

---

# When All We Need is a Piece of the Pie: A Generic Framework for Optimizing Two-way Partial AUC

---

Zhiyong Yang<sup>1,2</sup> Qianqian Xu<sup>3</sup> Shilong Bao<sup>1,2</sup> Yuan He<sup>4</sup> Xiaochun Cao<sup>1,2</sup> Qingming Huang<sup>3,5,6,7</sup>

## Abstract

The Area Under the ROC Curve (AUC) is a crucial metric for machine learning, which evaluates the average performance over all possible True Positive Rates (TPRs) and False Positive Rates (FPRs). Based on the knowledge that a skillful classifier should simultaneously embrace a high TPR and a low FPR, we turn to study a more general variant called Two-way Partial AUC (TPAUC), where only the region with  $\text{TPR} \geq \alpha, \text{FPR} \leq \beta$  is included in the area. Moreover, a recent work shows that the TPAUC is essentially inconsistent with the existing Partial AUC metrics where only the FPR range is restricted, opening a new problem to seek solutions to leverage high TPAUC. Motivated by this, we present the first trial in this paper to optimize this new metric. The critical challenge along this course lies in the difficulty of performing gradient-based optimization with end-to-end stochastic training, even with a proper choice of surrogate loss. To address this issue, we propose a generic framework to construct surrogate optimization problems, which supports efficient end-to-end training with deep-learning. Moreover, our theoretical analyses show that: 1) the objective function of the surrogate problems will achieve an upper bound of the original problem under mild conditions, and 2) optimizing the surrogate problems leads to good generalization performance in terms of TPAUC with a high probability. Finally, empirical studies over several benchmark datasets speak to the efficacy of our

framework.

## 1. Introduction

ROC (Receiver Operating Characteristics) curve is a well-known tool to evaluate classification performance at varying threshold levels. More precisely, as shown in Fig. 1-(a), it captures the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) as a function of the classification thresholds. AUC (Area Under the ROC Curve), summarizes the average performance of a given classifier by calculating its area. More intuitively, as shown in (Hanley & McNeil, 1982), AUC is equivalent to the possibility that a positive instance has a higher predicted score to be positive than a negative instance. Any skillful classifier that can produce well-separated scores for positive and negative instances will enjoy a high AUC value, no matter how skewed the class distribution is. As a natural result, AUC is more appropriate than accuracy for long-tail classification problems such as disease prediction (Hao et al., 2020; Zhou et al., 2020) and rare event detection (Liu et al., 2018; Wu et al., 2020; Liu et al., 2020a; Wang et al., 2019), due to this appealing property (Fawcett, 2006; Hand & Till, 2001).

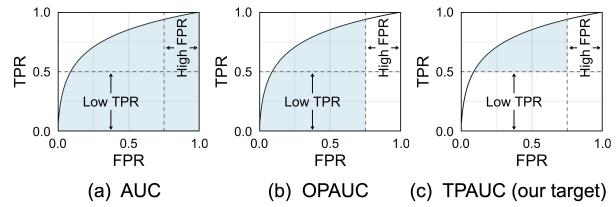


Figure 1. Comparisons of different AUC variants: (a) The entire area of ROC curve; (b) The One-way Partial AUC (OPAUC) which measures the area of a local region of ROC within an FPR range; (c) The Two-way Partial AUC (TPAUC).

Over the past two decades, the importance of AUC has raised a new wave to directly optimize AUC, which has achieved tremendous success. A partial list of the related studies includes (Alan & Raskutti, 2004; Joachims, 2005; 2006; Calders & Jaroszewicz, 2007; Narasimhan & Agarwal, 2013a; Gao et al., 2013; Narasimhan & Agarwal, 2017a).

<sup>1</sup>State Key Laboratory of Info. Security (SKLOIS), Inst. of Info. Engin., CAS, Beijing, China. <sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. <sup>3</sup>Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China. <sup>4</sup>Alibaba Group, Beijing, China. <sup>5</sup>School of Computer Science and Tech., University of Chinese Academy of Sciences, Beijing, China. <sup>6</sup>Peng Cheng Laboratory, Shenzhen, China. <sup>7</sup>BDKM, University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Qianqian Xu <xuqianqian@ict.ac.cn>, Qingming Huang <qmhuang@ucas.ac.cn>.

However, the vast majority of such studies only consider the area over the entire ROC curve. As argued by (Narasimhan & Agarwal, 2013b), for some applications, only the performance within a given range of False Positive Rate (FPR) is of interest, as shown in Fig.1-(b). In this sense, the standard AUC tends to provide a biased estimation of the performance by including unrelated regions. This key investigation has motivated a series of successful studies to optimize the One-way Partial AUC (OPAUC) with an FPR range  $[\alpha, \beta]$  (Narasimhan & Agarwal, 2013b;c; 2017b). Here we note that the choice to truncate FPR on the ROC curve is based on domain-specific prior knowledge for some specific fields such as biometric screening, and medical diagnosis (Narasimhan & Agarwal, 2013b).

**Taking a step further, what should be a general rule to select the target region under the ROC curve for classification problems?**

Since TPR and FPR evaluate complementary aspects of the model performance, we argue that a practical classifier in most applications must simultaneously have a high TPR and a low FPR. In other words, a high TPR is meaningless if the FPR is lower than a tolerance threshold, while a low FPR cannot compensate for a low TPR (*say, one can hardly consider a model with FPR higher than 0.8 even if its TPR is as high as 0.99, and vice versa for a low TPR model*). In this sense, we only need to pay attention to the upper-left head region under the ROC curve, as shown in Fig.1-(c).

A recent work (Yang et al., 2019) exactly realizes this idea, where a new metric called Two-Way Partial AUC (TPAUC) is proposed to measure the area of a partial region of the ROC curve with  $\text{TPR} \geq p$ ,  $\text{FPR} \leq q$ . Furthermore, (Yang et al., 2019) shows that the TPAUC is essentially inconsistent with one-way partial AUC. In other words, a higher OPAUC does not necessarily imply a higher TPAUC, posing a demand to seek new solutions to leverage high TPAUC.

**Inspired by this fact, we present the first trial to optimize the TPAUC metric with an end-to-end framework.**

The **major challenge** of this task is that the objective function is not differentiable even with a proper surrogate loss function, suggesting that there is no easy way to perform end-to-end training. Facing this challenge, we propose a generic framework to approximately optimize the TPAUC with the help of deep learning. Generally speaking, our contributions are as follows.

First, we reformulate the original optimization problem as a bi-level optimization problem, where the inner-level problem provides a sparse sample selection process and the outer-level problem minimizes the loss over the selected instances.

On top of the reformulation, we propose a generic frame-

work to construct surrogate optimization problems for the original problem. In the core of this framework lies the interplay of the surrogate penalty functions and surrogate weighting functions defined in this paper. Moreover, we construct a dual correspondence between these two classes of functions, such that we can easily find a standard single-level surrogate optimization problem whenever a surrogate penalty or a surrogate weighting function is obtained.

We then proceed to explore theoretical guarantees for the framework. On one hand, by comparing the surrogate problem and the original problem, we provide a mild sufficient condition under which the objective function surrogate problem becomes an upper bound of the original problem and further show that concave weighting function tends to be a better choice than their convex counterparts. On the other hand, we show that optimizing the surrogate problems could leverage reasonable generalization performance in terms of TPAUC with high probability.

## 2. Prior Art

**Partial AUC Optimization.** Comparing with existing studies to optimize partial AUC (Narasimhan & Agarwal, 2013b;c; 2017b), the key difference is two-fold. The previous studies only focus on a one-way partial AUC, where only the FPR is restricted within  $[\alpha, \beta]$ ; while we are the first to study TPAUC optimization, a new AUC metric where both TPR and FPR are truncated. Moreover, most related studies are based on the cutting plane algorithm, which do not fit to the end-to-end training framework in deep learning. In our work, getting rid of complicated combinatorial optimization techniques, we propose a general framework to construct much simpler surrogate optimization problems for TPAUC that supports end-to-end training. *Please see Appendix.A for a review of the general AUC optimization methods.*

## 3. Preliminaries

### 3.1. Standard AUC metric

Before showing the formal definition of the two-way partial AUC, we first provide a quick review of the standard AUC metric. Under the context of binary classification problems, an instance is denoted as  $(\mathbf{x}, y)$ , where  $\mathbf{x} \in \mathcal{X}$  is the input raw features and  $y \in \{0, 1\}$  is the label. Taking a step further, given a dataset  $\mathcal{D}$ , denote by  $\mathcal{X}_P$  the set of positive instances in our dataset, and  $\mathcal{X}_N$  the set of the negative ones, then the sampling process could be expressed as:

$$\begin{aligned}\mathcal{X}_P = \{\mathbf{x}_i^+\}_{i=1}^{n_+} &\stackrel{i.i.d}{\sim} \mathcal{P} : \mathbb{P} [\mathbf{x}^+ | y = 1], \\ \mathcal{X}_N = \{\mathbf{x}_j^-\}_{j=1}^{n_-} &\stackrel{i.i.d}{\sim} \mathcal{N} : \mathbb{P} [\mathbf{x}^- | y = 0],\end{aligned}$$

where  $n_+, n_-$  are the numbers of positive/negative instances, respectively; and  $\mathcal{P}, \mathcal{N}$  are the corresponding conditional distributions. For binary class problems, our goal is to learn a score function  $f_\theta : \mathcal{X} \rightarrow [0, 1]$ , such that  $f_\theta(\mathbf{x})$  is proportional to the possibility that  $\mathbf{x}$  belongs to the positive class. Based on the score function, we can further predict the label of an instance  $\mathbf{x}$  as  $\mathbf{1}[f_\theta(\mathbf{x}) > t]$ , where  $t$  is the decision threshold,  $\mathbf{1}[\cdot]$  is the indicator function. Given a threshold  $t$ , we can define two elementary metrics known as True Positive Rate (TPR) and False Positive Rate (FPR), which are the probabilities that a positive/negative instance is predicted as a positive instance, i.e.:

$$\begin{aligned} \text{TPR}_{f_\theta}(t) &= \mathbb{P}_{\mathbf{x}^+ \in \mathcal{P}} [f_\theta(\mathbf{x}^+) > t], \\ \text{FPR}_{f_\theta}(t) &= \mathbb{P}_{\mathbf{x}^- \in \mathcal{N}} [f_\theta(\mathbf{x}^-) > t]. \end{aligned} \quad (1)$$

Based on the label predictions, AUC is defined as the Area under the Receiver Operating Characteristic (ROC) curve plotted by True Positive Rate (TPR) against False Positive Rate (FPR) with varying thresholds, which could be expressed mathematically as follows:

$$\text{AUC}(f_\theta) = \int_0^1 \text{TPR}_{f_\theta} \left( \text{FPR}_{f_\theta}^{-1}(t) \right) dt. \quad (2)$$

When the possibility to observe a tied comparison is null, i.e.

$$\mathbb{P}_{\mathbf{x}^+ \in \mathcal{P}, \mathbf{x}^- \in \mathcal{N}} [f_\theta(\mathbf{x}^+) = f_\theta(\mathbf{x}^-)] = 0,$$

AUC is known (Hanley & McNeil, 1982) to enjoy a much simpler formulation as the possibility that correct ranking takes place between a positive and negative instance:

$$\text{AUC}(f_\theta) = 1 - \mathbb{E}_{\mathbf{x}^+ \sim \mathcal{P}} \left[ \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\ell_{0,1}(f_\theta(\mathbf{x}^+) - f_\theta(\mathbf{x}^-))] \right],$$

where  $\ell_{0,1}$  denotes the 0–1 loss with  $\ell_{0,1}(\mathbf{x}) = 1$  if  $\mathbf{x} < 0$ , and  $\ell_{0,1}(\mathbf{x}) = 0$ , otherwise. Given a finite dataset  $\mathcal{S} = \mathcal{X}_P \cup \mathcal{X}_N$ , the unbiased estimation of  $\text{AUC}(f_\theta)$  could be expressed as:

$$\hat{\text{AUC}}(f_\theta) = 1 - \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \frac{\ell_{0,1}(f_\theta(\mathbf{x}_i^+) - f_\theta(\mathbf{x}_j^-))}{n_+ n_-}.$$

### 3.2. Two-Way Partial AUC Metrics

**Definitions.** As presented in the introduction, instead of the complete area of ROC, we focus on the area of ROC in a partial region with  $\text{TPR}_{f_\theta}(t) \geq 1 - \alpha$ ,  $\text{FPR}_{f_\theta}(t) \leq \beta$ , which is called two-way partial AUC in (Yang et al., 2019). Here we define it as  $\text{AUC}_\alpha^\beta$ :

$$\begin{aligned} \text{AUC}_\alpha^\beta(f_\theta) &= \int_{\text{FPR}_{f_\theta}(\text{TPR}_{f_\theta}^{-1}(1-\alpha))}^{\beta} \text{TPR}_{f_\theta} \left( \text{FPR}_{f_\theta}^{-1}(t) \right) dt \\ &\quad - (1 - \alpha) \cdot \left( \beta - \text{FPR}_{f_\theta} \left( \text{TPR}_{f_\theta}^{-1}(1 - \alpha) \right) \right). \end{aligned}$$

