t$
6: Compute a gradient estimator ∇_t by

$$\nabla_t = \frac{1}{|\mathcal{B}_+|}\frac{1}{|\mathcal{B}_-|}\sum_{\mathbf{x}_i \in \mathcal{B}_+}\sum_{\mathbf{x}_j \in \mathcal{B}_-} p_{ij}\nabla L(\mathbf{w}_t;\mathbf{x}_i,\mathbf{x}_j)$$

7: Update $\mathbf{v}_t = (1-\gamma_1)\mathbf{v}_{t-1} + \gamma_1\nabla_t$
8: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{v}_t$ (or Adam-style)
9: **end for**

$1/(1+\exp(-\hat{h}(\mathbf{w};\mathbf{x}))$ is bounded and smooth. The Lipschitz continuity of $h(\mathbf{w};\mathbf{x})$ can be guaranteed if $\mathbf{w}$ is bounded.

We first consider the analysis of SOPA. Since $F(\mathbf{w},\mathbf{s})$ is non-smooth, for presenting the convergence result, we need to introduce a convergence measure based on the Moreau envelope of $F(\mathbf{w},\mathbf{s})$ given below for some $\hat{\rho} > \rho$:

$$F_{\hat{\rho}}(\mathbf{w},\mathbf{s}) = \min_{\mathbf{w},\mathbf{s}} F(\mathbf{w},\mathbf{s}) + \frac{\hat{\rho}}{2}(\|\mathbf{w}\|^2 + \|\mathbf{s}\|^2).$$

It is guaranteed that $F_{\hat{\rho}}(\mathbf{w},\mathbf{s})$ is a smooth function (Drusvyatskiy & Paquette, 2019). A point $(\mathbf{w},\mathbf{s})$ is called an ϵ-nearly stationary solution to $F(\mathbf{w},\mathbf{s})$ if $\|\nabla F_{\hat{\rho}}(\mathbf{w},\mathbf{s})\| \leq \epsilon$ for some $\hat{\rho} > \rho$, where ρ is the weak convexity parameter of F. This convergence measure has been widely used for weakly convex optimization problems (Davis & Drusvyatskiy, 2018; Rafique et al., 2020; Chen et al., 2019). Then we establish the following convergence guarantee for SOPA.

Theorem 3. *Under Assumption 2, Algorithm 1 ensures that after $T = O(1/(\beta\epsilon^4))$ iterations we can find an ϵ nearly stationary solution of $F(\mathbf{w},\mathbf{s})$, i.e., $\mathrm{E}\|\nabla F_{\hat{\rho}}(\mathbf{w}_\tau,\mathbf{s}_\tau)\|^2 \leq \epsilon^2$ for a randomly selected $\tau \in \{1,\ldots,T\}$ and $\hat{\rho} = 1.5\rho$.*

Next, we establish the convergence of SOPA-s. Under Assumption 2, we can show that $F(\mathbf{w})$ in (11) is smooth. Hence, we use the standard convergence measure in terms of gradient norm of $F(\mathbf{w})$.

Theorem 4. *Under Assumption 2, Algorithm 2 with $\gamma_0 = O(B_-\epsilon^2)$, $\gamma_1 = O(\min\{B_-, B_+\}\epsilon^2)$, $\eta = O(\min\{\gamma_0 B_1/n_+, \gamma_1\})$ ensures that after $T = O(\frac{1}{\min(B_+,B_-)\epsilon^4} + \frac{n_+}{B_+B_-\epsilon^4})$ iterations we can find an ϵ-stationary solution of $F(\mathbf{w})$, i.e., $\mathrm{E}[\|\nabla F(\mathbf{w}_\tau)\|^2] \leq \epsilon^2$ for a randomly selected $\tau \in \{1,\ldots,T\}$, where $B_+ = |\mathcal{B}_+|$ and $B_- = |\mathcal{B}_-|$.*

Remark: The convergence analysis of Algorithm 2 follows directly from that in (Wang & Yang, 2022). Compared with that in Theorem 3 for SOPA, the convergence of SOPA-s is stronger than that of SOPA in several aspects: (i) the convergence measure of SOPA-s is stronger than that of SOPA due to that Theorem 4 guarantees the convergence in terms of gradient norm of the objective, while Theorem 3 guarantees the convergence on a weaker convergence measure namely the gradient norm of the Moreau envelope of the objective; (ii) the complexity of SOPA-s enjoys a parallel speed-up by using a mini-batch of data. However, it is also notable that the complexity of SOPA does not depend on the number of positive data as that of SOPA-s.

5. AUC meets DRO for TPAUC Maximization

In this section, we propose estimators for the surrogate objective of TPAUC and stochastic algorithms with convergence guarantee for optimizing the estimators. To this end, we apply another level of DRO on top of $\hat{L}_\phi(\mathbf{x}_i,\mathbf{w}), \mathbf{x}_i \in \mathcal{S}_+$ and define the following estimator of TPAUC:

$$F(\mathbf{w};\phi,\phi') = \max_{\mathbf{p}\in\Delta}\sum_{\mathbf{x}_i\in\mathcal{S}_+} p_i\hat{L}_\phi(\mathbf{x}_i,\mathbf{w}) - \lambda' D_{\phi'}(\mathbf{p},\frac{1}{n_+}).$$

Next, we focus on optimizing the soft estimator of TPAUC defined by using $\phi = \phi' = \phi_{kl}$. First, we have the following form for the estimator.

Lemma 3. *When $\phi = \phi' = \phi_{kl}$, we have*

$$F(\mathbf{w};\phi_{kl},\phi_{kl}) = \lambda'\log \mathrm{E}_{\mathbf{x}_i\sim\mathcal{S}_+}\left(\mathrm{E}_{\mathbf{x}_j\sim\mathcal{S}_-}\exp(\frac{\ell(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)}{\lambda})\right)^{\frac{\lambda}{\lambda'}}.$$

For minimizing this function, we formulate the problem as a novel three-level compositional stochastic optimization:

$$\min_{\mathbf{w}} f_1(\frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+} f_2(g_i(\mathbf{w}))),$$

where $f_1(s) = \lambda'\log(s)$, $f_2(g) = g^{\lambda/\lambda'}$ and $g_i(\mathbf{w}) = \mathrm{E}_{\mathbf{x}_j\sim\mathcal{S}_-}\exp(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)/\lambda)$. We propose a novel stochastic algorithm for solving the above problem, which is shown in Algorithm 3, to which we refer as SOTA-s. Note that the problem is similar to multi-level compositional optimization (Balasubramanian et al., 2021) but also has subtle difference. The function inside f_1 has a form similar to that in (11). Hence, we use similar technique to SOPA-s by maintaining and updating u^i to track $g_i(\mathbf{w})$ in step 4. Besides, we need to maintain and update v_{t+1} to track $\frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+} f_2(g_i(\mathbf{w}_t))$ in step 5. Then the gradient estimator in step 7 is computed by $\nabla f_1(v_{t+1})\frac{1}{|\mathcal{B}_+|}\sum_{\mathbf{x}_i\in\mathcal{B}_+}\nabla\hat{g}_i(\mathbf{w}_t)\nabla f_2(u_t^i)$, where $\hat{g}_i(\mathbf{w}) = \mathrm{E}_{\mathbf{x}_j\sim\mathcal{B}_-}\exp(L(\mathbf{w}_t;\mathbf{x}_i,\mathbf{x}_j)/\lambda)$. Then we update the model parameter by the momentum-style or Adam-style update.

Theorem 5. *Under Assumption 2, Algorithm 3 with $\gamma_0 =$*

Algorithm 3 SOTA-s

Set $\mathbf{u}_0 = 0, v_0 = 0, \mathbf{m}_0 = 0$ and initialize $\mathbf{w}$
for $t = 1, \ldots, T$ **do**
  Sample two mini-batches $\mathcal{B}_+ \subset \mathcal{S}_+, \mathcal{B}_- \subset \mathcal{S}_-$
  For each $\mathbf{x}_i \in \mathcal{B}_+$ compute $u_t^i = (1-\gamma_0)u_{t-1}^i + \gamma_0 \frac{1}{|\mathcal{B}_-|}\sum_{\mathbf{x}_j \in \mathcal{B}_-} L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)$
  Let $v_t = (1-\gamma_1)v_{t-1} + \gamma_1 \frac{1}{|\mathcal{B}_+|}\sum_{\mathbf{x}_i \in \mathcal{B}_+} f_2(u_{t-1}^i)$
  Let $p_{ij} = (u_{t-1}^i)^{\lambda/\lambda'-1}\exp(L(\mathbf{w}_t, \mathbf{x}_i, \mathbf{x}_j)/\lambda)/v_t$
  Compute a gradient estimator ∇_t by

$$\nabla_t = \frac{1}{|\mathcal{B}_+|}\frac{1}{|\mathcal{B}_-|}\sum_{\mathbf{x}_i \in \mathcal{B}_+}\sum_{\mathbf{x}_j \in \mathcal{B}_-} p_{ij}\nabla L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)$$

  Update $\mathbf{m}_t = (1-\gamma_2)\mathbf{m}_{t-1} + \gamma_2\nabla_t$
  Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{m}_t$ (or Adam-style)
end for

$O(B_-\epsilon^2)$, $\gamma_1 = O(B_+\epsilon^2)$, $\gamma_2 = O(\min\{B_-, B_+\}\epsilon^2)$, $\eta = O(\min\{\gamma_0 B_1/n_+, \gamma_1, \gamma_2\})$ *ensures that after* $T = O(\frac{1}{\min(B_+,B_-)\epsilon^4} + \frac{n_+}{B_+B_-\epsilon^4})$ *iterations we can find an* ϵ *nearly stationary solution of* $F(\mathbf{w})$*, where* $B_+ = |\mathcal{B}_+|$ *and* $B_- = |\mathcal{B}_-|$.

Remark: It is notable that SOTA-s has an iteration complexity in the same order of SOPA-s for OPAUC maximization.

Finally, we discuss how to optimize the exact estimator of TPAUC defined by $F(\mathbf{w}; \phi_c, \phi_c')$, where $\phi_c(t) = \mathbb{I}(0 \le t \le 1/\beta)$ and $\phi_c'(t) = \mathbb{I}(0 \le t \le 1/\alpha)$ with $K_2 = n_-\beta$ and $K_1 = n_+\alpha$ being integers. Lemma 7 in the supplement shows that if $\ell(\cdot)$ is monotonically decreasing for $\ell(\cdot) > 0$

$$F(\mathbf{w}; \phi_c, \phi_c') = \frac{1}{K_1K_2}\sum_{\mathbf{x}_i \in \mathcal{S}_+^{\uparrow}[1,K_1]}\sum_{\mathbf{x}_j \in \mathcal{S}_-^{\downarrow}[1,K_2]} L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j),$$

is a consistent surrogate function of TPAUC for TPR $\ge \alpha$ and FPR $\le \beta$ in view of the estimator $\widehat{\text{TPAUC}}$ given in (3).

Similar to Theorem 1, we can show that $F(\mathbf{w}; \phi_c, \phi_c')$ is equivalent to:

$$\min_{s' \in \mathbb{R}, \mathbf{s} \in \mathbb{R}^{n_+}} s' + \frac{1}{n_+\alpha}\sum_{\mathbf{x}_i \in \mathcal{S}_+}(s_i + \frac{1}{\beta}\psi_i(\mathbf{w}; s_i) - s')_+.$$

Like $F(\mathbf{w}, \mathbf{s})$ in (9), we can prove the inner function is weakly convex in terms of $(\mathbf{w}, \mathbf{s}, s')$. However, computing an unbiased stochastic gradient in terms of $\mathbf{w}$ and s_i is also impossible due to that $\psi_i(\mathbf{w}; s_i)$ is inside a hinge function. To solve this problem, we can use the conjugate form of the hinge function to convert the minimization of $F(\mathbf{w}; \phi_c, \phi_c')$ into a weakly-convex concave min-max problem (Rafique et al., 2020) and we can develop a stochastic algorithm but only with $O(1/\epsilon^6)$ iteration complexity for finding a nearly stationary solution. We present the algorithm and analysis in the supplement for interested readers.

6. Experiments

Datasets. We consider binary classification tasks on two types of datasets, namely image datasets and molecular datasets. For image datasets, we use CIFAR-10, CIFAR-100, Melanoma for experiments. For CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), we construct imbalanced versions of the datasets by randomly removing some positive samples following (Yuan et al., 2020). Specifically, we take first half of classes as the negative class and last half of classes as the positive class, and then remove 80% samples from the positive class to make it imbalanced. The Melanoma dataset is a naturally imbalanced medical dataset which is released on Kaggle (Rotemberg et al., 2021). For molecular datasets, we use ogbg-moltox21 (the No.0 target), ogbg-molmuv (the No.1 target) and ogbg-molpbca (the No.0 target) for experiments, which are from the Stanford Open Graph Benchmark (OGB) website (Hu et al., 2020). The task on these molecular datasets is to predict certain property of molecules. The statistics for the datasets are presented in Table 5 in the supplement.

Deep Models. For image datasets, we learn convolutional neural network (CNN) and use ResNet18 (He et al., 2016) for CIFAR-10, CIFAR100 and Melanoma. For molecular datasets, we learn graph neural network (GNN) and use Graph Isomorphism Network (GIN) as the backbone model on all datasets (Xu et al., 2018), which has 5 mean-pooling layers with 64 number of hidden units and dropout rate 0.5.

Baselines. We will compare our methods with different baselines for both training performance and testing performance. For comparison of training convergence, we consider different methods for optimizing the same objective, i.e., partial AUC. We compare with 2 baselines, i.e., the naive mini-batch based method (Kar et al., 2014), to which we refer as MB, and a recently proposed ad-hoc weight based method (Yang et al., 2021), to which we refer as AW-poly. MB that optimizes OPAUC only considers the top negative samples in the mini-batch; and MB that optimizes TPAUC considers the top negative samples and bottom positive samples in the mini-batch. For AW-poly, we use the polynomial weight function according to their paper. It is notable that AW-poly was originally proposed for optimizing TPAUC. But it can be easily modified for optimizing OPAUC with FPR in $(0, \beta)$. For comparison of testing performance, we compare different methods for optimizing different objectives, including the cross-entropy loss (CE), the pair-wise squared hinge loss for AUC maximization (AUC-SH), the AUC min-max margin loss (AUC-M) (Yuan et al., 2021), p-norm push (P-push) (Rudin, 2009). For optimizing CE and AUC-SH, we use the standard Adam optimizer. For optimizing AUC-M, we use their proposed

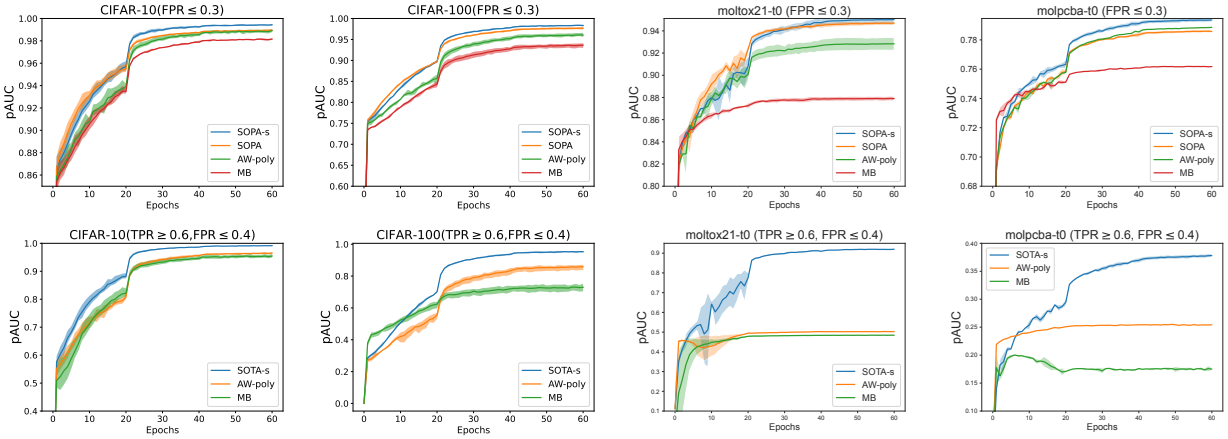

Figure 2. Training Convergence Curves on image and molecular datasets; Top for OPAUC maximization, bottom for TPAUC maximization.

optimizer PESG. For P-push, we use a stochastic algorithm with an Adam-style update similar to that proposed in (Qi et al., 2021). For our methods, we use $\ell(t) = (1-t)_+^2$ and also use the Adam-style update unless specified explicitly. Similar to (Yuan et al., 2021; Qi et al., 2021), we use a pre-training step that optimizes the base model by optimizing CE loss with an Adam optimizer, and then re-initialize the classifier layer and fine-tune all layers by different methods.

