Weakly Supervised AUC Optimization: A Unified Partial AUC Approach

Zheng Xie , Yu Liu , Hao-Yuan He , Ming Li , *Member, IEEE*, and Zhi-Hua Zhou , *Fellow, IEEE*

Abstract—Since acquiring perfect supervision is usually difficult, real-world machine learning tasks often confront inaccurate, incomplete, or inexact supervision, collectively referred to as weak supervision. In this work, we present WSAUC, a unified framework for weakly supervised AUC optimization problems, which covers noisy label learning, positive-unlabeled learning, multi-instance learning, and semi-supervised learning scenarios. Within the WSAUC framework, we first frame the AUC optimization problems in various weakly supervised scenarios as a common formulation of minimizing the AUC risk on contaminated sets, and demonstrate that the empirical risk minimization problems are consistent with the true AUC. Then, we introduce a new type of partial AUC, specifically, the reversed partial AUC (rpAUC), which serves as a robust training objective for AUC maximization in the presence of contaminated labels. WSAUC offers a universal solution for AUC optimization in various weakly supervised scenarios by maximizing the empirical rpAUC. Theoretical and experimental results under multiple settings support the effectiveness of WSAUC on a range of weakly supervised AUC optimization tasks.

Index Terms—AUC optimization, partial AUC, weakly supervised learning.

I. INTRODUCTION

THE ROC curve (receiver operating characteristic curve) depicts the performance of a classifier by illustrating the variation of its true positive rate (TPR) against its false positive rate (FPR) for different thresholds. Its widespread use extends to fields such as military, biology, medicine, and others. Within machine learning, the area under ROC curve [1], [2], [3], [4], commonly referred to as AUC, is frequently utilized as a measure of the model's classification ability, without the explicit setting of a threshold. Practically, AUC is robust to class imbalance problem and effectively depicts the ranking ability of the model. Its utility lead to AUC optimization [4], [5] becoming a prevalent solution for building models with high AUC performance. By optimizing AUC, the models can better deal with imbalanced data and achieve improved ranking performance. Research on AUC optimization covers topics ranging from effective batch

optimization [6], [7] to online optimization [8], [9], [10], theoretical properties [11], [12], and combination with deep learning [13], [14], etc.

Recently, studies on partial AUC (pAUC) have emerged, as researchers have argued that for certain tasks, only the TPR or FPR within a specific range is of interest [15]. Two major variations, One-way Partial AUC (OPAUC) [15] and Two-way Partial AUC (TPAUC) [16], have been proposed, followed by studies on efficient optimization of such metrics [17], [18], [19]. It has been demonstrated that the partial AUC is useful in many real-world tasks including gene detection [20], speech processing [21], etc.

Despite the success of AUC optimization, most of the studies in this area focus on learning from clean data. However, in real-world learning tasks, collecting enough clean data is often difficult, and it is therefore necessary to study AUC optimization in the context of weakly supervised learning (WSL) [22]. Common weakly supervised learning paradigms include semi-supervised learning [23], positive-unlabeled learning [24], noisy label learning [25], multi-instance learning [26], and more. According to [22], these paradigms can be categorized as learning from incomplete, inaccurate, or inexact supervision.

Due to the pairwise formulation of the AUC risk, accuracy-oriented weakly supervised learning approaches cannot be easily adapted for AUC maximization. Thus, AUC optimization under weak supervision demands dedicated research. Currently, studies in this area remain relatively limited, such as in the contexts of semi-supervised learning [27], [28], [29] and noisy label learning [30]. While these studies focus on specific scenarios within WSL, no research to date offers a general understanding of the series of problems in weakly supervised AUC optimization.

In this work, we present WSAUC, a unified framework for weakly supervised AUC optimization problems. The framework provides a unified view of multiple WSL scenarios including (1) noisy label learning, (2) positive-unlabeled learning, (3) multi-instance learning, and (4) semi-supervised learning with or without label noise. Within the WSAUC framework, we first frame the different types of weakly supervised AUC optimization problems into a unified formulation, which is in the form of AUC optimization with contaminated labels. By proper construct the weakly supervised AUC risks under various scenarios, the empirical risk minimization (ERM) problems are proven to be consistent with the true AUC risk, so that the problems can be solved with a unified solution.

To mitigate the effect of contamination in the unified formulation of WSAUC, we propose a new type of partial AUC,
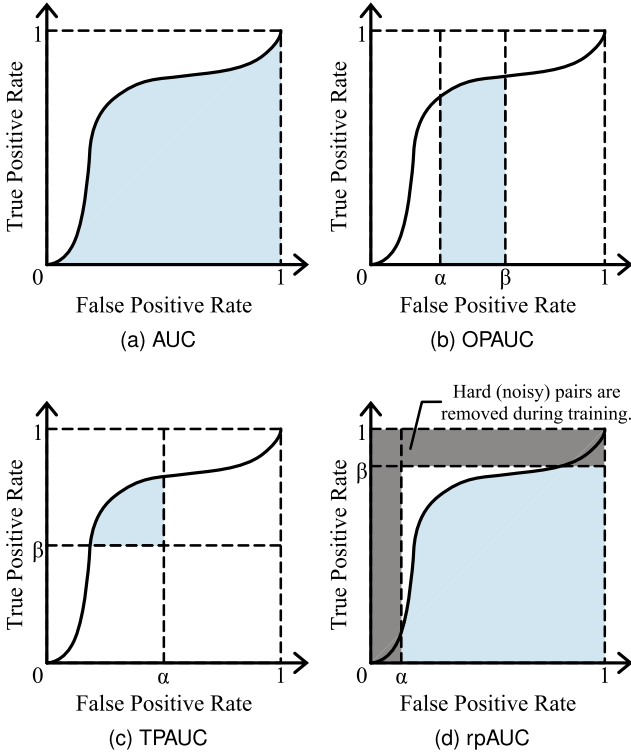
Fig. 1. Illustrations of ROC curves (black curves), AUC and Partial AUC variations (blue shades under the ROC curves). OPAUC and TPAUC are alternative performance measures specific to task needs. The proposed rpAUC aims for robust maximization of full AUC under weak supervision.

called reversed partial AUC (rpAUC, c.f. Fig. 1(d)). We show the connection between minimizing empirical rpAUC risk (or rpAUC maximization) and commonly used robust training approaches in noisy label learning, which suggests that minimizing empirical rpAUC risk leads to robust AUC optimization with contaminated labels. This method can be easily combined with the ERM problems of AUC optimization, and can be simply implemented by slightly modifying existing pAUC maximization algorithms. Based on rpAUC, the WSAUC framework offers a unified approach for AUC optimization in various weakly supervised scenarios. We evaluate the WSAUC framework through experiments under different WSL settings, and conclude its effectiveness.

The rest of this paper is organized as follows: We first discuss the related work in Section II, and then introduce preliminary knowledge in Section III. In Section IV, we present the unified view of weakly supervised AUC optimization, which is followed by the theoretical analysis in Section V. In Section VI, we introduce the novel concept of rpAUC for robust AUC optimization. Experimental results are presented in Section VII. Finally, Section VIII concludes with future work.

II. RELATED WORK

Weakly supervised learning (WSL) is a significant subfield of machine learning that tackles learning tasks with imperfect supervision. Unlike learning from fully and accurately labeled data, weak supervision can be incomplete, where only a subset of data is given with labels; inexact, where the data is given with only coarse-grind labels; or inaccurate, where the labels are not always true. For instance, semi-supervised learning [23] and positive-unlabeled learning [24] are common scenarios of learning with incomplete supervision. Multi-instance learning [26] and partial label learning [31] belong to the inexact supervision category, whereas learning from inaccurate supervision encompasses label noise learning [25] and crowdsourcing.

Among the topics mentioned, our work specifically relates to certain branches, i.e., unlabeled-unlabeled learning [32] and noisy label learning (NLL) [33]. The proposed unified view shares a similar problem setup of unlabeled-unlabeled learning, but with the different objective of maximizing AUC. The property of noisy label AUC optimization problem have been preliminarily studied in literature [34], [35], which is related partially to the theoretical analysis of our framework. The proposed robust method for weakly supervised AUC optimization through rpAUC maximization is closely related to the robust training approaches in noisy label learning. We demonstrate that it has a similar mechanism to techniques such as the small loss trick [36], [37], [38] and dynamic thresholding [39]. These techniques remove or lower the weight of instances inducing high risk, and belong to sample selection in the context of noisy label learning [40]. Other NLL approaches include robust surrogate losses [41], [42], robust model architectures [43], [44], [45], regularization methods [46], [47], [48], etc. It is noteworthy that a key information, namely noise transition matrix, is usually required by the NLL approaches for mitigate the label noise problem. Since such information is usually unknown in practice, its estimation becomes a essential step for noisy label learning [49], [50], [51]. In contrast, the proposed WSAUC does not require such knowledge for noisy label AUC optimization to achieve the consistency with the true risk.

