Appendices

Contents

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; Clémençon 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:

$$\mathsf{AUC}_{\alpha}^{\beta^{\mathsf{OP}}}(f_{\boldsymbol{\theta}}) = \int_{\alpha}^{\beta} \mathsf{TPR}_{f_{\boldsymbol{\theta}}}\left(\mathsf{FPR}_{f_{\boldsymbol{\theta}}}^{-1}(t)\right) dt.$$

Recall that TPAUC could be defined as:

$$\mathsf{AUC}_{\alpha}^{\beta^{\mathsf{TP}}}(f_{\boldsymbol{\theta}}) = \int_{\mathsf{FPR}_{f_{\boldsymbol{\theta}}}\left(\mathsf{TPR}_{f_{\boldsymbol{\theta}}}^{-1}(1-\alpha)\right)}^{\beta} \mathsf{TPR}_{f_{\boldsymbol{\theta}}}\left(\mathsf{FPR}_{f_{\boldsymbol{\theta}}}^{-1}(t)\right) dt - (1-\alpha)\cdot\left(\beta - \mathsf{FPR}_{f_{\boldsymbol{\theta}}}\left(\mathsf{TPR}_{f_{\boldsymbol{\theta}}}^{-1}(1-\alpha)\right)\right).$$

From the definitions, we can find that the TPAUC is intrinsically inconsistent with OPAUC. The source of the inconsistency is that both $\mathsf{FTR}_{f_{\boldsymbol{\theta}}}$ and $\mathsf{TPR}_{f_{\boldsymbol{\theta}}}$ are functions of $f_{\boldsymbol{\theta}}$. It is thus impossible to regard $\mathsf{FPR}_{f_{\boldsymbol{\theta}}}^{-1}(1-\alpha)$ and $\mathsf{TPR}_{f_{\boldsymbol{\theta}}}(\mathsf{FPR}_{f_{\boldsymbol{\theta}}}^{-1}(1-\alpha))$ as constants, even though α is fixed. Thus one cannot simply replace the FPR lower bound $\mathsf{FPR}_{f_{\boldsymbol{\theta}}}^{-1}(1-\alpha)$ with any constant c. Consequently, $\mathsf{AUC}_{\alpha}^{\beta}(f_{\boldsymbol{\theta}})$ 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 $v_i^\star = \mathbf{1}\left[t_i \geq t_{(k)}^{\downarrow}\right]$, 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 $\{v_i'\}_{i=1}^n \neq \{v_i^\star\}_{i=1}^n$. We can write down the difference between the objective functions as:

$$
\begin{aligned}
&\sum_{i=1}^{n}(v_i^\star - v_i') \cdot t_i \\
&\qquad = \sum_{i\le k}(1 - v_{(i)}^{'\downarrow}) \cdot t_{(i)}^{\downarrow} - \sum_{j>k} v_{(j)}^{'\downarrow} \cdot t_{(j)}^{\downarrow} \\
&\qquad \overset{(*)}{>} (k - \sum_{i\le k} v_{(i)}^{'\downarrow}) \cdot t_{(k)}^{\downarrow} - \sum_{j>k} v_{(j)}^{'\downarrow} \cdot t_{(j)}^{\downarrow} \\
&\qquad \overset{(**)}{>} (k - \sum_{i\le k} v_{(i)}^{'\downarrow}) \cdot t_{(k)}^{\downarrow} - \sum_{j>k} v_{(j)}^{'\downarrow} \cdot t_{(k)}^{\downarrow} \\
&\qquad = (k - \sum_{i=1}^{n} v_{(i)}^{'\downarrow}) \cdot t_{(k)}^{\downarrow} \\
&\qquad \ge 0,
\end{aligned}
$$