Target Measures. For OPAUC maximization, we evaluate OPAUC with two FPR upper bounds, i.e., FPR ≤ 0.3 and FPR≤ 0.5 separately. For TPAUC maximization, we evaluate TPAUC with two settings, i.e, FPR≤ 0.4 and TPR≥ 0.6, and FPR≤ 0.5 and TPR≥ 0.5.

Parameter Tuning. The learning rate of all methods is tuned in {1e-3, 1e-4, 1e-5}, except for PESG which is tuned at {1e-1, 1e-2, 1e-3} because it favors a larger learning rate. Weight decay is fixed as 2e-4. Each method is run 60 epochs in total and learning rate decays 10-fold after every 20 epochs. The mini-batch size is 64. For AUC-M, we tune the hyperparameter γ that controls consecutive epoch-regularization in {100, 500, 1000}. For P-push, we tune the polynomial power hyper-parameter in $\{2, 4, 6\}$. For MB that optimizes OPAUC, we tune the top proportion of negative samples in $\{10\%, 30\%, 50\%\}$, and for MB that optimizes TPAUC we tune the top proportion of negative samples in $\{30\%, 40\%, 50\%\}$, and tune the bottom proportion of positive samples in the range $\{30\%, 40\%, 50\%\}$. For AW-poly, we follow (Yang et al., 2021) and tune its parameter γ in {101, 34, 11}. For SOPA, we tune the truncated FPR i.e. β in {0.1, 0.3, 0.5}. For SOPA-s, we fix $\gamma_0 = 0.9$ and tune the KL-regularization parameter λ in {0.1, 1.0, 10}, and for SOTA-s, we fix $\gamma_0 = \gamma_1 = 0.9$, and tune both λ and λ' in {0.1, 1.0, 10}. The momentum parameter for updating $\mathbf{v}_t$ in SOPA-s (i.e., $1-\gamma_1$) and SOTA-s (i.e., $1-\gamma_2$) is set to the default value as in the Adam optimizer, i.e., 0.1. For comparison of training convergence, the parameters are tuned according to the training performance. For comparison of testing performance, the parameters are tuned according to the validation performance. For each experiment, we repeat multiple times with different train/validation splits and random seeds, then report average and standard deviation over multiple runs.

Results. We show the plots of training convergence in Figure 2 on two image datasets (CIFAR-10, -100) and on two molecular datasets (moltox21, molpcba). From the results, we can see that SOPA-s (SOTA-s) converge always faster than MB and AW-poly for OPAUC (TPAUC) maximization. For OPAUC maximization, SOPA-s is usually faster than SOPA. More results are included in the supplement on other datasets with similar observations. The testing performance on all six datasets are shown in Table 1, 2, 3 and 4. In most cases, the proposed methods are better than the baselines. In particular, dramatic improvements have been observed on Melanoma and ogbg-molmuv datasets, which are two datasets with the highest imbalance ratios. In addition, we see that AUC maximization methods (AUC-M, AUC-SH) are not necessarily good for pAUC maximization.

Accuracy of KLDRO-based estimator. Of independent interest, we conduct simple experiments to verify the accuracy of KLDRO-based estimator of OPAUC. To this end, we compute the relative error (RE) of KLDRO-based estimator compared with the exact estimator (i.e., CVaR-based estimator). For a given upper bound of FRP we vary λ for 100 independently randomly generated model parameters $\mathbf{w}$, and the results are shown in the following figure on moltox21-t0 data (please refer to the experiments section for more information of the dataset), which demonstrates that for a given FPR there exists λ such that KLDRO estimator is close to the exact estimator.

Ablation Study. We also conduct some ablation study to understand the proposed algorithm SOPA-s and SOTA-s. In particular for both algorithms, we verify that tuning γ_0 in SOPA-s and γ_0, γ_1 in SOTA-s can help further improve the performance. The results are included in the supplement.

Table 1. One way partial AUC on testing data of three image datasets

	CIFAR-10		CIFAR-100		Melanoma	
Methods	FPR≤0.3	FPR≤0.5	FPR≤0.3	FPR≤0.5	FPR≤0.3	FPR≤0.5
CE	0.8446(0.0018)	0.8777(0.0014)	0.7338(0.0047)	0.7787(0.0044)	0.7651(0.0135)	0.8151(0.0028)
AUC-SH	0.8657(0.0056)	0.8948(0.0036)	0.7467(0.0047)	0.7930(0.0027)	0.7824(0.0138)	0.8176(0.0160)
AUC-M	0.8678(0.0016)	0.8934(0.0022)	0.7371(0.0031)	0.7828(0.0005)	0.7788(0.0068)	0.8249(0.0141)
P-push	0.8610(0.0007)	0.8889(0.0021)	0.7445(0.0025)	0.7930(0.0029)	0.7440(0.0130)	0.8028(0.0170)
MB	0.8690(0.0016)	0.8931(0.0015)	0.7487(0.0017)	0.7930(0.0014)	0.7683(0.0303)	0.8184(0.0278)
AW-poly	0.8664(0.0052)	0.8915(0.0075)	0.7490(0.0058)	0.7909(0.0068)	0.7936(0.0238)	0.8355(0.0067)
SOPA	**0.8766(0.0034)**	**0.9028(0.0031)**	**0.7551(0.0044)**	**0.7999(0.0028)**	**0.8093(0.0248)**	**0.8585(0.0210)**
SOPA-s	0.8691(0.0036)	0.8961(0.0036)	0.7468(0.0056)	0.7877(0.0053)	0.7775(0.0076)	0.8401(0.0206)

Table 2. Two way partial AUC on testing data of three image datasets; (α,β) represents TPR$\geq$ α and FPR $\leq$ β.

	CIFAR-10		CIFAR-100		Melanoma	
Methods	(0.6,0.4)	(0.5,0.5)	(0.6,0.4)	(0.5,0.5)	(0.6,0.4)	(0.5,0.5)
CE	0.4981(0.0078)	0.6414(0.0080)	0.2178(0.0136)	0.4011(0.0118)	0.3399(0.0135)	0.5150(0.0038)
AUC-SH	0.5622(0.0064)	0.6923(0.0071)	0.2599(0.0061)	0.4397(0.0062)	0.3640(0.0354)	0.5291(0.0312)
AUC-M	0.5691(0.0021)	0.6907(0.0125)	0.2336(0.0041)	0.4153(0.0022)	0.3665(0.0646)	0.5404(0.0545)
P-push	0.5477(0.0077)	0.6781(0.0055)	0.2623(0.0042)	0.4417(0.0092)	0.3317(0.0304)	0.4870(0.0443)
MB	0.5404(0.0041)	0.6724(0.0011)	0.2207(0.0033)	0.4017(0.0149)	0.3330(0.0258)	0.4981(0.0252)
AW-poly	0.5536(0.0196)	0.6814(0.0203)	0.2489(0.0166)	0.4342(0.0112)	0.3878(0.0292)	0.5216(0.0288)
SOTA-s	**0.5799(0.0202)**	**0.7074(0.0145)**	**0.2708(0.0055)**	**0.4528(0.0069)**	**0.4198(0.0825)**	**0.5865(0.0664)**

Table 3. One way partial AUC on testing data of three molecular datasets

	moltox21(t0)		molmuv(t1)		molpcba(t0)	
Methods	FPR≤0.3	FPR≤0.5	FPR≤0.3	FPR≤0.5	FPR≤0.3	FPR≤0.5
CE	0.6671(0.0009)	0.6954(0.005)	0.8008(0.0090)	0.8201(0.0061)	0.6802(0.0002)	0.7169(0.0002)
AUC-SH	0.7161(0.0043)	0.7295(0.0036)	0.7880(0.0382)	0.8025(0.0437)	0.6939(0.0009)	0.7350(0.0015)
AUC-M	0.6866(0.0048)	0.7080(0.0020)	0.7960(0.0123)	0.8076(0.0175)	0.6985(0.0016)	0.7399(0.0005)
P-push	0.6946(0.0107)	0.7160(0.0073)	0.7832(0.0220)	0.7940(0.0321)	0.6841(0.0007)	0.7293(0.0043)
MB	**0.7398(0.0131)**	0.7329(0.0099)	0.7672(0.0563)	0.7772(0.0547)	0.6899(0.0002)	0.7253(0.0006)
AW-poly	0.7227(0.0024)	0.7271(0.0112)	0.7754(0.0372)	0.7883(0.0431)	0.6975(0.0006)	0.7350(0.0015)
SOPA	0.7209(0.0063)	0.7318(0.0084)	0.8187(0.0319)	0.8245(0.0312)	0.6989(0.0022)	0.7371(0.0011)
SOPA-s	0.7309(0.0151)	**0.7330(0.0073)**	**0.8449(0.0399)**	**0.8412(0.0447)**	**0.7027(0.0018)**	**0.7416(0.0006)**

Table 4. Two way partial AUC on testing data of three molecular datasets; (α,β) represents TPR$\geq$ α and FPR $\leq$ β.

	moltox21(t0)		molmuv(t1)		molpcba(t0)	
Methods	(0.6,0.4)	(0.5,0.5)	(0.6,0.4)	(0.5,0.5)	(0.6,0.4)	(0.5,0.5)
CE	0.0674(0.0014)	0.2082(0.0011)	0.1613(0.0337)	0.4691(0.0183)	0.0949(0.0006)	0.2639(0.0006)
AUC-SH	0.0640(0.0080)	0.2170(0.0140)	0.2600(0.1300)	0.4440(0.1280)	0.1400(0.0030)	0.3120(0.0030)
AUC-M	0.0660(0.0090)	0.2090(0.0100)	0.1140(0.0790)	0.4330(0.0530)	0.1420(0.0090)	0.3130(0.0030)
P-push	0.0610(0.0180)	0.2070(0.0120)	0.1860(0.1520)	0.4170(0.1080)	0.1350(0.0020)	0.3000(0.0120)
MB	0.0670(0.0150)	0.2150(0.0230)	0.1730(0.1530)	0.4260(0.1180)	0.0950(0.0020)	0.2620(0.0030)
AW-poly	0.0640(0.0100)	0.2060(0.0250)	0.1720(0.1440)	0.3930(0.1230)	0.1100(0.0010)	0.2810(0.0020)
SOTA-s	**0.0680(0.0180)**	**0.2300(0.0210)**	**0.3270(0.1640)**	**0.5260(0.1220)**	**0.1430(0.0010)**	**0.3140(0.0020)**

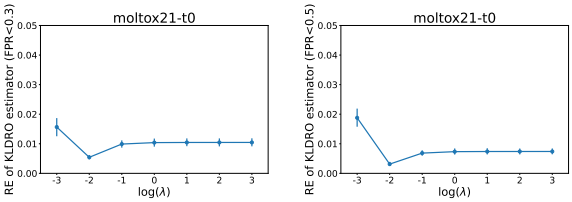

Figure 3. Relative error (RE) for KLDRO-based estimator for OPAUC on moltox21-t0 dataset with FPR=$\{0.3, 0.5\}$.

7. Conclusions

In this paper, we have proposed new formulations for partial AUC maximization by using distributionally robust optimization. We propose two formulations for both one-way and two-way partial AUC, and develop stochastic algorithms with convergence guarantee for solving the two formulations, respectively. Extensive experiments on image and molecular graph datasets verify the effectiveness of the proposed algorithms.

8. Acknowledgements

This work is partially supported by NSF Grant 2110545, NSF Career Award 1844403, and NSF Grant 1933212. D. Zhu and X. Wu was partially supported by NSF grant CCF-1733742. We also thank anonymous reviewers for constructive comments.

References

Agarwal, S. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SDM*, 2011.

Balasubramanian, K., Ghadimi, S., and Nguyen, A. Stochastic multi-level composition optimization algorithms with level-independent convergence rates. *ArXiv e-prints*, arXiv:2008.10526, 2021.

Chen, Z., Yuan, Z., Yi, J., Zhou, B., Chen, E., and Yang, T. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *CoRR*, /abs/1802.02988, 2018.

Dodd, L. and Pepe, M. Partial auc estimation and regression. *Biometrics*, 59:614–23, 10 2003. doi: 10.1111/1541-0420.00071.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, pp. 1–56, 2019.

Gao, W. and Zhou, Z.-H. On the consistency of auc pairwise optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 939–945. AAAI Press, 2015. ISBN 9781577357384.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. On stochastic moving-average estimators for non-convex optimization. *ArXiv e-prints*, arXiv:2104.14840, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Herbrich, R., Graepel, T., and Obermayer, K. Large Margin Rank Boundaries for Ordinal Regression. In *Advances in Large Margin Classifiers*, chapter 7, pp. 115–132. The MIT Press, 1999. URL http://www.herbrich.me/papers/nips98_ordinal.pdf.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Iwata, T., Fujino, A., and Ueda, N. Semi-supervised learning for maximizing the partial auc. In *AAAI*, 2020.

Kar, P., Narasimhan, H., and Jain, P. Online and stochastic gradient methods for non-decomposable loss functions. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 694–702, Cambridge, MA, USA, 2014. MIT Press.

Komori, O. and Eguchi, S. A boosting method for maximizing the partial area under the roc curve. *BMC Bioinformatics*, 11:314 – 314, 2010.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Li, N., Jin, R., and Zhou, Z.-H. Top rank optimization in linear time. In *NIPS*, 2014.

Liu, M., Yuan, Z., Ying, Y., and Yang, T. Stochastic AUC maximization with deep neural networks. In *8th International Conference on Learning Representations (ICLR)*, 2020.

Narasimhan, H. and Agarwal, S. Svmpauctight: 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*, KDD '13, pp. 167–175, New York, NY, USA, 2013a. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2487674. URL https://doi.org/10.1145/2487575.2487674.

Narasimhan, H. and Agarwal, S. A structural SVM based approach for optimizing partial auc. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 516–524, Atlanta, Georgia, USA, 17–19 Jun 2013b. PMLR. URL https://proceedings.mlr.press/v28/narasimhan13.html.

Narasimhan, H. and Agarwal, S. Support vector algorithms for optimizing the partial area under the roc curve. *Neural Computation*, 29:1919–1963, 2017.

Ogryczak, W. and Tamir, A. Minimizing the sum of the k largest functions in linear time. *Information Processing Letters*, 85:117–122, 02 2003. doi: 10.1016/S0020-0190(02)00370-8.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of area under precision-recall curve

for deep learning with provable convergence. In *Advances in neural information processing systems*, volume abs/2104.08736, 2021. URL https://arxiv.org/abs/2104.08736.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 2020.

