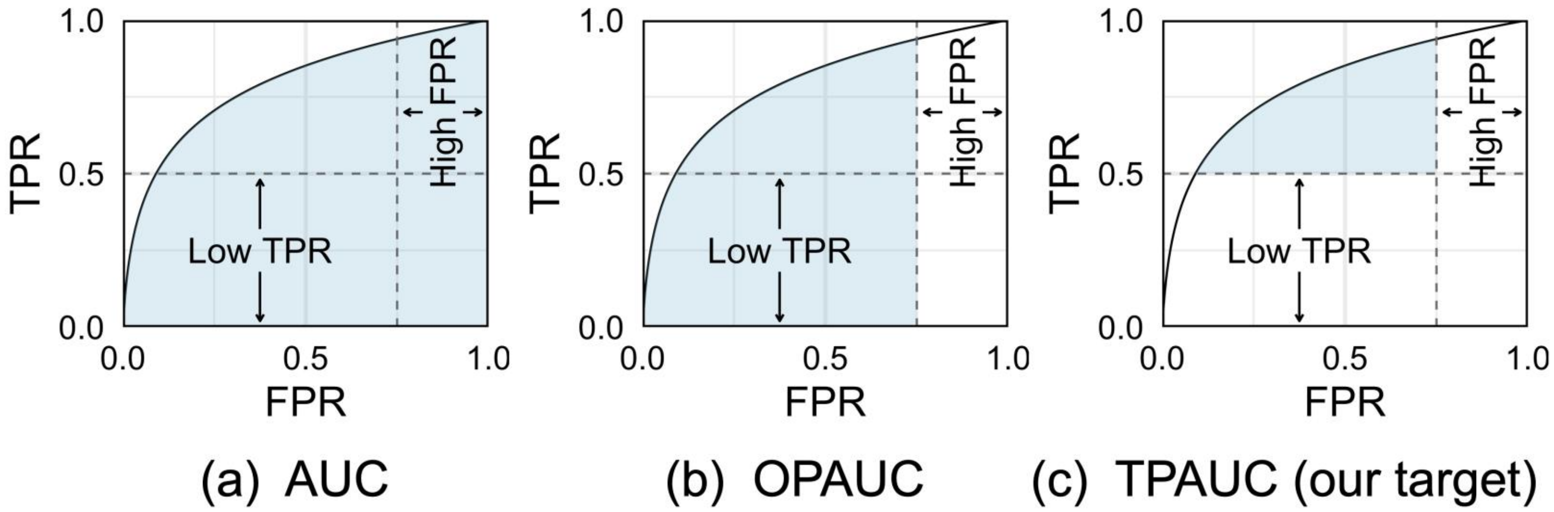
When All We Need is a Piece of the Pie:
A Generic Framework for Optimizing Two-way Partial AUC

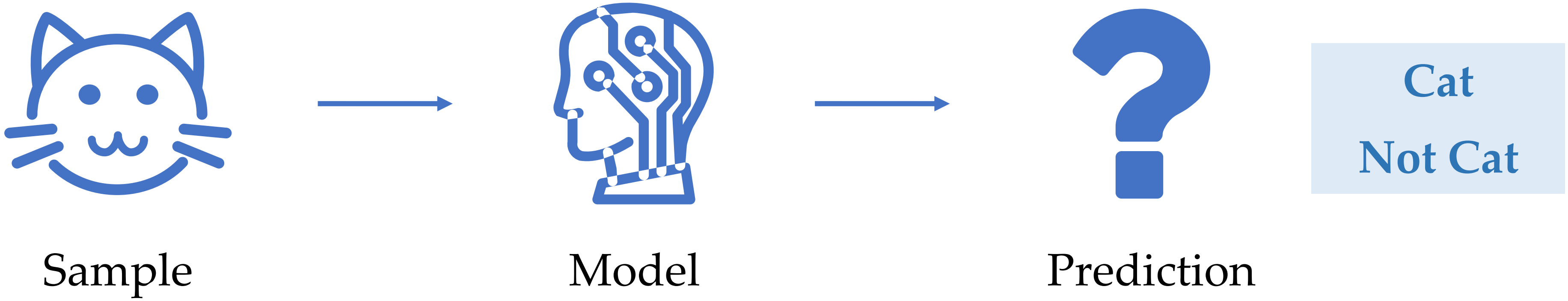


Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He,
Xiaochun Cao, Qingming Huang

2021.7.18

ICML International Conference On Machine Learning

Background



- Traditional classification methods adopt **error-rate-guided ERM**.

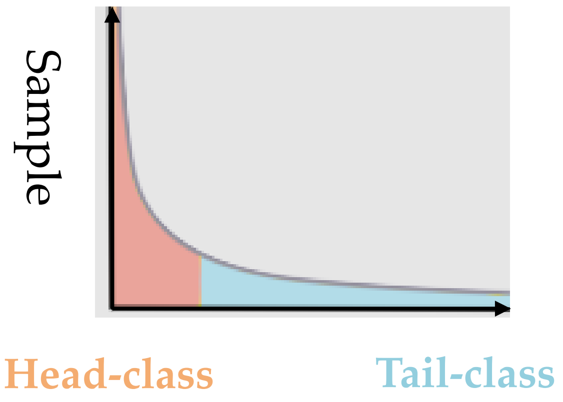
sensitive to class distribution

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[f(\boldsymbol{x}) \neq y_i\right] \xrightarrow[\ell]{\text{surrogate}} \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\left(f\left(x_i\right), y_i\right)$$

Background

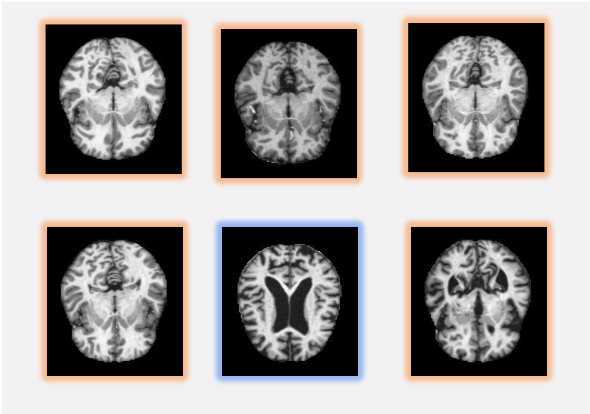
- Such ERM paradigm is problematic for imbalanced/long-tailed datasets

Long-tailed data



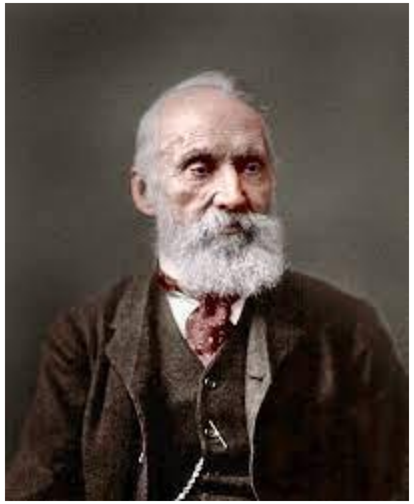
Spam Detection

Disease Diagnosis



- It is easy to get a high accuracy score by simply predicting all the samples as the majority class!