Additionally, semi-supervised learning (SSL) is another major field of WSL, which aims to learn from both labeled and unlabeled data. Traditional SSL approaches include generative models [52], [53], [54], low-density separation-based models [55], [56], [57], graph-based models [58], [59], and disagreement-based models [60], [61]. With the rise of deep neural networks in recent years, SSL approaches based on consistency regularization [46], [62], [63], entropy minimization [64], graph neural networks [65], [66], [67], and others have demonstrated strong capabilities.

Positive-unlabeled learning (PUL) is designed to learn from only positive and unlabeled data. Researchers have proposed various methods to address this problem, such as selecting reliable negative samples [68], [69], reweighting the unlabeled samples [70], [71], [72], and more. Recently, approaches based on the unbiased risk estimators of PUL have proven to be effective [73], [74], [75], [76].

Multi-instance learning (MIL) tackles the challenge of learning from inexact supervision. In MIL, training instances are organized into bags, and labels are only given at the bag level. Instance-level approaches typically assume the bag label depends on the presence of any positive instance [77], [78], [79], [80], while bag-level approaches assume the bag label is defined by multiple concepts or instances [81], [82], [83], [84]. In this work, we adopt the assumption of instance-level MIL.

Given the challenging nature of the task, only a limited number of studies have attempted to unify various weakly supervised learning scenarios. For instance, SAFEW [85] focused on the safety aspect of WSL tasks, while LIISP [86] tackled the challenge of learning from incomplete and inaccurate supervision within a single framework. Meanwhile, CEGE [87] proposed a general framework for WSL based on centroid estimation. DivideMix [88] is a holistic model that address semi-supervised learning and noisy label learning simultaneously. However, these studies mainly concentrate on accuracy-oriented classification problems and cannot be readily applied to the AUC optimization problem.

AUC is one of the most commonly used performance measure other than classification accuracy [1]. It measures how well the model ranks the positive instances before the negative ones without considering the threshold, and therefore, it does not suffer from the inefficiency seen in accuracy when dealing with data imbalance. AUC optimization [4], [5] is the learning paradigm that aims to maximize the AUC performance of models. This approach makes the models more robust to data imbalance and leads to improved ranking performance. Much effort has been devoted to AUC optimization problem from multiple aspects. In terms of theoretical properties, [11] have studied the consistency of commonly used surrogate losses of AUC, while [12] have studied the generalization bounds of AUC optimization models. In terms of online and stochastic optimization, [89] have proposed an AUC optimization algorithm based on the covariance matrix, while [10] have converted the AUC optimization problem into a stochastic saddle point problem that can be optimized using stochastic gradient-based methods. [90] have further proposed a stochastic proximal AUC maximization algorithm that achieves a linear convergence rate. Recently, AUC optimization with deep neural networks has gained increasing attention. [13] have proposed deep AUC maximization based on pairwise AUC loss, while [14] have proposed an end-to-end AUC optimization method that solves a compositional objective. These research efforts have motivated many real-world applications of AUC optimization, including software build outcome prediction [91], medical image classification [92], and abnormal behavior detection [93].

In contrast to the extensive research on accuracy-oriented weakly supervised learning, the study on the weakly supervised AUC optimization is rather few, and mostly restricted to specific situations of weakly supervised learning. [28] study the semi-supervised AUC optimization based on positive-unlabeled learning. [29] study unbiased risk estimator of the semi-supervised AUC optimization. [30] study the property of AUC optimization under label noise. So far, no research has provided a unified approach for solving the AUC optimization problem across various weakly supervised learning scenarios.

III. PRELIMINARY

In supervised learning, we are provided with a dataset containing labeled data that has been sampled from a certain distribution, i.e.:

$$\mathcal{X}_L := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y). \tag{1}$$

For convenience, we will refer to the subsets of positive and negative data in $\mathcal{X}_L$ as:

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1), \text{ and}$$

$$\mathcal{X}_N := \{\boldsymbol{x}_j'\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = -1),$$

so that $\mathcal{X}_L = \mathcal{X}_P \cup \mathcal{X}_N$.

Let $f : \mathcal{X} \to \mathcal{R}$ be a model that is expected to output larger score for positive instances than negative ones. Given a classification threshold t, we define the model's true positive rate $\text{TPR}_f(t) = \Pr(f(x) \geq t|y = 1)$, and the false positive rate $\text{FPR}_f(t) = \Pr(f(x) \geq t|y = 0)$. When t varies, the ROC curve (receiver operating characteristic curve) plots the TPR against FPR of the model f, as illustrated in Fig. 1(a). The AUC is defined as the area under the ROC curve:

$$\text{AUC} = \int_{-\infty}^{\infty} \text{TPR}_f(t) \, \text{d} \, \text{FPR}_f(t). \tag{2}$$

Since AUC is equivalent to the probability of a randomly drawn positive instance being ranked before a randomly drawn negative instance [1], [2], [3], the AUC of the model f can be formulated as:

$$\text{AUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_P(\boldsymbol{x})} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_N(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x}) - f(\boldsymbol{x}'))] \right]. \tag{3}$$

Here $\ell_{01}(z) = \mathbb{I}[z < 0]$. Without introducing any ambiguity, we will use $f(\boldsymbol{x}, \boldsymbol{x}')$ hereinafter to denote $f(\boldsymbol{x}) - f(\boldsymbol{x}')$ for clarity. Maximizing AUC is equivalent to minimizing the following AUC risk. To avoid confusion, we denote the true AUC risk as P-N AUC risk, since it measures the error rate of ranking positive instances before negative instances:

$$R_{PN}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_P(\boldsymbol{x})} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_N(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]. \tag{4}$$

In practice, we typically solve the following empirical risk minimization (ERM) problem with a finite sample:

$$\min_{f} \quad \hat{R}_{PN}(f) = \frac{1}{|\mathcal{X}_P||\mathcal{X}_N|} \sum_{\boldsymbol{x} \in \mathcal{X}_P} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(f(\boldsymbol{x}, \boldsymbol{x}')). \tag{5}$$

Recently, researchers have discovered that partial AUC (pAUC) is a useful metric for certain applications. Partial AUC focuses on specific part of the area under the ROC curve. There are two variations of pAUC: one-way pAUC (OPAUC [15], c.f. Fig. 1(b)), which limits FPR in a range $[\alpha, \beta]$, and two-way pAUC (TPAUC [16], c.f. Fig. 1(c)), which limits $\text{FPR} < \alpha$ and $\text{TPR} > \beta$. In this paper, we focus on the two-way pAUC. It can be formulated as follows:

$$\text{pAUC} = 1 - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_P^-(\boldsymbol{x})} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_N^+(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right], \tag{6}$$

where $p_P^-(\boldsymbol{x}) = p_P(\boldsymbol{x}|f(\boldsymbol{x}) \in (-\infty, \text{TPR}_f^{-1}(\beta)])$ and $p_N^+(\boldsymbol{x}) = p_N(\boldsymbol{x}|f(\boldsymbol{x}) \in [\text{FPR}_f^{-1}(\alpha), \infty))$ are difficult instances, i.e., positive instances with lower scores and negative instances with higher scores. In Section VI, we develop a variation of TPAUC, namely reversed partial AUC (rpAUC), to achieve robust AUC optimization under weak supervision.