Since the data distributions  $\mathcal{P}, \mathcal{N}$  are often unknown, it is necessary to study its empirical estimation based on an observed dataset  $\mathcal{S}$ . (Yang et al., 2019) derives an empirical version of  $\text{AUC}_\alpha^\beta(f_\theta)$  as the truncated AUC over the hard positive and negative instances, which is denoted as  $\hat{\text{AUC}}_\alpha^\beta(f_\theta, \mathcal{S})$  in our paper:

$$\hat{\text{AUC}}_\alpha^\beta(f_\theta, \mathcal{S}) = 1 - \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \frac{\ell_{0,1}(f_\theta(\mathbf{x}_{(i)}^+) - f_\theta(\mathbf{x}_{(j)}^-))}{n_+ n_-}$$

where  $\mathbf{x}_{(i)}^+$  denotes the hard positive instance that achieves **bottom- $i$**  score among all positive instances, and  $\mathbf{x}_{(j)}^-$  denotes the hard negative instance achieves **top- $j$**  score among all negative instances,  $n_+^\alpha = \lfloor n_+ \cdot \alpha \rfloor$ , and  $n_-^\beta = \lfloor n_- \cdot \beta \rfloor$ , are the numbers of the chosen hard positive and negative examples. *Please see Appendix.B for an analysis of the inconsistency between TPAUC and OPAUC.*

## 4. The Proposed Framework

### 4.1. A Generic Framework to Construct Surrogate Optimization Problems

Based on the empirical estimation shown in Sec.3.2, it is clear that optimizing TPAUC over a finite dataset  $\mathcal{S}$  requires minimizing the following quantity:

$$1 - \hat{\text{AUC}}_\alpha^\beta(f_\theta, \mathcal{S}) = \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \frac{\ell_{0,1}(f_\theta(\mathbf{x}_{(i)}^+) - f_\theta(\mathbf{x}_{(j)}^-))}{n_+ n_-}.$$

Following the framework of surrogate loss (Mohri et al., 2018), we replace the non-differential 0–1 loss with a convex loss function  $\ell$ , such that  $\ell(t)$  is an upper bound of  $\ell_{0,1}(t)$ . Note that if the scores live in  $[0, 1]$ , standard loss functions such as  $\ell_{\text{exp}}(t) = \exp(-t)$ ,  $\ell_{\text{sq}}(t) = (1 - t)^2$  often satisfy this constraint. Hence given a feasible surrogate loss  $\ell$ , our goal is then to solve the following problem:

$$(OP_0) \min_{\theta} \hat{\mathcal{R}}_{\alpha, \beta}^\ell(S, f_\theta) = \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \frac{\ell(f_\theta(\mathbf{x}_{(i)}^+) - f_\theta(\mathbf{x}_{(j)}^-))}{n_+ n_-}.$$

Unfortunately, even with the choice of differentiable surrogate losses, the objective function  $\hat{\mathcal{R}}_{\alpha, \beta}^\ell(S, f_\theta)$  is still not differentiable. This is because calculating  $\mathbf{x}_{(i)}^+$ ,  $\mathbf{x}_{(j)}^-$  requires sorting the scores of positive and negative instances. Nonetheless, the objective function is essentially a composition of a sparse sample selection operation and the original loss. This is shown in the following proposition, where  $(OP_0)$  is reformulated as a so-called bi-level optimization problem (Liu et al., 2020c;b). The inner-level problems provide a sparse sample selection process, and the outer-level problem performs the optimization based on the chosen instances. *Please see Appendix.C for the proof.*

**Proposition 1.** For any  $\alpha, \beta \in (0, 1)$ , if scores  $f_{\theta}(\mathbf{x}) \in [0, 1]$ , and there are no ties in the scores, the original optimization problem is equivalent to the following problem:

$$\begin{aligned} & \min_{\theta} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-) \\ \text{s.t. } & v_+ = \underset{v_i^+ \in [0, 1], \sum_{i=1}^{n_+} v_i^+ \leq n_+^\alpha}{\operatorname{argmax}} \sum_{i=1}^{n_+} (v_i^+ \cdot (1 - f_{\theta}(\mathbf{x}_i^+))) \\ & v_- = \underset{v_j^- \in [0, 1], \sum_{j=1}^{n_-} v_j^- \leq n_-^\beta}{\operatorname{argmax}} \sum_{j=1}^{n_-} (v_j^- \cdot f_{\theta}(\mathbf{x}_j^-)) \end{aligned}$$

where

$$\ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-) = \ell(f_{\theta}(\mathbf{x}_i^+) - f_{\theta}(\mathbf{x}_j^-)).$$

Based on the proposition, the source of the intractability of  $(OP_0)$  comes from the  $\ell_1$  ball constraints  $\sum_{i=1}^{n_+} v_i^+ \leq n_+^\alpha$ ,  $\sum_{j=1}^{n_-} v_j^- \leq n_-^\beta$  in the inner-level problem. To establish an efficient approximation of the original problem, we follow a standard trick to transform the  $\ell_1$  ball constraints to  $\ell_1$  penalty terms in the objective function (note that  $v_+, v_-$  are non-negative). In this way, the inner-level problems become:

$$\begin{aligned} v_+ &= \underset{v_i^+ \in [0, 1]}{\operatorname{argmax}} \sum_{i=1}^{n_+} (v_i^+ \cdot (1 - f_{\theta}(\mathbf{x}_i^+)) - \lambda^+ \cdot v_i^+) \\ v_- &= \underset{v_j^- \in [0, 1]}{\operatorname{argmax}} \sum_{j=1}^{n_-} (v_j^- \cdot f_{\theta}(\mathbf{x}_j^-) - \lambda^- \cdot v_j^-) \end{aligned}$$

Furthermore, to avoid sparsity, we replace the sparsity-inducing  $\ell_1$  penalty with a smooth surrogate  $\varphi_{\gamma}$ . This naturally leads to a smooth problem:

$$\begin{aligned} (OP_1) \min_{\theta} & \frac{1}{n_+^\alpha n_-^\beta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-) \\ \text{s.t. } & v_+ = \underset{v_i^+ \in [0, 1]}{\operatorname{argmax}} \sum_{i=1}^{n_+} (v_i^+ \cdot (1 - f_{\theta}(\mathbf{x}_i^+)) - \varphi_{\gamma}(v_i^+)) \\ & v_- = \underset{v_j^- \in [0, 1]}{\operatorname{argmax}} \sum_{j=1}^{n_-} (v_j^- \cdot f_{\theta}(\mathbf{x}_j^-) - \varphi_{\gamma}(v_j^-)) \end{aligned}$$

To ensure that the chosen  $\varphi_{\gamma}$  provides an effective approximation of the  $\ell_1$  penalty, we pose several regularities on such functions. In the following, we define this class of functions as the calibrated smooth penalty function.

**Definition 1.** A penalty function  $\varphi_{\gamma}(x) : \mathbb{R}_+ \rightarrow \mathbb{R}$  is called a **calibrated smooth penalty function**, if it satisfies the following regularities:

- (A)  $\varphi_{\gamma}$  has continuous third-order derivatives.
- (B)  $\varphi_{\gamma}$  is strictly increasing in the sense that  $\varphi'_{\gamma}(x) > 0$ .
- (C)  $\varphi_{\gamma}$  is strictly convex in the sense that  $\varphi''_{\gamma}(x) > 0$ .
- (D)  $\varphi_{\gamma}$  has positive third-order derivatives in the sense that  $\varphi'''_{\gamma}(x) > 0$ .

Note that the condition (B) is inherited from the  $\ell_1$  norm. While the other conditions improve the smoothness of the function. Moreover, the last condition is to ensure that the weighting function is strictly concave (see the arguments about the weights).

Given the penalty functions, we turn to explore a corresponding factor in the framework. According to the inner level problem, the sample weights  $v_i^+, v_j^-$  have a dual correspondence with the penalty functions. More precisely, given a fixed  $\phi_{\gamma}$ , one can derive the corresponding weighting function  $\psi_{\gamma}$  as a function of  $f_{\theta}(\mathbf{x})$  such that:

$$v_i^+ = \psi_{\gamma}(1 - f_{\theta}(\mathbf{x}_i^+)), \quad v_j^- = \psi_{\gamma}(f_{\theta}(\mathbf{x}_j^-)), \quad v_i^+, v_j^- \in [0, 1].$$

Moreover, if  $\psi_{\gamma}$  has a closed-form expression, then we can cancel the inner optimization problem and instead minimize the following weighted empirical risk  $\hat{\mathcal{R}}_{\psi}^{\ell}$ :

$$\hat{\mathcal{R}}_{\psi}^{\ell}(\mathcal{S}, f_{\theta}) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \psi_{\gamma}(1 - f_{\theta}(\mathbf{x}_i^+)) \cdot \psi_{\gamma}(f_{\theta}(\mathbf{x}_j^-)) \cdot \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-). \quad (3)$$

In this sense, adopting a smooth penalty function ends up with a dual soft weighting strategy over the hard instances. Again, to reach a proper weight function, we also require it to satisfy some necessary regularities. In the following, we define this class of functions as the calibrated weighting function.

**Definition 2.** A weighting function  $\psi_{\gamma}(x) : [0, 1] \rightarrow \text{Rng}$ , where  $\text{Rng} \subseteq [0, 1]$ , is called a **calibrated weighting function**, if it satisfies the following regularities:

- (A)  $\psi_{\gamma}$  has continuous second-order derivatives.
- (B)  $\psi_{\gamma}$  is strictly increasing in the sense that  $\psi'_{\gamma}(x) > 0$ .
- (C)  $\psi_{\gamma}$  is strictly concave in the sense that  $\psi''_{\gamma}(x) < 0$ .

In this definition, (B) is a natural requirement to make the weight proportional to the target instance's difficulty. Condition (C) is an interesting trait in our framework. To see why this is necessary, let us note that the weight functions  $v_i^+$ ,  $v_j^-$  are continuous surrogates residing in  $[0, 1]$  for threshold

function

$$\begin{aligned} & \mathbf{1} \left[ 1 - f_{\theta}(\mathbf{x}_i^+) > 1 - f_{\theta}(\mathbf{x}_{(n_+^\alpha)}^+) \right], \\ & \mathbf{1} \left[ f_{\theta}(\mathbf{x}_j^-) > f_{\theta}(\mathbf{x}_{(n_-^\beta)}^-) \right], \end{aligned}$$

respectively. To be simple, we continue our discussion with a general form  $\mathbf{1}[x > 0]$ . Obviously, weight decay for large  $x$  should be smooth such that the loss could attend at the top  $f_{\theta}(\mathbf{x}^+)$  and  $(1 - f_{\theta}(\mathbf{x}^-))$  scores. Moreover, to avoid overfitting, the model should as well have sufficient memory of the easy examples. Hence the weights for such examples should not be too close to zero. These observations are exactly typical traits for a concave function. As shown in Fig.2, we visualize the difference between a convex function  $y = x^2$  and  $y = \mathbf{1}[x > 0.5]$ , and the difference between concave function  $y = x^{0.05}$  and  $y = \mathbf{1}[x > 0.5]$ .

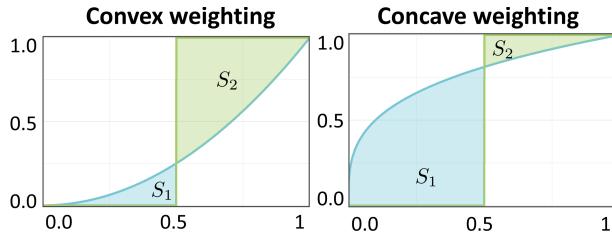


Figure 2. convex vs. concave weighting functions.

Another reason to choose concave functions is that they can benefit the optimization process. More precisely, we expect the loss function  $\hat{\mathcal{R}}_{\psi}^{\ell}$  in Eq. (3) to be an upper bound of  $\hat{\mathcal{R}}_{\alpha,\beta}^{\ell}$ , such that minimizing  $\hat{\mathcal{R}}_{\psi}^{\ell}$  could also minimize the original loss. Back to Fig.2, this is more likely to happen if  $S_1/S_2$  is large. In fact, this is a condition which is much easier for concave functions to satisfy. From a quantitative perspective, the following proposition provides a sufficient and a necessary condition under which  $\hat{\mathcal{R}}_{\psi}^{\ell} \geq \hat{\mathcal{R}}_{\alpha,\beta}^{\ell}$ . Moreover, it shows that it is generally more challenging for a convex function to realize an upper bound. **Please see Appendix.D for the proof.**

**Proposition 2.** Given a strictly increasing weighting function  $\psi_{\gamma} : [0, 1] \rightarrow [0, 1]$ , such that  $v_i^+ = \psi_{\gamma}(1 - f_{\theta}(\mathbf{x}_i^+))$ ,  $v_j^- = \psi_{\gamma}(f_{\theta}(\mathbf{x}_j^-))$ , denote:

$$\begin{aligned} \mathcal{I}_1^+ &= \left\{ x_+ : x_+ \in \mathcal{X}_P, f(x_+) \geq f(x_{(n_+^\alpha)}^+) \right\}, \\ \mathcal{I}_1^- &= \left\{ x_- : x_- \in \mathcal{X}_N, f(x_-) \leq f(x_{(n_-^\beta)}^-) \right\}, \end{aligned}$$

denote  $\mathcal{I}_2$  as  $(\mathcal{X}_P \times \mathcal{X}_N) \setminus (\mathcal{I}_1^+ \times \mathcal{I}_1^-)$ ; denote  $\bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+} [x]$  as the empirical expectation of  $x$  over the set  $\mathcal{I}_1^+$ , and  $\bar{\mathbb{E}}_{\mathbf{x}^- \in \mathcal{I}_1^-} [x]$ ,  $\bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-}$ ,  $\bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2}$  are defined sim-

ilarly; define  $l_{i,j} = \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-)$ . We assume that

$$n_+^\alpha \in \mathbb{N}, n_-^\beta \in \mathbb{N}, f_{\theta}(\mathbf{x}^+), f_{\theta}(\mathbf{x}^-) \in (0, 1),$$

then:

- (a) A sufficient condition for  $\hat{\mathcal{R}}_{\alpha,\beta}^{\ell}(\mathcal{S}, f_{\theta}) \leq \hat{\mathcal{R}}_{\psi}^{\ell}(\mathcal{S}, f_{\theta})$  is that:

$$\sup_{p \in (0,1), q = -\frac{p}{1-p}} [\rho_p - \xi_q] \geq 0,$$

where

$$\begin{aligned} \rho_p &= \frac{(\bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} [v_+^p \cdot v_-^p])^{1/p}}{\left( \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} [(1 - v_+ v_-)^2] \right)^{1/2}}, \\ \xi_q &= \frac{\alpha \beta}{1 - \alpha \beta} \cdot \frac{(\bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} (\ell_{i,j}^2))^{1/2}}{\left( \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} (\ell_{i,j}^q) \right)^{1/q}}. \end{aligned}$$

- (b) If there exists at least one strictly concave  $\psi_{\gamma}$  such that the  $\hat{\mathcal{R}}_{\alpha,\beta}^{\ell}(\mathcal{S}, f_{\theta}) > \hat{\mathcal{R}}_{\psi}^{\ell}(\mathcal{S}, f_{\theta})$ , then  $\hat{\mathcal{R}}_{\alpha,\beta}^{\ell}(\mathcal{S}, f_{\theta}) > \hat{\mathcal{R}}_{\psi}^{\ell}(\mathcal{S}, f_{\theta})$  holds for all convex  $\psi_{\gamma}$ .