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. □

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

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

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

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

Then the rest of the proof follows Lem.1 directly. □

D. Proof of Proposition 2

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

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

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

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

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

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

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

denote $\mathcal{I}_2$ as $(\mathcal{X}_P \times \mathcal{X}_N) \backslash (\mathcal{I}_1^+ \times \mathcal{I}_1^-)$; denote $\bar{\mathbb{E}}_{\boldsymbol{x}^+ \in \mathcal{I}_1^+}[x]$ as the empirical expectation of x over the set $\mathcal{I}_1^+$, and $\bar{\mathbb{E}}_{\boldsymbol{x}^- \in \mathcal{I}_1^-}[x]$, $\bar{\mathbb{E}}_{\boldsymbol{x}^+ \in \mathcal{I}_1^+, \boldsymbol{x}^- \in \mathcal{I}_1^-}$, $\bar{\mathbb{E}}_{\boldsymbol{x}^+, \boldsymbol{x}^- \in \mathcal{I}_2}$ 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 $\hat{\mathcal{R}}_{\alpha,\beta}^\ell(\mathcal{S}, f_{\boldsymbol{\theta}}) \leq \hat{\mathcal{R}}_\psi^\ell(\mathcal{S}, f_{\boldsymbol{\theta}})$ is that:

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

where

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

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

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

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

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

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

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

Meanwhile, we have:

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

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 $\psi_\gamma : [0,1] \to [0,1]$ and a convex function $\widetilde{\psi}_\gamma : [0,1] \to [0,1]$. We have that

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

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

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

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

□

E. Proof of Proposition 3

Reminder of Proposition 3. *Given a strictly convex function φ_γ, 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 φ_γ is a calibrated smooth penalty function, we have $\psi_\gamma(t) = \varphi_\gamma'^{-1}(t)$,which is a calibrated weighting function.

(b) If ψ_γ is a calibrated weighting function such that $v = \psi_\gamma(t)$, we have $\varphi_\gamma(v) = \int \psi_\gamma^{-1}(v)dv + const.$, which is a

calibrated smooth penalty function.

Proof.

(a) Since φ_γ is strictly convex, $v \cdot t - \varphi_\gamma(v)$ is strictly concave, then ψ_γ 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:

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

Moreover, we have:

$$\psi_\gamma'(t) = \frac{1}{\varphi''(\varphi_\gamma^{'-1}(t))} \ , \psi_\gamma''(t) = -\frac{\varphi^{'''}(\varphi_\gamma^{'-1}(t))}{(\varphi''(\varphi_\gamma^{'-1}(t)))^3}.$$

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:

$$\varphi_\gamma'(v) = \psi_\gamma^{-1}(v) \ , \varphi_\gamma''(v) = \frac{1}{\psi_\gamma'(\psi^{-1}(v))}, \ \ \varphi_\gamma'''(v) = -\frac{\psi_\gamma^{''}(\psi^{-1}(v))}{\left(\psi_\gamma'(\psi^{-1}(v))\right)^3}$$

Since $\psi_\gamma^{-1}(x) > 0, \psi_\gamma^{'}(x) > 0$ and $\psi_\gamma^{''}(x) < 0$, φ_γ 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:

$$\delta_\alpha = \operatorname*{argmin}_{\delta\in\mathbb{R}} \left[\delta\in\mathbb{R}: \mathop{\mathbb{E}}_{\boldsymbol{x}^+\sim\mathcal{P}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\le\delta\right]\right]=\alpha\right],\ \hat{\delta}_\alpha = \operatorname*{argmin}_{\delta\in\mathbb{R}} \left[\delta\in\mathbb{R}: \frac{1}{n_+}\sum_{i=1}^{n_+}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\le\delta\right]\right]=\alpha\right]$$

$$\delta_\beta = \operatorname*{argmin}_{\delta\in\mathbb{R}} \left[\delta\in\mathbb{R}: \mathop{\mathbb{E}}_{\boldsymbol{x}^-\sim\mathcal{N}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\ge\delta\right]\right]=\beta\right],\ \hat{\delta}_\beta = \operatorname*{argmin}_{\delta\in\mathbb{R}} \left[\delta\in\mathbb{R}: \frac{1}{n_-}\sum_{j=1}^{n_-}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\ge\delta\right]\right]=\beta\right]$$