IV. THE UNIFIED VIEW OF WSAUC

A. The Unified Formulation

The unified view of weakly supervised AUC optimization can be formulated based on the problem of minimizing AUC risk with two contaminated instance sets that have different class proportions. This learning problem can also be referred to as unlabeled-unlabeled learning [32], which focus on building classification models with minimal supervision, i.e., two unlabeled instances set with different class priors. Suppose we have two contaminated instance sets $\mathcal{X}_A$ and $\mathcal{X}_B$, which can be regarded as samples from the mixture distributions of the positive and negative distributions with different proportions:

$$\mathcal{X}_A := \{\boldsymbol{x}_i\}_{i=1}^{n_A} \overset{\text{i.i.d.}}{\sim} p_A(\boldsymbol{x}) := \theta_A p_P(\boldsymbol{x}) + (1-\theta_A)p_N(\boldsymbol{x});$$

$$\mathcal{X}_B := \{\boldsymbol{x}_j'\}_{j=1}^{n_B} \overset{\text{i.i.d.}}{\sim} p_B(\boldsymbol{x}) := \theta_B p_P(\boldsymbol{x}) + (1-\theta_B)p_N(\boldsymbol{x}),$$

Without loss of generality, here we assume $\theta_A > \theta_B$. The AUC risk defined on the distribution p_A and p_B is:

$$R_{AB}(f) := \underset{\boldsymbol{x}\sim p_A(\boldsymbol{x})}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\sim p_B(\boldsymbol{x})}{\mathbb{E}} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right], \quad (7)$$

where $f(\boldsymbol{x}, \boldsymbol{x}')$ is the shorthand for $f(\boldsymbol{x}) - f(\boldsymbol{x}')$. Then, the model can be obtained by solving the empirical risk minimization (ERM) problem:

$$\min_f \quad \hat{R}_{AB}(f) = \frac{1}{|\mathcal{X}_A||\mathcal{X}_B|} \sum_{\boldsymbol{x}\in\mathcal{X}_A} \sum_{\boldsymbol{x}'\in\mathcal{X}_B} \ell(f(\boldsymbol{x}, \boldsymbol{x}')). \quad (8)$$

The optimization problem is akin to treating both sets as positive and negative instance sets without accounting for their impurity. However, it can be shown that the impure AUC risk can be rewritten in a unified formulation which the true risk can be obtained with a linear transformation.

Theorem 1 (unified formulation): The impure AUC risk R_{AB} over two contaminated distributions p_A and p_B can be rewrote in the unified formulation as:

$$R_{AB} = aR_{PN} + b, \quad (9)$$

where the bias term $b = {(1-a)}/{2}$. Based on this formulation, the true risk R_{PN} can be obtained from R_{AB} with a linear transformation.

Proof: For any f, we can rewrite the risk:

$$R_{AB}(f) = \underset{\boldsymbol{x}\in\mathcal{X}_A}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\in\mathcal{X}_B}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$= \underset{\boldsymbol{x}\in\mathcal{X}_A}{\mathbb{E}} \left[(1-\theta_B)\underset{\boldsymbol{x}'\in\mathcal{X}_N}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] + \theta_B\underset{\boldsymbol{x}'\in\mathcal{X}_P}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$= \theta_A(1-\theta_B)\underset{\boldsymbol{x}\in\mathcal{X}_P}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\in\mathcal{X}_N}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$+ \theta_A\theta_B\underset{\boldsymbol{x}\in\mathcal{X}_P}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\in\mathcal{X}_P}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$+ (1-\theta_A)(1-\theta_B)\underset{\boldsymbol{x}\in\mathcal{X}_N}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\in\mathcal{X}_N}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$+ (1-\theta_A)\theta_B\underset{\boldsymbol{x}\in\mathcal{X}_N}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\in\mathcal{X}_P}{\mathbb{E}} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$= \theta_A(1-\theta_B)R_{PN}(f) + \theta_A\theta_B R_{PP}(f)$$

$$+ (1-\theta_A)(1-\theta_B)R_{NN}(f) + (1-\theta_A)\theta_B R_{NP}(f)$$

$$= \theta_A(1-\theta_B)R_{PN}(f) + (1-\theta_A)\theta_B(1-R_{PN})$$

$$+ \frac{1}{2}(2\theta_A\theta_B + 1 - \theta_A - \theta_B)$$

$$= (\theta_A - \theta_B)R_{PN}(f) + \frac{1-(\theta_A-\theta_B)}{2}.$$

This completes the proof. □

Corollary 2 (consistency of inaccurate case): In the AUC optimization problem described above, suppose f^* is a minimizer of the impure AUC risk R_{AB} over two contaminated distributions p_A and p_B, i.e., $f^* = \arg\min R_{AB}$, then f^* is also a minimizer of the true AUC risk R_{PN}, i.e., R_{AB} is consistent with R_{PN}.

Proof. According to Theorem 1, for any f we have

$$R_{AB}(f) = aR_{PN}(f) + \frac{1-a}{2},$$

where $a > 0$. Thus, for any f,

$$R_{PN}(f^*) - R_{PN}(f) = \frac{R_{AB}(f^*) - R_{AB}(f)}{a} \leq 0$$

because $a > 0$ and f^* is a minimizer of R_{AB}. This proves that f^* is also a minimizer of R_{PN}. □

In the rest of this section, we will demonstrate how different weakly supervised AUC optimization problems can be connected with this unified formulation. This provides a simple way for handling AUC optimization problem in almost all common case of weakly supervised AUC optimization tasks by simply adopting any pairwise risk minimization algorithm. An overview is provided in Table I.

It is noteworthy that when optimizing for accuracy, it is required to know the mixing proportions of the distributions (i.e., θ_A and θ_B) in order to correct estimation bias and achieve consistency. In AUC optimization, however, we can maintain statistical consistency even without knowledge of the mixing proportions. Nevertheless, the presence of impurity can still have an impact on model learning with finite data, which is depicted in the coefficient a in (9). Directly solving the optimization problem above may be less robust when there is significant impurity in the instance sets. In Section VI, we address this issue by minimizing the empirical risk of a novel variety of partial AUC.

B. Specialization to Common Cases

1) Noisy Label AUC Optimization: We begin by discussing AUC optimization under label noise, which is the most common case of inaccurate supervision. In binary classification with label noise, an instance label may be flipped with a certain probability. Consider the asymmetrical noise for two classes, a positive instance may be mislabeled with a probability of η_P, and a negative instance may be mislabeled with a probability of η_N.

TABLE I
SUMMARY OF THE CONNECTIONS BETWEEN WEAKLY SUPERVISED AUC OPTIMIZATION RISKS AND SUPERVISED AUC RISK R_{PN}

Section	Setting	Data sets	R	a	b	Risk def.
§ III	Supervised AUC Optimization	$\mathcal{X}_P, \mathcal{X}_N$	R_{PN}	1	0	Eq. (4)
§ IV-B.1	Noisy Label AUC Optimization	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_{\tilde{N}}$	$R_{\tilde{P}\tilde{N}}$	$1-\eta_P-\eta_N$	$(\eta_P+\eta_N)/2$	Eq. (10)
§ IV-B.2	Positive-Unlabeled AUC Optimization	$\mathcal{X}_P, \mathcal{X}_U$	R_{PU}	π_N	$\pi_P/2$	Eq. (12)
§ IV-B.3	Multi-Instance AUC Optimization	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_N$	$R_{\tilde{P}N}$	$1-\eta_P$	$\eta_P/2$	Eq. (14)
§ IV-B.4	Semi-Supervised AUC Optimization	$\mathcal{X}_P, \mathcal{X}_U, \mathcal{X}_N$	R_{PNU}	1	0 (corrected)	Eq. (19)
§ IV-B.5	Semi-Supervised AUC Optimization with Label Noise	$\mathcal{X}_{\tilde{P}}, \mathcal{X}_U, \mathcal{X}_{\tilde{N}}$	$R_{\tilde{P}\tilde{N}U}$	$1-\eta_P-\eta_N$	$(\eta_P+\eta_N)/2$	Eq. (20)
§ IV-A	Unified Risk Formulation		R	$=$	$a \ R_{PN}$	$+ \ b$

Such a noisy label AUC optimization problem can be easily converted to the problem defined in (7) by setting the noise ratio as the mixture proportion:

$$\mathcal{X}_{\tilde{P}} := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_{\tilde{P}}(\boldsymbol{x}) := (1-\eta_P)p_P(\boldsymbol{x}) + \eta_P p_N(\boldsymbol{x}) ;$$

$$\mathcal{X}_{\tilde{N}} := \{\boldsymbol{x}'_j\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_{\tilde{N}}(\boldsymbol{x}) := \eta_N p_P(\boldsymbol{x}) + (1-\eta_N)p_N(\boldsymbol{x}) .$$

The problem of learning with noisy labels can be seen as a variant of learning from two contaminated sets or unlabeled-unlabeled learning. The distinctions between these two problem formulations are discussed in [32]. Specifically, learning with noisy labels assumes a noisy rate of $\eta_P + \eta_N < 0.5$, while the class prior remains the same regardless of the presence of noise. Without this assumption, the marginal distribution, $p(\boldsymbol{x})$, may change, which would require addressing the problem under a covariate shift assumption.