According to Prop.2,  $\hat{\mathcal{R}}_{\psi}^{\ell}$  can achieve the upper bound of  $\hat{\mathcal{R}}_{\alpha,\beta}^{\ell}$  if  $\alpha, \beta$  are small, and the empirical distribution has significant masses at instances with moderate difficulty.

**Dual Correspondence Theory.** Now with the penalty function and weighting function clarified, we establish their dual correspondence with the following proposition. **Please see Appendix.E for the proof.**

**Proposition 3.** Given a strictly convex function  $\varphi_{\gamma}$ , and define  $\psi_{\gamma}(t)$  as:

$$\psi_{\gamma}(t) = \operatorname{argmax}_{v \in [0,1]} v \cdot t - \varphi_{\gamma}(v),$$

then we can draw the following conclusions:

- (a) If  $\varphi_{\gamma}$  is a calibrated smooth penalty function, we have  $\psi_{\gamma}(t) = \varphi_{\gamma}'^{-1}(t)$ , which is a calibrated weighting function.  
 (b) If  $\psi_{\gamma}$  is a calibrated weighting function such that  $v = \psi_{\gamma}(t)$ , we have  $\varphi_{\gamma}(v) = \int \psi_{\gamma}^{-1}(v) dv + \text{const.}$ , which is a calibrated smooth penalty function.

According to Prop.3, given a calibrated smooth penalty function, one can obtain an implicit soft weighting strategy via Prop.3-(a). Likewise, given a calibrated weighting function, one can find an implicit regularizer over the sample weights via Prop.3-(b). Based on the regularities of the two

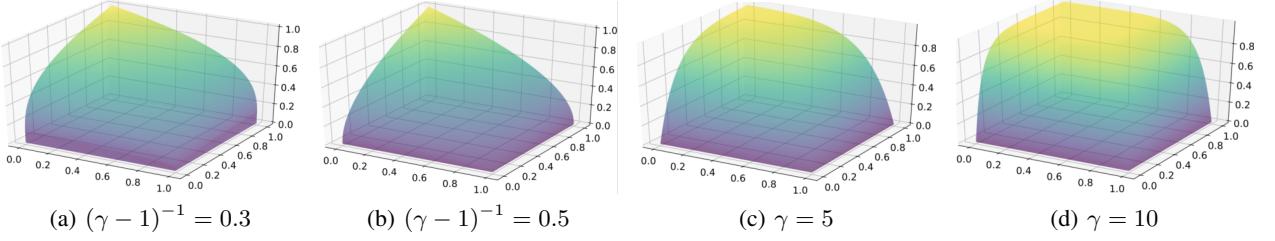


Figure 3. Visualization of the Landscape of the pairwise weights  $\psi_\gamma(x) \cdot \psi_\gamma(y)$ . Here, (a) and (b) plot  $\psi_\gamma^{\text{poly}}$ , while (c) and (d) plot  $\psi_\gamma^{\text{exp}}$ .

components, both  $\psi$  and  $\varphi$  have closed-form formulation if the other one is known. This means that we can solve the bi-level optimization framework in  $(OP_1)$  by simply minimizing the resulting  $\hat{\mathcal{R}}_\psi^\ell(\mathcal{S}, f_\theta)$ , leading to a much simpler optimization that can be solved directly by an end-to-end training framework. *Consequently, the dual correspondence theory provides a simple way to establish a surrogate optimization problem of TPAUC, once a weighting function or a penalty function is at hand.*

#### 4.2. Two Instantiations of the Generic Framework

Based on the generic framework, in this subsection, we provide two practical instantiations.

**Polynomial Surrogate Model.** From the penalty function perspective, the original  $\ell_1$  penalty realizes  $\psi_\gamma = \gamma \cdot t$ . In this way, it is a natural choice to adopt a polynomial penalty  $\varphi_\gamma^{\text{poly}}(t) = C \cdot t^\gamma$  as a dense surrogate for  $\ell_1$ . Inspired by this, we propose a polynomial surrogate model as example 1.

**Example 1** (Polynomial Surrogate Model). *In the polynomial surrogate model, we set:*

$$\varphi_\gamma^{\text{poly}}(t) = \frac{1}{\gamma} \cdot t^\gamma, \quad \psi_\gamma^{\text{poly}}(t) = t^{\frac{1}{\gamma-1}}, \quad \gamma > 1$$

The visualizations of the weights are shown in Fig.3.

**Example 2** (Exponential Surrogate Model). *In the exponential surrogate model, we set:*

$$\varphi_\gamma^{\text{exp}}(t) = \frac{(1-t)(\log(1-t)-1)+1}{\gamma}, \quad \psi_\gamma^{\text{exp}}(t) = 1 - e^{-\gamma t}$$

**Exponential Surrogate Model.** Considering the properties of the weighting functions, we expect that  $\psi_\gamma$  will have a flat landscape for large  $t$ . Motivated by this, we adopt an exponential weighting function  $\psi_\gamma^{\text{exp}}(t) = 1 - e^{-\gamma t}$  (the landscape is shown in Fig.3 (c)-(d)). The resulting model is then shown as Exp.2. The visualizations of the weights are shown in Fig.3.

#### 4.3. Generalization Analysis

In this subsection, we turn to explore how generalization error behaves away from the training error in terms of the TPAUC metric. In other words, we will show when a well-trained model will lead to a reasonable generalization performance. Our analysis is based on a standard assumption that the classifiers are chosen from a hypothesis class  $\mathcal{F}$  (e.g. the class of a specific type of deep neural networks). The key challenge here is that  $\hat{\mathcal{R}}_{\alpha,\beta}^\ell(\mathcal{S}, f_\theta)$  is not an unbiased estimation of  $\text{AUC}_\alpha^\beta(f_\theta, \mathcal{S})$ , making standard generalization analysis (Mohri et al., 2018) unavailable. Here we extend the error decomposition technique for OPAUC (Narasimhan & Agarwal, 2017b) and employ the result in Prop.2 to reach the following theorem. *Please see Appendix.F for the proof.*

**Theorem 1 (Informal).** *Assume that there are no ties in the datasets, and the surrogate loss function  $\ell$  with range  $[0, 1]$ , is an upper bound of the 0-1 loss, then, for all  $f_\theta \in \mathcal{F}$ , and all  $\alpha, \beta \in (0, 1)$  such that condition (a) of Proposition 2 holds, the following inequality holds with high probability:*

$$\mathcal{R}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) \leq \hat{\mathcal{R}}_\psi^\ell(f_\theta, \mathcal{S}) + \tilde{\mathcal{O}} \left( \left( \frac{VC}{n_+} \right)^{1/2} + \left( \frac{VC}{n_-} \right)^{1/2} \right),$$

where  $\tilde{\mathcal{O}}$  is the big-O complexity notation hiding the logarithm factors,  $\mathcal{R}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) = 1 - \text{AUC}_\alpha^\beta(f_\theta, \mathcal{S})$ , and  $VC$  is the VC dimension of the hypothesis class:

$$\mathcal{T}(\mathcal{F}) \triangleq \{\text{sign}(f_\theta(\cdot) - \delta) : f_\theta \in \mathcal{F}, \delta \in \mathbb{R}\}.$$

According to the theorem, for all  $\alpha, \beta$  satisfying condition (a) of Prop.2 and any model in  $\mathcal{F}$ , the generalization error represented by the loss version of TPAUC  $\mathcal{R}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) = 1 - \text{AUC}_\alpha^\beta(f_\theta, \mathcal{S})$  is no larger than the empirical loss  $\hat{\mathcal{R}}_\psi^\ell(f_\theta, \mathcal{S})$  plus a complexity term. The complexity term is affected by two factors. On one hand, it vanishes with large enough training datasets. On the other hand, it remains moderate if the model hypothesis class's VC dimension is not too large. Moreover, moderate upper bounds for the VC dimension are now available for typical models ranging from linear models to deep neural

networks. Finally, for a well-trained model, the empirical loss  $\hat{\mathcal{R}}_\psi^\ell(f_\theta, \mathcal{S})$  is restricted to be small in our framework; one can then reach reasonable generalization results with high probability.

## 5. Experiments

In this section, we present our empirical results and some of the details of the experiments. *Please see Appendix G for more details on the settings and results.*

### 5.1. Competitors

To validate the effectiveness of our proposed methods, we consider two types of competitors in our experiments. On one hand, we compare our proposed methods with other methods dealing with imbalanced data. The competitors include class-reweighted **CE**, **Focal** (Lin et al., 2017), **CB-CE** (Cui et al., 2019), and **CE-Focal** (Cui et al., 2019). On the other hand, we also include a standard AUC optimization method as our baseline. Here we use the square function  $\ell_{sq}(t) = (1-t)^2$  as the surrogate loss, which is widely-adopted in AUC optimization studies. The resulting competitor is named **SqAUC**. Finally, we implement our polynomial surrogate model and the exponential surrogate model on top of **SqAUC**, which are denoted as **Poly** and **Exp** in the rest of this section.

### 5.2. Evaluation Metrics

Aiming at optimizing the TPAUC metrics, we consider TPAUC with  $\alpha = 0.3, \beta = 0.3$ ,  $\alpha = 0.4, \beta = 0.4$ ,  $\alpha = 0.5, \beta = 0.5$ , respectively. Moreover, to normalize the range of their magnitude to  $[0, 1]$ , we adopt the following variant of the TPAUC metric:

$$\text{TPAUC}(\alpha, \beta) = 1 - \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \frac{\ell_{0,1}\left(f_\theta(\mathbf{x}_{(i)}^+) - f(\mathbf{x}_{(j)}^-)\right)}{n_+^\alpha n_-^\beta}.$$

### 5.3. Dataset Description

Note that AUC is aimed at dealing with binary classification problems, hence we construct long-tail binary datasets as follows.

**Binary CIFAR-10-LT Dataset.** We create a long-tailed CIFAR-10 dataset, where the sample sizes across different classes decay exponentially, and the ratio of sample sizes of the least frequent to the most frequent class  $\rho$  is set to 0.01. Afterwards, we create 3 binary long-tailed datasets based on CIFAR-10-LT by selecting one category as positive examples and the others as negative examples.

**Binary CIFAR-100-LT Dataset.** We create 3 CIFAR-100-LT subsets in the same way as CIFAR-10-LT, where a superclass is selected as positive examples each time.

**Binary Tiny-ImageNet-200-LT Dataset.** The original Tiny-ImageNet-200 dataset contains 100,000 colour images sourced from 200 different categories, with 500 images for each category. Similar to the CIFAR-100-LT dataset, we choose 3 positive superclasses to construct binary subsets.

### 5.4. Warm-Up Training Phase With Delay Epochs

Focusing on the hard examples at the beginning of the training process brings a high risk of over-fitting. It is thus necessary to focus on the entire dataset to capture the global information. Inspired by this investigation, we adopt a warm-up training strategy. Specifically, the model will go through a warm-up phase with  $E_k$  Epochs of ordinary AUC optimization training. Afterward, we start the TPAUC training phase by optimizing our proposed surrogate problems. We will show its effect in the next subsection.

### 5.5. Overall Performance

The performances of all the involved methods on three subsets of CIFAR-10-LT, CIFAR-100-LT, and Tiny-Imagenet-200-LT are shown in Tab.1. For each method here, the results for different TPAUC metrics are tuned independently. Consequently, we have the following observations: 1) The best performance of our proposed methods consistently surpasses all the competitors significantly on all the datasets, except the result for TPAUC(0.4, 0.4) and TPAUC(0.5, 0.5) on subset 2 of Tiny-Imagenet-200-LT and TPAUC(0.3, 0.3) for subset 3 of Tiny-Imagenet-200-LT. Our proposed methods attain fairly competitive results compared with the competitors even for the three failure results. 2) The performance improvement is especially sharp on TPAUC(0.3, 0.3). This suggests the ability of our proposed methods to optimize the head region under the ROC curve.

### 5.6. Sensitivity Analysis

Our proposed framework evolves two hyperparameters:  $\gamma$  for loss functions and  $E_k$  for the warm-up strategy. Next, we analyze their effect respectively. Our analysis is based on a 2d grid search over  $E_k, \gamma$ . When investigating the effect of  $E_k$  ( $\gamma$  resp.), we will show the performance variation in terms of  $\gamma$  ( $E_k$  resp.) given each fixed  $E_k$  ( $\gamma$  resp.).