Rakotomamonjy, A. Sparse support vector infinite push. In *ICML*, 2012.

Ricamato, M. T. and Tortorella, F. Partial auc maximization in a linear combination of dichotomizers. *Pattern Recognit.*, 44:2669–2677, 2011.

Rockafellar, R. T., Uryasev, S., et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

Rudin, C. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10(78):2233–2271, 2009. URL http://jmlr.org/papers/v10/rudin09b.html.

Takenouchi, T., Komori, O., and Eguchi, S. An extension of the receiver operating characteristic curve and auc-optimal classification. *Neural computation*, 24:2789–824, 06 2012. doi: 10.1162/NECO_a_00336.

Ueda, N. and Fujino, A. Partial auc maximization via nonlinear scoring functions. *ArXiv*, abs/1806.04838, 2018.

Wang, B. and Yang, T. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *Proceedings of the International Conference on Machine Learning*, pp. –, 2022.

Wang, Z. and Chang, Y.-C. I. Marker selection via maximizing the partial area under the roc curve of linear risk scores. *Biostatistics*, 12 2:369–85, 2011.

Wang, Z., Liu, M., Luo, Y., Xu, Z., Xie, Y., Wang, L., Cai, L., Qi, Q., Yuan, Z., Yang, T., and Ji, S. Advanced graph and sequence neural networks for molecular property prediction and drug discovery, 2021.

Wu, S.-H., Lin, K.-P., Chen, C.-M., and Chen, M.-S. Asymmetric support vector machines: Low false-positive learning under the user tolerance. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 749–757, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401980. URL https://doi.org/10.1145/1401890.1401980.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yang, H., Lu, K., Lyu, X., and Hu, F. Two-way partial auc and its properties. *Statistical methods in medical research*, 28(1):184–195, 2019.

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., and Huang, Q. When all we need is a piece of the pie: A generic framework for optimizing two-way partial auc. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11820–11829. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21k.html.

Yao, Y., Lin, Q., and Yang, T. Large-scale optimization of partial auc in a range of false positive rates. *arXiv preprint*, 2022.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *Interntional Conference on Computer Vision*, volume abs/2012.03173, 2020. URL https://arxiv.org/abs/2012.03173.

Yuan, Z., Guo, Z., Xu, Y., Ying, Y., and Yang, T. Federated deep AUC maximization for hetergeneous data with a constant communication complexity. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12219–12229. PMLR, 2021. URL http://proceedings.mlr.press/v139/yuan21a.html.

A. More Experimental Results

Table 5. Datasets Statistics. The percentage in parenthesis represents the proportion of positive samples.

Dataset	Train	Validation	Test
CIFAR-10	24000 (16.67%)	6000 (16.67%)	6000 (16.67%)
CIFAR-100	24000 (16.67%)	6000 (16.67%)	6000 (16.67%)
Melanoma	26500 (1.76%)	3313 (1.78%)	3313 (1.75%)
moltox21(t0)	5834 (4.25%)	722 (4.01%)	709 (4.51%)
molmuv(t1)	11466 (0.18%)	1559 (0.13%)	1709 (0.35%)
molpcba(t0)	120762 (9.32%)	19865 (11.74%)	20397 (11.61%)

A.1. Additional plots for training convergence

We present more training convergence plots on Melanoma dataset and molmuv dataset at Figure 4. For OPAUC maximization, We can observe that both our proposed SOPA-s and SOPA converge much better than AW-poly and MB method under different settings, i.e., FPR≤ 0.3 and FPR≤ 0.5. And our proposed SOTA-s converge higher by a noticeable margin than AW-poly and MB method for TPAUC maximization all the time.

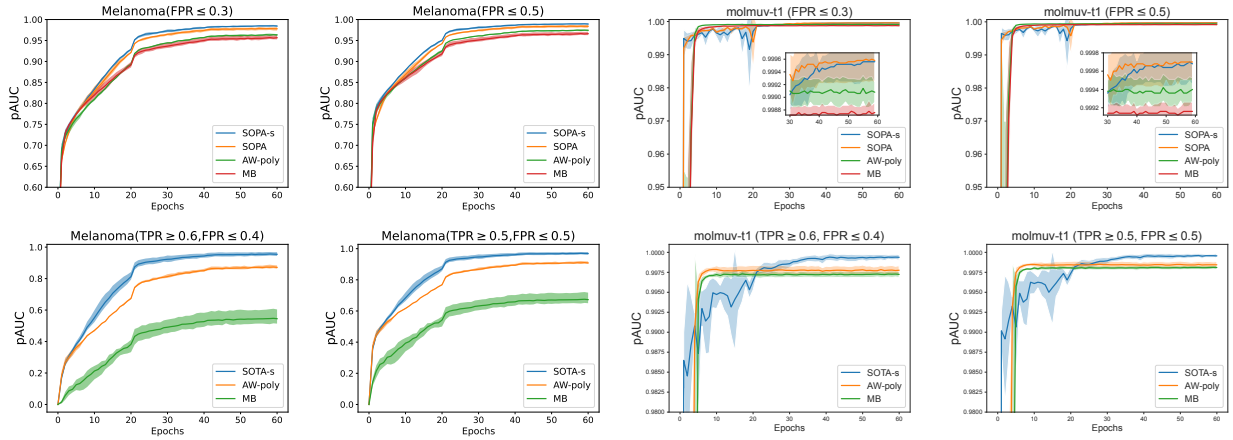

Figure 4. Training Convergence Curves on Melanoma and molmuv datasets; Top for OPAUC maximization, bottom for TPAUC maximization.

A.2. Ablation study for γ_0 in SOPA-s and γ_0, γ_1 in SOTA-s

We conduct extensive ablation study for understanding the extra hyper-parameters γ_0 in SOPA-s and γ_0, γ_1 in SOTA-s algorithms. We fix it as 0.9 for all of our experiments in the main content. But in practice, the performance would be further improved if we tune those hyper-parameters as well.

For image datasets, we conduct experiments on CIFAR-10 and CIFAR-100; for molecule datasets, we conduct experiments on ogbg-moltox21 and ogbg-molmuv. For each dataset, we first fix the best learning rate and other hyper-parameters based on our previous results in the paper. Then, for SOPA-s, we investigate γ_0 at $\{1.0, 0.9, 0.7, 0.5, 0.3, 0.1\}$; for SOTA-s, we investigate both γ_0 and γ_1 at $\{1.0, 0.9, 0.7, 0.5, 0.3, 0.1\}$

For training aspect, we include the comparisons for SOPA-s at Figure 5; we include the comparisons for SOTA-s at Figure 6. From Figure 5, we can see that better training performance could be achieved by tuning the parameter γ_0 in SOPA-s, compared with fixing it as 0.9; the similar result for SOTA-s can be also observed from Figure 6.

For testing aspect, we include the testing pAUC results at Table 6 for SOPA-s; Table 7 for SOTA-s. Both verify that tuning these parameters can further improve the performance.

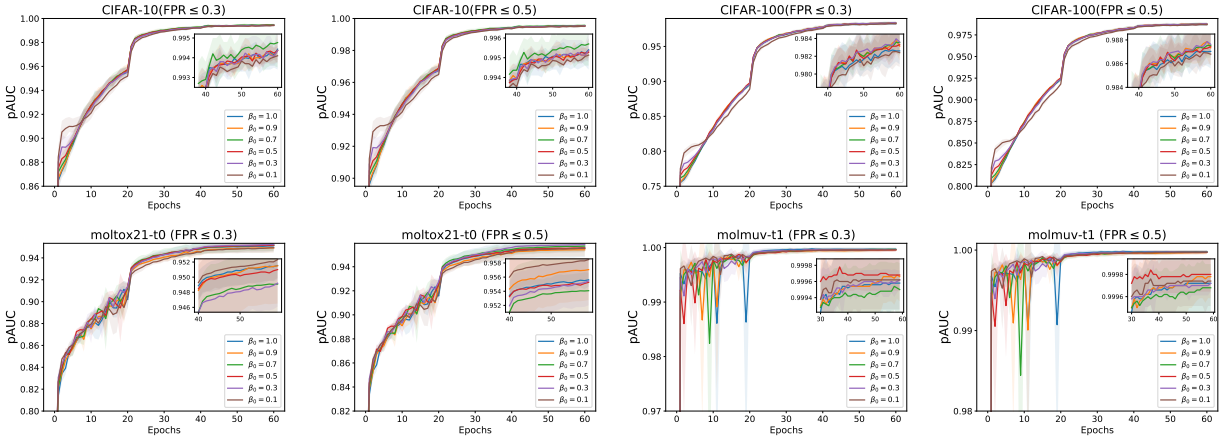

Figure 5. Training convergence for SOPA-s on different γ_0.

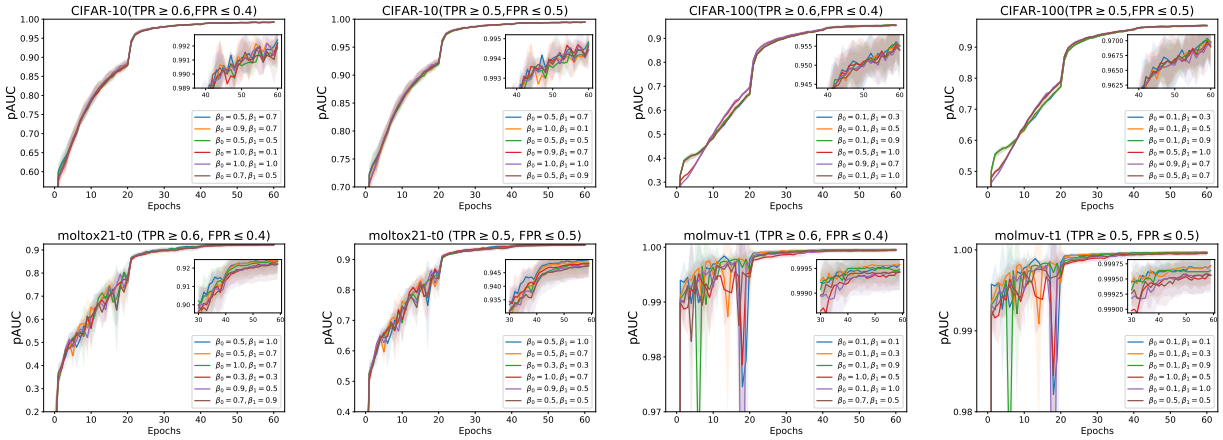

Figure 6. Top-6 choices of γ_0 and γ_1 for SOTA-s from training perspective.

Table 6. Test pAUC for SOPA-s on different γ_0.

Dataset	Metric	$\gamma_0 = 1.0$	$\gamma_0 = 0.9$	$\gamma_0 = 0.7$	$\gamma_0 = 0.5$	$\gamma_0 = 0.3$	$\gamma_0 = 0.1$
CIFAR-10	FPR≤0.3	0.8721(0.0049)	0.8691(0.0036)	0.8682(0.0048)	0.8697(0.0032)	0.8674(0.0045)	**0.8725(0.0012)**
	FPR≤0.5	0.8989(0.0051)	0.8961(0.0036)	0.8946(0.0040)	0.8980(0.0037)	0.8947(0.0028)	**0.8996(0.0021)**
CIFAR-100	FPR≤0.3	0.7464(0.0012)	0.7460(0.0068)	0.7482(0.0031)	0.7508(0.0038)	0.7494(0.0048)	**0.7514(0.0018)**
	FPR≤0.5	0.7888(0.0016)	0.7877(0.0053)	0.7936(0.0040)	0.7961(0.0063)	0.7922(0.0059)	**0.7954(0.0015)**
moltox21	FPR≤0.3	0.7242(0.0170)	0.7288(0.0125)	0.7274(0.0087)	0.7274(0.0062)	0.7158(0.0069)	**0.7340(0.0079)**
	FPR≤0.5	0.7245(0.0194)	0.7266(0.0111)	0.7280(0.0066)	0.7358(0.0079)	0.7249(0.0090)	**0.7360(0.0123)**
molmuv	FPR≤0.3	0.8692(0.0116)	0.8376(0.0340)	0.8642(0.0214)	**0.8735(0.0070)**	0.8732(0.0104)	0.8496(0.0392)
	FPR≤0.5	0.8773(0.0186)	0.8616(0.0355)	0.8747(0.0053)	**0.8996(0.0228)**	0.8918(0.0232)	0.8622(0.0415)

B. Proofs

We next present several lemmas. The first lemma is straightforward.

Lemma 4. *For $\hat{L}_{kl}(\cdot;\lambda)$, when $\lambda = 0$ it reduces to the maximal value of $\{\ell_1(\cdot),\ldots,\ell_n(\cdot)\}$, i.e., $\hat{L}_{kl}(\cdot,0) = \max_i \ell_i(\cdot)$; and when $\lambda = \infty$, it reduces to the average value of $\{\ell_1,\ldots,\ell_n\}$, i.e., $\hat{L}_{kl}(\cdot;\infty) = \frac{1}{n}\sum_{i=1}^n \ell_i(\cdot)$.*

Lemma 5. *[Lemma 1 (Ogryczak & Tamir, 2003)] Assume $\gamma = k/n$ for some integer $k \in [n]$, we have $\hat{L}_{cvar}(\cdot;\gamma) = \min_s s + \frac{1}{n\gamma}\sum_{i=1}^n [\ell_i(\cdot) - s]_+$.*

A major difference between $\hat{L}_{kl}(\cdot;\lambda)$ and $\hat{L}_{cvar}(\cdot;\gamma)$ that has an impact on optimization is that $\hat{L}_{cvar}(\cdot;\gamma)$ is a non-smooth function, and $\hat{L}_{kl}(\cdot;\lambda)$ is a smooth function for $\lambda > 0$ when $\ell_i(\cdot)$ is smooth and bounded as indicated by the following lemma.

Table 7. Test pAUC for SOTA-s on different γ_0 and γ_1.