Background



If you can not **measure** it,
you can not improve it

~Lord Kelvin

Seek out a suitable metric for imbalanced datasets

Receiver Operating Characteristic Curve (ROC)

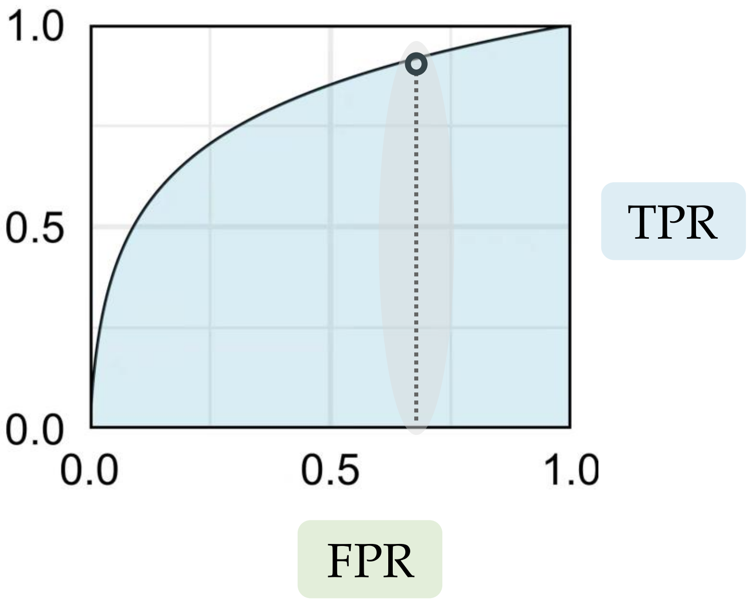
- **ROC curve**: True Positive Rate (TPR) vs. False Positive Rate (FPR).

Decision with a fixed threshold

label: $y \in \{0, 1\}$

classifier: $f(\boldsymbol{x})$, threshold: t

prediction: $\hat{y} = \mathbb{1}\left[f(\boldsymbol{x}) > t\right]$

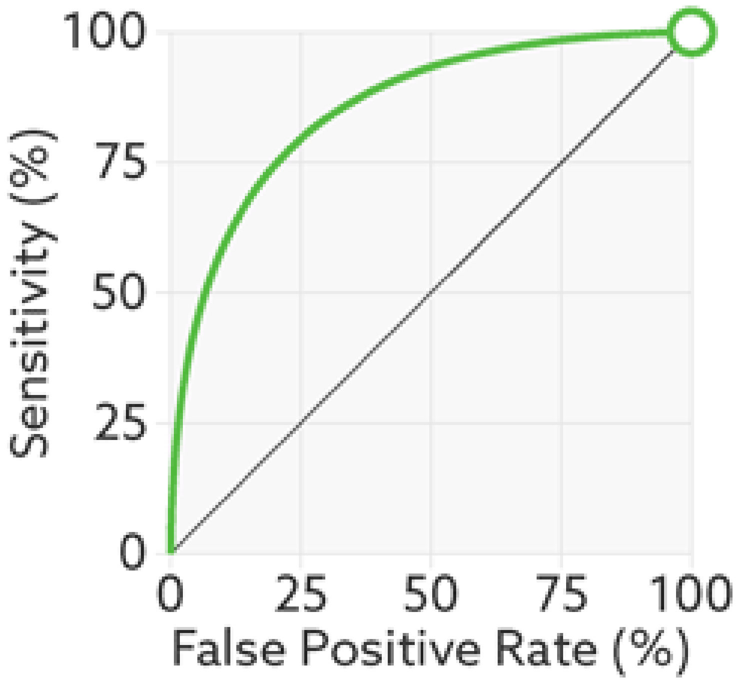


$$TPR = \mathbb{P}\left[f(\boldsymbol{x}) > t | y = 1\right]$$

$$FPR = \mathbb{P}\left[f(\boldsymbol{x}) > t | y = 0\right]$$

Area Under the ROC Curve (AUC)

- AUC is the area under the ROC curve **(over all possible thresholds)**



$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(\theta)\right)\, d\theta$$

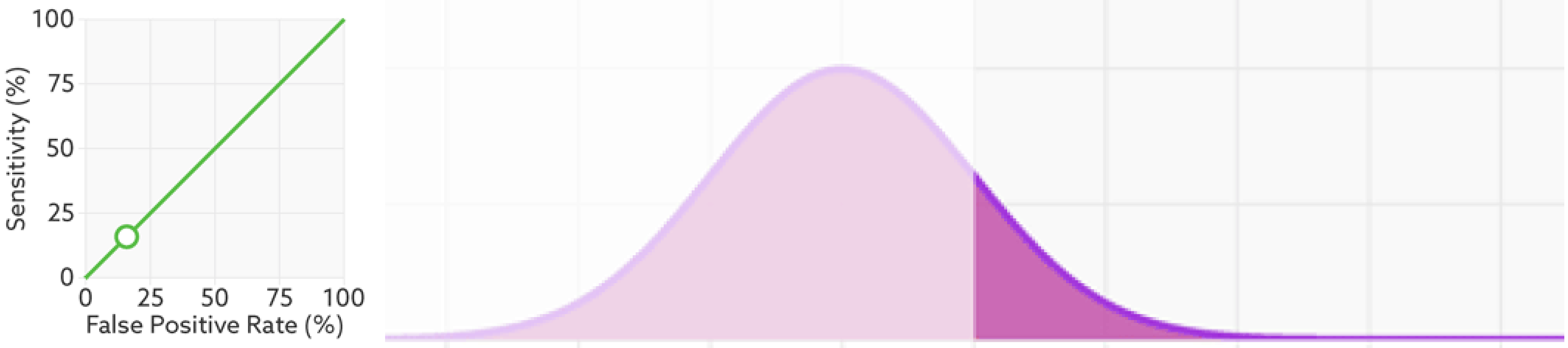
Involves a non-trivial integral

Area Under the ROC Curve (AUC)

- A much simpler reformulation:

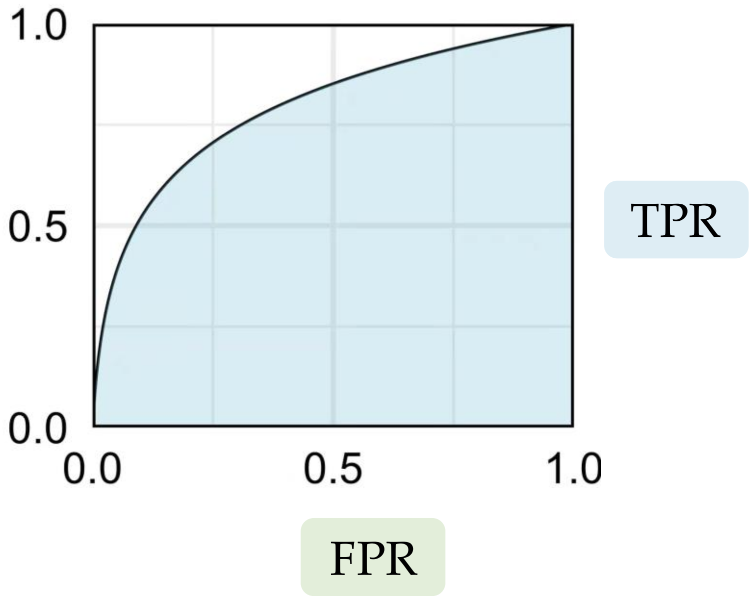
$$\text{AUC} = \mathbb{P}\left[f(\boldsymbol{x}) > f(\boldsymbol{x}')|y = 1, y' = -1\right]$$