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

$$\mathcal{R}_{AUC}^{\alpha,\beta}(f_{\boldsymbol{\theta}},\mathcal{S}) = \mathop{\mathbb{E}}_{\boldsymbol{x}^-\sim\mathcal{N}}\mathop{\mathbb{E}}_{\boldsymbol{x}^+\sim\mathcal{P}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)>f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)<\delta_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)>\delta_\beta\right]\right]$$

$$\hat{\mathcal{R}}_{AUC}^{\alpha,\beta}(f_{\boldsymbol{\theta}},\mathcal{S}) = \frac{1}{n_+n_-}\cdot\sum_{j=1}^{n_-}\sum_{i=1}^{n_+}\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\le\hat{\delta}_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\hat{\delta}_\beta\right]$$

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

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

where

$$\Delta_+ = \sup_{\delta\in\mathbb{R}}\left|\frac{1}{n_+}\cdot\sum_{i=1}^{n_+}\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\le\delta\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}^+\sim\mathcal{P}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\le\delta\right]\right]\right|$$

$$\Delta_- = \sup_{\delta\in\mathbb{R}}\left|\frac{1}{n_-}\cdot\sum_{j=1}^{n_-}\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\delta\right] - \mathop{\mathbb{E}}_{\boldsymbol{x}^-\sim\mathcal{N}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\ge\delta\right]\right]\right|$$

Proof. First, we define some intermediate variables:

$$\ell_+(f_{\boldsymbol{\theta}},\boldsymbol{x}_j^-) = \mathop{\mathbb{E}}_{\boldsymbol{x}^+\sim\mathcal{P}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\le\delta_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\right]\right]$$

$$R_1 = \mathcal{R}_{AUC}^{\alpha,\beta}(f_{\boldsymbol{\theta}},\mathcal{S}) = \mathop{\mathbb{E}}_{\boldsymbol{x}^-\sim\mathcal{N}}\mathop{\mathbb{E}}_{\boldsymbol{x}^+\sim\mathcal{P}}\left[\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\ge f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^+)\le\delta_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}^-)\ge\delta_\beta\right]\right]$$

$$R_2 = \frac{1}{n_-}\cdot\sum_{j=1}^{n_-}\ell_+(f_{\boldsymbol{\theta}},\boldsymbol{x}_j^-)\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\delta_\beta\right]$$

$$R_3 = \frac{1}{n_-}\cdot\sum_{j=1}^{n_-}\ell_+(f_{\boldsymbol{\theta}},\boldsymbol{x}_j^-)\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\hat{\delta}_\beta\right]$$

$$R_4 = \frac{1}{n_+n_-}\cdot\sum_{j=1}^{n_-}\sum_{i=1}^{n_+}\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\le\delta_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\hat{\delta}_\beta\right]$$

$$R_5 = \hat{\mathcal{R}}_{AUC}^{\alpha,\beta}(f_{\boldsymbol{\theta}},\mathcal{S}) = \frac{1}{n_+n_-}\cdot\sum_{j=1}^{n_-}\sum_{i=1}^{n_+}\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\le\hat{\delta}_\alpha\right]\cdot\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-)\ge\hat{\delta}_\beta\right]$$

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

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

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

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

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

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

Here, (a_1) follows from the fact that $\mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-) \geq \delta_\beta\right] - \mathbf{1}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^-) \geq \hat{\delta}_\beta\right]$ must be simultaneously ≥ 0 or ≤ 0; (a_2) and (a_3) are based on the definition of δ_β and $\hat{\delta}_\beta$ and the assumption that no tie occurs in the dataset.