By simply substituting the two noisy distributions for $\mathcal{X}_A$ and $\mathcal{X}_B$ in (7), we obtain the noisy AUC risk:

$$R_{\tilde{P}\tilde{N}}(f) := \underset{\boldsymbol{x}\sim p_{\tilde{P}}(\boldsymbol{x})}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\sim p_{\tilde{N}}(\boldsymbol{x})}{\mathbb{E}} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right] . \quad (10)$$

And by letting $\theta_A = 1 - \eta_P$ and $\theta_B = \eta_N$, its easy to show the following corollary:

Corollary 3: $R_{\tilde{P}\tilde{N}}$ is consistent with the AUC risk over the true distribution R_{PN}, and

$$R_{\tilde{P}\tilde{N}} = \tilde{a}R_{PN} + \frac{1-\tilde{a}}{2} ,$$

$$\tilde{a} = 1 - \eta_P - \eta_N .$$

This indicates that the noisy AUC risk is still consistent with the clean AUC risk. Empirically, we can solve the following ERM problem:

$$\min_f \ \hat{R}_{\tilde{P}\tilde{N}}(f) = \frac{1}{|\mathcal{X}_{\tilde{P}}||\mathcal{X}_{\tilde{N}}|} \sum_{\boldsymbol{x}\in\mathcal{X}_{\tilde{P}}} \sum_{\boldsymbol{x}'\in\mathcal{X}_{\tilde{N}}} \ell(f(\boldsymbol{x}, \boldsymbol{x}')). \quad (11)$$

2) Positive-Unlabeled AUC Optimization: We next discuss positive-unlabeled AUC optimization, in which only supervision of one class is available. Suppose the underlying class prior probabilities of positive and negative class are π_P and π_N, in this case, the instances with positive labels can be regarded as a pure positive set $\mathcal{X}_P$ with $\theta_A = 1$, and the unlabeled data consists an contaminated set $\mathcal{X}_U$ with $\theta_B = \pi_P$. Then the two sets can be formulated as:

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}) , \text{ and}$$

$$\mathcal{X}_U := \{\boldsymbol{x}'_j\}_{j=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p_U(\boldsymbol{x}) := \pi_P p_P(\boldsymbol{x}) + \pi_N p_N(\boldsymbol{x}) .$$

Then we define P-U AUC risk on the positive and unlabeled set:

$$R_{PU}(f) := \underset{\boldsymbol{x}\sim p_P(\boldsymbol{x})}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\sim p_U(\boldsymbol{x})}{\mathbb{E}} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right] . \quad (12)$$

And it is easy to show that:

Corollary 4: The P-U AUC risk R_{PU} is consistent with the true AUC risk R_{PN}, and

$$R_{PU} = \pi_N R_{PN} + \frac{\pi_P}{2} .$$

Such a formulation connects the positive-unlabeled AUC optimization problem with the noisy label case by treating one set as having the smallest noise rate and the other set as having the largest noise ratio (as noisy as the marginal distribution $p(\boldsymbol{x})$). It shows that the AUC optimization with one-sided supervision can be solved by treating the unlabeled data as negative data. Practically, we solve the positive-unlabeled AUC optimization problem by minimizing the following empirical risk:

$$\min_f \ \hat{R}_{PU}(f) = \frac{1}{|\mathcal{X}_P||\mathcal{X}_U|} \sum_{\boldsymbol{x}\in\mathcal{X}_P} \sum_{\boldsymbol{x}'\in\mathcal{X}_U} \ell(f(\boldsymbol{x}, \boldsymbol{x}')). \quad (13)$$

3) Multi-Instance AUC Optimization: We demonstrate how our unified view can be applied to multi-instance AUC optimization, a learning problem with inexact, bag-level supervision. In this case, we have a set of positive bags $S_P = \{B_i^+\}_{i=1}^{N_P}$ and a set of negative bags $S_N = \{B_j^-\}_{j=1}^{N_N}$. Each positive bag B_i^+ contains at least one positive instance, while the negative bag B_j^- contains none of positive instances. Following the formulation of the prior research [87], we regard the instance occurred in the negative bags as they are sampled from the pure negative distribution p_N, while the instances occurred in the positive bags as from some mixture distribution $p_{\tilde{P}}$ of the positive and negative distributions with some unknown proportion $1 - \eta_P$ and η_P. Then, the AUC risk over the positive bag instances and negative bag instances can be defined as:

$$R_{\tilde{P}N}(f) := \underset{\boldsymbol{x}\sim p_{\tilde{P}}(\boldsymbol{x})}{\mathbb{E}} \left[\underset{\boldsymbol{x}'\sim p_N(\boldsymbol{x})}{\mathbb{E}} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right] . \quad (14)$$

Corollary 5: The multi-instance AUC risk $R_{\tilde{P}N}$ is consistent with the true AUC risk R_{PN}, and

$$R_{\tilde{P}N} = (1 - \eta_P)R_{PN} + \frac{\eta_P}{2}.$$

The above corollary provides us a way to handle the AUC optimization in multi-instance learning. Practically, we need to first construct the instance sets $\mathcal{X}_{\tilde{P}}$ and $\mathcal{X}_N$ by taking the union of the instance bags:

$$\mathcal{X}_{\tilde{P}} = \bigcup_{i=1}^{N_P} B_i^+, \quad \mathcal{X}_N = \bigcup_{j=1}^{N_N} B_j^-, \quad (15)$$

and solve the following ERM problem:

$$\min_f \quad \hat{R}_{\tilde{P}N}(f) = \frac{1}{|\mathcal{X}_{\tilde{P}}||\mathcal{X}_N|} \sum_{\boldsymbol{x} \in \mathcal{X}_{\tilde{P}}} \sum_{\boldsymbol{x}' \in \mathcal{X}_N} \ell(f(\boldsymbol{x}, \boldsymbol{x}')). \quad (16)$$

This approach produces scores on instance level. To obtain bag-level score, one can simply calculate the maximum of the instance scores in the bags.

4) Semi-Supervised AUC Optimization: Building a model with limited labeled data and a relatively large amount of unlabeled data is a common scenario in semi-supervised learning. In contrast to the previous cases, the data can be divided into three sets based on their labels:

$$\mathcal{X}_P := \{\boldsymbol{x}_i\}_{i=1}^{n_P} \overset{\text{i.i.d.}}{\sim} p_P(\boldsymbol{x}),$$

$$\mathcal{X}_N := \{\boldsymbol{x}_j'\}_{j=1}^{n_N} \overset{\text{i.i.d.}}{\sim} p_N(\boldsymbol{x}), \text{ and}$$

$$\mathcal{X}_U := \{\boldsymbol{x}_k''\}_{k=1}^{n_U} \overset{\text{i.i.d.}}{\sim} p_U(\boldsymbol{x}) := \pi_P p_P(\boldsymbol{x}) + \pi_N p_N(\boldsymbol{x}).$$

Symmetry to the P-U AUC risk (12), we can define U-N AUC risk over unlabeled and negative data:

$$R_{UN}(f) := \mathbb{E}_{\boldsymbol{x} \sim p_U(\boldsymbol{x})} \left[\mathbb{E}_{\boldsymbol{x}' \sim p_N(\boldsymbol{x})} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]. \quad (17)$$

By combining each two the three data sets, we have three risk terms: R_{PN}, R_{PU}, and R_{UN}. Minimizing each risk term corresponds to a subproblem in the form of (8). However, the risk terms calculated with unlabeled data, i.e., R_{PU} and R_{UN}, are biased. Next, we show that, with a proper combination of all the three terms, the unbiasedness can be maintained while the unlabeled data is fully utilized in the learning process.

First, it can be shown that that by combining the P-U AUC risk and U-N AUC risk, the summation risk yields an unbiased risk estimator even if the class prior probabilities are unknown [29].