- A measure of how well the two class conditional p.d.fs are separated



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

AUC is too informative



Global integration

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(\theta)\right) \, d\theta$$

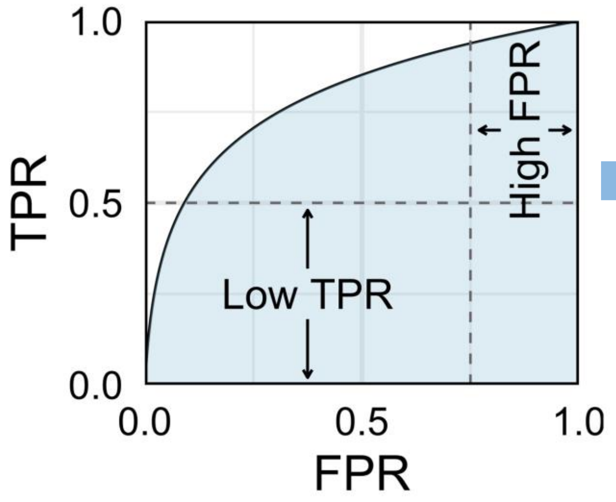
- Considers **all possible** TPR and FPR
- Real-world problems have performance **constraints** (*e.g.*, TPR>0.5, FPR < 0.1)

Consider Local analog of AUC

One Way Partial AUC (OPAUC)

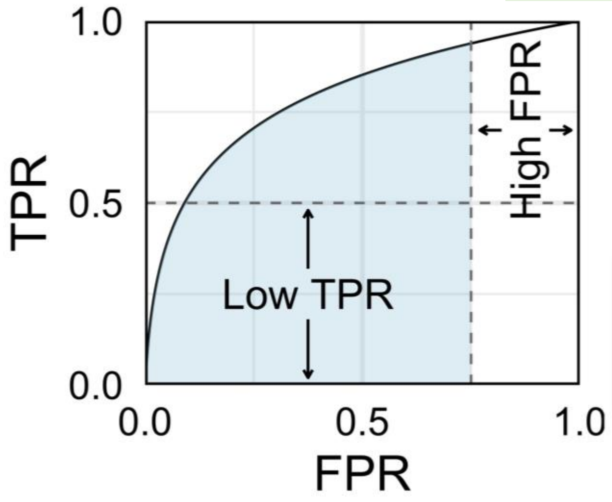
- AUC involves **all possible** TPRs and FPRs
- Many real-world applications have specific requirement on **FPR**
- Solution: Measure the partial area of ROC

AUC



One way Partial AUC (OPAUC)

[Dobb and Pepe 2003]



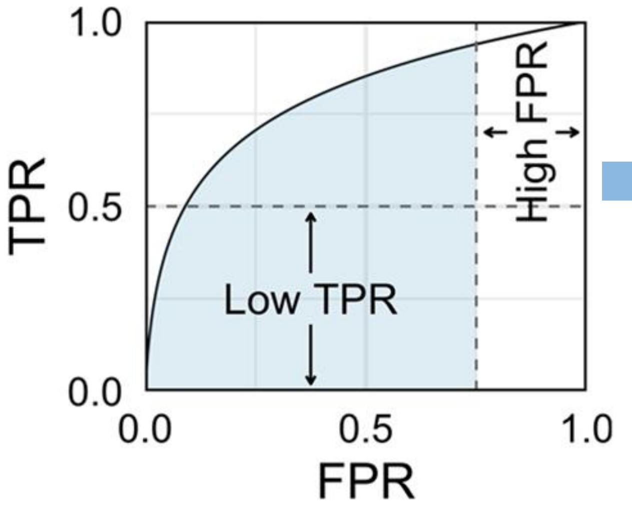
Only consider $\alpha \leq \text{FPR} \leq \beta$

$$\text{OPAUC} = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} \text{TPR}\left(\text{FPR}^{-1}(\theta)\right) \, d\theta$$

Two Way Partial AUC (TPAUC)

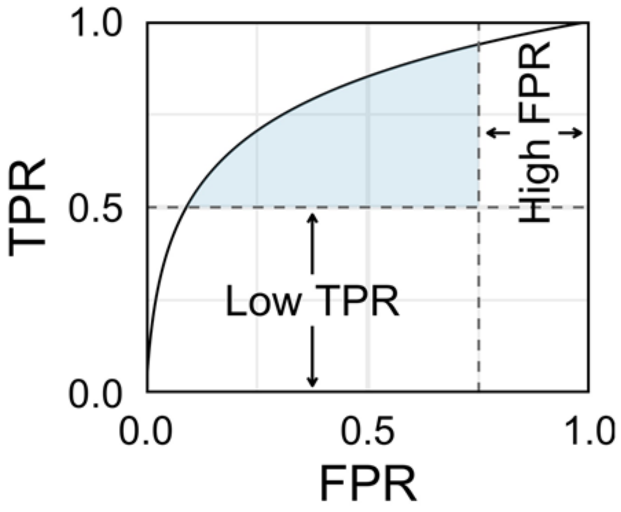
- A reasonable case should simultaneously enjoy a **low FPR and a high TPR**
- TPAUC measures the AUC within such a local area

OPAUC



Two way Partial AUC (TPAUC)

[Yang *et.al.* 2019]



$\mathsf{FPR} \leq \beta$

$\mathsf{TPR} \geq 1 - \alpha$

$$\mathsf{TPAUC} = \int_{\xi}^{\beta} \mathsf{TPR}\left(\mathsf{FPR}^{-1}(\theta)\right) d\theta - (1-\alpha)\cdot(\beta-\xi)$$

$$\xi = \mathsf{FPR}\left(\mathsf{TPR}^{-1}(1-\alpha)\right)$$

How to optimize local AUCs

- OPAUC
  - Cutting Plane Solvers [Narasimhan *et.al.* 2013; Narasimhan *et.al.* 2017; Tomoharu *et.al.* 2020]
  - Projected Sub-gradient Descent [Narasimhan *et.al.* 2013; Narasimhan *et.al.* 2017; Yamaguchi *et.al.* 2020]
  - Evolutionary Algorithms [Fan *et.al.* 2019]
  - Sampling Algorithms [Bai *et.al.* 2020a, b]

Not support the **end-to-end** training!

- Requires a **sampling** process
- Do not have **theoretical guarantee**

- TPAUC
  - ?

Optimize TPAUC in an end-to-end fashion

Can We estimate TPAUC from OPAUC ?