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

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

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

Here (b_1) and (b_2) follow a similar argument to (a_1)-(a_3). □

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

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

where VC *is the VC dimension of the hypothesis class:*

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

and

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

Proof. First, we have:

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

Following Lem.1 in (Narasimhan & Agarwal, 2017b), we have that, for all $f_{\boldsymbol{\theta}} \in \mathcal{T}(\mathcal{F})$, 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 $1-\delta$:

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

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

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

□

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:

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

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

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

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**, γ is searched in $\{0.03, 0.05, 0.08, 0.1, 1, 3, 5\}$. For **Exp**, γ 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 32×32 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 ρ 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.

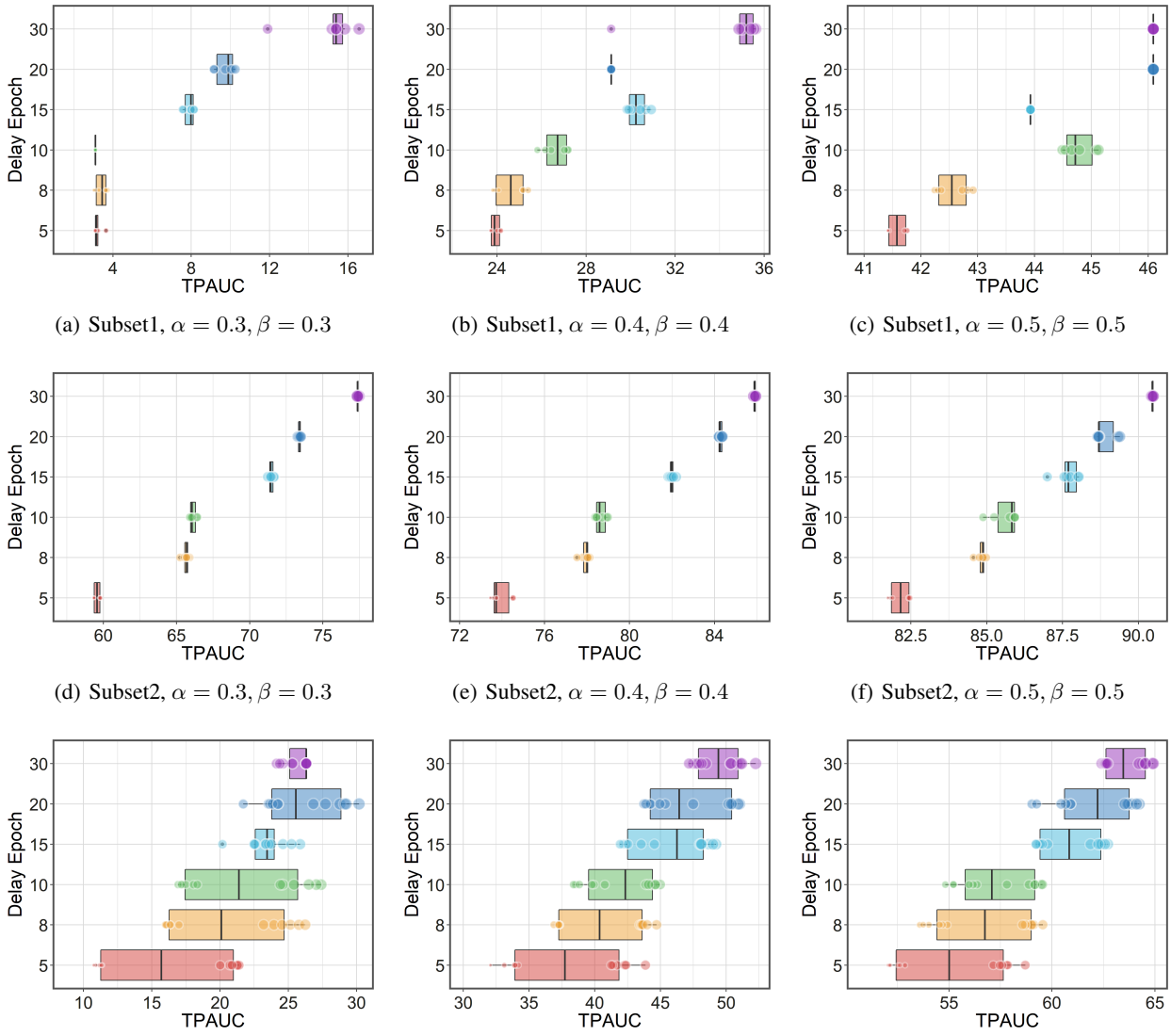
Implementation details On CIFAR Datasets. We utilize the ResNet-20 (He et al., 2015) as the backbone, which takes images with size $32 \times 32 \times 3$ as input and outputs 64-d features. Then the features are mapped into $[0, 1]$ with an FC layer and Sigmoid function. During the training phase, we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