CIFAR-10	(0.4, 0.6)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.5731(0.0069)	0.5744(0.0083)	0.5668(0.0048)	0.5686(0.0089)	0.5678(0.0070)	0.5739(0.0135)
	$\gamma_0 = 0.9$	0.5802(0.0081)	**0.5839(0.0116)**	0.5693(0.0072)	0.5743(0.0131)	0.5688(0.0048)	0.5760(0.0095)
	$\gamma_0 = 0.7$	0.5755(0.0098)	0.5715(0.0032)	0.5684(0.0131)	0.5797(0.0036)	0.5718(0.0095)	0.5695(0.0111)
	$\gamma_0 = 0.5$	0.5743(0.0058)	0.5706(0.0025)	0.5641(0.0094)	0.5819(0.0077)	0.5725(0.0109)	0.5739(0.0068)
	$\gamma_0 = 0.3$	0.5767(0.0121)	0.5638(0.0144)	0.5617(0.0075)	0.5644(0.0099)	0.5595(0.0028)	0.5768(0.0100)
	$\gamma_0 = 0.1$	0.5608(0.0075)	0.5689(0.0134)	0.5699(0.0096)	0.5685(0.0056)	0.5583(0.0101)	0.5756(0.0135)
CIFAR-10	(0.5, 0.5)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.7022(0.0054)	0.6999(0.0076)	0.6922(0.0076)	0.6969(0.0087)	0.6982(0.0046)	0.6981(0.0074)
	$\gamma_0 = 0.9$	**0.7051(0.0083)**	0.7047(0.0072)	0.6987(0.0034)	0.7008(0.0103)	0.6964(0.0048)	0.7003(0.0064)
	$\gamma_0 = 0.7$	0.7027(0.0046)	0.6999(0.0031)	0.6959(0.0099)	0.7043(0.0056)	0.6988(0.0099)	0.6914(0.0110)
	$\gamma_0 = 0.5$	0.7012(0.0050)	0.6988(0.0008)	0.6944(0.0067)	0.7044(0.0067)	0.7014(0.0047)	0.7031(0.0088)
	$\gamma_0 = 0.3$	0.7024(0.0099)	0.6940(0.0085)	0.6901(0.0060)	0.6981(0.0047)	0.6908(0.0025)	0.6977(0.0129)
	$\gamma_0 = 0.1$	0.6934(0.0068)	0.7009(0.0060)	0.6972(0.0088)	0.6946(0.0037)	0.6891(0.0098)	0.7027(0.0082)
CIFAR-100	(0.4, 0.6)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.2710(0.0144)	0.2650(0.0052)	0.2528(0.0061)	0.2573(0.0129)	0.2587(0.0096)	0.2613(0.0046)
	$\gamma_0 = 0.9$	0.2666(0.0144)	0.2708(0.0055)	0.2698(0.0117)	0.2586(0.0116)	0.2562(0.0032)	0.2663(0.0056)
	$\gamma_0 = 0.7$	0.2634(0.0086)	0.2624(0.0065)	0.2577(0.0045)	0.2664(0.0051)	0.2634(0.0090)	0.2626(0.0144)
	$\gamma_0 = 0.5$	0.2633(0.0010)	0.2591(0.0026)	0.2552(0.0068)	0.2530(0.0069)	0.2656(0.0089)	0.2594(0.0052)
	$\gamma_0 = 0.3$	0.2572(0.0063)	0.2542(0.0076)	0.2517(0.0111)	0.2599(0.0184)	0.2580(0.0191)	0.2631(0.0081)
	$\gamma_0 = 0.1$	0.2532(0.0050)	**0.2745(0.0029)**	0.2529(0.0060)	0.2573(0.0115)	0.2568(0.0029)	0.2671(0.0187)
CIFAR-100	(0.5, 0.5)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.4489(0.0122)	0.4454(0.0084)	0.4337(0.0097)	0.4418(0.0083)	0.4361(0.0093)	0.4405(0.0063)
	$\gamma_0 = 0.9$	0.4416(0.0115)	**0.4528(0.0069)**	0.4494(0.0073)	0.4449(0.0171)	0.4337(0.0046)	0.4421(0.0058)
	$\gamma_0 = 0.7$	0.4426(0.0119)	0.4426(0.0098)	0.4367(0.0092)	0.4414(0.0072)	0.4455(0.0046)	0.4430(0.0108)
	$\gamma_0 = 0.5$	0.4404(0.0058)	0.4388(0.0085)	0.4341(0.0066)	0.4332(0.0079)	0.4426(0.0107)	0.4421(0.0057)
	$\gamma_0 = 0.3$	0.4384(0.0090)	0.4313(0.0081)	0.4377(0.0104)	0.4392(0.0185)	0.4401(0.0095)	0.4384(0.0125)
	$\gamma_0 = 0.1$	0.4325(0.0006)	0.4500(0.0015)	0.4361(0.0085)	0.4316(0.0083)	0.4355(0.0053)	0.4461(0.0213)
moltox21	(0.4, 0.6)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.0769(0.0403)	0.0800(0.0345)	0.0594(0.0377)	0.0409(0.0241)	0.0774(0.0342)	0.0607(0.0139)
	$\gamma_0 = 0.9$	0.0668(0.0304)	0.0733(0.0198)	0.0676(0.0357)	0.0761(0.0316)	0.0737(0.0245)	0.0588(0.0249)
	$\gamma_0 = 0.7$	0.0657(0.0280)	0.0849(0.0451)	0.0840(0.0162)	0.0807(0.0253)	0.0665(0.0171)	0.0573(0.0262)
	$\gamma_0 = 0.5$	0.0795(0.0537)	0.0513(0.0140)	0.0704(0.0301)	0.0367(0.0164)	0.0754(0.0190)	0.0676(0.0350)
	$\gamma_0 = 0.3$	**0.1009(0.0133)**	0.0695(0.0256)	0.0806(0.0289)	0.0805(0.0358)	0.0943(0.0453)	0.0546(0.0409)
	$\gamma_0 = 0.1$	0.0746(0.0473)	0.0610(0.0224)	0.0694(0.0303)	0.0578(0.0458)	0.0768(0.0574)	0.0485(0.0254)
moltox21	(0.5, 0.5)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.2474(0.0154)	0.2489(0.0282)	0.2355(0.0508)	0.2127(0.0326)	**0.2798(0.0640)**	0.2254(0.0223)
	$\gamma_0 = 0.9$	0.2483(0.0214)	0.2476(0.0222)	0.2373(0.0552)	0.2457(0.0296)	0.2341(0.0137)	0.2198(0.0233)
	$\gamma_0 = 0.7$	0.2306(0.0286)	0.2354(0.0360)	0.2531(0.0217)	0.2546(0.0174)	0.2415(0.0181)	0.2233(0.0225)
	$\gamma_0 = 0.5$	0.2536(0.0531)	0.2251(0.0163)	0.2276(0.0241)	0.2153(0.0196)	0.2488(0.0121)	0.2230(0.0228)
	$\gamma_0 = 0.3$	0.2773(0.0207)	0.2482(0.0256)	0.2434(0.0224)	0.2603(0.0323)	0.2574(0.0455)	0.2208(0.0494)
	$\gamma_0 = 0.1$	0.2368(0.0193)	0.2395(0.0267)	0.2323(0.0239)	0.1973(0.0383)	0.2543(0.0611)	0.2121(0.0369)
molmuv	(0.4, 0.6)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.5021(0.1222)	0.4508(0.1263)	0.3878(0.0776)	0.5399(0.1575)	0.3972(0.0768)	0.3904(0.1404)
	$\gamma_0 = 0.9$	0.5651(0.0833)	0.4524(0.0290)	**0.6060(0.2187)**	0.5711(0.1495)	0.4250(0.0543)	0.4313(0.1174)
	$\gamma_0 = 0.7$	0.4120(0.1026)	0.4338(0.0407)	0.5375(0.1339)	0.5765(0.1719)	0.3251(0.0889)	0.4744(0.1379)
	$\gamma_0 = 0.5$	0.5432(0.1177)	0.4934(0.1065)	0.4812(0.0744)	0.3890(0.0976)	0.4507(0.0387)	0.3859(0.0488)
	$\gamma_0 = 0.3$	0.4680(0.1079)	0.4559(0.1037)	0.4811(0.1237)	0.4221(0.1322)	0.4297(0.2263)	0.3828(0.0606)
	$\gamma_0 = 0.1$	0.4777(0.0591)	0.4264(0.1414)	0.5322(0.1342)	0.4122(0.0260)	0.3836(0.0368)	0.3903(0.0325)
molmuv	(0.5, 0.5)	$\gamma_1 = 1.0$	$\gamma_1 = 0.9$	$\gamma_1 = 0.7$	$\gamma_1 = 0.5$	$\gamma_1 = 0.3$	$\gamma_1 = 0.1$
	$\gamma_0 = 1.0$	0.6780(0.0880)	0.6384(0.0964)	0.6175(0.0447)	0.7135(0.1074)	0.6188(0.0483)	0.6152(0.0965)
	$\gamma_0 = 0.9$	0.7405(0.0621)	0.6399(0.0276)	**0.7481(0.1464)**	0.7230(0.1020)	0.6136(0.0321)	0.6277(0.0854)
	$\gamma_0 = 0.7$	0.6201(0.0698)	0.6445(0.0430)	0.7007(0.1013)	0.7272(0.1178)	0.5588(0.0482)	0.6496(0.1030)
	$\gamma_0 = 0.5$	0.7071(0.0869)	0.6854(0.0853)	0.6542(0.0646)	0.5867(0.0657)	0.6495(0.0470)	0.5996(0.0431)
	$\gamma_0 = 0.3$	0.6626(0.0737)	0.6561(0.0778)	0.6658(0.0910)	0.6408(0.0825)	0.6264(0.1791)	0.5801(0.0399)
	$\gamma_0 = 0.1$	0.6643(0.0562)	0.6362(0.1103)	0.7142(0.0904)	0.6271(0.0188)	0.5746(0.0283)	0.5877(0.0449)

Lemma 6. *If $\ell_i(\cdot)$ is a smooth function, and has a bounded value and bounded gradient for a bounded input, then for $\lambda > 0$ the function $\hat{L}_{kl}(\cdot;\lambda)$ is also a smooth and bounded function.*

Lemma 7. *When $\phi_c(t) = \mathbb{I}(0 \le t \le 1/\beta)$ and $\phi_c'(t) = \mathbb{I}(0 \le t \le 1/\alpha)$ with $K_2 = n_-\beta$ and $K_1 = n_+\alpha$ being integers, if $\ell(\cdot)$ is monotonically decreasing for $\ell(\cdot) > 0$, we can show that*

$$F(\mathbf{w};\phi_c,\phi_c') = \frac{1}{K_1K_2}\sum_{\mathbf{x}_i\in\mathcal{S}_+^{\uparrow}[1,K_1]}\sum_{\mathbf{x}_j\in\mathcal{S}_-^{\downarrow}[1,K_2]} L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j),$$

which is also equivalent to

$$F(\mathbf{w};\phi_c,\phi_c') = \min_{s'\in\mathbb{R},\mathbf{s}\in\mathbb{R}^{n_+}} s' + \frac{1}{n_+\alpha}\sum_{\mathbf{x}_i\in\mathcal{S}_+}(s_i + \frac{1}{\beta}\psi_i(\mathbf{w};s_i) - s')_+.$$

B.1. Proof of Lemma 6

It is easy to see if $\ell(\mathbf{w}) \in [0, C]$ is bounded and smooth, we have $\exp(\frac{\ell(\mathbf{w})}{\lambda})$ is bounded and smooth due to its second order gradient is upper bounded. Then $\log \mathrm{E}_i \exp(\frac{\ell_i(\mathbf{w})}{\lambda})$ is bounded. Its smoothness due to that is a composition of $f = \log(\cdot)$ and $g = \mathrm{E}_i \exp(\frac{\ell_i(\mathbf{w})}{\lambda}) \in [1, C']$ and both f, g are smooth and Lipschitz continuous for their inputs.

B.2. Proof of Lemma 7

Proof. First, following Lemma 1, we have $F(\mathbf{w};\phi,\phi')$ is equivalent to $\frac{1}{K_1}\sum_{i=1}^{K_1}\hat{L}_\phi(\mathbf{x}_{\pi_i},\mathbf{w})$, where π_i denote the index of the positive example whose $\hat{L}_\phi(\mathbf{x}_{\pi_i},\mathbf{w})$ is the i-th largest among all positive examples. We prove that this is equivalent to $\frac{1}{K_1}\sum_{\mathbf{x}_i\in\mathcal{S}_+^{\uparrow}[1,K_1]}\hat{L}_\phi(\mathbf{x}_i,\mathbf{w})$. To this end, we just need to show that if $h_\mathbf{w}(\mathbf{x}) \ge h_\mathbf{w}(\mathbf{x}')$ then $\hat{L}_\phi(\mathbf{x}',\mathbf{w}) \ge \hat{L}_\phi(\mathbf{x},\mathbf{w})$, which is true due to $\hat{L}_\phi(\mathbf{x}_i,\mathbf{w}) = \frac{1}{K_2}\sum_{\mathbf{x}_j\in\mathcal{S}_-^{\downarrow}[1,K_2]}\ell(h_\mathbf{w}(\mathbf{x}_i) - h_\mathbf{w}(\mathbf{x}_j))$ and ℓ is monotonically decreasing function. The second equation in the lemma is applying Lemma 5 twice and by noting that $(\min_x f(x) - s)_+ = \min_x(f(x) - s)_+$. □

B.3. Proof of Theorem 1

Proof. Let us consider for a particular $\mathbf{x}_i \in \mathcal{S}_+$. When $\phi(\cdot) = \phi_c(\cdot) = \mathbb{I}(\cdot \in (0, 1/\beta])$, then $\hat{L}(\mathbf{w};\mathbf{x}_i)$ becomes the CVaR estimator, i.e., the average of top $K = n_-\beta$ losses of $\ell(h_\mathbf{w}(\mathbf{x}_i) - h_\mathbf{w}(\mathbf{x}_j))$ among $\mathbf{x}_j \in \mathcal{S}_-$. Since $\ell(\cdot)$ is monotonically decreasing when $\ell(\cdot)) > 0$, the top $K = n_-\beta$ losses of $\ell(h_\mathbf{w}(\mathbf{x}_i) - h_\mathbf{w}(\mathbf{x}_j))$ among all $\mathbf{x}_j \in \mathcal{S}_-$ correspond to negative samples with top K prediction scores. Hence, $\hat{L}_{\phi_c}(\mathbf{w};\mathbf{x}_i) = \frac{1}{K}\sum_{\mathbf{x}_j\in\mathcal{S}_-^{\downarrow}[1,K]}\ell(h_\mathbf{w}(\mathbf{x}_i) - h_\mathbf{w}(\mathbf{x}_j))$. Then the equivalent problem in (9) follows from Lemma 5. □

B.4. Proof of Theorem 2

Proof. When choosing $\phi(\cdot) = \phi_{kl}(\cdot)$, the $\hat{L}_\phi(\mathbf{w};\mathbf{x}_i)$ in problem (8) becomes

$$\max_{\mathbf{p}\in\Delta}\sum_{\mathbf{x}_j\in\mathcal{S}_-} p_j L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - \lambda\sum_j p_j\log(np_j).$$

With Karush-Kuhn-Tucker(KKT) conditions, it is not difficult to have $p_j^\star = \frac{\exp(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)/\lambda}{\sum_j\exp(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)/\lambda}$. By plugging in this back we obtain the claimed objective (10). When $\lambda = 0$, the above becomes the maximal one among $\{L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j), \mathbf{x}_j \in \mathcal{S}_+\}$ for each $\mathbf{x}_i$. Then the object is $\frac{1}{n_+}\max_{\mathbf{x}_j\in\mathcal{S}_-} L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)$, which is the surrogate of pAUC with FPR$\le 1/n_-$. When $\lambda = \infty$, the above becomes the average of $\{L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j), \mathbf{x}_j \in \mathcal{S}_+\}$, which gives the standard surrogate of AUC. □

B.5. Proof of Lemma 2

First note that $F(\mathbf{w},\mathbf{s}) = \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\Big(s_i + \frac{1}{\beta}g_i(\mathbf{w},s_i)\Big)$, and $g_i(\mathbf{w},s_i) = \frac{1}{n_-}\sum_{\mathbf{x}_j\in\mathcal{S}_-}(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - s_i)_+$. We prove that $(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - s_i)_+$ is weakly convex in terms of $(\mathbf{w},\mathbf{s}_i)$, i.e. there exists $\rho > 0$ such that $(L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - s_i)_+ + \frac{\rho}{2}\|\mathbf{w}\|^2 + \frac{\rho}{2}s_i^2$ is jointly convex in terms of $\mathbf{w},\mathbf{s}$. To this end, let $\psi(\cdot) = [\cdot]_+$ which is convex and Lipchitz continuous, and $q(\mathbf{w},\mathbf{s}_i) = L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j) - s_i$, which is L_s-smooth function with respect to $(\mathbf{w},s_i)$ due to that $L(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j)$ is L_s-smooth

function. Then for any $\omega \in \phi'(\psi(\mathbf{w}', s_i'))$ we have

$$\begin{aligned}
&\psi(q(\mathbf{w}, s_i)) \geq \psi(q(\mathbf{w}', s_i')) + \omega(q(\mathbf{w}, s_i) - q(\mathbf{w}', s_i')) \\
&\geq \psi(q(\mathbf{w}', s_i')) + \omega(\nabla q(\mathbf{w}', s_i') - \frac{L_s}{2}(\|\mathbf{w} - \mathbf{w}'\|^2 + |s_i - s_i'|^2)) \\
&\geq \psi(q(\mathbf{w}', s_i')) + \partial\psi(q(\mathbf{w}', s_i')) - \frac{L_s}{2}(\|\mathbf{w} - \mathbf{w}'\|^2 + |s_i - s_i'|^2)
\end{aligned}$$

where we use $0 \leq \omega \leq 1$. The above inequality implies that $[L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - s_i]_+$ is L_s-weakly convex in terms of $(\mathbf{w}, s_i)$ (Davis & Drusvyatskiy, 2018), i.e., $\frac{1}{n_-}\sum_{\mathbf{x}_j \in \mathcal{S}_-}\left\{(L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - s_i)_+ + \frac{L_s}{2}(\|\mathbf{w}\|^2 + |s_i|^2)\right\}$ is convex. As a result $\frac{1}{n_-}\sum_{\mathbf{x}_j \in \mathcal{S}_-}(L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j) - s_i)_+ + \frac{L_s}{2}(\|\mathbf{w}\|^2 + \|s_i\|^2)$ is jointly convex in $(\mathbf{w}, s_i)$. Then $F(\mathbf{w}, \mathbf{s}) + \frac{L_s}{2\beta}(\|\mathbf{w}\|^2 + \sum_i |s_i|^2)$ is jointly convex in terms of $(\mathbf{w}, \mathbf{s})$.