Theorem 6: The sum of R_{PU} and R_{UN} is consistent with the true AUC risk R_{PN}, and the bias is always $\frac{1}{2}$.

$$R_{PU} + R_{UN} - \frac{1}{2} = R_{PN}. \quad (18)$$

Proof: For any f, we have

$$R_{PU}(f) + R_{UN}(f)$$

$$= \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}_P} \left[\mathbb{E}_{\boldsymbol{x}' \in \mathcal{X}_U} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right] + \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}_U} \left[\mathbb{E}_{\boldsymbol{x}' \in \mathcal{X}_N} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$= \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}_P} \left[\pi_N \mathbb{E}_{\boldsymbol{x}' \in \mathcal{X}_N} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] + \pi_P \mathbb{E}_{\boldsymbol{x}' \in \mathcal{X}_P} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$+ \mathbb{E}_{\boldsymbol{x}' \in \mathcal{X}_N} \left[\pi_P \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}_P} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] + \pi_N \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}_N} [\ell(f(\boldsymbol{x}, \boldsymbol{x}'))] \right]$$

$$= (\pi_P + \pi_N)R_{PN}(f) + \pi_P R_{PP}(f) + \pi_N R_{NN}(f)$$

$$= R_{PN}(f) + \frac{1}{2}.$$

Thus we obtain the theorem. $\square$

This also indicates that in the semi-supervised scenario, we can achieve unbiased AUC risk estimation by subtracting a $\frac{1}{2}$ without knowing the class prior.

Second, to fully exploit the data for reducing the estimation variance, instead of directly minimizing (18) through ERM, we define the following risk:

$$R_{PNU} = \gamma R_{PN} + (1 - \gamma) \left(R_{PU} + R_{UN} - \frac{1}{2} \right), \quad (19)$$

where γ is the weighting coefficient.

To calculate the empirical risk $\hat{R}_{PNU}$, we need to sum up the pairwise losses over three data set pairs: $\mathcal{X}_P \times \mathcal{X}_N$, $\mathcal{X}_P \times \mathcal{X}_U$, and $\mathcal{X}_U \times \mathcal{X}_N$. The bias induced by $(R_{PU} + R_{UN})$ is constantly $1/2$, which can be compensated by subtracting it from the empirical risk. By doing so, the empirical risk $\hat{R}_{PNU}$ becomes an unbiased risk estimator of the true AUC risk. Practically, whether to compensate the bias does not affect the training of the model, so we can simply ignore it.

5) Semi-Supervised AUC Optimization With Label Noise: We further consider the scenario where the available supervision is inaccurate and incomplete. Such a situation typically involves learning from a relatively small number of inaccurately labeled instances and a set of unlabeled instances. Let η_P and η_N denote the probabilities of a positive and a negative label being incorrect, respectively, and let π_P and π_N denote the class prior probabilities of the true distribution, with $\pi_P + \pi_N = 1$. It is noteworthy that we do not require knowledge of these values. The instances can then be categorized into three sets: a noisy positive set $\mathcal{X}_{\tilde{P}} \sim p_{\tilde{P}}$, a noisy negative set $\mathcal{X}_{\tilde{N}} \sim p_{\tilde{N}}$, and an unlabeled set $\mathcal{X}_U \sim p_U$.

Similar to the former case, by combining the noisy version of the P-U AUC risk and U-N AUC risk, the biased risk yields the same attenuation coefficient as in the noisy case.

Corollary 7: The sum of $R_{\tilde{P}U}$ and $R_{U\tilde{N}}$ is consistent with the true AUC risk R_{PN}.

$$R_{\tilde{P}U} + R_{U\tilde{N}} - \frac{1}{2} = \tilde{a}R_{PN} + \frac{1 - \tilde{a}}{2},$$

$$\tilde{a} = 1 - \eta_P - \eta_N.$$

To reduce the variance of the risk estimator, we define the following risk:

$$R_{\tilde{P}\tilde{N}U} = \gamma R_{\tilde{P}\tilde{N}} + (1-\gamma)\left(R_{\tilde{P}U} + R_{U\tilde{N}} - \frac{1}{2}\right). \quad (20)$$

And the ERM problem can be solved like the previous semi-supervised case.

6) Summary: In this section, we show that various scenarios of weakly supervised AUC optimization problems can be addressed by minimizing the sum or weighted average of one or more AUC risk terms over instance set pairs. This type of optimization problem can be easily solved by incorporating additional instance pairs into a standard pairwise AUC optimization algorithm. The settings discussed are summarized in Table I. In Section V we introduce the theoretical analysis of the framework. In Section VI, we further propose a robust learning solution for the problems.

V. THEORETICAL ANALYSES

In this section, we theoretically analyze the proposed risk functions, which consistent with the AUC risk over the true distribution. Briefly, we (1) prove the excess risk bounds for the general case and inaccurate and incomplete case, which can be easily applied to all of the WSL scenarios discussed above, and (2) discuss the variance reduction of the incomplete supervised learning scenario, showing that by introducing the unlabeled data, we can achieve better risk estimation with lower variance.

Here, we consider a kernel K over $\mathcal{X}^2$, a strictly positive real number C_w. Let $\mathcal{F}_K$ be a class of functions:

$$\mathcal{F}_K = \{f_w : \mathcal{X} \to R, f_w(x) = K(w,x) \| w \|_K \le C_w\},$$

where $\|x\|_K = \sqrt{K(x,x)}$. We also assume that the surrogate loss ℓ is L-Lipschitz continuous, bounded by a strictly positive real number C_ℓ, and satisfies inequality $\ell \ge \ell_{01}$. For example, the squared loss and exponential loss satisfy these condition.

A. Excess Risk

In this part, we prove the excess risk bounds when we minimize the proposed risk functions.

Denote by $\hat{f}^*_{AB}$ the minimizer of empirical risk $\hat{R}_{AB}(f)$, we introduce the following excess risk bound, showing that the risk of $\hat{f}^*_{AB}$ converges to risk of the optimal decision function in the function family $\mathcal{F}_K$.

Theorem 8 (Excess Risk of General Case): Assume that $\hat{f}^*_{AB} \in \mathcal{F}_K$ is the minimizer of empirical risk $\hat{R}_{AB}(f)$, $f^*_{PN} \in \mathcal{F}_K$ is the minimizer of true risk $R_{PN}(f)$. For any $\delta > 0$, with the probability at least $1-\delta$, we have

$$R_{PN}(\hat{f}^*_{AB}) - R_{PN}(f^*_{PN}) \le \frac{h(\delta)}{a}\sqrt{\frac{n_A + n_B}{n_A n_B}},$$

where $h(\delta) = 8\sqrt{2}C_\ell C_w C_x + 5\sqrt{2\ln(2/\delta)}$, and n_A, n_B is the size of sampled contaminated instance set.

Proof: Let $R'_{AB}(f) = \frac{R_{AB} - \frac{1}{2}}{a}$ denote the linear transformation of R_{AB} to estimate R_{PN}, and $\hat{R}'_{AB}(f)$ denote its empirical estimation. The excess risk of optimizing $\hat{R}_{AB}(f)$ can

be write as

$$R_{PN}(\hat{f}^*_{AB}) - R_{PN}(f^*_{PN})$$
$$= R_{PN}(\hat{f}^*_{AB}) - \hat{R}'_{AB}(\hat{f}^*_{AB}) + \hat{R}'_{AB}(\hat{f}^*_{AB})$$
$$- \hat{R}'_{AB}(f^*_{PN}) + \hat{R}'_{AB}(f^*_{PN}) - R_{PN}(f^*_{PN})$$
$$\le 2\max_{f\in\mathcal{F}}|\hat{R}'_{AB}(f) - R_{PN}(f)|. \quad (21)$$

Due to Theorem 1, the right term can be write as

$$\max_{f\in\mathcal{F}}|\hat{R}'_{AB}(f) - R_{PN}(f)| = \max_{f\in\mathcal{F}}|\hat{R}'_{AB}(f) - R'_{AB}(f)|. \quad (22)$$