Binary Tiny-ImageNet-200-LT Dataset. The Tiny-ImageNet-200 dataset contains 100,000 256×256 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.

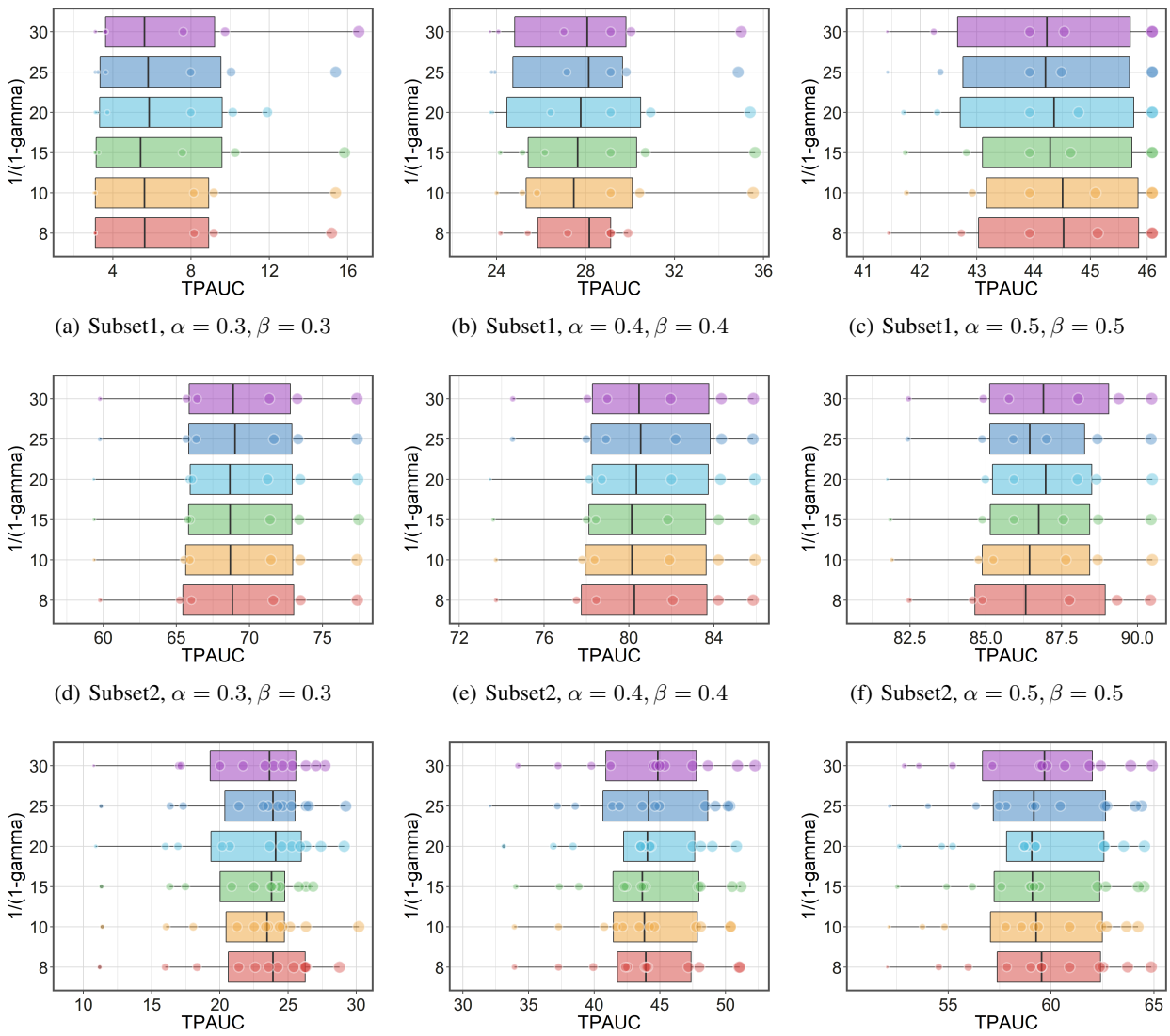


(a) Subset1, $\alpha = 0.3, \beta = 0.3$ (b) Subset1, $\alpha = 0.4, \beta = 0.4$ (c) Subset1, $\alpha = 0.5, \beta = 0.5$

(d) Subset2, $\alpha = 0.3, \beta = 0.3$ (e) Subset2, $\alpha = 0.4, \beta = 0.4$ (f) Subset2, $\alpha = 0.5, \beta = 0.5$

(g) Subset3, $\alpha = 0.3, \beta = 0.3$ (h) Subset3, $\alpha = 0.4, \beta = 0.4$ (i) Subset3, $\alpha = 0.5, \beta = 0.5$

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 $(\gamma - 1)^{-1}$.