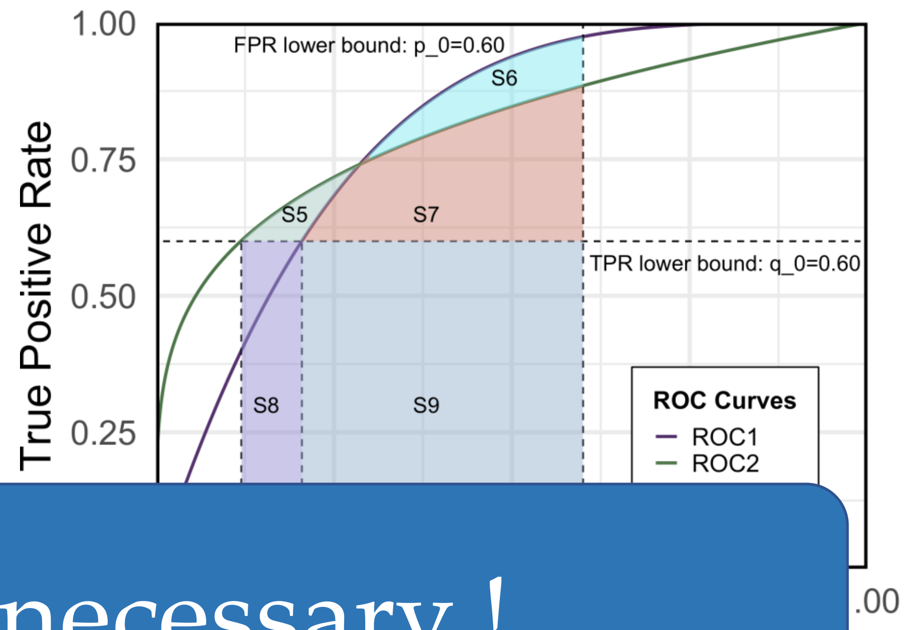
$$\text{OPAUC} = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} \text{TPR}\left(\text{FPR}^{-1}(\theta)\right) \, d\theta$$

approximate ξ with a fixed α ?

TPAU

Direct optimization is necessary !

ξ is a function of the scoring function f



TPAUC1 > TPAUC2: S5+S7< S6+S7

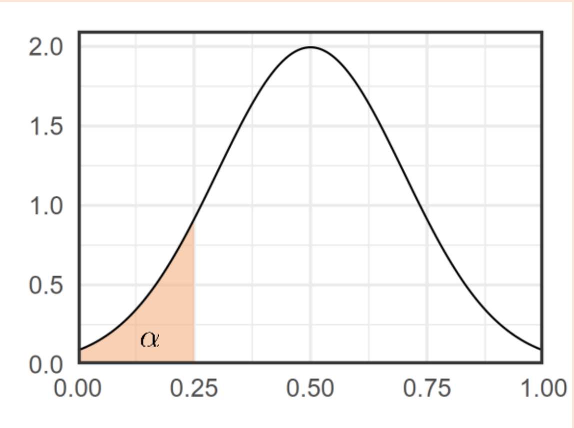
OPAUC2 > OPAUC1: S5+S7+S8+S9 > S7+S9+S6

ICML
International Conference On Machine Learning

Reformulation of TPAUC

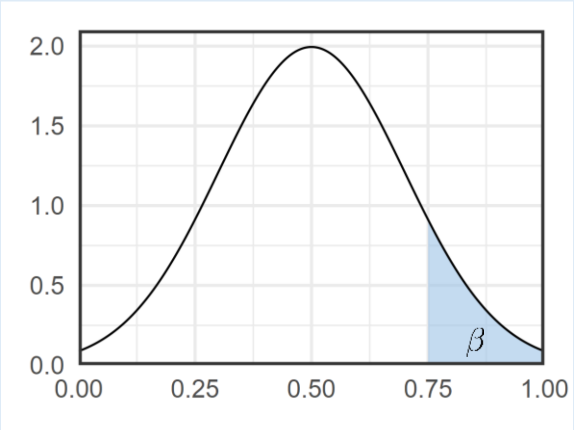
$$\mathrm{AUC}_{\beta}^{\alpha}(f_\theta, \mathcal{S}) = \mathbb{E}\left[\mathbf{1}\left[f(\boldsymbol{x}) > f(\boldsymbol{x}'), f(\boldsymbol{x}) \le t_{1-\alpha}, f(\boldsymbol{x}') \ge t_\beta \middle| y = 1, y' = -1\right]\right]$$

score quantile of the positive class



$$t_{1-\alpha} = \underset{\delta \in \mathbb{R}}{\mathrm{argmin}} \left[\delta \in \mathbb{R} : \ \mathbb{E}\left[\mathbf{1}\left[f(\boldsymbol{x}^+) \le \delta\right]\right] = \alpha\right]$$

score quantile of the negative class



$$t_{\beta} = \underset{\delta \in \mathbb{R}}{\mathrm{argmin}} \left[\delta \in \mathbb{R} : \ \mathbb{E}\left[\mathbf{1}\left[f(\boldsymbol{x}^-) \ge \delta\right]\right] = \beta\right]$$

- Requires empirical estimation of the **expectation**
- Requires empirical estimation of the **quantiles**

Empirical Estimation of TPAUC

$$\hat{\mathrm{AUC}}_{\beta}^{\alpha}(f_\theta, \mathcal{S}) = \frac{1}{n_+^\alpha n_-^\beta} \sum_{i=1}^{n_+^\alpha} \sum_{j=1}^{n_-^\beta} \mathbf{1}\left[f(\boldsymbol{x}) > f(\boldsymbol{x}')\right] \cdot \mathbf{1}\left[f(\boldsymbol{x}) \le \hat{t}_\alpha, f(\boldsymbol{x}') \ge \hat{t}_\beta | y = 1, y' = -1\right]$$

empirical expectation

empirical quantile

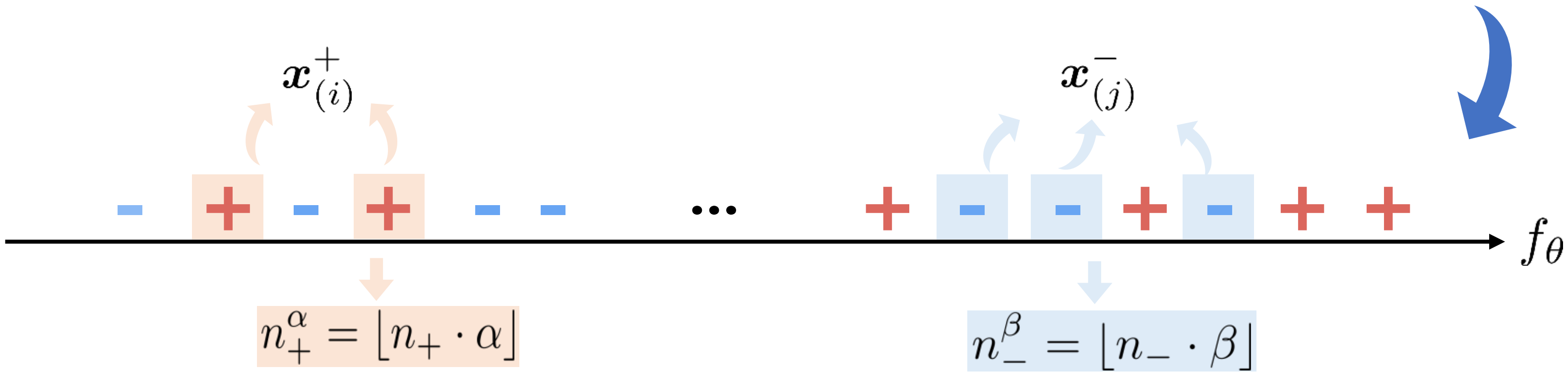
Theorem 1 Asymptotic Normality of the Bias (Informal) [Yang-Lu-Lyu-Hu 2019]

$$\hat{\mathrm{AUC}}_{\alpha}^{\beta} - \mathrm{AUC}_{\alpha}^{\beta} \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad n_+, n_- \to \infty$$

$$\sigma = \mathcal{O}\left(\sqrt{\frac{1}{n_+} + \frac{1}{n_-}}\right)$$

Empirical Estimation of TPAUC

$$\mathrm{AUC}_{\alpha}^{\beta}(f_{\boldsymbol{\theta}}, \mathcal{S}) = 1 - \sum_{i=1}^{n_+^{\alpha}} \sum_{j=1}^{n_-^{\beta}} \frac{\ell_{0,1}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{(i)}^{+}\right) - f\left(\boldsymbol{x}_{(j)}^{-}\right)\right)}{n_+^{\alpha} n_-^{\beta}}$$