**Effect of  $E_k$ .** In Fig.4-(a), (c), we show the sensitivity in terms of  $E_k$  on subset 2 of CIFAR-10-LT for **Exp** and **Poly**, respectively. For **Exp**, we see that increasing  $E_k$  from 5 to 30 leads to a significantly increasing trend of performance. This shows that a warm-up phase is necessary for **Exp**. For **Poly**, we observe that the increasing trend of average performance is much weaker. This is probably because  $\gamma$  has a strong influence on the performance so that the variances become much larger in general.

**Effect of  $\gamma$ .** In Fig.4-(b), (d), we show the sensitivity in

Table 1. Performance Comparisons over different metrics and datasets, where  $(x, y)$  stands for  $\text{TPAUC}(x, y)$  in short.

dataset	type	methods	Subset1			Subset2			Subset3		
			(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)	(0.3,0.3)	(0.4,0.4)	(0.5,0.5)
CIFAR-10-LT	Competitors	CE-RW	9.09	30.86	47.99	72.83	83.33	88.71	23.47	44.44	59.69
		Focal	9.84	30.89	50.83	75.72	85.10	90.06	21.47	45.88	59.09
		CBCE	3.29	27.30	43.95	69.48	80.80	86.87	12.94	34.06	51.09
		CBFocal	9.04	31.73	48.13	77.99	86.75	91.13	21.32	43.03	59.11
		SqAUC	18.05	40.74	57.94	80.09	87.78	91.87	31.52	50.00	64.42
	Ours	Poly	<b>21.43</b>	<b>44.41</b>	<b>59.10</b>	<b>80.66</b>	<b>88.07</b>	<b>92.15</b>	<b>36.54</b>	<b>54.48</b>	<b>67.19</b>
		Exp	<b>20.86</b>	<b>41.78</b>	<b>58.38</b>	<b>81.22</b>	<b>87.88</b>	<b>91.93</b>	<b>32.47</b>	<b>53.86</b>	<b>67.32</b>
CIFAR-100-LT	Competitors	CE-RW	31.43	52.60	66.21	79.70	88.06	92.64	3.09	21.32	40.75
		Focal	36.51	61.71	73.25	83.08	90.35	93.76	8.09	28.88	49.89
		CBCE	17.53	38.79	55.19	67.91	79.32	85.82	1.84	18.46	37.04
		CBFocal	41.85	62.41	73.13	82.75	89.57	92.89	7.10	29.12	44.84
		SqAUC	63.24	76.62	84.68	91.02	93.69	94.73	41.60	60.36	70.86
	Ours-TPAUC	Poly	<b>68.02</b>	<b>79.11</b>	<b>85.17</b>	<b>91.13</b>	<b>93.78</b>	<b>95.69</b>	<b>47.07</b>	<b>65.89</b>	<b>75.08</b>
		Exp	63.24	77.94	84.62	90.69	93.74	95.41	44.54	64.58	73.02
Tiny-ImageNet-200-LT	Competitors	CE-RW	80.90	87.76	91.54	93.30	96.15	97.53	90.37	94.34	96.75
		Focal	81.18	88.06	91.72	93.23	96.08	97.59	91.35	94.87	96.63
		CBCE	80.64	87.58	91.17	93.77	<b>96.52</b>	<b>97.77</b>	91.66	95.19	96.79
		CBFocal	80.44	87.95	91.91	93.46	<b>96.43</b>	97.64	91.06	94.82	96.62
		SqAUC	80.16	87.99	91.67	93.10	96.07	97.32	<b>92.15</b>	95.16	96.75
	Ours-TPAUC	Poly	80.44	<b>88.21</b>	91.98	93.00	95.61	97.47	<b>92.02</b>	<b>95.25</b>	<b>96.84</b>
		Exp	<b>82.61</b>	<b>89.13</b>	<b>92.62</b>	<b>93.82</b>	96.12	97.38	91.25	94.78	96.57

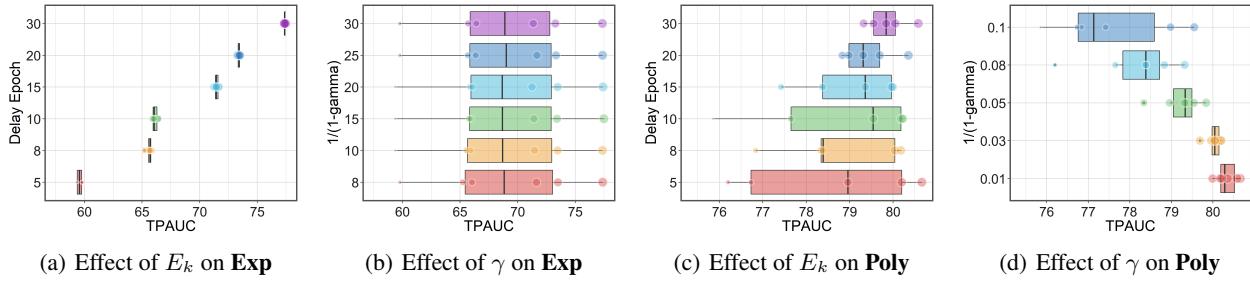


Figure 4. Sensitivity analysis on subset 2 of CIFAR-10-LT where TPAUC is measured with  $\alpha = 0.3, \beta = 0.3$ . For each box in (a) and (c),  $E_k$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $\gamma$ . For each Box in (b) and (d),  $(\gamma - 1)^{-1}$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $E_k$ .

terms of  $\gamma$  on subset 2 of CIFAR-10-LT for **Exp** and **Poly**, respectively. One can observe very different trends on these two methods. This is because that **Exp** and **Poly** have different characteristics in terms of the landscape of the weight function. As shown in Fig.3-(c), (d), the weight landscape of **Exp** is flat within a large subset of the domain. In this sense, it does not have a strong dependency on  $\gamma$ . As shown in Fig.3-(a), (b), the weight landscape of **Poly** is more sensitive toward  $\gamma$ .

## 6. Conclusion

In this paper, we initiate the study on TPAUC optimization. Since the original optimization problem could not be solved directly with an end-to-end framework, we propose a general framework to construct surrogate optimization problems for TPAUC. Following our dual correspondence theory, we can establish a surrogate problem once a calibrated penalty

function or a calibrated weighting function is found. To see how and when our framework could provide efficient approximations of the original problem, we show that the surrogate objective function could reach the upper bound of the original one and that concave weighting functions are better choices than their convex counterparts. Moreover, we also provide high probability uniform upper bounds for the generalization error. The experiments on three datasets consistently show the advantage of our framework.

## 7. Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China: 61931008, 61620106009, 61836002, U2001202, U1736219, and 61976202, in part by Youth Innovation Promotion Association CAS, and in part by the Strategic Priority Re-

search Program of Chinese Academy of Sciences, Grant No. XDB28000000.

## References

- Agarwal, S. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15(1):1653–1674, 2014.
- Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.
- Alan, A. H. and Raskutti, B. Optimising area under the roc curve using gradient descent. *International Conference on Machine Learning*, pp. 49–56, 2004.
- Calders, T. and Jaroszewicz, S. Efficient auc optimization for classification. *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 42–53, 2007.
- Clémençon, S., Lugosi, G., Vayatis, N., et al. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.
- Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. *Advances in Neural Information Processing Systems*, pp. 313–320, 2003.
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.
- Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.
- Gao, W. and Zhou, Z. On the consistency of AUC pairwise optimization. *International Joint Conference on Artificial Intelligence*, pp. 939–945, 2015.
- Gao, W., Wang, L., Jin, R., Zhu, S., and Zhou, Z. One-pass auc optimization. *International Conference on Machine Learning*, pp. 906–914, 2013.
- Hand, D. J. and Till, R. J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
- Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- Hao, H., Fu, H., Xu, Y., Yang, J., Li, F., Zhang, X., Liu, J., and Zhao, Y. Open-narrow-synechiae anterior chamber angle classification in AS-OCT sequences. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- Joachims, T. A support vector method for multivariate performance measures. *International Conference on Machine Learning*, pp. 377–384, 2005.
- Joachims, T. Training linear svms in linear time. *the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 217–226, 2006.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- Liu, C., Zhong, Q., Ao, X., Sun, L., Lin, W., Feng, J., He, Q., and Tang, J. Fraud transactions detection via behavior tree with local intention calibration. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020a.
- Liu, R., Li, Z., Zhang, Y., Fan, X., and Luo, Z. Bi-level probabilistic feature learning for deformable image registration. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020b.
- Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020c.
- Liu, W., Luo, W., Lian, D., and Gao, S. Future frame prediction for anomaly detection - A new baseline. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545, 2018.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. Foundations of machine learning. 2018.
- Narasimhan, H. and Agarwal, S. A structural SVM based approach for optimizing partial AUC. *International Conference on Machine Learning*, pp. 516–524, 2013a.
- 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*, pp. 167–175, 2013b.

- Narasimhan, H. and Agarwal, S. A structural svm based approach for optimizing partial auc. In *International Conference on Machine Learning*, pp. 516–524, 2013c.
- Narasimhan, H. and Agarwal, S. Support vector algorithms for optimizing the partial area under the ROC curve. *Neural Computation*, 29(7):1919–1963, 2017a.
- Narasimhan, H. and Agarwal, S. Support vector algorithms for optimizing the partial area under the roc curve. *Neural computation*, 29(7):1919–1963, 2017b.
- Natole, M., Ying, Y., and Lyu, S. Stochastic proximal algorithms for auc maximization. *International Conference on Machine Learning*, pp. 3707–3716, 2018.
- Natole, M. A., Ying, Y., and Lyu, S. Stochastic auc optimization algorithms with linear convergence. *Frontiers in Applied Mathematics and Statistics*, 5:30, 2019.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. pp. 8024–8035, 2019.
- Ralaivola, L., Szafranski, M., and Stempfel, G. Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary  $\beta$ -mixing processes. *Journal of Machine Learning Research*, 11(Jul):1927–1956, 2010.
- Sutskever, I., Martens, J., Dahl, G. E., and Hinton, G. E. On the importance of initialization and momentum in deep learning. In *ICML 2013*, pp. 1139–1147, 2013.
- Usunier, N., Amini, M.-R., and Gallinari, P. A data-dependent generalisation error bound for the auc. *ICML Workshop on ROC Analysis in Machine Learning*, 2005.
- Usunier, N., Amini, M. R., and Gallinari, P. Generalization error bounds for classifiers trained with interdependent data. *Advances in Neural Information Processing Systems*, pp. 1369–1376, 2006.
- Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., and Qi, Y. A semi-supervised graph attentive network for financial fraud detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 2019.
- Wu, H., Hu, Z., Jia, J., Bu, Y., He, X., and Chua, T.-S. Mining unfollow behavior in large-scale online social networks via spatial-temporal interaction. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 254–261, 2020.
- 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.
- Ying, Y., Wen, L., and Lyu, S. Stochastic online auc maximization. *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.
- Zhang, X., Saha, A., and Vishwanathan, S. Smoothing multivariate performance measures. *Journal of Machine Learning Research*, 13(1):3623–3680, 2012.
- Zhao, P., Hoi, S. C., Jin, R., and Yang, T. Online auc maximization. In *International Conference on Machine Learning*, pp. 233–240, 2011.
- Zhou, K., Gao, S., Cheng, J., Gu, Z., Fu, H., Tu, Z., Yang, J., Zhao, Y., and Liu, J. Sparse-gan: Sparsity-constrained generative adversarial network for anomaly detection in retinal oct image. In *IEEE International Symposium on Biomedical Imaging*, pp. 1227–1231, 2020.

# Appendices

## Contents

<b>A Review of General AUC Optimization Methods</b>	<b>12</b>
<b>B Inconsistency between OPAUC and TPAUC</b>	<b>12</b>
<b>C Proof of Proposition 1</b>	<b>12</b>
<b>D Proof of Proposition 2</b>	<b>13</b>
<b>E Proof of Proposition 3</b>	<b>15</b>
<b>F Proof of Theorem 1</b>	<b>17</b>
<b>G Experiments</b>	<b>20</b>
G.1 Competitors . . . . .	20
G.2 General Implementation Details . . . . .	20
G.3 Dataset Description . . . . .	21
G.4 Sensitivity Analysis . . . . .	22

## A. Review of General AUC Optimization Methods

As a motivating early study, (Cortes & Mohri, 2003) points out that maximizing AUC should not be replaced with minimizing the error rate, which shows the necessity to study direct AUC optimization methods. After that, a series of algorithms have been designed for optimizing AUC. At the early stage, the majority of studies focus on a full-batch off-line setting. (Alan & Raskutti, 2004; Calders & Jaroszewicz, 2007) optimize AUC based on a logistic surrogate loss function and ordinary gradient descent method. RankBoost (Freund et al., 2003) provides an efficient ensemble-based AUC learning method based on a ranking extension of the AdaBoost algorithm. The work of (Joachims, 2006; Zhang et al., 2012) constructs  $SVM^{struct}$ -based frameworks that optimize a direct upper bound of the  $0 - 1$  loss version AUC metric instead of its surrogates. Later on, to accommodate big data analysis, researchers start to explore online extensions of AUC optimization methods. (Zhao et al., 2011) provides an early trial for this direction based on the reservoir sampling technique. (Gao et al., 2013) provides a completely one-pass AUC optimization method for streaming data based on the squared surrogate loss. Most recently, (Ying et al., 2016) reformulates the squared-loss-based stochastic AUC maximization problem as a stochastic saddle point problem. The new saddle point problem's objective function only involves summations of instance-wise loss terms, which significantly reduces the burden from the pairwise formulation. (Natole et al., 2018; 2019) further accelerate this framework with tighter convergence rates. Beyond optimization methods, a substantial amount of researches also provide theoretical support for this learning framework from different dimensions, including generalization analysis (Agarwal et al., 2005; Cléménçon et al., 2008; Usunier et al., 2005; 2006; Ralaivola et al., 2010) and consistency analysis (Agarwal, 2014; Gao & Zhou, 2015). In this paper, we take a further step to optimize the two-way partial AUCs.