According to Theorem 6 in [94], for any $\delta > 0$, with probability at least $1-\delta$ for any $f \in \mathcal{F}_K$:

$$\max_{f\in\mathcal{F}}|\hat{R}_{AB}(f) - R_{AB}(f)|$$
$$\le 4\sqrt{2}C_\ell C_w C_x\sqrt{\frac{n_A+n_B}{n_A n_B}} + 5\sqrt{\frac{n_A+n_B}{2n_A n_B}\ln(2/\delta)}, \quad (23)$$

where $C_x = \max(\max_i \|x_i\|, \max_j \|x'_j\|)$. For convenience, we define

$$h(\delta) = 8\sqrt{2}C_\ell C_w C_x + 5\sqrt{2\ln(2/\delta)}.$$

So we have

$$\max_{f\in\mathcal{F}}|\hat{R}'_{AB}(f) - R'_{AB}(f)| \le \frac{h(\delta)}{2a}\sqrt{\frac{n_A+n_B}{n_A n_B}}. \quad (24)$$

Applying (22) and (24) to the right term in (21), we obtain the theorem. □

Theorem 8 guarantees that the excess risk of general case can be bounded plus the confidence term of order

$$\mathcal{O}\left(\frac{1}{a\sqrt{n_A}} + \frac{1}{a\sqrt{n_B}}\right).$$

Denote by $\hat{f}^*_{\tilde{P}\tilde{N}U}$ the minimizer of the empirical risk $\hat{R}_{\tilde{P}\tilde{N}U}(f)$, similarly, we have a theorem to show that the risk of $\hat{f}^*_{\tilde{P}\tilde{N}U}$ converges to risk of the optimal decision function in the function family $\mathcal{F}_K$.

Theorem 9 (Excess Risk of Inaccurate and Incomplete Case): Assume that $\hat{f}^*_{\tilde{P}\tilde{N}U} \in \mathcal{F}_K$ is the minimizer of empirical risk $\hat{R}_{\tilde{P}\tilde{N}U}(f)$, $f^*_{PN} \in \mathcal{F}_K$ is the minimizer of true risk $R_{PN}(f)$. For any $\delta > 0$, with the probability at least $1-\delta$, we have

$$R_{PN}(\hat{f}^*_{\tilde{P}\tilde{N}U}) - R_{PN}(f^*_{PN}) \le \frac{h(\frac{\delta}{3})}{\tilde{a}}\left(\gamma\sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}} + (1-\gamma)\right.$$
$$\left. \times \left(\sqrt{\frac{n_{\tilde{P}}+n_U}{n_{\tilde{P}}n_U}} + \sqrt{\frac{n_U+n_{\tilde{N}}}{n_U n_{\tilde{N}}}}\right)\right),$$

where $h(\delta) = 8\sqrt{2}C_\ell C_w C_x + 5\sqrt{2\ln(2/\delta)}$, and $n_{\tilde{P}}, n_{\tilde{N}}$ is the size of sampled contaminated instance set.

Proof: Let $R'_{\tilde{P}\tilde{N}U}(f) = \gamma\frac{R_{\tilde{P}\tilde{N}} - \frac{1-\tilde{a}}{2}}{\tilde{a}} + (1-\gamma)$ $\frac{R_{\tilde{P}U}+R_{U\tilde{N}} - \frac{1}{2} - \frac{1-\tilde{a}}{2}}{\tilde{a}}$ denote the linear transformation of $R_{\tilde{P}\tilde{N}U}$ to estimate R_{PN}, and $\hat{R}'_{\tilde{P}\tilde{N}U}(f)$ denote its empirical estimation.

Similarly to (21), the excess risk of optimizing $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ can be write as

$$
\begin{aligned}
R_{PN}&(\hat{f}^*_{\tilde{P}\tilde{N}U}) - R_{PN}(f^*_{PN}) \\
&= R_{PN}(\hat{f}^*_{\tilde{P}\tilde{N}U}) - \hat{R}'_{\tilde{P}\tilde{N}U}(\hat{f}^*_{\tilde{P}\tilde{N}U}) + \hat{R}'_{\tilde{P}\tilde{N}U}(\hat{f}^*_{\tilde{P}\tilde{N}U}) \\
&\quad - \hat{R}'_{\tilde{P}\tilde{N}U}(f^*_{PN}) + \hat{R}'_{\tilde{P}\tilde{N}U}(f^*_{PN}) - R_{PN}(f^*_{PN}) \\
&\leq 2 \max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}U}(f) - R_{PN}(f)|.
\end{aligned} \tag{25}
$$

Due to Corollary 7, the right term can be write as

$$
\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}U}(f) - R_{PN}(f)| = \max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}U}(f) - R'_{\tilde{P}\tilde{N}U}(f)|. \tag{26}
$$

Respectively, replace $x\in\mathcal{X}_A, x'\in\mathcal{X}_B$ by $x\in\mathcal{X}_{\tilde{P}}, x'\in\mathcal{X}_{\tilde{N}}$, $x\in\mathcal{X}_{\tilde{P}}, x'\in\mathcal{X}_U$, and $x\in\mathcal{X}_U, x'\in\mathcal{X}_{\tilde{N}}$ in (24), we have that for any $\delta > 0$, with probability at least $1-\delta$ for any $f\in\mathcal{F}_K$:

$$
\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}}(f) - R'_{\tilde{P}\tilde{N}}(f)| \leq \frac{h(\delta)}{2\tilde{a}} \sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}},
$$

$$
\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}U}(f) - R'_{\tilde{P}U}(f)| \leq \frac{h(\delta)}{2\tilde{a}} \sqrt{\frac{n_{\tilde{P}}+n_U}{n_{\tilde{P}}n_U}},
$$

$$
\max_{f\in\mathcal{F}} |\hat{R}'_{U\tilde{N}}(f) - R'_{U\tilde{N}}(f)| \leq \frac{h(\delta)}{2\tilde{a}} \sqrt{\frac{n_U+n_{\tilde{N}}}{n_U n_{\tilde{N}}}}.
$$

Simple calculation showed that for any $\delta' > 0$, with probability at least $1-\delta'$, we have

$$
\begin{aligned}
\max_{f\in\mathcal{F}} & |\hat{R}'_{\tilde{P}\tilde{N}U}(f) - R'_{\tilde{P}\tilde{N}U}(f)| \\
&\leq \gamma \left(\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}\tilde{N}}(f) - R'_{\tilde{P}\tilde{N}}(f)| \right) \\
&\quad + (1-\gamma) \left(\max_{f\in\mathcal{F}} |\hat{R}'_{\tilde{P}U}(f) - R'_{\tilde{P}U}(f)| \right) \\
&\quad + (1-\gamma) \left(\max_{f\in\mathcal{F}} |\hat{R}'_{U\tilde{N}}(f) - R'_{U\tilde{N}}(f)| \right) \\
&\leq \frac{h(\frac{\delta'}{3})}{2\tilde{a}} \Bigg(\gamma \sqrt{\frac{n_{\tilde{P}}+n_{\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}}} + (1-\gamma) \\
&\quad \times \left(\sqrt{\frac{n_{\tilde{P}}+n_U}{n_{\tilde{P}}n_U}} + \sqrt{\frac{n_U+n_{\tilde{N}}}{n_U n_{\tilde{N}}}} \right) \Bigg).
\end{aligned} \tag{27}
$$

Applying (26) and (24) to the right term in (25), we obtain the theorem. □

Theorem 9 guarantees that the excess risk of inaccurate and incomplete case can be bounded plus the confidence term of order

$$
\mathcal{O}\left(\frac{1}{\tilde{a}\sqrt{n_{\tilde{P}}}} + \frac{1}{\tilde{a}\sqrt{n_{\tilde{N}}}} + \frac{1}{\tilde{a}\sqrt{n_U}} \right).
$$

It can be easily seen that Theorem 9 degenerates into Theorem 8 for $\gamma = 1$.

B. Variance Reduction

It shows that our empirical risk estimators proposed is unbiased, and the excess risk can be bounded. Similar to [28], the next question is that, in the incomplete scenario i.e., when $\gamma < 1$, whether the variance of $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ can be smaller than it of $\hat{R}_{\tilde{P}\tilde{N}}(f)$, or whether $\mathcal{X}_U$ can help reduce the variance in estimating R_{PN}. To answer this question, pick any f of interest. For simplicity, we assume that $n_U \to \infty$, to illustrate the maximum variance reduction that could be achieved.