Optimize Empirical TPAUC approximately

score

mong

all negative instances.

Step1 Surrogate Loss Minimization

- Replace $\ell_{0,1}$ with a continuous surrogate ℓ

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

$$\ell_{\exp}(t) = \exp(-t), \ell_{sq}(t) = (1-t)^2$$

- $\hat{\mathcal{R}}^{\ell}_{\alpha,\beta}(S, f_{\boldsymbol{\theta}})$ is still not differentiable!
- Calculating $\boldsymbol{x}^+_{(i)}, \boldsymbol{x}^-_{(j)}$ requires sorting the scores of positive and negative instances.

Step2 Bi-level optimization

The original optimization problem is equivalent to the following problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell\left(f_{\boldsymbol{\theta}}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-\right)$$

$$\text{s.t. } v_+ = \underset{v_i^+ \in [0,1], \sum_{i=1}^{n_+} v_i^+ \leq n_+^\alpha}{\operatorname{argmax}} \sum_{i=1}^{n_+} \left(v_i^+ \cdot \left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right)\right)\right)$$

$$v_- = \underset{v_j^- \in [0,1], \sum_{j=1}^{n_-} v_j^- \leq n_-^\beta}{\operatorname{argmax}} \sum_{j=1}^{n_-} \left(v_j^- \cdot f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right)\right)$$

where

$$\ell\left(f_{\boldsymbol{\theta}}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-\right) = \ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right) - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right)\right)$$

Outer-level problem
optimization based on the chosen instances

The ball constraints makes the optimization intractable

Inner-level problem
a sparse sample selection process

Step2 Bi-level optimization

- Transform the ℓ_1 ball constraints to ℓ_1 penalty terms (note that v_+, v_- are non-negative):

$$v_+ = \underset{v_i^+ \in [0,1], \sum_{i=1}^{n_+} v_i^+ \le n_+^\alpha}{\operatorname{argmax}} \sum_{i=1}^{n_+} \left(v_i^+ \cdot \left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right)\right)\right)$$

$$v_- = \underset{v_j^- \in [0,1], \sum_{j=1}^{n_-} v_j^- \le n_-^\beta}{\operatorname{argmax}} \sum_{j=1}^{n_-} \left(v_j^- \cdot f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right)\right)$$

$$v_+ = \underset{v_i^+ \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{n_+} \left(v_i^+ \cdot \left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right)\right) - \lambda^+ \cdot v_i^+ \right)$$

$$v_- = \underset{v_j^- \in [0,1]}{\operatorname{argmax}} \sum_{j=1}^{n_-} \left(v_j^- \cdot f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right) - \lambda^- \cdot v_j^- \right)$$

Step2 Bi-level optimization

- Replace the sparsity-inducing ℓ_1 penalty with a smooth surrogate φ_γ

$$(OP_1)\ \min_{\boldsymbol{\theta}} \frac{1}{n_+^\alpha n_-^\beta} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} v_i^+ \cdot v_j^- \cdot \ell\left(f_{\boldsymbol{\theta}}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-\right)$$

$$\text{s.t } v_+ = \underset{v_i^+ \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{n^+} \left(v_i^+ \cdot \left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right)\right) - \varphi_\gamma\left(v_i^+\right)\right)$$

$$v_- = \underset{v_j^- \in [0,1]}{\operatorname{argmax}} \sum_{j=1}^{n_-} \left(v_j^- \cdot f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right) - \varphi_\gamma\left(v_j^-\right)\right)$$

Sample Weights
Choose what to learn in the outer level problem

Penalty Function
Choose the weighting strategy

The connection between weight and the penalty is the key

Step3 Dual Correspondence

$$v_{+} = \underset{v_i^{+} \in [0,1]}{\operatorname{argmax}} \sum_{i=1}^{n^{+}} \left(v_i^{+} \cdot \left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^{+}\right)\right) - \varphi_{\gamma}\left(v_i^{+}\right)\right)$$

$$v_{-} = \underset{v_j^{-} \in [0,1]}{\operatorname{argmax}} \sum_{j=1}^{n_{-}} \left(v_j^{-} \cdot f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^{-}\right) - \varphi_{\gamma}\left(v_j^{-}\right)\right)$$

With a Closed-form Solution

$$v_i^{+} = \psi_{\gamma}\left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^{+}\right)\right) \qquad v_j^{-} = \psi_{\gamma}\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^{-}\right)\right)$$

weighting function

Under what condition can we realize such a simplification ?

Step 3 Dual Correspondence

Definition 1 Calibrated Smooth Penalty Function

A penalty function $\varphi_\gamma(x) : \mathbb{R}_+ \to \mathbb{R}$ *satisfies the following regularities:*

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

Definition 2 Calibrated Weighting Function

A weighting function $\psi_\gamma(x) : [0,1] \to \mathrm{Rng}$, *where* $\mathrm{Rng} \subseteq [0,1]$, *satisfies the following regularities:*

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

Step 3 Dual Correspondence

Proposition 1

Given a strictly convex function φ_γ *, and define* $\psi_\gamma(t)$ *as*

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

Then we can draw the following conclusions:

(a) If φ_γ *is a calibrated smooth penalty function, we have* $\psi_\gamma(t) = \varphi_\gamma'^{-1}(t)$.

penalty to weight

(b) If ψ_γ *is a calibrated weighting function such that* $v = \psi_\gamma(t)$, *we have*

$$\varphi_\gamma(v) = \int \psi_\gamma^{-1}(v) dv + \text{const.}$$

weight to penalty

This provides a simple way to establish a surrogate optimization problem of TPAUC

Step 3 Dual Correspondence

- Given the penalty functions φ_γ ,

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

If ψ_γ has a closed-form expression

Cancel the inner optimization problem

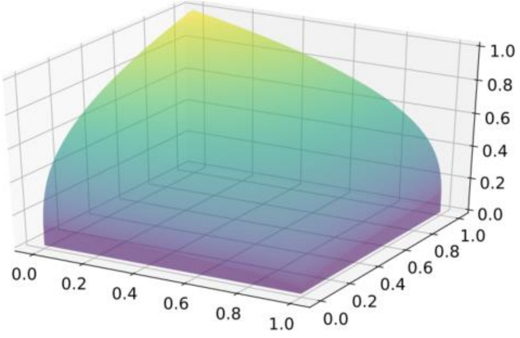
- Weighted empirical risk:

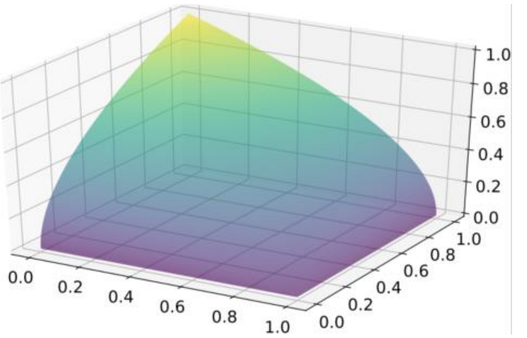
$$\hat{\mathcal{R}}_\psi^\ell\left(\mathcal{S}, f_{\boldsymbol{\theta}}\right) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \psi_\gamma\left(1 - f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i^+\right)\right) \psi_\gamma\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_j^-\right)\right) \cdot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

Instantiations of the Generic Framework

Example 1 (Polynomial Surrogate Model).

$$\varphi_\gamma^{\text{poly}}(t) = \frac{1}{\gamma} \cdot t^\gamma, \psi_\gamma^{\text{poly}}(t) = t^{\frac{1}{\gamma-1}}, \gamma > 2$$

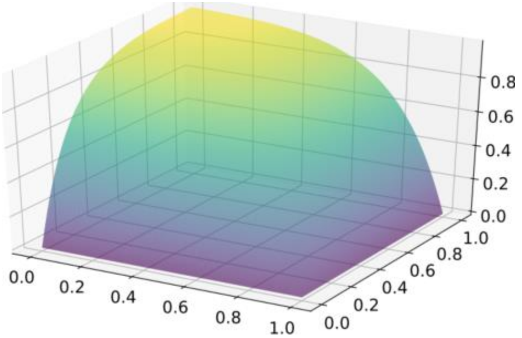


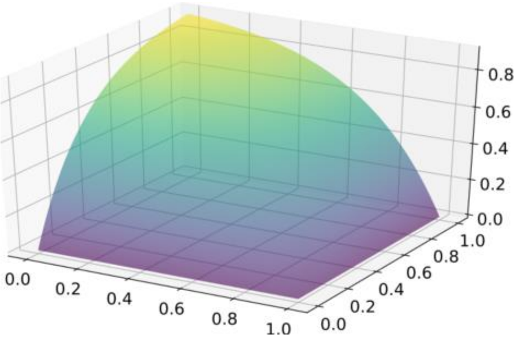


Example 2 (Exponential Surrogate Model).

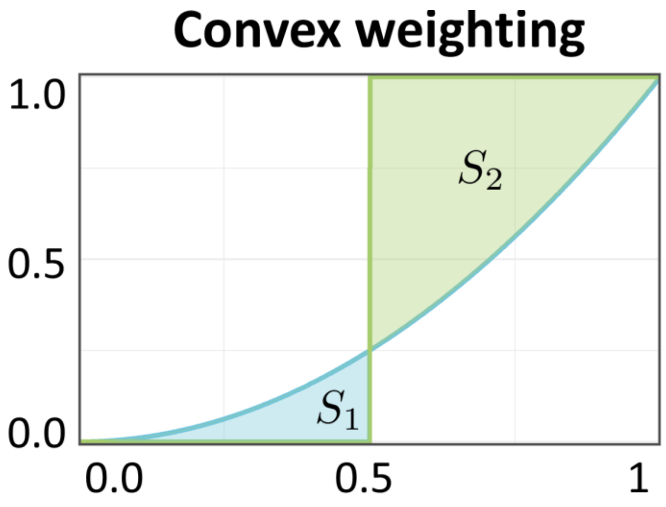
$$\varphi_\gamma^{\exp}(t) = \frac{(1-t)(\log(1-t)-1)+1}{\gamma}$$

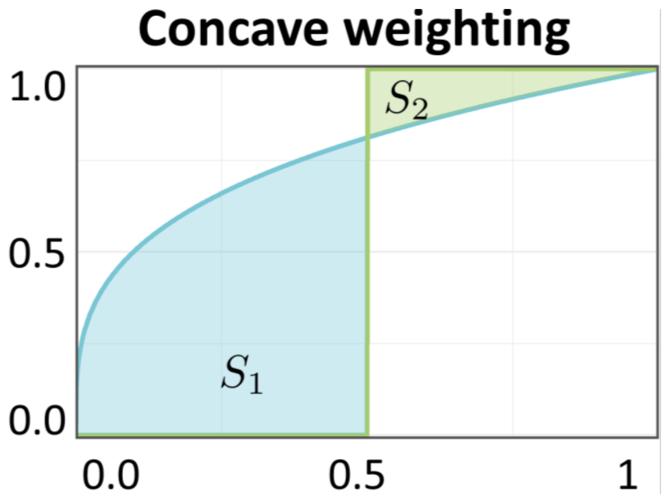
$$\psi_\gamma^{\exp}(t) = 1 - e^{-\gamma t}$$





Theoretical Analysis: concave vs. convex weighting function





S_1/S_2 should be **large**

Proposition 2 (Informal).

- **Concave** functions ψ are always **easier** to induces an upper bound of the original objective function

surrogate risk $\hat{\mathcal{R}}^{\ell}_{\psi}(\mathcal{S}, f_{\boldsymbol{\theta}}) > \hat{\mathcal{R}}^{\ell}_{\alpha,\beta}(\mathcal{S}, f_{\boldsymbol{\theta}})$ true risk

- A **sufficient** condition for achieving the **upper** bound:

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

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

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

The empirical distribution has **significant** mass over instances with **moderate difficulty**