## B. Inconsistency between OPAUC and TPAUC

In this section, we show the inconsistency between TPAUC metric and the OPAUC metric. Mathematically, OPAUC calculates the partial AUC within FPR range  $[\alpha, \beta]$ , which could be defined as:

$$AUC_{\alpha}^{\beta OP}(f_{\theta}) = \int_{\alpha}^{\beta} TPR_{f_{\theta}}(FPR_{f_{\theta}}^{-1}(t)) dt.$$

Recall that TPAUC could be defined as:

$$AUC_{\alpha}^{\beta TP}(f_{\theta}) = \int_{FPR_{f_{\theta}}(TPR_{f_{\theta}}^{-1}(1-\alpha))}^{\beta} TPR_{f_{\theta}}(FPR_{f_{\theta}}^{-1}(t)) dt - (1-\alpha) \cdot (\beta - FPR_{f_{\theta}}(TPR_{f_{\theta}}^{-1}(1-\alpha))).$$

From the definitions, we can find that the TPAUC is intrinsically inconsistent with OPAUC. The source of the inconsistency is that both  $FTR_{f_{\theta}}$  and  $TPR_{f_{\theta}}$  are functions of  $f_{\theta}$ . It is thus impossible to regard  $FPR_{f_{\theta}}^{-1}(1-\alpha)$  and  $TPR_{f_{\theta}}(FPR_{f_{\theta}}^{-1}(1-\alpha))$  as constants, even though  $\alpha$  is fixed. Thus one cannot simply replace the FPR lower bound  $FPR_{f_{\theta}}^{-1}(1-\alpha)$  with any constant  $c$ . Consequently,  $AUC_{\alpha}^{\beta}(f_{\theta})$  is in general not consistent with any OPAUC with FPR range  $[c, \beta]$ . The readers are also referred to (Yang et al., 2019) for illustrative analysis of why OPAUC could not be used to approximate TPAUC.

## C. Proof of Proposition 1

First we need the following lemma to finish the proof:

**Lemma 1.** For  $\{t_i\}_{i=1}^n$  with  $t_i \geq 0$ , assume that  $\min_{i \neq j} |t_i - t_j| > 0$ . Then for the problem:

$$\max_{v_i \in [0,1], \sum_{i=1}^n v_i \leq k} \sum_{i=1}^{n+} v_i \cdot t_i,$$

the unique solution is  $v_i^* = \mathbf{1}[t_i \geq t_{(k)}^{\downarrow}]$ , where  $k < n$ ,  $k \in \mathbb{N}_+$ ,  $t_{(k)}^{\downarrow}$  is top  $k$ -th element in  $\{t_i\}_{i=1}^n$ .

*Proof.* For a set of weights  $\{v_i\}_{i=1}^n$ , let us denote  $v_{(i)}^{\downarrow}$  as the weight for  $t_i^{\downarrow}$ . For any  $\{v'_i\}_{i=1}^n \neq \{v_i^*\}_{i=1}^n$ . We can write down the difference between the objective functions as:

$$\begin{aligned}
 & \sum_{i=1}^n (v_i^* - v'_i) \cdot t_i \\
 &= \sum_{i \leq k} (1 - v'_{(i)} \downarrow) \cdot t_{(i)}^\downarrow - \sum_{j > k} v'_{(j)} \downarrow \cdot t_{(j)}^\downarrow \\
 &\stackrel{(*)}{>} (k - \sum_{i \leq k} v'_{(i)} \downarrow) \cdot t_{(k)}^\downarrow - \sum_{j > k} v'_{(j)} \downarrow \cdot t_{(j)}^\downarrow \\
 &\stackrel{(**)}{>} (k - \sum_{i \leq k} v'_{(i)} \downarrow) \cdot t_{(k)}^\downarrow - \sum_{j > k} v'_{(j)} \downarrow \cdot t_{(k)}^\downarrow \\
 &= (k - \sum_{i=1}^n v'_{(i)} \downarrow) \cdot t_{(k)}^\downarrow \\
 &\geq 0,
 \end{aligned}$$

where  $(*)$ ,  $(**)$ , follows the assumption that  $\min_{i \neq j} |t_i - t_j| > 0$ . Note that since the  $\{v'_i\}_{i=1}^n$  is arbitrarily chosen, the proof is thus completed.  $\square$

**Reminder of Proposition 1.** For any  $\alpha, \beta \in (0, 1)$ , if scores  $f_{\theta}(\mathbf{x}) \in [0, 1]$ , and there are no ties in the scores, the original optimization problem is equivalent to the following problem:

$$\begin{aligned}
 & \min_{\theta} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-) \\
 \text{s.t. } & v_+ = \operatorname{argmax}_{v_i^+ \in [0, 1], \sum_{i=1}^{n_+} v_i^+ \leq n_+^\alpha} \sum_{i=1}^{n_+} (v_i^+ \cdot (1 - f_{\theta}(\mathbf{x}_i^+))) \\
 & v_- = \operatorname{argmax}_{v_j^- \in [0, 1], \sum_{j=1}^{n_-} v_j^- \leq n_-^\beta} \sum_{j=1}^{n_-} (v_j^- \cdot f_{\theta}(\mathbf{x}_j^-))
 \end{aligned}$$

*Proof.* First it is easy to see that  $(OP_0)$  could be formulated as follows:

$$\begin{aligned}
 & \min_{\theta} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-) \\
 \text{s.t. } & v_i^+ = \begin{cases} 1, & 1 - f_{\theta}(\mathbf{x}_i^+) \geq 1 - f_{\theta}(\mathbf{x}_{(n_+^\alpha)}^+) \\ 0, & \text{otherwise} \end{cases} \\
 & v_j^- = \begin{cases} 1, & f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}_{(n_-^\beta)}^-) \\ 0, & \text{otherwise} \end{cases}
 \end{aligned}$$

Then the rest of the proof follows Lem.1 directly.  $\square$

## D. Proof of Proposition 2

**Lemma 2** (Hölder's Inequality).  $\forall p > 1, q > 1$  such that  $1/p + 1/q = 1$ , we have:

$$\bar{\mathbb{E}} [|XY|] \leq (\bar{\mathbb{E}} [|X|^p])^{1/p} \cdot (\bar{\mathbb{E}} [|Y|^q])^{1/q}$$

**Lemma 3.**  $\forall 0 < p < 1, q = -p/(1-p)$ , we have:

$$\bar{\mathbb{E}} [|X'Y'|] \geq (\bar{\mathbb{E}} [|X'|^p])^{1/p} \cdot (\bar{\mathbb{E}} [|Y'|^q])^{1/q}$$

*Proof.* It could be proved by applying Lem.2 to  $X = |X'Y'|^p$  and  $Y = |Y'|^{-p}$ .  $\square$

**Reminder of Proposition 2.** Given a strictly increasing weighting function  $\psi_\gamma : [0, 1] \rightarrow [0, 1]$ , such that  $v_i^+ = \psi_\gamma(1 - f_{\theta}(x_i^+)), v_j^- = \psi_\gamma(f_{\theta}(x_j^-)), \psi_\gamma(0) = 0, \psi_\gamma(1) = 1$  denote:

$$\begin{aligned}\mathcal{I}_1^+ &= \left\{ x_+ : x_+ \in \mathcal{X}_P, f(x_+) \geq f(x^{(n_+^\alpha)}) \right\}, \\ \mathcal{I}_1^- &= \left\{ x_- : x_- \in \mathcal{X}_N, f(x_-) \leq f(x^{(n_-^\beta)}) \right\},\end{aligned}$$

denote  $\mathcal{I}_2$  as  $(\mathcal{X}_P \times \mathcal{X}_N) \setminus (\mathcal{I}_1^+ \times \mathcal{I}_1^-)$ ; denote  $\bar{\mathbb{E}}_{x^+ \in \mathcal{I}_1^+}[x]$  as the empirical expectation of  $x$  over the set  $\mathcal{I}_1^+$ , and  $\bar{\mathbb{E}}_{x^- \in \mathcal{I}_1^-}[x], \bar{\mathbb{E}}_{x^+ \in \mathcal{I}_1^+, x^- \in \mathcal{I}_1^-}, \bar{\mathbb{E}}_{x^+, x^- \in \mathcal{I}_2}$  are defined similarly. Without loss of generality, we assume that  $n_+^\alpha \in \mathbb{N}, n_-^\beta \in \mathbb{N}$ . We have:

(a) A sufficient condition for  $\hat{\mathcal{R}}_{\alpha,\beta}^\ell(\mathcal{S}, f_{\theta}) \leq \hat{\mathcal{R}}_\psi^\ell(\mathcal{S}, f_{\theta})$  is that:

$$\sup_{p \in (0,1), q = -\frac{p}{1-p}} [\rho_p - \xi_q] \geq 0$$

where

$$\begin{aligned}\rho_p &= \frac{(\bar{\mathbb{E}}_{x^+, x^- \in \mathcal{I}_2} [v_+^p \cdot v_-^p])^{1/p}}{\left( \bar{\mathbb{E}}_{x^+ \in \mathcal{I}_1^+, x^- \in \mathcal{I}_1^-} [(1 - v_+ v_-)^2] \right)^{1/2}} \\ \xi_q &= \frac{\alpha \beta}{1 - \alpha \beta} \cdot \frac{(\bar{\mathbb{E}}_{x^+, x^- \in \mathcal{I}_2} (\ell_{i,j}^2))^{1/2}}{\left( \bar{\mathbb{E}}_{x^+ \in \mathcal{I}_1^+, x^- \in \mathcal{I}_1^-} (\ell_{i,j}^q) \right)^{1/q}}\end{aligned}$$

(b) If there exists at least one strictly concave  $\psi_\gamma$  such that the  $\hat{\mathcal{R}}_{\alpha,\beta}^\ell(\mathcal{S}, f_{\theta}) > \hat{\mathcal{R}}_\psi^\ell(\mathcal{S}, f_{\theta})$ , then  $\hat{\mathcal{R}}_{\alpha,\beta}^\ell(\mathcal{S}, f_{\theta}) > \hat{\mathcal{R}}_\psi^\ell(\mathcal{S}, f_{\theta})$  holds for all convex  $\psi_\gamma$ .

*Proof.* First,  $l_{i,j} = \ell(f_{\theta}, \mathbf{x}_i^+, \mathbf{x}_j^-)$ , we can reformulate  $\hat{\mathcal{R}}_{\psi}^{\ell} - \hat{\mathcal{R}}_{\alpha,\beta}^{\ell}$  as follows.

$$\begin{aligned}
 \hat{\mathcal{R}}_{\psi}^{\ell} - \hat{\mathcal{R}}_{\alpha,\beta}^{\ell} &= \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell_{i,j} - \frac{1}{n_+ n_-} \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \ell_{i,j} \\
 &= \frac{1}{n_+ \cdot n_-} \cdot \sum_{\mathbf{x}_i^+, \mathbf{x}_j^- \in \mathcal{I}_2} v_i^+ \cdot v_j^- \cdot \ell_{i,j} - \frac{1}{n_+ \cdot n_-} \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} (1 - v_i^+ v_j^-) \ell_{i,j} \\
 &= (1 - \alpha\beta) \cdot \frac{1}{|\mathcal{I}_2|} \cdot \sum_{\mathbf{x}_i^+, \mathbf{x}_j^- \in \mathcal{I}_2} v_i^+ \cdot v_j^- \cdot \ell_{i,j} - (\alpha \cdot \beta) \cdot \frac{1}{|\mathcal{I}_1|} \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} (1 - v_i^+ v_j^-) \ell_{i,j} \\
 &= (1 - \alpha\beta) \cdot \bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} [v_+ \cdot v_- \cdot \ell] - \alpha \cdot \beta \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} [(1 - v_+ \cdot v_-) \cdot \ell]
 \end{aligned} \tag{4}$$

Now we prove (a)-(b) based on this result.

(a) According to Lem.3,  $\forall 1 > p > 0$ ,  $q = -p/(1-p)$ , we have:

$$(1 - \alpha\beta) \cdot \bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} [v_+ \cdot v_- \cdot \ell] \geq \underbrace{(1 - \alpha\beta) \cdot \left( \bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} [v_+^p \cdot v_-^p] \right)^{1/p} \cdot \left( \bar{\mathbb{E}}_{\mathbf{x}^+, \mathbf{x}^- \in \mathcal{I}_2} [\ell^q] \right)^{1/q}}_{(a)}$$

Meanwhile, we have:

$$\alpha \cdot \beta \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} [(1 - v_+ \cdot v_-) \cdot \ell] \leq \underbrace{\alpha \cdot \beta \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} [(1 - v_+ \cdot v_-)^2]^{1/2} \cdot \bar{\mathbb{E}}_{\mathbf{x}^+ \in \mathcal{I}_1^+, \mathbf{x}^- \in \mathcal{I}_1^-} [\ell^2]^{1/2}}_{(b)}$$

This shows that  $(a) - (b) \geq 0$  implies  $\hat{\mathcal{R}}_{\psi}^{\ell} \geq \hat{\mathcal{R}}_{\alpha,\beta}^{\ell}$ . Moreover,  $(a) - (b) \geq 0$  is equivalent to  $\rho_p - \xi_q \geq 0$ . The proof of (a) is ended since  $p$  and  $q$  are arbitrarily chosen within their domain.