B.6. Proof of Theorem 3

Let $F(\mathbf{w}, \mathbf{s})$ denote the objective function, and let $\mathbf{v} = (\mathbf{w}, \mathbf{s})$. Define $F_{1/\hat{\rho}}(\mathbf{v}) = \min_{\mathbf{u}} F(\mathbf{u}) + \frac{\hat{\rho}}{2}\|\mathbf{u} - \mathbf{v}\|^2$ for some $\hat{\rho} > \rho$ and the minimizer is denoted by $\text{prox}_{F/\hat{\rho}}(\mathbf{v})$. Let $\widehat{\mathbf{v}}_t = \text{prox}_{F/\hat{\rho}}(\mathbf{v}_t)$. Define $\|\mathbf{v} - \mathbf{v}'\|^2 = \|\mathbf{w} - \mathbf{w}'\|^2 + \|\mathbf{s} - \mathbf{s}'\|^2$.

$$\begin{aligned}
&\mathrm{E}_t[F_{1/\hat{\rho}}(\mathbf{v}_{t+1})] = \mathrm{E}_t[F(\widehat{\mathbf{v}}_t) + \frac{\hat{\rho}}{2}\|\mathbf{v}_{t+1} - \widehat{\mathbf{v}}_t\|^2] \\
&\leq F(\widehat{\mathbf{v}}_t) + \frac{\hat{\rho}}{2}\mathrm{E}_t[\|\mathbf{w}_t - \eta_1\nabla_{\mathbf{w}}F(\mathbf{w}_t, \mathbf{s}_t, \xi_t) - \widehat{\mathbf{w}}_t\|^2 + \|\mathbf{s}_{t+1} - \hat{\mathbf{s}}_t\|^2] \\
&\leq F(\widehat{\mathbf{v}}_t) + \frac{\hat{\rho}}{2}\mathrm{E}_t[\|\mathbf{w}_t - \eta_1\nabla_{\mathbf{w}}F(\mathbf{w}_t, \mathbf{s}_t, \xi_t) - \widehat{\mathbf{w}}_t\|^2] + \frac{\hat{\rho}}{2}\mathrm{E}_t[\|\mathbf{s}_{t+1} - \hat{\mathbf{s}}_t\|^2] \\
&\leq F(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|^2 + \hat{\rho}\eta_t\mathrm{E}_t[(\widehat{\mathbf{w}}_t - \mathbf{w}_t)^\top\nabla_{\mathbf{w}}F(\mathbf{w}_t, \mathbf{s}_t)] + \frac{\hat{\rho}\eta_1^2G^2}{2} + \frac{\hat{\rho}}{2}\mathrm{E}_t[\|\mathbf{s}_{t+1} - \hat{\mathbf{s}}_t\|^2]
\end{aligned}$$

where we assume $\mathrm{E}[\|\nabla_{\mathbf{w}}F(\mathbf{w}_t, \mathbf{s}_t, \xi_t)\|^2] \leq G^2 = \frac{C^2}{\beta^2}$. According to the analysis of stochastic coordinate descent method, for any $\mathbf{s} = (s_1, \ldots, s_{n_+})$ we have

$$2\eta_2(s_{t,i} - s_i)^\top\nabla_{s_i}F(\mathbf{w}_t, \mathbf{s}_t; \xi_t) \leq \eta_2^2\|\nabla_{s_i}F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + (\|s_i - s_{t,i}\|^2 - \|s_i - s_{t+1,i}\|^2)$$

Summing the above inequality over $i \in \mathcal{B}_+$, we have

$$2\eta_2\sum_{i\in\mathcal{B}_+}(s_{t,i} - s_i)^\top\nabla_{s_i}F(\mathbf{w}_t, \mathbf{s}_t; \xi_t) \leq \eta_2^2\sum_{i\in\mathcal{B}_+}\|\nabla_{s_i}F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + (\|\mathbf{s} - \mathbf{s}_t\|^2 - \|\mathbf{s} - \mathbf{s}_{t+1}\|^2)$$

Taking expectation and re-arrange, we have

$$\mathrm{E}[\frac{1}{2}\|\mathbf{s}_{t+1} - \widehat{\mathbf{s}}_t\|^2] \leq \mathrm{E}[\frac{1}{2}\|\mathbf{s}_t - \widehat{\mathbf{s}}_t\|^2 + \eta_2\frac{B_+}{n_+}(\widehat{\mathbf{s}}_t - \mathbf{s}_t)^\top\nabla_{\mathbf{s}}F(\mathbf{w}_t, \mathbf{s}_t) + \frac{\eta_2^2B_+C_2^2}{2}]$$

where we use the fact $\mathrm{E}[|\nabla_{s_i}F(\mathbf{w}, \mathbf{s})|^2] \leq C_2^2 = \frac{1}{n_+^2}(1 + 1/\beta)^2$. Let $\eta_2B_+/n_+ = \eta_1$, we have

$$\begin{aligned}
&\mathrm{E}_t[F_{1/\hat{\rho}}(\mathbf{v}_{t+1})] \\
&\leq F(\widehat{\mathbf{x}}_t) + \frac{\hat{\rho}}{2}\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|^2 + \hat{\rho}\eta_1\mathrm{E}_t[(\widehat{\mathbf{w}}_t - \mathbf{w}_t)^\top\nabla_{\mathbf{w}}F(\mathbf{w}_t, \mathbf{s}_t)] + \frac{\hat{\rho}}{2}\mathrm{E}_t[\|\mathbf{s}_{t+1} - \widehat{\mathbf{s}}_t)] + \frac{\hat{\rho}\eta_1^2G^2}{2} \\
&\leq F(\widehat{\mathbf{v}}_t) + \frac{\hat{\rho}}{2}\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2 + \hat{\rho}\eta_1\mathrm{E}_t[(\widehat{\mathbf{v}}_t - \mathbf{v}_t)^\top\partial F(\mathbf{w}_t, \mathbf{s}_t)] + \frac{\hat{\rho}\eta_1^2(G^2 + n_+^2C_2^2/B_+)}{2} \\
&\leq F_{1/\hat{\rho}}(\mathbf{v}_t) + \hat{\rho}\eta_1(F(\widehat{\mathbf{v}}_t) - F(\mathbf{v}_t) + \frac{\rho}{2}\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2) + \frac{\hat{\rho}\eta_1^2(G^2 + n_+^2C_2^2/B_+)}{2}
\end{aligned}$$

As a result, we have

$$\hat{\rho}\eta_1(F(\mathbf{v}_t) - F(\widehat{\mathbf{v}}_t) - \rho\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2) \leq F_{1/\hat{\rho}}(\mathbf{v}_t) - \mathrm{E}_t[F_{1/\hat{\rho}}(\mathbf{v}_{t+1})] + \frac{\hat{\rho}\eta_1^2(G^2 + n_+^2 C_2^2/B_+)}{2}$$

Since we have

$$F(\mathbf{v}_t) - F(\widehat{\mathbf{v}}_t) - \rho\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2 = (F(\mathbf{v}_t) + \hat{\rho}\|\mathbf{v}_t - \mathbf{v}_t\|^2) - (F(\widehat{\mathbf{v}}_t) + \hat{\rho}\|\widehat{\mathbf{v}}_t - \mathbf{v}_t\|^2) + (\hat{\rho} - \rho)\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2$$
$$\geq \frac{(2\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{v}}_t - \mathbf{v}_t\|^2 + (\hat{\rho} - \rho)\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2 = (2\hat{\rho} - 3/2\rho)\|\mathbf{v}_t - \widehat{\mathbf{v}}_t\|^2 = \frac{(2\hat{\rho} - 3/2\rho)}{\hat{\rho}^2}\|\nabla F_{1/\hat{\rho}}(\mathbf{v}_t)\|^2$$

Let $\hat{\rho} = 3\rho/2$. As a result, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\hat{\rho}}(\mathbf{v}_t)\|^2 &\leq \frac{(F_{1/\hat{\rho}}(\mathbf{v}_1) - \min F)}{\eta_1 T} + \frac{\hat{\rho}\eta_1(G^2 + n_+^2 C_2^2/B_+)}{2} \\ &\leq \frac{(F_{1/\hat{\rho}}(\mathbf{v}_1) - \min F)}{\eta_1 T} + \frac{\hat{\rho}\eta_1(C^2/\beta^2 + (1 + 1/\beta)^2/B_+)}{2}\end{aligned}$$

By setting $\eta_1 = O(\beta\epsilon^2)$ and $T = O(\frac{1}{\epsilon^2\eta_1}) = O(1/(\beta\epsilon^4))$ we have $\mathrm{E}[\|\nabla F_{1/\hat{\rho}}(\mathbf{v}_\tau)\|^2] \leq \epsilon^2$ for a randomly selected $\tau \in [T]$.

B.7. Proof of Theorem 4

Note that the SOPA-s algorithm is just a special case of the SOX algorithm (Wang & Yang, 2022) for the more general problem $\min_{\mathbf{w}} \frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g_i(\mathbf{w}))$ and the convergence proof just follows the proof of Theorem 1 in Wang & Yang (2022).

B.8. Proof of Theorem 5

We consider the following problem:

$$\min_{\mathbf{w}} F(\mathbf{w}), \quad F(\mathbf{w}) = f_1\left(\frac{1}{n}\sum_{i \in \mathcal{S}} f_2(g_i(\mathbf{w}))\right). \tag{12}$$

Lemma 8. *If g_i is C_g-Lipschitz, L_g-smooth and f_1, f_2 are C_f-Lipschitz, L_g-smooth, F in (12) is L_F-smooth and $L_F = L_f C_f^2 C_g^2 + C_f^2 L_g + C_f C_g L_f$.*

Proof. Based on the definition of F, we have

$$\begin{aligned}&\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| \\ &= \left\|\nabla f_1\left(\frac{1}{n}\sum_{i \in \mathcal{S}} f_2(g_i(\mathbf{w}))\right)\left(\frac{1}{n}\sum_{i \in \mathcal{S}} \nabla f_2(g_i(\mathbf{w}))\nabla g_i(\mathbf{w})\right)\right. \\ &\qquad \left. -\nabla f_1\left(\frac{1}{n}\sum_{i \in \mathcal{S}} f_2(g_i(\mathbf{w}'))\right)\left(\frac{1}{n}\sum_{i \in \mathcal{S}} \nabla f_2(g_i(\mathbf{w}'))\nabla g_i(\mathbf{w}')\right)\right\| \\ &\leq L_f\left\|\frac{1}{n}\sum_{i \in \mathcal{S}} \nabla f_2(g_i(\mathbf{w}))\nabla g_i(\mathbf{w})\right\|\left\|\frac{1}{n}\sum_{i \in \mathcal{S}}(f_2(g_i(\mathbf{w})) - f_2(g_i(\mathbf{w}')))\right\| \\ &\qquad + \left\|\nabla f_1\left(\frac{1}{n}\sum_{i \in \mathcal{S}} f_2(g_i(\mathbf{w}'))\right)\right\|\left\|\frac{1}{n}\sum_{i \in \mathcal{S}}(\nabla f_2(g_i(\mathbf{w}))\nabla g_i(\mathbf{w}) - \nabla f_2(g_i(\mathbf{w}'))\nabla g_i(\mathbf{w}'))\right\|.\end{aligned}$$

We can show that $\left\|\frac{1}{n}\sum_{i \in \mathcal{S}}(\nabla f_2(g_i(\mathbf{w}))\nabla g_i(\mathbf{w}) - \nabla f_2(g_i(\mathbf{w}'))\nabla g_i(\mathbf{w}'))\right\| \leq (C_f L_g + C_g L_f)\|\mathbf{w} - \mathbf{w}'\|$. Thus,

$$\begin{aligned}\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\| &\leq L_f C_f^2 C_g^2\|\mathbf{w} - \mathbf{w}'\| + C_f(C_f L_g + C_g L_f)\|\mathbf{w} - \mathbf{w}'\| \\ &= (L_f C_f^2 C_g^2 + C_f^2 L_g + C_f C_g L_f)\|\mathbf{w} - \mathbf{w}'\|.\end{aligned}$$

□

We propose SOTA-s to solve (12).

Lemma 9. *Consider a sequence $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{m}_t$ and the L_F-smooth function F and the step size $\eta L_F \leq 1/2$.*

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \frac{\eta}{2}\left\|\Delta_t\right\|^2 - \frac{\eta}{2}\left\|\nabla F(\mathbf{w}_t)\right\|^2 - \frac{\eta}{4}\left\|\mathbf{m}_t\right\|^2, \tag{13}$$

where $\Delta_t \coloneqq \mathbf{m}_t - \nabla F(\mathbf{w}_t)$

Lemma 10. *For the gradient estimator $\mathbf{m}_t$ in SOTA-s and $\Delta_t = \mathbf{m}_t - \nabla F(\mathbf{w}_t)$,*

$$\begin{aligned}\mathrm{E}\left[\left\|\Delta_{t+1}\right\|^2\right] \leq (1-\gamma_2)\mathrm{E}\left[\left\|\Delta_t\right\|^2\right] + \frac{2L_F^2\eta^2}{\gamma_2}\mathrm{E}\left[\left\|\mathbf{m}_t\right\|^2\right] + 10\gamma_2 C_f^2 C_1^2 L_f^2 \mathrm{E}\left[\left\|\Psi_{t+1}\right\|^2\right] \\ + 20\gamma_2 C_f^2 L_f^2 C_1^2 \mathrm{E}\left[\frac{1}{n}\left\|\Xi_{t+1}\right\|^2\right] + 20\gamma_2 C_f^2 L_f^2 C_1^2 \mathrm{E}\left[\frac{1}{n}\left\|\mathbf{u}_{t+1} - \mathbf{u}_t\right\|^2\right] + 2\gamma_2^2 C_f^4\left(\frac{\zeta^2}{B_-} + \frac{C_g^2}{B_+}\right),\end{aligned} \tag{14}$$

where we denote $\Delta_t \coloneqq \mathbf{m}_{t+1} - \nabla F(\mathbf{w}_t)$, $\Xi_t \coloneqq \mathbf{u}_{t+1} - \mathbf{g}(\mathbf{w}_t)$ and $\Psi_t \coloneqq v_{t+1} - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_t))$.