The variances and covariances are defined below:

$$
\begin{aligned}
\sigma^2_{\tilde{P}\tilde{N}}(f) &= \text{Var}_{\tilde{P}\tilde{N}}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}}))], \\
\sigma^2_{\tilde{P}U}(f) &= \text{Var}_{\tilde{P}U}[\ell(f(x_{\tilde{P}}, x_U))], \\
\sigma^2_{U\tilde{N}}(f) &= \text{Var}_{U\tilde{N}}[\ell(f(x_U, x_{\tilde{N}}))], \\
\tau_{\tilde{P}\tilde{N},\tilde{P}U}(f) &= \text{Cor}_{\tilde{P}\tilde{N},\tilde{P}U}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}})), \ell(f(x_{\tilde{P}}, x_U))], \\
\tau_{\tilde{P}\tilde{N},U\tilde{N}}(f) &= \text{Cor}_{\tilde{P}\tilde{N},U\tilde{N}}[\ell(f(x_{\tilde{P}}, x_{\tilde{N}})), \ell(f(x_U, x_{\tilde{N}}))], \\
\tau_{\tilde{P}U,U\tilde{N}}(f) &= \text{Cor}_{\tilde{P}U,U\tilde{N}}[\ell(f(x_{\tilde{P}}, x_U)), \ell(f(x_U, x_{\tilde{N}}))].
\end{aligned}
$$

Then the following theorem can be obtained:

Theorem 10: Assume $n_U \to \infty$, for any fixed f, the minimizers of the variance of the empirical risk $\hat{R}_{\tilde{P}\tilde{N}U}(f)$ is obtained by

$$
\gamma_{\tilde{P}\tilde{N}} = \arg\min_\gamma \text{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] = \frac{\psi_{\tilde{P}\tilde{N}U}}{\psi_{\tilde{P}\tilde{N}U} - \psi_{\tilde{P}\tilde{N}}},
$$

where

$$
\psi_{\tilde{P}\tilde{N}} = \frac{1}{n_{\tilde{P}}n_{\tilde{N}}} \sigma^2_{\tilde{P}\tilde{N}}(f), \text{ and}
$$

$$
\psi_{\tilde{P}\tilde{N}U} = \frac{1}{n_{\tilde{P}}} \tau_{\tilde{P}\tilde{N},\tilde{P}U}(f) + \frac{1}{n_{\tilde{N}}} \tau_{\tilde{P}\tilde{N},U\tilde{N}}(f).
$$

In addition, we have $\text{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] \leq \text{Var}[\hat{R}_{\tilde{P}\tilde{N}}(f)]$ for any $\gamma \in (2\gamma_{\tilde{P}\tilde{N}} - 1, 1)$ if $\psi_{\tilde{P}\tilde{N}U} > \psi_{\tilde{P}\tilde{N}}$.

Proof: The empirical risk can be expressed as

$$
\begin{aligned}
\hat{R}_{\tilde{P}\tilde{N}U}(f) &= \gamma \hat{R}_{\tilde{P}\tilde{N}}(f) + (1-\gamma)\left(\hat{R}_{\tilde{P}U}(f) + \hat{R}_{U\tilde{N}}(f) - \frac{1}{2} \right) \\
&= \frac{\gamma}{n_{\tilde{P}}n_{\tilde{N}}} \sum_{i=1}^{n_{\tilde{P}}} \sum_{j=1}^{n_{\tilde{N}}} \ell\left(f\left(x_i^{\tilde{P}}, x_j^{\tilde{N}}\right) \right) \\
&\quad + \frac{\gamma}{n_{\tilde{P}}n_U} \sum_{i=1}^{n_{\tilde{P}}} \sum_{j=1}^{n_U} \ell\left(f\left(x_i^{\tilde{P}}, x_j^U\right) \right) \\
&\quad + \frac{\gamma}{n_U n_{\tilde{N}}} \sum_{i=1}^{n_U} \sum_{j=1}^{n_{\tilde{N}}} \ell\left(f\left(x_i^U, x_j^{\tilde{N}}\right) \right) + \frac{1-\gamma}{2}.
\end{aligned}
$$

Assuming $n_U \to \infty$, it follows that

$$
\begin{aligned}
\text{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)] &= \gamma^2 \frac{\sigma^2_{\tilde{P}\tilde{N}}}{n_{\tilde{P}}n_{\tilde{N}}} + 2\gamma(1-\gamma) \frac{\tau_{\tilde{P}\tilde{N},\tilde{P}U}(f)}{n_{\tilde{P}}} \\
&\quad + 2\gamma(1-\gamma) \frac{\tau_{\tilde{P}\tilde{N},U\tilde{N}}(f)}{n_{\tilde{N}}}
\end{aligned}
$$

$$= \gamma^2 \psi_{\tilde{P}\tilde{N}} + 2\gamma(1-\gamma)\psi_{\tilde{P}\tilde{N}U},$$

where the terms divided by n_U are disappeared.

To minimize the variance, the derivative is set to zero with respect to γ,

$$\frac{\text{Var}[\hat{R}_{\tilde{P}\tilde{N}U}(f)]}{\gamma} = 2\gamma\psi_{\tilde{P}\tilde{N}} + (2-2\gamma)\psi_{\tilde{P}\tilde{N}U}$$

$$= (2\psi_{\tilde{P}\tilde{N}} - 2\psi_{\tilde{P}\tilde{N}U})\gamma + 2\psi_{\tilde{P}\tilde{N}U}$$

$$= 0.$$

Solving the equation then the minimizer of variance is obtained. □

Theorem 10 implies that the proposed risk estimator $\hat{R}_{\tilde{P}\tilde{N}U}$ have smaller variance than the standard supervised risk estimator $\hat{R}_{\tilde{P}\tilde{N}}$ if γ is chosen properly, showing that the unlabeled data is helpful for building the model.

VI. RPAUC FOR ROBUST AUC OPTIMIZATION

In previous sections, we developed a unified view of the weakly supervised AUC optimization problems from the perspective of AUC optimization with two contaminated sets. Our findings provide a statistical consistent way to solve the AUC optimization problem in various weakly supervised scenarios. However, with finite training data, sorely training the models through above ERM problems still suffers from the contaminated supervision, the negative impact is reflected in the coefficients in the bounds (c.f. a in Theorem 8 and $\tilde{a}$ in Theorem 9).

To mitigate this problem, in this section, we introduce a novel partial AUC, namely two-way reversed partial AUC (rpAUC), and demonstrate that by maximizing rpAUC through empirical risk minimization on contaminated data sets, WSAUC achieves robust AUC optimization. We will first introduce the definition of rpAUC, which is obtained by reversing the restrictions of two-way pAUC. Then, we show the equivalency between the rpAUC optimization and the small loss trick, which is a commonly used technique for learning with noisy labels [36], [37], [38], [95]. This results in a straightforward solution to solve weakly supervised AUC optimization by simply replacing the full AUC risk with partial AUC risk.

Regarding the contaminated distribution p_A and p_B as positive and negative, the two-way reversed partial AUC (rpAUC) is defined as follows.

Definition 11: Two-way Reverse Partial AUC with FPR threshold α and TPR threshold β of model f can be defined as:

$$\text{rpAUC}(f; \alpha, \beta) = 1 - \underset{\boldsymbol{x} \sim p_A^+(\boldsymbol{x})}{\mathbb{E}} \left[\underset{\boldsymbol{x}' \sim p_B^-(\boldsymbol{x})}{\mathbb{E}} [\ell_{01}(f(\boldsymbol{x}, \boldsymbol{x}'))] \right],$$

where $p_A^+(\boldsymbol{x}) = p_A(\boldsymbol{x}|f(\boldsymbol{x}) \in [\text{TPR}_f^{-1}(\beta), \infty))$,

$$p_B^-(\boldsymbol{x}) = p_B(\boldsymbol{x}|f(\boldsymbol{x}) \in (-\infty, \text{FPR}_f^{-1}(\alpha)]).$$

An illustration of rpAUC is shown in Fig. 1(d). By reversing the constraints of TPR and FPR, rpAUC trims the leftmost and the uppermost margin of the ROC curve. This is equivalent

---

Algorithm 1: rpAUC for Robust WSAUC Optimization.