(b) Given a strictly concave function  $\psi_{\gamma} : [0, 1] \rightarrow [0, 1]$  and a convex function  $\tilde{\psi}_{\gamma} : [0, 1] \rightarrow [0, 1]$ . We have that

$$\forall y \in [0, 1], \psi_{\gamma}(y) = \psi_{\gamma}(0 \cdot (1-y) + y \cdot 1) > y \cdot \psi_{\gamma}(1) = y$$

$$\forall y \in (0, 1), \tilde{\psi}_{\gamma}(y) = \tilde{\psi}_{\gamma}(0 \cdot (1-y) + y \cdot 1) \leq y \cdot \tilde{\psi}_{\gamma}(1) = y$$

This implies that  $\psi_{\gamma}(y) > \tilde{\psi}_{\gamma}(y)$ ,  $\forall y \in (0, 1)$ . The proof then follows that

$$\hat{\mathcal{R}}_{\psi}^{\ell} - \hat{\mathcal{R}}_{\alpha,\beta}^{\ell} \propto \min_{i,j} [v_i^+ \cdot v_j^-] = \min_{i,j} [\psi(f_{\theta}(\mathbf{x}^+)) \cdot \psi(1 - f_{\theta}(\mathbf{x}^-))]$$

and  $f_{\theta}(\mathbf{x}^+), f_{\theta}(\mathbf{x}^-) \in (0, 1)$ .

□

## E. Proof of Proposition 3

**Reminder of Proposition 3.** Given a strictly convex function  $\varphi_{\gamma}$ , and define  $\psi_{\gamma}(t)$  as:

$$\psi_{\gamma}(t) = \operatorname{argmax}_{v \in [0,1]} v \cdot t - \varphi_{\gamma}(v)$$

then we can draw the following conclusions:

- (a) If  $\varphi_{\gamma}$  is a calibrated smooth penalty function, we have  $\psi_{\gamma}(t) = \varphi_{\gamma}'^{-1}(t)$ , which is a calibrated weighting function.
- (b) If  $\psi_{\gamma}$  is a calibrated weighting function such that  $v = \psi_{\gamma}(t)$ , we have  $\varphi_{\gamma}(v) = \int \psi_{\gamma}^{-1}(v) dv + \text{const.}$ , which is a

calibrated smooth penalty function.

*Proof.*

- (a) Since  $\varphi_\gamma$  is strictly convex,  $v \cdot t - \varphi_\gamma(v)$  is strictly concave, then  $\psi_\gamma$  has a unique global optimal solution. To reach the optimal solution, we have:

$$(v \cdot t - \varphi_\gamma(v))' = t - \varphi'_\gamma(v) = 0$$

Note that  $v = \varphi_\gamma(t)$ , we have:

$$t - \varphi'_\gamma(\psi_\gamma(t)) = 0$$

Equivalently, note that  $\varphi'_\gamma(t)$  is invertible since it is strictly increasing ( $\varphi''_\gamma(t) > 0$ ), we have:

$$\psi_\gamma(t) = \varphi'^{-1}_\gamma(t)$$

Moreover, we have:

$$\psi'_\gamma(t) = \frac{1}{\varphi''(\varphi'^{-1}_\gamma(t))}, \quad \psi''_\gamma(t) = -\frac{\varphi'''(\varphi'^{-1}_\gamma(t))}{(\varphi''(\varphi'^{-1}_\gamma(t)))^3}.$$

Since  $\varphi''_\gamma(x) > 0, \varphi'''_\gamma(x) > 0$ , we know that  $\psi'_\gamma(t)$  is a calibrated weighting function according to the definition.

- (b) Assume that  $\psi_\gamma(t)$  is the solution of the optimization problem, recall the optimal condition:

$$t - \varphi'_\gamma(v) = 0$$

Since  $t = \psi^{-1}(v)$ , we have:

$$\psi^{-1}(v) = \varphi'_\gamma(v)$$

leading to the fact that

$$\int \psi^{-1}(v) dv = \varphi_\gamma(v)$$

Moreover, we have:

$$\varphi'_\gamma(v) = \psi^{-1}_\gamma(v), \quad \varphi''_\gamma(v) = \frac{1}{\psi'_\gamma(\psi^{-1}(v))}, \quad \varphi'''_\gamma(v) = -\frac{\psi''_\gamma(\psi^{-1}(v))}{(\psi'_\gamma(\psi^{-1}(v)))^3}$$

Since  $\psi^{-1}_\gamma(x) > 0, \psi'_\gamma(x) > 0$  and  $\psi''_\gamma(x) < 0, \varphi_\gamma$  is then a calibrated weighting function according to the definition.

□

## F. Proof of Theorem 1

First, we need the following definitions about the population and empirical quantile of the scores:

$$\begin{aligned}\delta_\alpha &= \underset{\delta \in \mathbb{R}}{\operatorname{argmin}} \left[ \delta \in \mathbb{R} : \underset{\mathbf{x}^+ \sim \mathcal{P}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) \leq \delta]] = \alpha \right], \quad \hat{\delta}_\alpha = \underset{\delta \in \mathbb{R}}{\operatorname{argmin}} \left[ \delta \in \mathbb{R} : \frac{1}{n_+} \sum_{i=1}^{n_+} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) \leq \delta]] = \alpha \right] \\ \delta_\beta &= \underset{\delta \in \mathbb{R}}{\operatorname{argmin}} \left[ \delta \in \mathbb{R} : \underset{\mathbf{x}^- \sim \mathcal{N}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) \geq \delta]] = \beta \right], \quad \hat{\delta}_\beta = \underset{\delta \in \mathbb{R}}{\operatorname{argmin}} \left[ \delta \in \mathbb{R} : \frac{1}{n_-} \sum_{j=1}^{n_-} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) \geq \delta]] = \beta \right]\end{aligned}$$

Furthermore, we denote the loss version population-level  $1 - \text{AUC}_\alpha^\beta(f_{\boldsymbol{\theta}})$  and empirical TPAUC  $1 - \hat{\text{AUC}}_\alpha^\beta(f_{\boldsymbol{\theta}})$  as:

$$\begin{aligned}\mathcal{R}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) &= \underset{\mathbf{x}^- \sim \mathcal{N}}{\mathbb{E}} \underset{\mathbf{x}^+ \sim \mathcal{P}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) > f_{\boldsymbol{\theta}}(\mathbf{x}^-)] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) < \delta_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) > \delta_\beta]] \\ \hat{\mathcal{R}}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) &= \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq f_{\boldsymbol{\theta}}(\mathbf{x}_i^+)] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_i^+) \leq \hat{\delta}_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \hat{\delta}_\beta]\end{aligned}$$

**Lemma 4.** For  $\forall f \in \mathcal{F}$ , we have:

$$\mathcal{R}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) - \hat{\mathcal{R}}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) \leq 2(\Delta_+ + \Delta_-)$$

where

$$\begin{aligned}\Delta_+ &= \sup_{\delta \in \mathbb{R}} \left| \frac{1}{n_+} \cdot \sum_{i=1}^{n_+} \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_i^+) \leq \delta] - \underset{\mathbf{x}^+ \sim \mathcal{P}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) \leq \delta]] \right| \\ \Delta_- &= \sup_{\delta \in \mathbb{R}} \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \delta] - \underset{\mathbf{x}^- \sim \mathcal{N}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) \geq \delta]] \right|\end{aligned}$$

*Proof.* First, we define some intermediate variables:

$$\begin{aligned}\ell_+(f_{\boldsymbol{\theta}}, \mathbf{x}_j^-) &= \underset{\mathbf{x}^+ \sim \mathcal{P}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) \leq \delta_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq f_{\boldsymbol{\theta}}(\mathbf{x}^+)]] \\ R_1 &= \mathcal{R}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) = \underset{\mathbf{x}^- \sim \mathcal{N}}{\mathbb{E}} \underset{\mathbf{x}^+ \sim \mathcal{P}}{\mathbb{E}} [\mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) \geq f_{\boldsymbol{\theta}}(\mathbf{x}^+)] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^+) \leq \delta_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}^-) \geq \delta_\beta]] \\ R_2 &= \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \ell_+(f_{\boldsymbol{\theta}}, \mathbf{x}_j^-) \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \delta_\beta] \\ R_3 &= \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \ell_+(f_{\boldsymbol{\theta}}, \mathbf{x}_j^-) \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \hat{\delta}_\beta] \\ R_4 &= \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq f_{\boldsymbol{\theta}}(\mathbf{x}_i^+)] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_i^+) \leq \delta_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \hat{\delta}_\beta] \\ R_5 &= \hat{\mathcal{R}}_{AUC}^{\alpha, \beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) = \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq f_{\boldsymbol{\theta}}(\mathbf{x}_i^+)] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_i^+) \leq \hat{\delta}_\alpha] \cdot \mathbf{1} [f_{\boldsymbol{\theta}}(\mathbf{x}_j^-) \geq \hat{\delta}_\beta]\end{aligned}$$

In this sense, we can decompose  $R_1 - R_5$  as:

$$|R_1 - R_5| \leq |R_1 - R_2| + |R_2 - R_3| + |R_3 - R_4| + |R_4 - R_5|$$

Now, we bound each term in the equation above successively. For  $|R_1 - R_2|$ , we have:

$$\begin{aligned}
 |R_1 - R_2| &= \left| \mathbb{E}_{\mathbf{x}^+ \sim \mathcal{P}} \left[ \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq f_{\theta}(\mathbf{x}^+)] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}^+) \leq \delta_{\alpha}] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq \delta_{\beta}]] \right. \right. \\
 &\quad \left. \left. - \frac{1}{n_-} \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}^+)] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}^+) \leq \delta_{\alpha}] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] \right] \right| \\
 &\leq \left| \sup_{\mathbf{x}^+} \left[ \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq \max\{f_{\theta}(\mathbf{x}^+), \delta_{\beta}\}]] - \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \max\{f_{\theta}(\mathbf{x}_i^+), \delta_{\beta}\}] \right] \right| \\
 &\leq \sup_{\delta \in \mathbb{R}} \left| \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq \delta]] - \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta] \right|
 \end{aligned}$$

For  $|R_2 - R_3|$ , we have:

$$\begin{aligned}
 |R_2 - R_3| &= \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \ell_+(f_{\theta}, \mathbf{x}_j^-) \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] - \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \ell_+(f_{\theta}, \mathbf{x}_j^-) \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right| \\
 &\stackrel{(a_1)}{\leq} \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] - \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right| \\
 &\stackrel{(a_2)}{=} \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] - \beta \right| \\
 &\stackrel{(a_3)}{=} \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] - \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq \delta_{\beta}]] \right| \\
 &\leq \sup_{\delta \in \mathbb{R}} \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta] - \mathbb{E}_{\mathbf{x}^- \sim \mathcal{N}} [\mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq \delta]] \right|
 \end{aligned}$$

Here, (a<sub>1</sub>) follows from the fact that  $\mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \delta_{\beta}] - \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}]$  must be simultaneously  $\geq 0$  or  $\leq 0$ ; (a<sub>2</sub>) and (a<sub>3</sub>) are based on the definition of  $\delta_{\beta}$  and  $\hat{\delta}_{\beta}$  and the assumption that no tie occurs in the dataset.

For  $|R_3 - R_4|$ , we have:

$$\begin{aligned}
 |R_3 - R_4| &= \left| \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \ell_+(f_{\theta}, \mathbf{x}_j^-) \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right. \\
 &\quad \left. - \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}_i^+)] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta_{\alpha}] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right| \\
 &\leq \frac{1}{n_-} \cdot \sum_{j=1}^{n_-} \left| \ell_+(f_{\theta}, \mathbf{x}_j^-) - \frac{1}{n_+} \cdot \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \min\{f_{\theta}(\mathbf{x}_j^-), \delta_{\alpha}\}] \right| \\
 &\leq \sup_{\delta \in \mathbb{R}} \left| \frac{1}{n_+} \cdot \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta] - \mathbb{E}_{\mathbf{x}^+ \sim \mathcal{P}} [\mathbf{1}[f_{\theta}(\mathbf{x}^+) \leq \delta]] \right|
 \end{aligned}$$

For  $|R_4 - R_5|$ , we have:

$$\begin{aligned}
 |R_4 - R_5| &= \left| \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}_i^+)] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta_{\alpha}] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right. \\
 &\quad \left. - \frac{1}{n_+ n_-} \cdot \sum_{j=1}^{n_-} \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}_i^+)] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \hat{\delta}_{\alpha}] \cdot \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right| \\
 &\leq \frac{1}{n_-} \cdot \left( \sum_{j=1}^{n_-} \left| \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq \hat{\delta}_{\beta}] \right| \cdot \left| \frac{1}{n_+} \cdot \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_j^-) \geq f_{\theta}(\mathbf{x}_i^+)] \cdot \left( \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta_{\alpha}] - \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \hat{\delta}_{\alpha}] \right) \right| \right) \\
 &\leq \frac{1}{n_+} \sup_{\mathbf{x}^-} \left[ \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}^-) \geq f_{\theta}(\mathbf{x}_i^+)] \cdot \left( \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta_{\alpha}] - \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \hat{\delta}_{\alpha}] \right) \right] \\
 &\stackrel{(b_1)}{\leq} \frac{1}{n_+} \left| \sum_{i=1}^{n_+} \left( \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta_{\alpha}] - \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \hat{\delta}_{\alpha}] \right) \right| \\
 &\stackrel{(b_2)}{\leq} \sup_{\delta \in \mathbb{R}} \left| \frac{1}{n_+} \cdot \sum_{i=1}^{n_+} \mathbf{1}[f_{\theta}(\mathbf{x}_i^+) \leq \delta] - \mathbb{E}_{\mathbf{x}^+ \sim \mathcal{P}} [\mathbf{1}[f_{\theta}(\mathbf{x}^+) \leq \delta]] \right|
 \end{aligned}$$