Proof. Based on the update rule $\mathbf{m}_{t+1} = (1-\gamma_2)\mathbf{m}_t + \gamma_2 G(\mathbf{w}_{t+1})$, we have

$$\begin{aligned}&\left\|\mathbf{m}_{t+1} - \nabla F(\mathbf{w}_{t+1})\right\|^2 \\ &= \left\|(1-\gamma_2)\mathbf{m}_t + \gamma_2\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1(v_{t+1})\nabla f_2(u_t^i)\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) - \nabla F(\mathbf{w}_{t+1})\right\|^2 \\ &= \Bigg\|\underbrace{(1-\gamma_2)(\mathbf{m}_t - \nabla F(\mathbf{w}_t))}_{ⓐ} + \underbrace{(1-\gamma_2)(\nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t+1}))}_{ⓑ} + \gamma_2\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1(v_{t+1})\nabla f_2(u_t^i)\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) \\ &\quad - \gamma_2\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) \\ &\quad + \underbrace{\gamma_2\left(\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) - \nabla F(\mathbf{w}_{t+1})\right)}_{ⓓ}\Bigg\|^2.\end{aligned}$$

We define that

$$\begin{aligned}ⓒ \coloneqq \gamma_2\Bigg(&\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1(v_{t+1})\nabla f_2(u_t^i)\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) \\ &- \frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\Bigg).\end{aligned}$$

We define that $\Delta_t \coloneqq \mathbf{m}_t - \nabla F(\mathbf{w}_t)$. Note that $\mathrm{E}\left[\langle ⓐ, ⓓ\rangle\right] = \mathrm{E}\left[\langle ⓑ, ⓓ\rangle\right] = 0$. Then,

$$\begin{aligned}\mathrm{E}_t\left[\left\|ⓐ + ⓑ + ⓒ + ⓓ\right\|^2\right] = \left\|ⓐ\right\|^2 + \left\|ⓑ\right\|^2 + \mathrm{E}_t\left[\left\|ⓒ\right\|^2\right] + \mathrm{E}_t\left[\left\|ⓓ\right\|^2\right] + 2\left\langle ⓐ, ⓑ\right\rangle \\ + 2\mathrm{E}_t\left[\langle ⓐ, ⓒ\rangle\right] + 2\mathrm{E}_t\left[\langle ⓑ, ⓒ\rangle\right] + 2\mathrm{E}_t\left[\langle ⓒ, ⓓ\rangle\right].\end{aligned}$$

Based on the Young's inequality for products, we have $2\left\langle \mathbf{a}, \mathbf{b}\right\rangle \leq \frac{\|\mathbf{a}\|^2 c}{2} + \frac{2\|\mathbf{b}\|^2}{c}$ for $c > 0$.

$$\begin{aligned}&\mathrm{E}_t\left[\left\|ⓐ + ⓑ + ⓒ + ⓓ\right\|^2\right] \\ &\leq (1+\gamma_2)\left\|ⓐ\right\|^2 + 2(1+1/\gamma_2)\left\|ⓑ\right\|^2 + \frac{2+3\gamma_2}{\gamma_2}\mathrm{E}_t\left[\left\|ⓒ\right\|^2\right] + 2\mathrm{E}_t\left[\left\|ⓓ\right\|^2\right].\end{aligned}$$

Thus, we have

$$\mathrm{E}_t[\left\|\Delta_{t+1}\right\|^2] \leq (1-\gamma_2)\left\|\Delta_t\right\|^2 + \frac{2(1+\gamma_2)}{\gamma_2}\left\|ⓑ\right\|^2 + \frac{5}{\gamma_2}\mathrm{E}_t[\|ⓒ\|^2] + 2\mathrm{E}_t[\|ⓓ\|^2]. \tag{15}$$

Moreover, we have

$$\|ⓑ\|^2 = (1-\gamma_2)^2 \|\nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t+1})\|^2 \leq (1-\gamma_2)^2 \eta^2 L_F^2 \|\mathbf{m}_t\|^2. \tag{16}$$

On the other hand,

$$\begin{aligned}
\|ⓒ\|^2 \leq{}& \frac{2\gamma_2^2}{B_+} \sum_{i\in\mathcal{B}_+} \|\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\|^2 \left\|\nabla f_1(v_{t+1}) - \nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\right\|^2 \left\|\nabla f_2(u_t^i)\right\|^2 \\
&+ \frac{2\gamma_2^2}{B_+} \sum_{i\in\mathcal{B}_+} \|\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\|^2 \left\|\nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\right\|^2 \left\|\nabla f_2(u_t^i) - \nabla f_2(g_i(\mathbf{w}_{t+1}))\right\|^2 \\
\leq{}& \frac{2\gamma_2^2 L_f^2 C_f^2}{B_+} \sum_{i\in\mathcal{B}_+} \|\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\|^2 \left\|v_{t+1} - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right\|^2 \\
&+ \frac{2\gamma_2^2 L_f^2 C_f^2}{B_+} \sum_{i\in\mathcal{B}_+} \|\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\|^2 \left\|u_t^i - g_i(\mathbf{w}_{t+1})\right\|^2
\end{aligned}$$

Due to $\mathrm{E}_{\mathcal{B}_-}\left[\|\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\|^2\right] \leq C_g^2 + \zeta^2/\mathcal{B}_- \coloneqq C_1^2$, we have

$$\mathrm{E}_t\left[\|ⓒ\|^2\right] \leq \frac{2\gamma_2^2 L_f^2 C_f^2 C_1^2}{n} \sum_{i\in\mathcal{S}} \mathrm{E}_t\left[\left\|v_{t+1} - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right\|^2 + \left\|u_t^i - g_i(\mathbf{w}_{t+1})\right\|^2\right]$$

Besides, we also have

$$\begin{aligned}
&\mathrm{E}_t\left[\|ⓓ\|^2\right] \\
&\leq \gamma_2^2 \mathrm{E}_t\left[\left\|\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} \nabla f_1\left(\frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right) \nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) - \nabla F(\mathbf{w}_{t+1})\right\|^2\right] \\
&= \gamma_2^2 C_f^2 \mathrm{E}_t\left[\left\|\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} (\nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) - \nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1}))\right\|^2\right] \\
&+ \gamma_2^2 C_f^2 \mathrm{E}_t\left[\left\|\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} \nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-) - \frac{1}{n}\sum_{i\in\mathcal{S}} \nabla f_2(g_i(\mathbf{w}_{t+1}))\nabla g_i(\mathbf{w}_{t+1};\mathcal{B}_-)\right\|^2\right] \\
&\leq \frac{\gamma_2^2 C_f^4 \zeta^2}{B_-} + \frac{\gamma_2^2 C_f^4 C_g^2}{B_+}.
\end{aligned} \tag{17}$$

Then,

$$\begin{aligned}
\mathrm{E}\left[\|\Delta_{t+1}\|^2\right] \leq{}& (1-\gamma_2)\mathrm{E}\left[\|\Delta_t\|^2\right] + \frac{2L_F^2\eta^2}{\gamma_2}\mathrm{E}\left[\|\mathbf{m}_t\|^2\right] + 10\gamma_2 C_f^2 C_1^2 L_f^2 \mathrm{E}\left[\|\Psi_{t+1}\|^2\right] \\
&+ 20\gamma_2 C_f^2 L_f^2 C_1^2 \mathrm{E}\left[\frac{1}{n}\|\Xi_{t+1}\|^2\right] + 20\gamma_2 C_f^2 L_f^2 C_1^2 \mathrm{E}\left[\frac{1}{n}\|\mathbf{u}_{t+1} - \mathbf{u}_t\|^2\right] + 2\gamma_2^2 C_f^4\left(\frac{\zeta^2}{B_-} + \frac{C_g^2}{B_+}\right),
\end{aligned}$$

where we denote $\Xi_t \coloneqq \mathbf{u}_t - \mathbf{g}(\mathbf{w}_t)$ and $\Psi_t \coloneqq v_t - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_t))$. □

Lemma 11. *For the function value estimator v_{t+1} in SOTA-s and $\Psi_t \coloneqq v_t - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_t))$,*

$$\begin{aligned}
\mathrm{E}_t\left[\|\Psi_{t+1}\|^2\right] \leq{}& (1-\gamma_1)\|\Psi_t\|^2 + \frac{2C_f^2 C_g^2 \eta^2}{\gamma_1}\|\mathbf{m}_t\|^2 + 5\gamma_1 C_f^2 \mathrm{E}_t\left[\frac{1}{n}\|\Xi_{t+1}\|^2\right] \\
&+ \frac{10\gamma_1 C_f^2}{n}\mathrm{E}_t\left[\|\mathbf{u}_t - \mathbf{u}_{t+1}\|^2\right] + \frac{2\gamma_1^2 C_f^2}{B_+}.
\end{aligned} \tag{18}$$

Proof. According to the update of v in SOTA-s, we have

$$\begin{aligned}
&\left\|v_{t+1} - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right\|^2 = \left\|(1-\gamma_1)v_t + \gamma_1\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} f_2(u_t^i) - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right\|^2 \\
&= \left\|(1-\gamma_1)\left(v_t - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_t))\right) + (1-\gamma_1)\left(\frac{1}{n}\sum_{i\in\mathcal{S}}\left(f_2(g_i(\mathbf{w}_t)) - f_2(g_i(\mathbf{w}_{t+1}))\right)\right)\right. \\
&\qquad \left. +\gamma_1\left(\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} f_2(u_t^i) - \frac{1}{B_+}\sum_{i\in\mathcal{B}_+} f_2(g_i(\mathbf{w}_{t+1}))\right) + \gamma_1\left(\frac{1}{B_+}\sum_{i\in\mathcal{B}_+} f_2(g_i(\mathbf{w}_{t+1})) - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_{t+1}))\right)\right\|^2.
\end{aligned}$$

Denoting $\Psi_t \coloneqq \left\|v_t - \frac{1}{n}\sum_{i\in\mathcal{S}} f_2(g_i(\mathbf{w}_t))\right\|^2$, we have

$$\begin{aligned}
&\mathrm{E}_t\left[\|\Psi_{t+1}\|^2\right] \\
&\le (1-\gamma_1)\|\Psi_t\|^2 + \frac{2(1-\gamma_1)C_f^2C_g^2\eta^2}{\gamma_1}\|\mathbf{m}_t\|^2 + \frac{(2+3\beta)C_f^2}{\gamma_1}\gamma_1^2\mathrm{E}_t\left[\frac{1}{B_+}\sum_{i\in\mathcal{B}_+}\left\|u_t^i - g_i(\mathbf{w}_{t+1})\right\|^2\right] + \frac{2\gamma_1^2C_f^2}{B_+} \\
&\le (1-\gamma_1)\|\Psi_t\|^2 + \frac{2C_f^2C_g^2\eta^2}{\gamma_1}\|\mathbf{m}_t\|^2 + \frac{5\gamma_1C_f^2}{n}\mathrm{E}_t\left[\|\mathbf{u}_t - \mathbf{g}(\mathbf{w}_{t+1})\|^2\right] + \frac{2\gamma_1^2C_f^2}{B_+} \\
&\le (1-\gamma_1)\|\Psi_t\|^2 + \frac{2C_f^2C_g^2\eta^2}{\gamma_1}\|\mathbf{m}_t\|^2 + \frac{10\gamma_1C_f^2}{n}\mathrm{E}_t\left[\|\mathbf{u}_{t+1} - \mathbf{g}(\mathbf{w}_{t+1})\|^2\right] + \frac{10\gamma_1C_f^2}{n}\mathrm{E}_t\left[\|\mathbf{u}_t - \mathbf{u}_{t+1}\|^2\right] + \frac{2\gamma_1^2C_f^2}{B_+}.
\end{aligned}$$

□

Proof of Theorem 5. Based on (13), we have

$$\mathrm{E}\left[F(\mathbf{w}_{t+1}) - F_{\inf}\right] \le \mathrm{E}\left[F(\mathbf{w}_t) - F_{\inf}\right] + \frac{\eta}{2}\mathrm{E}\left[\|\Delta_t\|^2\right] - \frac{\eta}{2}\mathrm{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] - \frac{\eta}{4}\mathrm{E}\left[\|\mathbf{m}_t\|^2\right] \tag{19}$$

Re-arranging the terms and telescoping (19) from $t=1$ to T leads to

$$\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \le \frac{2\mathrm{E}\left[F(\mathbf{w}_1) - F_{\inf}\right]}{\eta T} + \underbrace{\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\Delta_t\|^2\right]}_{\coloneqq ⓔ} - \frac{1}{2T}\sum_{t=1}^T \mathrm{E}\left[\|\mathbf{m}_t\|^2\right]. \tag{20}$$

Based on (14), the term ⓔ can be upper bounded as

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\Delta_t\|^2\right] &\le \frac{\mathrm{E}\left[\|\Delta_1\|^2\right]}{\gamma_2 T} + \frac{2L_F^2\eta^2}{\gamma_2^2}\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\mathbf{m}_t\|^2\right] + 2\gamma_2C_f^4\left(\frac{\zeta^2}{B_-} + \frac{C_g^2}{B_+}\right) \\
&\quad + 20C_f^2C_1^2L_f^2\sum_{t=1}^T\frac{1}{n}\mathrm{E}[\|\mathbf{u}_{t+1} - \mathbf{u}_t\|^2] + 20C_f^2C_1^2L_f^2\left(\underbrace{\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\Psi_{t+1}\|^2\right]}_{\coloneqq ⓕ} + \underbrace{\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\frac{1}{n}\|\Xi_{t+1}\|^2\right]}_{\coloneqq ⓖ}\right).
\end{aligned}$$

Based on Lemma 2 in Wang & Yang (2022), the term ⓖ can be upper bounded as

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\frac{1}{n}\|\Xi_{t+1}\|^2\right] &\le \frac{4n\mathrm{E}\left[\frac{1}{n}\|\Xi_2\|^2\right]}{\gamma_0 B_+ T} + 20C_g^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^T \mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] \\
&\quad + \frac{8\gamma_0\sigma^2}{B_-} - \frac{1}{\gamma_0 B_+}\sum_{t=1}^T \mathrm{E}\left[\left\|\mathbf{u}^{t+1} - \mathbf{u}^t\right\|^2\right].
\end{aligned}$$

With $1/(\gamma_0 B_+) \geq \max(10C_f^2/n, 1/n)$, based on (18), the term ⓕ can be bounded as

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\Psi_{t+1}\|^2\right] &\leq \frac{\mathrm{E}\left[\|\Psi_2\|^2\right]}{\gamma_1 T} + \frac{2C_f^2C_g^2\eta^2}{\gamma_1^2}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] + 5C_f^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\frac{1}{n}\|\Xi_{t+2}\|^2\right] + \frac{2\gamma_1 C_f^2}{B_+} \\
&\leq \frac{\mathrm{E}\left[\|\Psi_2\|^2\right]}{\gamma_1 T} + \frac{2C_f^2C_g^2\eta^2}{\gamma_1^2}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] + \frac{2\gamma_1 C_f^2}{B_+} + \frac{20C_f^2 n\mathrm{E}\left[\frac{1}{n}\|\Xi_3\|^2\right]}{\gamma_0 B_+ T} \\
&\quad + 100C_f^2C_g^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+2}\|^2\right] + \frac{40C_f^2\gamma_0\sigma^2}{B_-}.
\end{aligned}$$

Plug the upper bounds of ⓕ and ⓖ into (20).