Theoretical Analysis: concave vs. convex weighting function

Validation on simulated Dataset

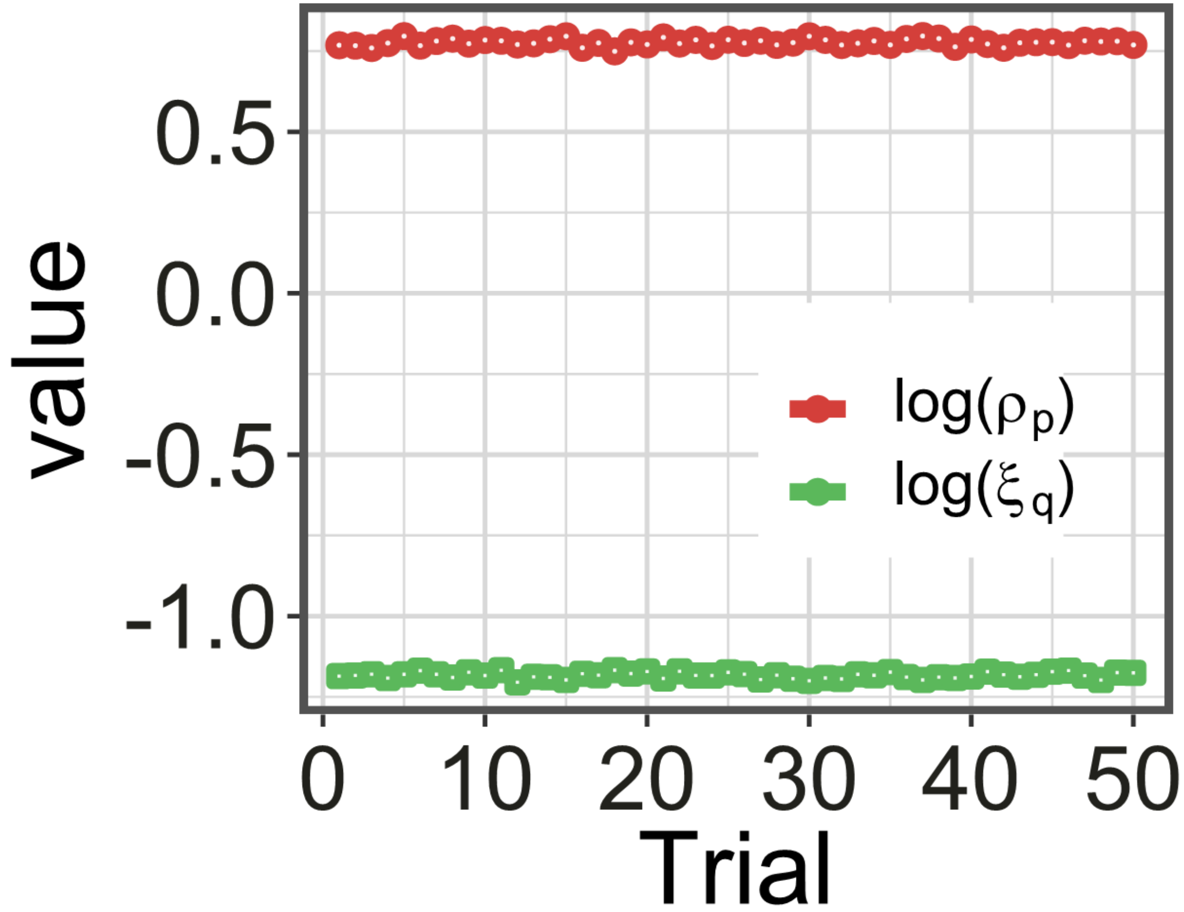
$$f(x^+) \sim \mathcal{N}(0.5, 0.08)$$

$$f(x^-) \sim \mathcal{N}(0.3, 0.08)$$

Generate 100 points for each class

plot for 50 such trails

$$\rho_p > \xi_q$$



Theoretical Analysis: Excess risk bound

ICML International Conference On Machine Learning

Theorem 2 (Informal).

The following inequality holds with high probability:

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

The empirical risk is biased

where $\tilde{\mathcal{O}}$ *is the big-O complexity notation hiding the logarithm factors,*

$$\mathcal{R}_{AUC}^{\alpha,\beta}(f_{\boldsymbol{\theta}},\mathcal{S}) = 1 - \mathrm{AUC}_{\alpha}^{\beta}(f_{\boldsymbol{\theta}},\mathcal{S}),$$

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

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

Empirical Results

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

- We construct long-tail binary datasets with different subsets:
  - ✓ Binary CIFAR-10-LT Dataset
  - ✓ Binary CIFAR-100-LT Dataset
  - ✓ Binary Tiny-ImageNet-200-LT Dataset

- We adopt the following variant of the TPAUC metric:

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

Empirical Results

- We consider TPAUC with

$$\alpha = 0.3, \beta = 0.3$$
$$\alpha = 0.4, \beta = 0.4$$
$$\alpha = 0.5, \beta = 0.5$$

- Table 1 shows the performance comparison against other methods dealing with imbalanced data.

- The empirical results demonstrate the superiority of our proposed TPAUC algorithm.

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

dataset	type	methods	Subset1 (0.3,0.3)	Subset1 (0.4,0.4)	Subset1 (0.5,0.5)	Subset2 (0.3,0.3)	Subset2 (0.4,0.4)	Subset2 (0.5,0.5)	Subset3 (0.3,0.3)	Subset3 (0.4,0.4)	Subset3 (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	**21.43**	**44.41**	**59.10**	80.66	**88.07**	**92.15**	**36.54**	**54.48**	67.19
		Exp	20.86	41.78	58.38	**81.22**	87.88	91.93	32.47	53.86	**67.32**
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	**68.02**	**79.11**	**85.17**	**91.13**	**93.78**	**95.69**	**47.07**	**65.89**	**75.08**
		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	**96.52**	**97.77**	91.66	95.19	96.79
		CBFocal	80.44	87.95	91.91	93.46	96.43	97.64	91.06	94.82	96.62
		SqAUC	80.16	87.99	91.67	93.10	96.07	97.32	**92.15**	95.16	96.75
	Ours-TPAUC	Poly	80.44	88.21	91.98	93.00	95.61	97.47	92.02	**95.25**	**96.84**
		Exp	**82.61**	**89.13**	**92.62**	**93.82**	96.12	97.38	91.25	94.78	96.57