Here  $(b_1)$  and  $(b_2)$  follow a similar argument to  $(a_1)$ - $(a_3)$ .  $\square$

---

**Reminder of Theorem 1.** Assume that there are no ties in the datasets, and the surrogate loss function  $\ell$  with range  $[0, 1]$ , is an upper bound of the 0-1 loss, then, for all  $f_{\theta} \in \mathcal{F}$ , and all  $(\alpha, \beta) \in \mathcal{I}_{suff}(\mathcal{S})$ , the following inequality holds with probability at least  $1 - \delta$  over the choice of  $\mathcal{S}$ :

$$\mathcal{R}_{AUC}^{\alpha, \beta}(f_{\theta}, \mathcal{S}) \leq \hat{\mathcal{R}}_{\psi}^{\ell}(f_{\theta}, \mathcal{S}) + C \left( \sqrt{\frac{\text{VC} \cdot \log(n_+) + \log(1/\delta)}{n_+}} + \sqrt{\frac{\text{VC} \cdot \log(n_-) + \log(1/\delta)}{n_-}} \right),$$

where  $\text{VC}$  is the VC dimension of the hypothesis class:

$$\mathcal{T}(\mathcal{F}) \triangleq \{\text{sign}(f_{\theta}(\cdot) - \delta) : f_{\theta} \in \mathcal{F}, \delta \in \mathbb{R}\}$$

and

$$\mathcal{I}_{suff}(\mathcal{S}) = \left\{ (\alpha, \beta) : \alpha \in (0, 1), \beta \in (0, 1), n_+^{\alpha} \in \mathbb{N}_+, n_-^{\beta} \in \mathbb{N}_+, \text{condition (a) in Prop.2 holds} \right\},$$

---

*Proof.* First, we have:

$$\begin{aligned}
 &\mathbb{P} \left[ \sup_{f \in \mathcal{F}, (\alpha, \beta) \in \mathcal{I}_{suff}(\mathcal{S})} [|\mathcal{R}_{AUC}^{\alpha, \beta}(f_{\theta}, \mathcal{S}) - \hat{\mathcal{R}}_{AUC}^{\alpha, \beta}(f_{\theta}, \mathcal{S})|] > \epsilon \right] \\
 &\leq \mathbb{P} \left[ \sup_{f \in \mathcal{F}, (\alpha, \beta) \in \mathcal{I}_{suff}(\mathcal{S}), \delta \in \mathbb{R}} [\Delta_+] > \epsilon/4 \right] + \mathbb{P} \left[ \sup_{f \in \mathcal{F}, (\alpha, \beta) \in \mathcal{I}_{suff}(\mathcal{S}), \delta \in \mathbb{R}} [\Delta_-] > \epsilon/4 \right] \\
 &= \mathbb{P} \left[ \sup_{f \in \mathcal{F}, \delta \in \mathbb{R}} [\Delta_+] > \epsilon/4 \right] + \mathbb{P} \left[ \sup_{f \in \mathcal{F}, \delta \in \mathbb{R}} [\Delta_-] > \epsilon/4 \right]
 \end{aligned}$$

Following Lem.1 in (Narasimhan & Agarwal, 2017b), we have that, for all  $f_{\theta} \in \mathcal{T}(\mathcal{F})$ , and all  $\alpha, \beta \in (0, 1)$  s.t.

$n_+^\alpha \in \mathbb{N}_+$ ,  $n_-^\beta \in \mathbb{N}_+$ , the following inequality holds with probability at least  $1 - \delta$ :

$$\mathcal{R}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) \leq \hat{\mathcal{R}}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) + C \left( \sqrt{\frac{\text{VC} \cdot \log(n_+) + \log(1/\delta)}{n_+}} + \sqrt{\frac{\text{VC} \cdot \log(n_-) + \log(1/\delta)}{n_-}} \right).$$

Since  $\alpha, \beta \in \mathcal{I}_{suff}(\mathcal{S})$ ,  $\hat{\mathcal{R}}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) \leq \hat{\mathcal{R}}_{\alpha,\beta}^\ell(f_\theta, \mathcal{S}) \leq \hat{\mathcal{R}}_\psi^\ell(f_\theta, \mathcal{S})$ , we have the following inequality holds with probability at least  $1 - \delta$  under the same condition:

$$\mathcal{R}_{AUC}^{\alpha,\beta}(f_\theta, \mathcal{S}) \leq \hat{\mathcal{R}}_\psi^\ell(f_\theta, \mathcal{S}) + C \left( \sqrt{\frac{\text{VC} \cdot \log(n_+) + \log(1/\delta)}{n_+}} + \sqrt{\frac{\text{VC} \cdot \log(n_-) + \log(1/\delta)}{n_-}} \right).$$

□

## G. Experiments

### G.1. Competitors

To validate the effectiveness of our proposed methods, we consider two types of competitors in our experiments. On one hand, we compare our proposed methods with other methods dealing with imbalanced data:

1. **CE**: Here use a class-wise reweighted version of the CE loss as one of our competitors, the sample weight is set to  $1/n_y$ , where  $n_y$  the frequency of the class the sample belongs to.
2. **Focal**: (Lin et al., 2017) It tackles the imbalance problem by adding a modulating factor to the cross-entropy loss to highlight the hard and minority samples during the training process.
3. **CB-CE**: It refers to the loss function that applies the reweighting scheme proposed in (Cui et al., 2019) on the cross-entropy loss.
4. **CB-Focal**: It refers to the loss function that applies the reweighting scheme proposed in (Cui et al., 2019) on the Focal loss.

On the other hand, we also include standard AUC optimization methods as our baseline.

1. **SqAUC**: Perform a standard AUC optimization with the surrogate loss function  $\ell_{sq}(t) = (1 - t)^2$ .

Finally, we implement our proposed methods on top of SqAUC:

1. **Poly**: Perform TPAUC optimization with the objective function:

$$\frac{1}{n_+^\alpha n_-^\beta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \psi_\gamma^{\text{poly}}(1 - f_\theta(\mathbf{x}_i^+)) \cdot \psi_\gamma^{\text{poly}}(f_\theta(v_j^-)) \cdot \ell(f_\theta, \mathbf{x}_i^+, \mathbf{x}_j^-)$$

2. **Exp**: Perform TPAUC optimization with the objective function:

$$\frac{1}{n_+^\alpha n_-^\beta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \psi_\gamma^{\text{Exp}}(1 - f_\theta(\mathbf{x}_i^+)) \cdot \psi_\gamma^{\text{Exp}}(f_\theta(v_j^-)) \cdot \ell(f_\theta, \mathbf{x}_i^+, \mathbf{x}_j^-)$$

## G.2. General Implementation Details

All the experiments are carried out on a ubuntu 16.04.1 server equipped with Intel(R) Xeon(R) Silver 4110 CPU and a TITAN RTX GPU, and all codes are implemented with PyTorch (v-1.4.0) (Paszke et al., 2019) under python 3.7 environment. Stochastic Gradient Descent (SGD) (Sutskever et al., 2013) with Nesterov momentum is adopted to optimize the objective function. Empirically, for all datasets, the learning rate is  $10^{-3}$ ; the  $l_2$  regularization term is set as  $10^{-5}$ , and the Nesterov momentum is 0.9. We also employ an exponential learning rate decay scheduler to adjust the learning rate after each training epoch, where the learning rate decay rate is set as 0.99 for all methods. The training batch size is 128, and we restrict the ratio of positive and negative samples by 1 : 10 in each batch. The batch size of validation/test examples is 256. Specifically,  $E_k$  is searched in {3, 5, 8, 10, 12, 15, 18, 20}. For **Poly**,  $\gamma$  is searched in {0.03, 0.05, 0.08, 0.1, 1, 3, 5}. For **Exp**,  $\gamma$  is searched in {8, 10, 15, 20, 25, 30}. Finally, we select the model based on the best validation performance and report the test set results.

## G.3. Dataset Description

Table 2. Details on the datasets.

Dataset	Pos. Class ID	Pos. Class Name	# Pos. Examples	# Neg. Examples
CIFAR-10-LT-1	2	birds	1,508	8,907
CIFAR-10-LT-2	1	automobiles	2,517	7,898
CIFAR-10-LT-3	3	cats	904	9,511
CIFAR-100-LT-1	6, 7, 14, 18, 24	insects	1,928	13,218
CIFAR-100-LT-2	0, 51, 53, 57, 83	fruits and vegetables	885	14,261
CIFAR-100-LT-3	15, 19, 21, 32, 38	large omnivores and herbivores	1,172	13,974
Tiny-ImageNet-200-LT-1	24, 25, 26, 27, 28, 29	dogs	2,100	67,900
Tiny-ImageNet-200-LT-2	11, 20, 21, 22	birds	1,400	68,600
Tiny-ImageNet-200-LT-3	70, 81, 94, 107, 111, 116, 121, 133, 145, 153, 164, 166	vehicles	4,200	65,800

**Binary CIFAR-10-LT Dataset.** The original CIFAR-10 dataset consists of 60,000  $32 \times 32$  colour images in 10 classes, with 6,000 images per class. There are 50,000 and 10,000 images in the training set and the test set, respectively. We create a long-tailed CIFAR-10 where the sample sizes across different classes decay exponentially, and the ratio of sample sizes of the least frequent to the most frequent class  $\rho$  is set to 0.01. We then create binary long-tailed datasets based on CIFAR-10-LT by selecting one category as positive examples and the others as negative examples. We construct three binary subsets, in which the positive categories are **1**) birds, **2**) automobiles, and **3**) cats, respectively. The datasets are split into training, validation and test sets according to the ratio of 0.7 : 0.15 : 0.15. More details are provided in Tab. 2.

**Binary CIFAR-100-LT Dataset.** The original CIFAR-100 dataset is similar to CIFAR-10, except it has 100 classes with each containing 600 images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. We create CIFAR-100-LT in the same way as CIFAR-10-LT, and transform it into three binary long-tailed datasets by selecting a superclass as positive class examples each time. Specifically, the positive superclasses are **1**) fruits and vegetables, **2**) insects and **3**) large omnivores and herbivores, respectively. More details are provided in Tab. 2.

**Implementation details On CIFAR Datasets.** We utilize the ResNet-20 (He et al., 2015) as the backbone, which takes images with size  $32 \times 32 \times 3$  as input and outputs 64-d features. Then the features are mapped into  $[0, 1]$  with an FC layer and Sigmoid function. During the training phase, we apply data augmentation including random horizontal flipping (50%), random rotation (from  $-15^\circ$  to  $15^\circ$ ) and random cropping ( $32 \times 32$ ).

**Binary Tiny-ImageNet-200-LT Dataset.** The Tiny-ImageNet-200 dataset contains 100,000  $256 \times 256$  color images from 200 different categories, with 500 images per category. Similar to the CIFAR-100-LT dataset, we choose three positive superclasses to construct binary subsets: **1**) dogs, **2**) birds and **3**) vehicles. The datasets are further split into training, validation and test sets according to the ratio of 0.7 : 0.15 : 0.15. See Tab. 2 for more details.

**Implementation details On Tiny-ImageNet-200 .** The implementation details are basically the same with CIFAR-10-LT and CIFAR-100-LT datasets, except the backbone network is implemented with ResNet-18 (He et al., 2015), which takes images with size  $224 \times 224 \times 3$  as input and outputs 512-d features.

#### G.4. Sensitivity Analysis

In this subsection, we show the sensitivity analysis results for all subsets on CIFAR-10-LT. The results show similar trends as the analysis shown in the main paper.

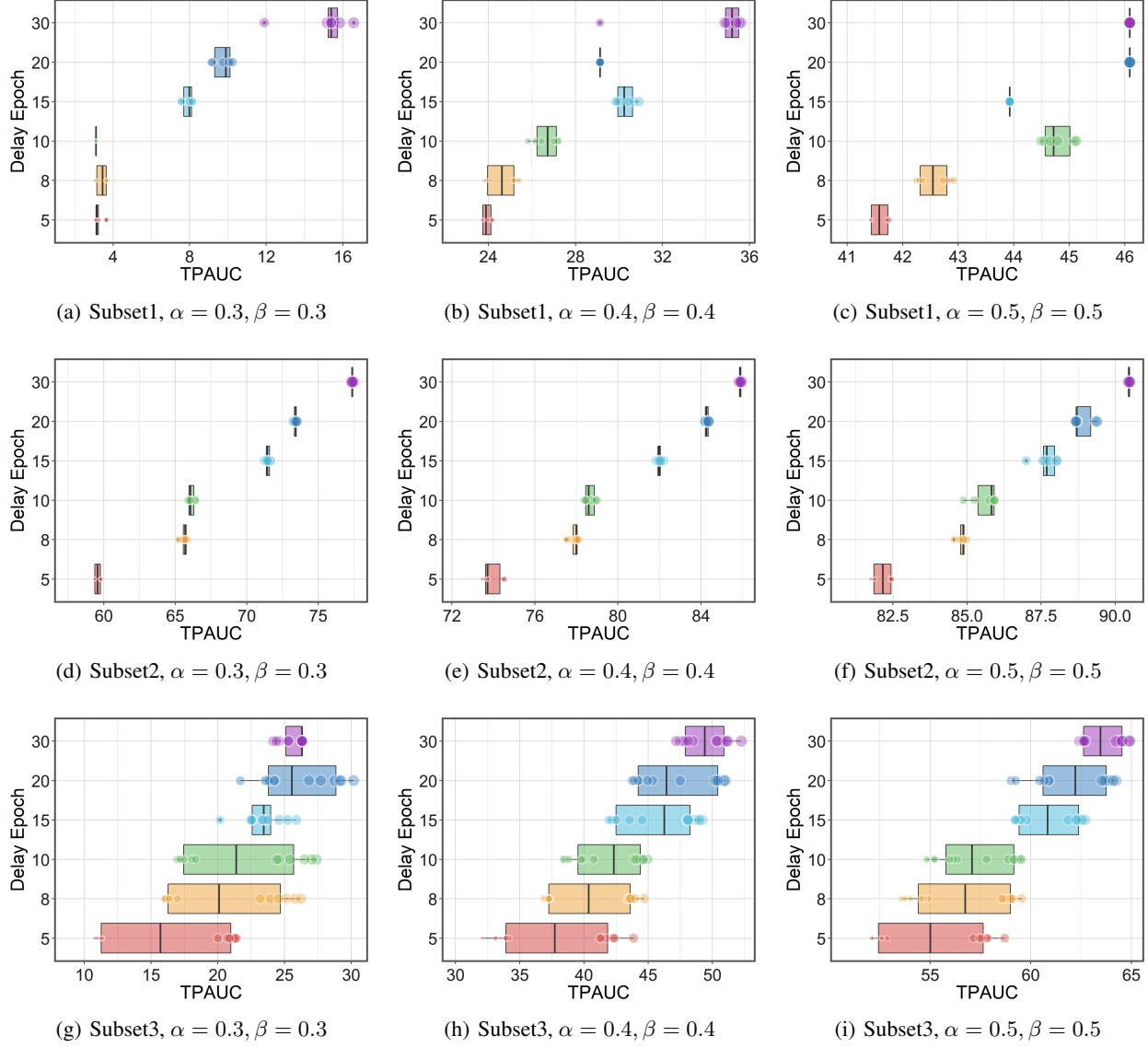


Figure 5. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Exp** with respect to  $E_k$ . For each Box in the plots,  $E_k$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $(\gamma - 1)^{-1}$ .