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \\
&\leq \frac{2\mathrm{E}\left[F(\mathbf{w}_1) - F_{\inf}\right]}{\eta T} + \frac{\mathrm{E}\left[\|\Delta_1\|^2\right]}{\gamma_2 T} + 2\gamma_2 C_f^4\left(\frac{\zeta^2}{B_-} + \frac{C_g^2}{B_+}\right) \\
&+ \frac{40nC_f^2C_1^2L_f^2\mathrm{E}\left[\frac{1}{n}\|\Xi_2\|^2\right]}{\gamma_0 B_+ T} - \left(\frac{1}{2} - \frac{2L_F^2\eta^2}{\gamma_2^2}\right)\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_t\|^2\right] \\
&+ 200C_f^2C_g^2C_1^2L_f^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] + \frac{80\gamma_0 C_f^2C_1^2L_f^2\sigma^2}{B_-} \\
&+ \frac{10C_f^2C_1^2L_f^2\mathrm{E}\left[\|\Psi_2\|^2\right]}{\gamma_1 T} + \frac{20C_f^4C_g^2C_1^2L_f^2\eta^2}{\gamma_1^2}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] + \frac{20\gamma_1 C_f^4C_1^2L_f^2}{B_+} \\
&+ \frac{200nC_f^4C_1^2L_f^2\mathrm{E}\left[\frac{1}{n}\|\Xi_3\|^2\right]}{\gamma_0 B_+ T} + 1000C_f^4C_1^2C_g^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+2}\|^2\right] + \frac{400\gamma_0 C_f^4C_1^2L_f^2\sigma^2}{B_-}.
\end{aligned}$$

If we choose $\eta \leq \min\left\{\frac{\gamma_2}{4L_F}, \frac{\gamma_0 B_+}{40nC_fC_gC_1\sqrt{L_f^2+5C_f^2}}, \frac{\gamma_1}{15C_f^2C_gC_1L_f}\right\}$, we have

$$\begin{aligned}
&-\left(\frac{1}{2} - \frac{2L_F^2\eta^2}{\gamma_2^2}\right)\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_t\|^2\right] + 200C_f^2C_g^2C_1^2L_f^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] \\
&\quad + \frac{20C_f^4C_g^2C_1^2L_f^2\eta^2}{\gamma_1^2}\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+1}\|^2\right] + 1000C_f^4C_1^2C_g^2\left(\frac{n\eta}{\gamma_0 B_+}\right)^2\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\mathbf{m}_{t+2}\|^2\right] \\
&\leq \frac{\mathrm{E}\left[\|\mathbf{m}_{T+1}\|^2\right] + \mathrm{E}\left[\|\mathbf{m}_{T+2}\|^2\right]}{8T}.
\end{aligned}$$

Besides, Lemma 2 in Wang & Yang (2022) and Lemma 11 imply that

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{n}\|\Xi_2\|^2\right] &\leq \mathrm{E}\left[\frac{1}{n}\|\Xi_1\|^2\right] + \frac{5n\eta^2C_g^2}{\gamma_0 B_+}\mathrm{E}\left[\|\mathbf{m}_1\|^2\right] + \frac{2\gamma_0^2\sigma^2 B_+}{nB_-} \\
&\leq \mathrm{E}\left[\frac{1}{n}\|\Xi_1\|^2\right] + \frac{\eta C_g}{8C_fC_1\sqrt{L_f^2+5C_f^2}}\mathrm{E}\left[\|\mathbf{m}_1\|^2\right] + \frac{2\gamma_0^2\sigma^2 B_+}{nB_-}, \\
\mathrm{E}\left[\frac{1}{n}\|\Xi_3\|^2\right] &\leq \mathrm{E}\left[\frac{1}{n}\|\Xi_1\|^2\right] + \frac{\eta C_g}{8C_fC_1\sqrt{L_f^2+5C_f^2}}\left(\mathrm{E}\left[\|\mathbf{m}_1\|^2 + \|\mathbf{m}_2\|^2\right]\right) + \frac{4\gamma_0^2\sigma^2 B_+}{nB_-}, \\
\mathrm{E}\left[\|\Psi_2\|^2\right] &\leq \mathrm{E}\left[\|\Psi_1\|^2\right] + \frac{2C_g\eta}{15C_1L_f}\|\mathbf{m}_1\|^2 + 5\gamma_1 C_f^2\mathrm{E}\left[\frac{1}{n}\|\Xi_{t+1}\|^2\right] + \frac{2\gamma_1^2C_f^2}{B_+}.
\end{aligned}$$

If we initialize $\mathbf{m}_1 = 0$, then $\mathrm{E}\left[\|\mathbf{m}_t\|^2\right] \leq C_f^2 C_1^2$ for any $t \geq 1$. We define that $\Lambda_{F,1} = \mathrm{E}\left[F(\mathbf{w}_1) - F_{\inf}\right] < +\infty$, $\Lambda_{\Delta,1} = \mathrm{E}\left[\|\Delta_1\|^2\right] < +\infty$, $\Lambda_{\Xi,2} = \mathrm{E}\left[\frac{1}{n}\|\Xi_2\|^2\right] < +\infty$, $\Lambda_{\Xi,3} = \mathrm{E}\left[\frac{1}{n}\|\Xi_2\|^2\right] < +\infty$, $\Lambda_{\Psi}^2 = \mathrm{E}\left[\|\Psi_2\|^2\right] < +\infty$. Then,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \\
&\leq \frac{2\Lambda_{F,1}}{\eta T} + \frac{\Lambda_{\Delta,1}}{\gamma_2 T} + \frac{40nC_f^2C_1^2L_f^2\Lambda_{\Xi,2}}{\gamma_0 B_+ T} + \frac{10C_f^2C_1^2L_f^2\Lambda_{\Psi}^2}{\gamma_1 T} + \frac{200nC_f^4C_1^2L_f^2\Lambda_{\Xi,3}}{\gamma_0 B_+ T} \\
&\quad + 2\beta C_f^4\left(\frac{\zeta^2}{B_-} + \frac{C_g^2}{B_+}\right) + \frac{80\gamma_0 C_f^2C_1^2L_f^2\sigma^2}{B_-} + \frac{20\gamma_1 C_f^4C_1^2L_f^2}{B_+} + \frac{400\gamma_0 C_f^4C_1^2L_f^2\sigma^2}{B_-} + \frac{C_f^2C_1^2}{4T}.
\end{aligned}
$$

Set $\gamma_0 = \frac{B_-\epsilon^2}{400C_f^2C_1^2L_f^2\sigma^2(1+5C_f^2)}$, $\gamma_1 = \frac{B_+\epsilon^2}{200C_f^4C_1^2L_f^2}$, $\gamma_2 = \frac{\min\{B_-,B_+\}\epsilon^2}{20C_f^4(\zeta^2+C_g^2)}$, and

$$
\begin{aligned}
\eta &\leq \min\left\{\frac{\gamma_2}{4L_F}, \frac{\gamma_0 B_+}{40nC_fC_gC_1\sqrt{L_f^2+5C_f^2}}, \frac{\gamma_1}{15C_f^2C_gC_1L_f}\right\}, \\
T &= \max\left\{\frac{1600\Lambda_{F,1}L_FC_f^4(\zeta^2+C_g^2)}{\min\{B_-,B_+\}\epsilon^4}, \frac{320000n\Lambda_{F,1}C_f^3C_gC_1^3L_f^2(1+5C_f^2)\sqrt{L_f^2+5C_f^2}\sigma^2}{B_-B_+\epsilon^4},\right. \\
&\quad \frac{60000\Lambda_{F,1}C_f^6C_gC_1^3L_f^3}{B_+\epsilon^4}, \frac{200C_f^4(\zeta^2+C_g^2)\Lambda_{\Delta,1}}{\min\{B_-,B_+\}\epsilon^4}, \frac{160000nC_f^4C_1^4L_f^4\sigma^2(1+5C_f^2)\Lambda_{\Xi}^2}{B_-B_+\epsilon^4}, \\
&\quad \left.\frac{20000C_f^6C_1^4L_f^4\Lambda_{\Psi}^2}{B_+\epsilon^4}, \frac{800000nC_f^6C_1^4L_f^4(1+5C_f^2)\Lambda_{\Xi,3}\sigma^2}{B_-B_+\epsilon^4}\right\}.
\end{aligned}
$$

Then, we have $\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left[\|\nabla F(\mathbf{w}_t)\|^2\right] \leq \epsilon^2$.

□

C. Optimization of CVaR-estimator of TPAUC

We have the following estimator

$$
F(\mathbf{w}) = \min_{s'\in\mathbb{R}} s' + \frac{1}{n_+\alpha}\sum_{\mathbf{x}_i\in\mathcal{S}_+}(\min_{s_i} s_i + \frac{1}{\beta}g_i(\mathbf{w};s_i) - s')_+.
$$

It is not difficult to show that the above estimator is equlvalent to

$$
F(\mathbf{w}) = \min_{s'\in\mathbb{R},\mathbf{s}\in\mathbb{R}^{n_+}} s' + \frac{1}{n_+\alpha}\sum_{\mathbf{x}_i\in\mathcal{S}_+}(s_i + \frac{1}{\beta}g_i(\mathbf{w};s_i) - s')_+.
$$

The reason is that $(\min_x f(x) - s)_+ = \min_x (f(x) - s)_+$. Using the conjugate of $[\cdot]_+$, we have

$$
\min_{\mathbf{w},\mathbf{s}\in\mathbb{R}^{n_+},s'} \max_{\mathbf{u}\in[0,1]^{n_+}} \underbrace{s' + \frac{1}{n_+\alpha}\sum_{\mathbf{x}_i\in\mathcal{S}_+} u_i(s_i + \frac{1}{\beta}g_i(\mathbf{w};s_i) - s')}_{F(\mathbf{w},\mathbf{s},s',\mathbf{u})}.
$$

Let $\bar{\mathbf{w}} = (\mathbf{w}, \mathbf{s}, s')$. We consider the function $F(\bar{\mathbf{w}}, \mathbf{u})$, which can be proved to be weakly convex and concave. Hence, we can use the stagewise proximal point method to solve the problem (Rafique et al., 2020). At the k-th stage, we solve the following problem approximately:

$$
\min_{\bar{\mathbf{w}}}\max_{\mathbf{u}\in[0,1]} F_k(\bar{\mathbf{w}},\mathbf{u}) = F(\bar{\mathbf{w}},\mathbf{u}) + \frac{1}{2\gamma}\|\bar{\mathbf{w}} - \bar{\mathbf{w}}_1^k\|^2
$$

We will use stochastic primal-dual algorithm for solving F_k. However, for $\mathbf{s}$ we use stochastic coordinate gradient descent update, and for $\mathbf{u}$ we also use stochastic coordinate gradient descent update. Let $\nabla_1 F$, $\nabla_2 F$, $\nabla_3 F$, $\nabla_4 F$ denote the partial

Algorithm 4 SOTA

Set $\mathbf{s}_0 = 0, s'_0 = 0, \mathbf{u}_0 = 1$ and initialize $\mathbf{w}_0$
for $k = 1, \ldots, K$ **do**
  Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}, \mathbf{s}_1^k = \mathbf{s}_{k-1}, s'^k_1 = s'_{k-1}, \mathbf{u}_1^k = \mathbf{u}_{k-1}$
  for $t = 1, \ldots, T_k$ **do**
    Sample $\mathcal{B}_+ \subset \mathcal{S}_+$ and $\mathcal{B}_- \subset \mathcal{S}_-$
    Update $\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, s'_{t+1}, \mathbf{u}_{t+1}$ according to (21)
  end for
  Let $\mathbf{w}_k, \mathbf{s}_k, s'_k, \mathbf{u}_k$ be the average of $\mathbf{w}_t, \mathbf{s}_t, s'_t, \mathbf{u}_t$, respectively
end for

gradient of F in terms of $\mathbf{w}, \mathbf{s}, s', \mathbf{u}$, respectively. We consider the following update:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}} \mathbf{w}^\top \nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + \frac{1}{2\eta_1}\|\mathbf{w} - \mathbf{w}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{w} - \mathbf{w}_1^k\|^2 \\
s_{t+1,i} &= \arg\min_{s} s U_i \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + \frac{1}{2\eta_2}|s - s_{t,i}|^2 + \frac{1}{2\gamma}|s - s_{1,i}^k|^2, \forall \mathbf{x}_i \in \mathcal{B}_+ \\
s'_{t+1} &= \arg\min_{s'} \nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + \frac{1}{2\eta_3}|s' - s'_t|^2 + \frac{1}{2\gamma}|s' - s'_{k,1}|^2 \\
u_{t+1,i} &= [u_{t,i} + \frac{n_+}{B_+}\eta_4 U_i \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)]_{[0,1]}, \forall \mathbf{x}_i \in \mathcal{B}_+,
\end{aligned} \tag{21}$$

where $U_i \in \mathbb{R}^{1\times d}$ denotes an operation that chooses the i-th coordinate of a vector. Note that different from (Rafique et al., 2020), we use stochastic coordinate descent (ascent) to update $\mathbf{s}_{t+1}$ ($\mathbf{u}_{t+1}$).

Next, we present the stochastic gradients.

$$\begin{aligned}
\nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) &= \frac{1}{B_+ B_- \alpha\beta} \sum_{\mathbf{x}_i \in \mathcal{B}_+} \sum_{\mathbf{x}_j \in \mathcal{B}_-} u_i \mathbb{I}(L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j) - s_{t,i} > 0) \nabla L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j) \\
U_i \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) &= \frac{1}{n_+\alpha} u_{t,i} (1 - \frac{1}{B_-\beta} \sum_{\mathbf{x}_j \in \mathcal{B}_-} \mathbb{I}(L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)) > s_{t,i}) \\
\nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) &= 1 - \frac{1}{B_+\alpha} \sum_{\mathbf{x}_i \in \mathcal{B}_+} u_{t,i} \\
U_i \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) &= \frac{1}{n_+\alpha}(s_{t,i} - s'_t + \frac{1}{B_-\beta} \sum_{\mathbf{x}_j \in \mathcal{B}_-} (L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j) - s_{t,i})_+)
\end{aligned}$$

We have

$$\begin{aligned}
&\mathrm{E}[\nabla_1 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)] = \nabla_1 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t), \quad \mathrm{E}[\|\nabla_1 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2] \leq \frac{C}{\alpha^2\beta^2} \\
&\mathrm{E}[n_+ U_i^\top U_i \nabla_2 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)] = \nabla_2 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t), \quad \mathrm{E}[\|U_i \nabla_2 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2] \leq \frac{1}{n_+^2\alpha^2}(1 + \frac{1}{\beta})^2 \\
&\mathrm{E}[\nabla_3 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)] = \nabla_3 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t), \quad \mathrm{E}[\|\nabla_3 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2] \leq (1 + \frac{1}{\alpha})^2 \\
&\mathrm{E}[n_+ U_i^\top U_i \nabla_4 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)] = \nabla_4 F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t), \quad \mathrm{E}[\|U_i \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2] \leq \frac{4C^2}{n_+^2\alpha^2}(1 + \frac{1}{\beta})^2
\end{aligned}$$

where we assume $\max(|s'_t|, |s_{t,i}|, L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j), \|\nabla L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)\|) \leq C$. The algorithm is shown in Algorithm 4.