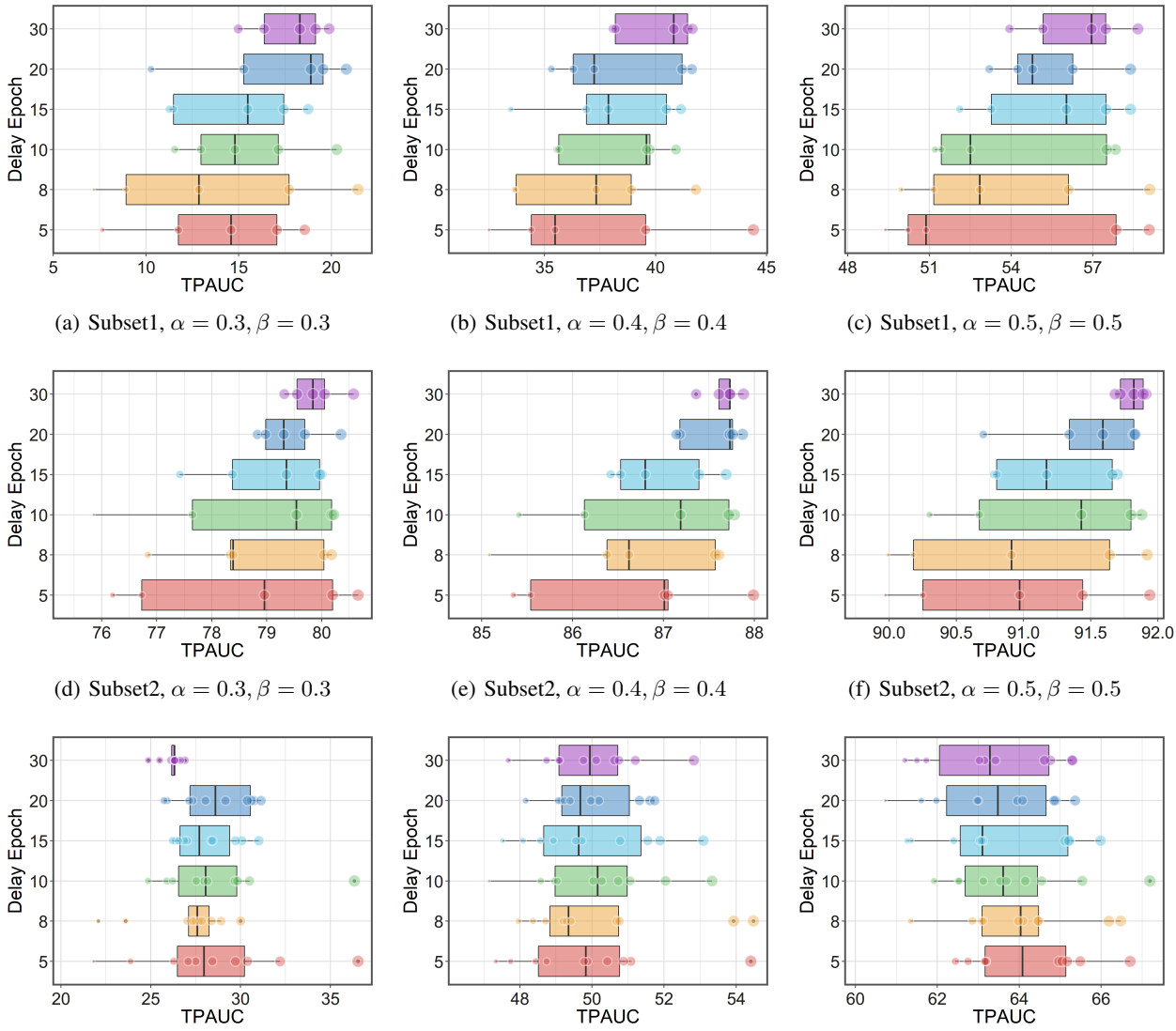


(a) Subset1, $\alpha = 0.3, \beta = 0.3$ (b) Subset1, $\alpha = 0.4, \beta = 0.4$ (c) Subset1, $\alpha = 0.5, \beta = 0.5$

(d) Subset2, $\alpha = 0.3, \beta = 0.3$ (e) Subset2, $\alpha = 0.4, \beta = 0.4$ (f) Subset2, $\alpha = 0.5, \beta = 0.5$

(g) Subset3, $\alpha = 0.3, \beta = 0.3$ (h) Subset3, $\alpha = 0.4, \beta = 0.4$ (i) Subset3, $\alpha = 0.5, \beta = 0.5$

Figure 6. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Exp** with respect to γ. 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.