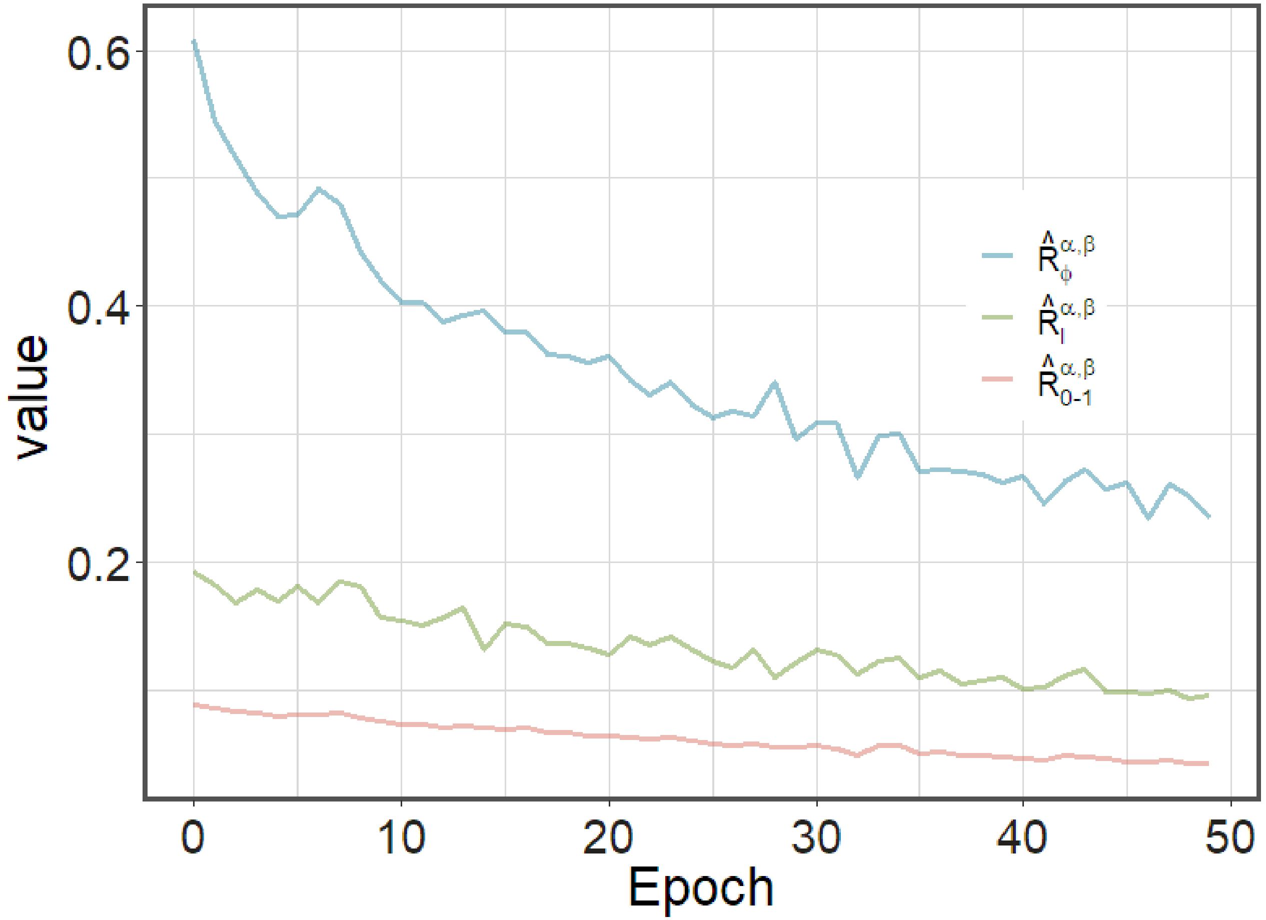
Validation of the upper bound

- We show the training curve of different losses, where we consistently observe that :

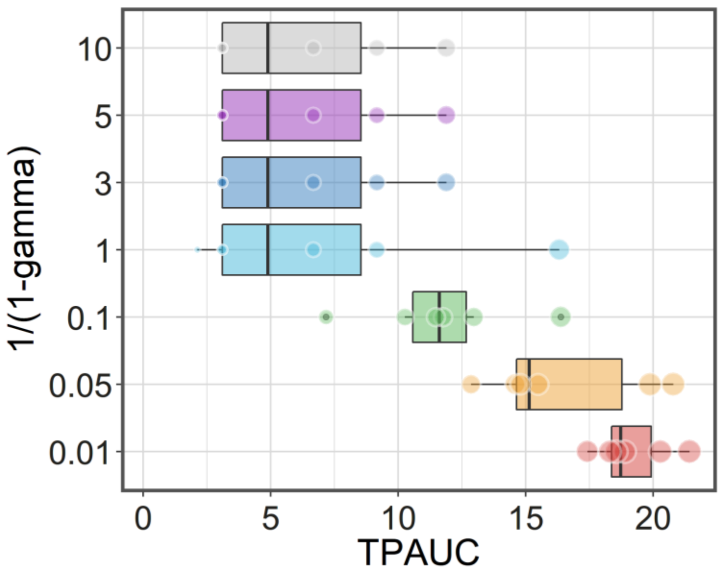
$$\hat{\mathcal{R}}_{\psi}^{\ell}\left(\mathcal{S}, f_{\boldsymbol{\theta}}\right) > \hat{\mathcal{R}}_{\alpha,\beta}^{\ell}\left(\mathcal{S}, f_{\boldsymbol{\theta}}\right) > \hat{\mathcal{R}}_{\alpha,\beta}^{0-1}\left(\mathcal{S}, f_{\boldsymbol{\theta}}\right)$$

App. Surr. Loss > Surr. Loss > emp. TPAUC

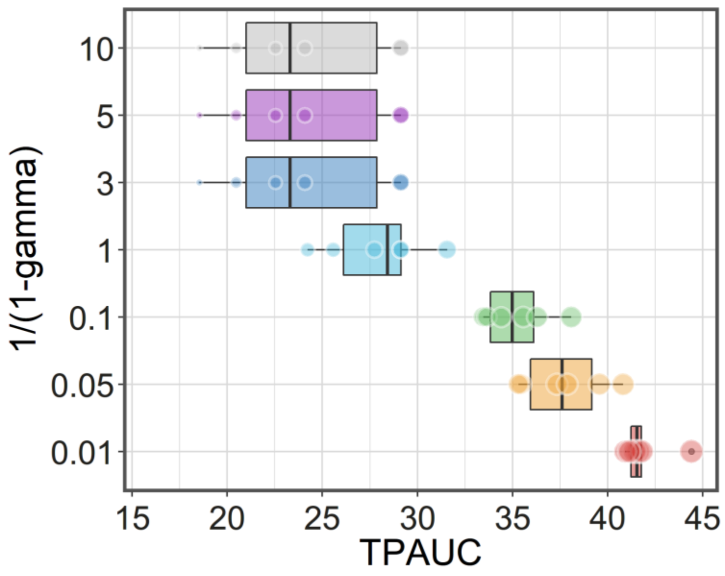
- This validate the proposed proposition about concave weights.



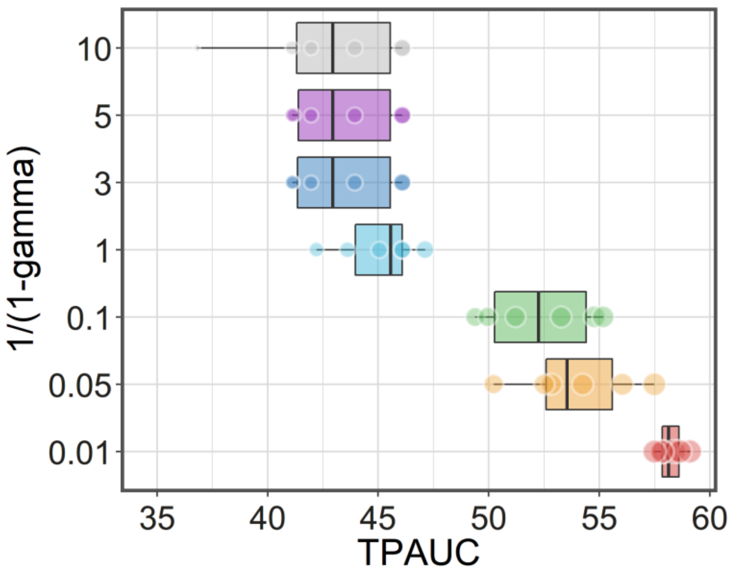
Convex vs. Concave Weighting



$\alpha = 0.3, \beta = 0.3.$



$\alpha = 0.4, \beta = 0.4.$



$\alpha = 0.5, \beta = 0.5$

- We analyze the effect of γ on CIFAR-10-Subset-1 with poly model
- The results shows that the concave function $((1-\gamma)^{-1} < 1)$ significantly outperforms convex functions $((1-\gamma)^{-1} \geq 1)$

Conclusion

Problem	How to optimize TPAUC (AUC with FPR upper bound and a TPR lower bound) in an **end-to-end** manner?
Method	A Bi-level reformulation of ERM framework for TPAUC A relaxation scheme for sample selection of the inner-level problem A generic surrogate objective function based on the dual correspondence
Theory	A sufficient condition for achieving the upper bound of the objective Concave weighting functions are easier to achieve the upper bound An $\mathcal{O}((VC/n_+ + VC/n_-)^{1/2})$ excess risk bound for the approximated ERM

Q&A