C.1. Analysis

Theorem 6. *Assume there exists $C > 0$ such that $\max(|s'_t|, |s_{t,i}|, L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j), \|\nabla L(\mathbf{w}_t; \mathbf{x}_i, \mathbf{x}_j)\|) \leq C$ at every stage. Let $1/\gamma \geq \rho, \eta_1^k = \eta_3^k = \eta_4^k = \eta_2^k B_+/n_+ = \eta_k \propto 1/k, T_k \propto k^2$. SOTA ensures that after $T = O(1/\epsilon^6)$ iterations we can find an ϵ-nearly stationary solution for $\min_{\mathbf{w},\mathbf{s},s'} F(\mathbf{w}, \mathbf{s}, s')$.*

We first show that $F(\bar{\mathbf{w}}, \mathbf{u})$ is weakly convex in terms of $\bar{\mathbf{w}}$ for any $\mathbf{u}$.

Lemma 12. *Under Assumption 2, then $F(\bar{\mathbf{w}}, \mathbf{u})$ is $\rho/(\alpha\beta)$-weakly convex in terms of $\bar{\mathbf{w}}$ for any $\mathbf{u}$, where ρ is equal to the smoothness parameter of $L(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)$.*

Proof. Following similar analysis of Lemma 2, we can show that $F(\bar{\mathbf{w}}, \mathbf{u}) + \frac{\rho}{2\alpha\beta}\|\bar{\mathbf{w}}\|^2 = F(\bar{\mathbf{w}}, \mathbf{u}) + \frac{1}{n_+\alpha}\sum_i u_i(\frac{\rho}{2\beta}\|\mathbf{w}\|^2 + \frac{\rho}{2\beta}|s_i|^2 + \frac{\rho}{2\beta}|s'|^2) + \frac{\rho}{2n_+\alpha\beta}\sum_i(1-u_i)(\|\mathbf{w}\|^2 + |s_i|^2 + |s'|^2) + \frac{(n_+-1)}{2n_+\alpha\beta}\|\mathbf{s}\|^2$ is jointly convex in terms $\mathbf{w}, \mathbf{s}, s'$ for any $\mathbf{u} \in [0,1]$. Then $F(\bar{\mathbf{w}}, \mathbf{u})$ is $\rho' = \frac{\rho}{\alpha\beta}$-weakly convex in terms of $\bar{\mathbf{w}}$ for any $\mathbf{u}$. □

We need the following lemma for analysis.

Lemma 13. *Consider the proximal gradient update*

$$\mathbf{x}_{t+1} = \arg\min \mathbf{x}^\top G_t + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_t\|^2 + g(\mathbf{x}),$$

we have

$$\begin{aligned}
&(\mathbf{x}_t - \mathbf{x})^\top G_t + g(\mathbf{x}_t) - g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \\
&\leq \eta\|G_t\|^2 + \frac{1}{2\eta}(\|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2) + g(\mathbf{x}_t) - g(\mathbf{x}_{t+1}) - \frac{1}{4\eta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2
\end{aligned}$$

Proof. Due to the update of $\mathbf{x}_{t+1}$, we have

$$\begin{aligned}
&\mathbf{x}_{t+1}^\top G_t + \frac{1}{2\eta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + g(\mathbf{x}_{t+1}) + (\frac{1}{2\eta} + \frac{1}{2\gamma})\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \\
&\leq \mathbf{x}^\top G_t + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_t\|^2 + g(\mathbf{x})
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&(\mathbf{x}_t - \mathbf{x})^\top G_t + g(\mathbf{x}_t) - g(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \\
&\leq (\mathbf{x}_t - \mathbf{x}_{t+1})^\top G_t + \frac{1}{2\eta}(\|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2) + g(\mathbf{x}_t) - g(\mathbf{x}_{t+1}) - \frac{1}{2\eta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\leq \eta\|G_t\|^2 + \frac{1}{2\eta}(\|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2) + g(\mathbf{x}_t) - g(\mathbf{x}_{t+1}) - \frac{1}{4\eta}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2
\end{aligned}$$

□

Below, we let $g_1(\mathbf{w}) = \frac{1}{2\gamma}\|\mathbf{w} - \mathbf{w}_1^k\|^2$, $g_2(\mathbf{s}) = \frac{1}{2\gamma}\|\mathbf{s} - \mathbf{s}_1^k\|^2$, and $g_3(s') = \frac{1}{2\gamma}\|s' - s'_{k,1}\|^2$. Applying the above Lemma to $\mathbf{w}_{t+1}$, we have

$$\begin{aligned}
&(\mathbf{w}_t - \mathbf{w})^\top \nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + g_1(\mathbf{w}_t) - g_1(\mathbf{w}) \\
&\leq \eta_1\|\nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_1}(\|\mathbf{w} - \mathbf{w}_t\|^2 - \|\mathbf{w} - \mathbf{w}_{t+1}\|^2) + g_1(\mathbf{w}_t) - g_1(\mathbf{w}_{t+1})
\end{aligned}$$

Assume $\mathbf{w}$ is independent of noise and taking expectation on both sides, we have

$$\begin{aligned}
&\mathrm{E}[(\mathbf{w}_t - \mathbf{w})^\top \nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) + g_1(\mathbf{w}_t) - g_1(\mathbf{w})] \\
&\leq \mathrm{E}[\eta_1\|\nabla_1 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_1}(\|\mathbf{w} - \mathbf{w}_t\|^2 - \|\mathbf{w} - \mathbf{w}_{t+1}\|^2) + g_1(\mathbf{w}_t) - g_1(\mathbf{w}_{t+1})]
\end{aligned}$$

Applying the above Lemma to $s_{t+1,i}$, we have

$$\begin{aligned}
&(s_{t,i} - s_i)^\top U_i \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + g_2(s_{t,i}) - g_2(s_i) \\
&\leq \eta_2\|U_i \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_2}(\|s_i - s_{t,i}\|^2 - \|s_i - s_{t+1,i}\|^2) + g_2(s_{t,i}) - g_2(s_{t+1,i})
\end{aligned}$$

Summing the above inequality over $\mathbf{x}_i \in \mathcal{B}_+$, assuming $\mathbf{s}$ is independent of noise and taking expectation on both sides, we have

$$\begin{aligned}
&\frac{B_+}{n_+}\mathrm{E}(\mathbf{s}_t - \mathbf{s})^\top \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) + g_2(\mathbf{s}_t) - g_2(\mathbf{s})] \\
&= \mathrm{E}[\eta_2 \sum_{i\in\mathcal{B}_+} \|U_i \nabla_2 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_2}(\|\mathbf{s} - \mathbf{s}_t\|^2 - \|\mathbf{s} - \mathbf{s}_{t+1}\|^2) + g_2(\mathbf{s}_t) - g_2(\mathbf{s}_{t+1})]
\end{aligned}$$

Applying the above lemma to s'_{t+1}, we have

$$\begin{aligned}
&(s'_t - s')^\top \nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + g_3(s'_t) - g_3(s') \\
&\leq \eta_2 \|\nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_2}(\|s' - s'_t\|^2 - \|s' - s'_{t+1}\|^2) + g_3(s'_t) - g_3(s'_{t+1})
\end{aligned}$$

Assume $\mathbf{s}'$ is independent of noise and taking expectation on both sides, we have

$$\begin{aligned}
&\mathrm{E}[(s'_t - s')^\top \nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) + g_3(s'_t) - g_3(s')] \\
&\leq \mathrm{E}[\eta_2 \|\nabla_3 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_2}(\|s' - s'_t\|^2 - \|s' - s'_{t+1}\|^2) + g_3(s'_t) - g_3(s'_{t+1})]
\end{aligned}$$

Adding the above inequalities for $\mathbf{w}, \mathbf{s}, s'$ together we have

$$\begin{aligned}
&\mathrm{E}[\sum_t (\bar{\mathbf{w}}_t - \bar{\mathbf{w}})^\top \nabla_{\bar{\mathbf{w}}} F(\bar{\mathbf{w}}_t, \mathbf{u}_t) + g(\bar{\mathbf{w}}_t) - g(\bar{\mathbf{w}})] \\
&\leq \eta_1 C_1^2 T + \eta_2 n_+ C_2^2 T + \eta_3 C_3^2 T \\
&+ \frac{1}{2\eta_1}(\|\mathbf{w} - \mathbf{w}_1\|^2) + \frac{n_+}{2\eta_2 B_+}(\|\mathbf{s} - \mathbf{s}_1\|^2) + \frac{1}{2\eta_3}(\|s' - s'_1\|^2) + g_1(\mathbf{w}_1) + \frac{n_+}{B_+} g_2(\mathbf{s}_1) + g_3(s'_1)
\end{aligned}$$

As a result,

$$\begin{aligned}
&\mathrm{E}[\sum_t F_k(\bar{\mathbf{w}}_t, \mathbf{u}_t) - F_k(\bar{\mathbf{w}}, \mathbf{u}_t)] \leq \eta_1 C_1^2 T + \eta_2 n_+ C_2^2 T + \eta_3 C_3^2 T \\
&+ \frac{1}{2\eta_1}(\|\mathbf{w} - \mathbf{w}_1\|^2) + \frac{n_+}{2\eta_2 B_+}(\|\mathbf{s} - \mathbf{s}_1\|^2) + \frac{1}{2\eta_3}(\|s' - s'_1\|^2)
\end{aligned}$$

For the update of $\mathbf{u}$, it is equivalent to

$$\mathbf{u}_{t+1} = \arg\min_{\mathbf{u}\in[0,1]} \mathbf{u}^\top \widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) + \frac{1}{2\eta_4}\|\mathbf{u} - \mathbf{u}_t\|^2$$

where $\widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) = \frac{n_+}{B_+}\sum_{i\in\mathcal{B}_+} U_i^\top U_i \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)$. It is easy to show that $\mathrm{E}[\widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)] = \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t)$. Applying the same analysis to the update of $\mathbf{u}_t$, we have

$$\begin{aligned}
&(\mathbf{u} - \mathbf{u}_t)^\top \widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t) \\
&\leq \eta_4 \|\widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_4}(\|\mathbf{u} - \mathbf{u}_t\|^2 - \|\mathbf{u} - \mathbf{u}_{t+1}\|^2)
\end{aligned}$$

We do not assume $\mathbf{u}$ is independent of the randomness in order to derive the primal objective gap. As a result,

$$\begin{aligned}
&(\mathbf{u} - \mathbf{u}_t)^\top \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) \\
&= \eta_4 \|\widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_4}(\|\mathbf{u} - \mathbf{u}_t\|^2 - \|\mathbf{u} - \mathbf{u}_{t+1}\|^2) \\
&+ (\mathbf{u} - \mathbf{u}_t)^\top (\nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) - \widetilde{\nabla}_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t))
\end{aligned}$$

Following previous analysis (e.g., Proposition A.1 in (Rafique et al., 2020)), we have

$$\begin{aligned}
&(\mathbf{u} - \tilde{\mathbf{u}}_t)^\top (\nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) - \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)) \\
&\leq \eta_4 \|\nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t) - \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t; \xi_t)\|^2 + \frac{1}{2\eta_4}(\|\mathbf{u} - \tilde{\mathbf{u}}_t\|^2 - \|\mathbf{u} - \tilde{\mathbf{u}}_{t+1}\|^2)
\end{aligned}$$

Hence for any $\mathbf{u}$, we have

$$\mathrm{E}(\mathbf{u} - \mathbf{u}_t)^\top \nabla_4 F(\bar{\mathbf{w}}_t, \mathbf{u}_t)] \leq \frac{2n_+^2}{B_+}\eta_4 C_4^2 + \frac{1}{\eta_4}\|\mathbf{u} - \mathbf{u}_1\|^2$$

As a result, for any $\mathbf{u}$ we have

$$\mathrm{E}[\sum_t F_k(\mathbf{u}, \bar{\mathbf{w}}_t) - F_k(\mathbf{u}_t, \bar{\mathbf{w}}_t)] \leq \frac{2n_+^2}{B_+}\eta_4 C_4^2 + \frac{1}{\eta_4}\|\mathbf{u} - \mathbf{u}_1\|^2$$

As a result,

$$\mathrm{E}[\sum_t F_k(\bar{\mathbf{w}}_t, \mathbf{u}) - F_k(\bar{\mathbf{w}}, \mathbf{u}_t)] \leq \eta_1 C_1^2 T + \eta_2 n_+ C_2^2 T + \eta_3 C_3^2 T + \frac{2n_+^2}{B_+}\eta_4 C_4^2$$
$$+ \frac{1}{2\eta_1}(\|\mathbf{w} - \mathbf{w}_1\|^2) + \frac{n_+}{2\eta_2 B_+}(\|\mathbf{s} - \mathbf{s}_1\|^2) + \frac{1}{2\eta_3}(\|s' - s_1'\|^2) + \frac{1}{\eta_4}\|\mathbf{u} - \mathbf{u}_1\|^2$$

Let $\eta_1 = \eta_3 = \eta_2 B_+/n_+ = \eta, \eta_4 = \eta$. Then we have

$$\mathrm{E}[\sum_t F_k(\bar{\mathbf{w}}_t, \mathbf{u}) - F_k(\bar{\mathbf{w}}, \mathbf{u}_t)] \leq \eta C_1^2 T + \eta \frac{n_+^2}{B_+} C_2^2 T + \eta C_3^2 T + \frac{2n_+^2}{B_+}\eta C_4^2$$
$$+ \frac{1}{2\eta}(\|\bar{\mathbf{w}} - \bar{\mathbf{w}}_1\|^2) + \frac{1}{\eta}$$

Then we have

$$\mathrm{E}[\max_{\mathbf{u}} F_k(\widehat{\mathbf{w}}_T, \mathbf{u}) - F_k(\bar{\mathbf{w}}^*, \widehat{\mathbf{u}}_t)] \leq \eta C_1^2 T + \eta \frac{n_+^2}{B_+} C_2^2 + \eta C_3^2 + 2\frac{n_+^2}{B_+}\eta C_4^2$$
$$+ \frac{1}{2\eta T}(\|\bar{\mathbf{w}}^* - \bar{\mathbf{w}}_1\|^2) + \frac{1}{\eta T}$$

where $\widehat{\mathbf{w}}_T$ is the average of $\bar{\mathbf{w}}_t, t = 1, \ldots, T$ and $\widehat{\mathbf{u}}_T$ is the averaged solution of $\mathbf{u}_t$, $\bar{\mathbf{w}}^*$ is the optimal solution to $F_k(\bar{\mathbf{w}})$. The above implies that

$$\mathrm{E}[F_k(\widehat{\mathbf{w}_T}) - \min_{\bar{\mathbf{w}}} F_k(\bar{\mathbf{w}})] \leq \eta(\frac{C}{\alpha^2\beta^2} + \frac{n_+^2}{B_+}\frac{1}{n_+^2\alpha^2}(1 + \frac{1}{\beta})^2 + (1 + \frac{1}{\alpha})^2 + 2\frac{n_+^2}{B_+}\frac{4C^2}{n_+^2\alpha^2}(1 + \frac{1}{\beta})^2)$$
$$+ \frac{1}{2\eta}(\|\bar{\mathbf{w}}^* - \bar{\mathbf{w}}_1\|^2) + \frac{1}{\eta T}$$
$$\leq \eta O\left(\frac{C}{\alpha^2\beta^2}\right) + \frac{1}{2\eta}(\|\bar{\mathbf{w}}^* - \bar{\mathbf{w}}_1\|^2) + \frac{1}{\eta T}$$

It remains to apply the analysis in (Rafique et al., 2020, Theorem 4.1) to derive the convergence for the Moreau envelope of $F(\mathbf{w}, \mathbf{s}, s')$ with a complexity in the order of $O(1/\epsilon^6)$ by setting $\eta_k \propto 1/k$ and $T_k \propto k^2$ and the total number of stages $K = O(1/\epsilon^2)$.