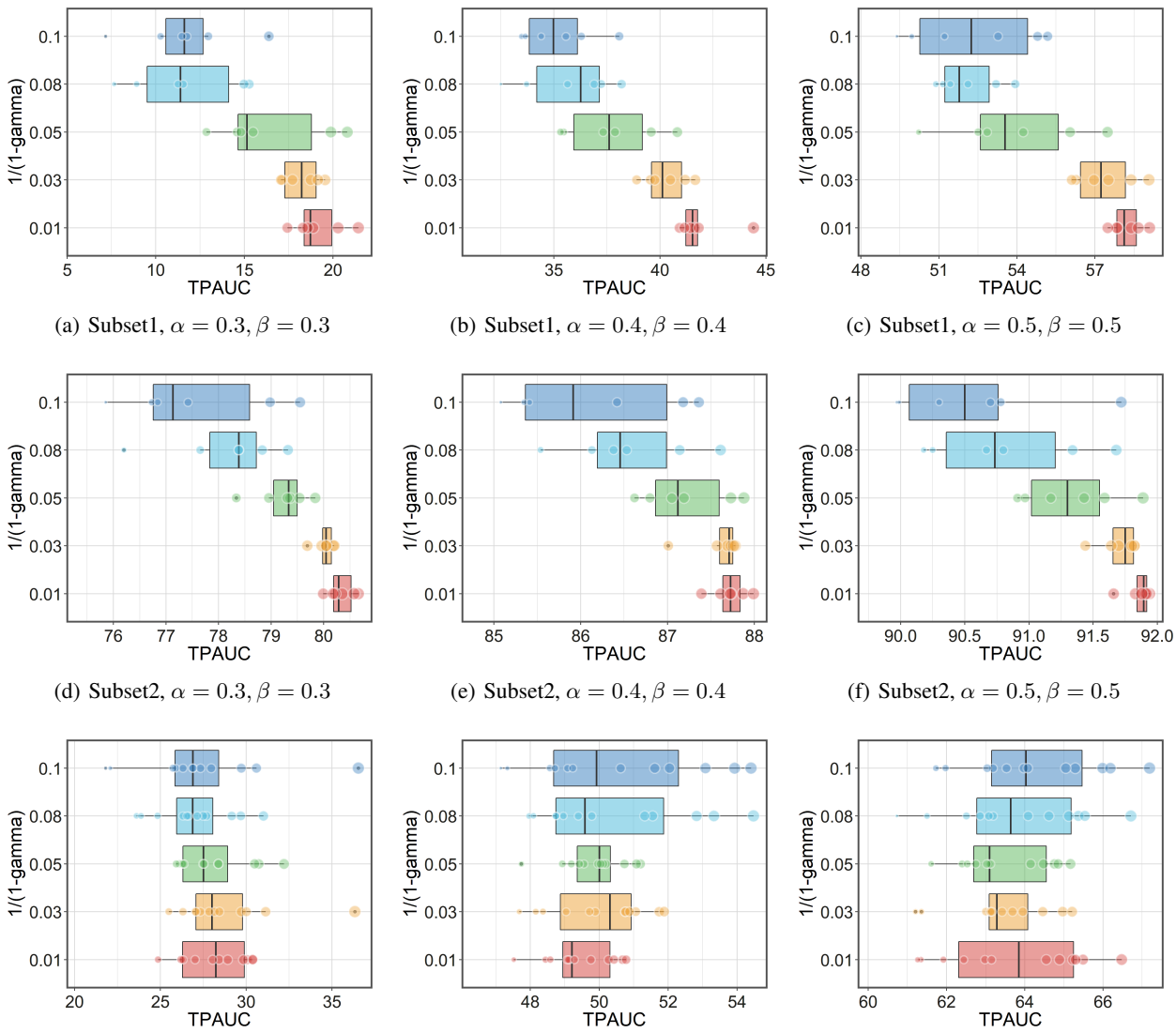


(a) Subset1, $\alpha = 0.3, \beta = 0.3$ (b) Subset1, $\alpha = 0.4, \beta = 0.4$ (c) Subset1, $\alpha = 0.5, \beta = 0.5$

(d) Subset2, $\alpha = 0.3, \beta = 0.3$ (e) Subset2, $\alpha = 0.4, \beta = 0.4$ (f) Subset2, $\alpha = 0.5, \beta = 0.5$

(g) Subset3, $\alpha = 0.3, \beta = 0.3$ (h) Subset3, $\alpha = 0.4, \beta = 0.4$ (i) Subset3, $\alpha = 0.5, \beta = 0.5$

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 $(\gamma - 1)^{-1}$.



(a) Subset1, $\alpha = 0.3, \beta = 0.3$ (b) Subset1, $\alpha = 0.4, \beta = 0.4$ (c) Subset1, $\alpha = 0.5, \beta = 0.5$

(d) Subset2, $\alpha = 0.3, \beta = 0.3$ (e) Subset2, $\alpha = 0.4, \beta = 0.4$ (f) Subset2, $\alpha = 0.5, \beta = 0.5$

(g) Subset3, $\alpha = 0.3, \beta = 0.3$ (h) Subset3, $\alpha = 0.4, \beta = 0.4$ (i) Subset3, $\alpha = 0.5, \beta = 0.5$

Figure 8. Sensitivity analysis on CIFAR-10-LT where TPAUC for **Poly** with respect to γ. 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.