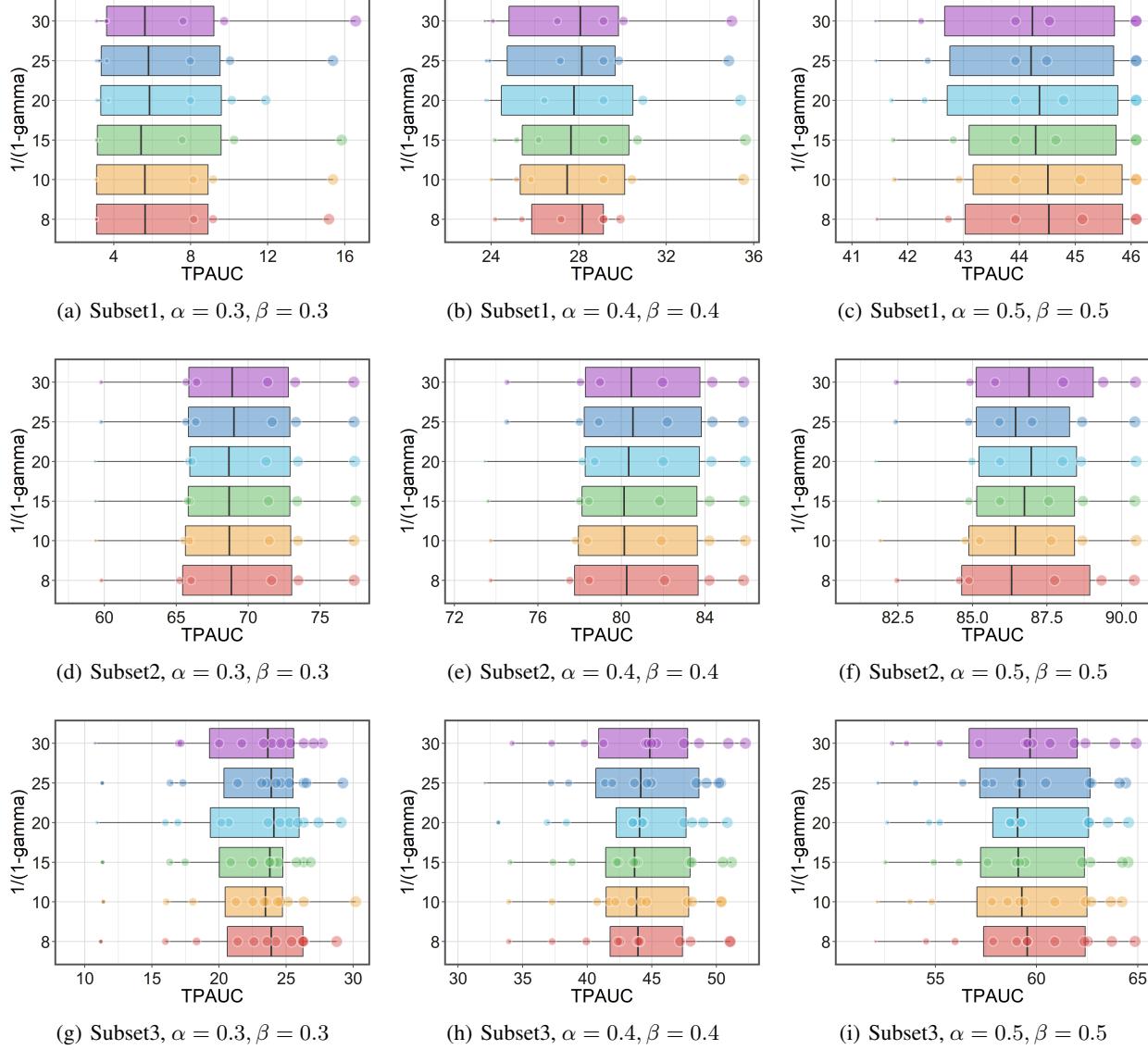


Figure 6. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Exp** with respect to  $\gamma$ . For each Box in the plots,  $(\gamma - 1)^{-1}$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $E_k$ .

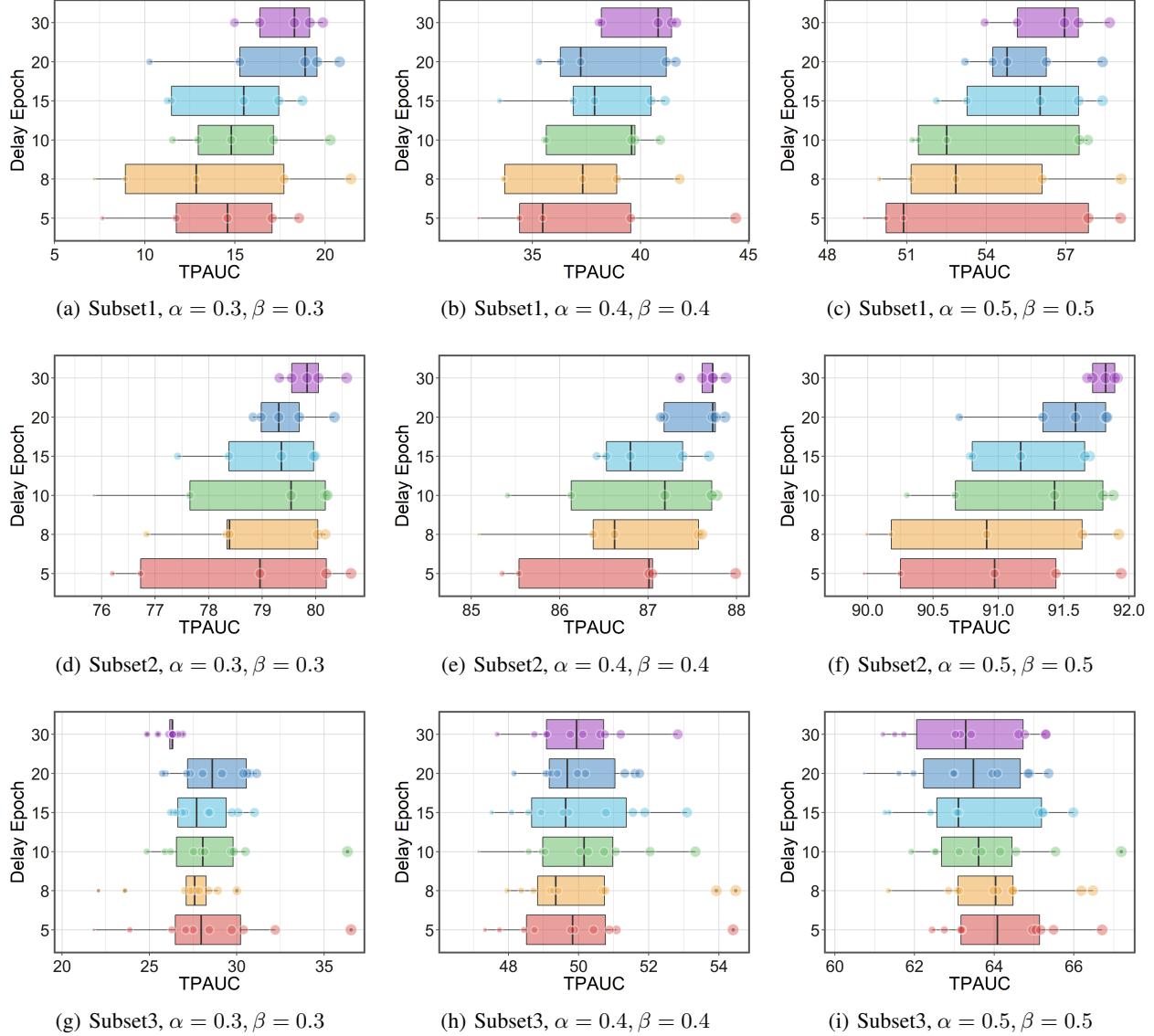


Figure 7. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Poly** with respect to  $E_k$ . For each Box in the plots,  $E_k$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $(\gamma - 1)^{-1}$ .

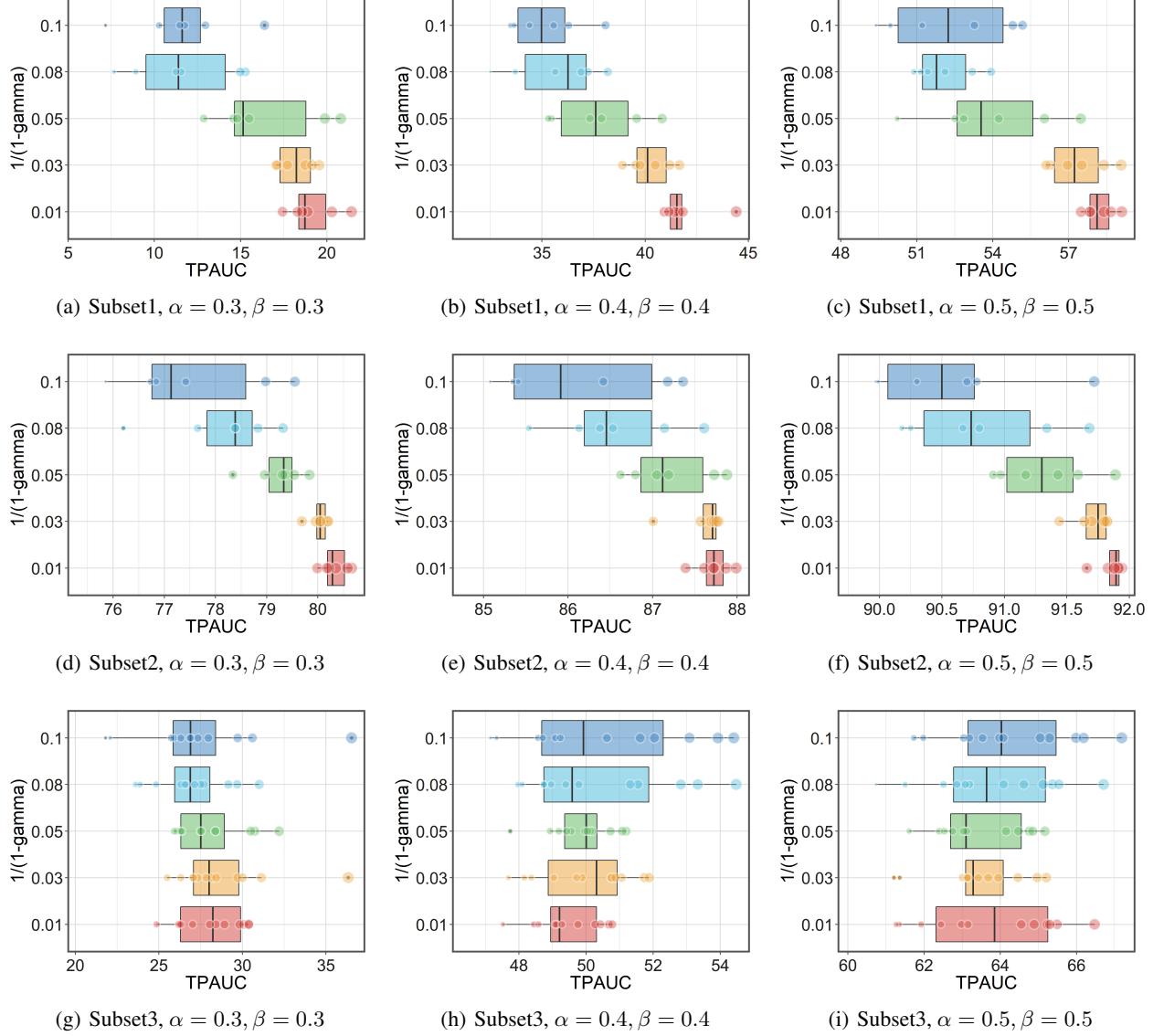


Figure 8. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Poly** with respect to  $\gamma$ . For each Box in the plots,  $(\gamma - 1)^{-1}$  is fixed as the y-axis value, and the scattered points along the box show the variation of  $E_k$ .