Input: Clean or noisy instance sets: $\{\mathcal{X}_P[, \mathcal{X}_U][, \mathcal{X}_N]\}$;
Input: hyper-parameters α, β.
Initialize model f.
for $t = 1 \to T$ **do**
  for $k = 1 \to K$ **do**
    for each pair of input sets $(\mathcal{X}_A, \mathcal{X}_B)$ **do**
      Remove instances with top α / β proportion losses.
      Sample batches $B_A \in \mathcal{X}_A^+, B_B \in \mathcal{X}_B^-$.
      Calculate rpAUC risk $\hat{R}_{AB}^{(\text{rp})}(f)$ on the batch.
      Update the model f by back propagation.
    end for
  end for
end for
Output: The trained model f.

---

to eliminating β proportion of positive instances with bottom scores and α proportion of negative instances with top scores by its definition. Next, to show that maximizing rpAUC is de facto removing instances include largest losses when maximizing AUC, we prove that the instances that induce largest losses are those who has bottom scores in $\mathcal{X}_A$ and top scores in $\mathcal{X}_B$.

The instance loss can be defined as follows:

$$L(x) = \begin{cases} \frac{1}{|\mathcal{X}_B|} \sum_{\boldsymbol{x}' \in \mathcal{X}_B} \ell(f(\boldsymbol{x}, \boldsymbol{x}')), & \text{if } \boldsymbol{x} \in \mathcal{X}_A, \\ \frac{1}{|\mathcal{X}_A|} \sum_{\boldsymbol{x}' \in \mathcal{X}_A} \ell(f(\boldsymbol{x}', \boldsymbol{x})), & \text{if } \boldsymbol{x} \in \mathcal{X}_B. \end{cases}$$

Suppose the surrogate loss $l(z)$ is monotonically nonincreasing, it is easy to show that $L(x)$ varies monotonically with $f(x)$. Hence we have the following proposition.

Proposition 12: Suppose f is a scoring function. Then its rpAUC risk $\hat{R}_{AB}^{(\text{rp})}(f; \alpha, \beta)$ equals to full AUC risk $\hat{R}_{AB}(f)$ with α proportion of instances induce top losses $L(x)$ from $\mathcal{X}_B$ and β proportion of instances induce top losses from $\mathcal{X}_A$ removed.

This proposition bridges the minimization of empirical rpAUC risk with clean label selection, which has been shown effective for learning from contaminated labels. To achieve robust AUC optimization, in training phase we minimize the following empirical rpAUC risk:

$$\hat{R}_{AB}^{(\text{rp})}(f; \alpha, \beta) = \frac{1}{|\mathcal{X}_A^+||\mathcal{X}_B^-|} \sum_{\boldsymbol{x} \in \mathcal{X}_A^+} \sum_{\boldsymbol{x}' \in \mathcal{X}_B^-} \ell(f(\boldsymbol{x}, \boldsymbol{x}')), \quad (28)$$

where $\mathcal{X}_A^+$ is a set of instances in $\mathcal{X}_A$ with top $\lfloor (1-\beta)|\mathcal{X}_A| \rfloor$ scores, and $\mathcal{X}_B^-$ is a set of instances in $\mathcal{X}_B$ with bottom $\lfloor (1 - \alpha)|\mathcal{X}_B| \rfloor$ scores. This can be achieved by applying any existing two-way pAUC optimization algorithm, e.g., [15], [19], [96], but reverses the instance selection in two sets. Algorithm 1 shows a possible procedure for rpAUC optimization, which is simply based on pairwise gradient.

VII. EXPERIMENTS

In this section, we empirically evaluate the proposed WSAUC for robust weakly supervised AUC optimization. We compare WSAUC with multiple baselines under different WSL scenarios to show that WSAUC provides a unified yet effective solution for

building AUC-optimized models under these scenarios. Furthermore, we compare rpAUC against normal AUC as the training objective under varying noise ratios to demonstrate the benefits of using rpAUC in WSAUC when the labels are not perfectly clean.

A. Experimental Setup

The experiments under various weakly supervised learning scenarios are conducted with commonly used benchmark datasets. For AUC optimization with inaccurate or incomplete supervision, we utilize image benchmark MNIST, FashionMNIST, CIFAR10, and CIFAR100 to synthetic multiple datasets with varying setups per task. The datasets are transformed into binary classification datasets, i.e., odd versus even numbers for MNIST, upper versus lower garments for FashionMNIST, and animals versus non-animals for CIFAR datasets. Such a dataset transformation is widely used in th related literature, e.g., [32], [97], [98]. For AUC optimization with inexact supervision, we adopt several multi-instance learning datasets[1] that are commonly used in the related literature, including Musk1, Musk2, fox, tiger, and elephant.

Our method is compared with multiple baselines designed for different problem settings, which will be introduced in the following subsections for each setting. All methods are implemented with the same backbone and optimizer where applicable. Specifically, for CIFAR10 and CIFAR100 datasets, the backbone is Mini-VGG; for MNIST, FashionMNIST, and the MIL datasets, the backbone is a simple fully-connected network with one hidden layer. The hyper-parameters of the baselines are tuned through grid search or selected according to their original papers. The hyper-parameter γ of WSAUC is set as 0.45 for all tasks. All of the experiments are repeated for 10 times to eliminate the impact of the randomness, the AUC performance and the standard deviation are reported. For more details please refer to the released code.

B. WSAUC in Various Weakly Supervised Settings

1) Noisy Label AUC Optimization: For noisy label AUC optimization, we compare WSAUC with several noisy label learning losses/methods for AUC optimization: AUC-B [30], a noisy AUC optimization method based on the barrier hinge loss, which is shown to be robust to the corrupted labels. AUC-H, AUC-R, AUC-U are variations of AUC-B with different surrogate loss functions: hinge loss, ramp loss, and unhinged loss, respectively. DRAUC [99], a robust self-paced AUC optimization model against noisy and adversarial examples. We compare the baselines on the task of AUC optimization with label noise, with varying noise ratios. The noise ratios of the positive data and negative data are selected from $\{20\%, 30\%, 40\%\}$, respectively, and all different combinations of the positive and negative ratios are tested for showing the performances under symmetrical and asymmetrical noise.

The results are shown in Table II. The results indicate that the performance of the multiple methods are close on the easy

datasets MNIST and FashionMNIST, while WSAUC achieves relatively larger improvement on the challenging datasets CIFAR10 and CIFAR100. DRAUC shows unstable performance on CIFAR datasets, which may be due to the assumption of the generative models not holding on those datasets.

2) Positive-Unlabeled AUC Optimization: We compare the following baselines on the task of positive-unlabeled AUC optimization: PU-AUC [28] and SAMULT^{P+U} [29]. PU-AUC minimizes an unbiased AUC risk by compensating the excess risk induced by the positive data in the unlabeled set. SAMULT^{P+U} is the degenerate version of SAMULT. It achieves PU AUC optimization by showing the consistency of risk obtained by treating the unlabeled data as negative. The positive-unlabeled task is known to be sensitive to the class prior probability and label ratio. Thus, we test the situations where the positive ratios are chosen in $\{20\%, 30\%, 40\%\}$, and the labeled ratios are chosen in $\{5\%, 10\%\}$. We compare WSAUC with PU-AUC and SAMULT^{P+U}, which are the variations of PNU-AUC and SAMULT, respectively.

The results are shown in Table III. It can be observed that WSAUC shows its advantage on most of the datasets and scenarios, especially when the label ratio is relatively low. This indicates that the advantage of WSAUC is greater when the amount of labeled data is small.

3) Multi-Instance AUC Optimization: For multi-instance AUC optimization, we compare WSAUC with multi-label learning approaches: MI-SVM and mi-SVM [78], two margin based approach for multi-instance learning. MissSVM [100], which uses a semi-supervised learning approach, treating the instances in positive bags as unlabeled data. SIL [101], a multi-instance learning approach that learns from the bag label. The multi-instance learning approach sbMIL [102], which is specially designed for the situation that the positive bags are sparse. And two varieties of MIDAM [103], which is a multi-instance AUC optimization method based on non-convex min-max optimization with smoothed-max (smx) and attention-based (att) pooling. Due to the challenge of synthesizing multi-instance tasks from image benchmark datasets, we conduct experiments on several widely used multi-instance learning datasets, i.e., Musk1, Musk2, fox, tiger, and elephant. The AUC is calculated based on the scores of the instance bags. For WSAUC, the model is trained at the instance level as described previously, and the bag score is predicted as the maximum instance score in the bag.

The results are shown in Table IV. It is indicated that WSAUC achieves relatively large improvement, compared with miSVM, MISVM, MissSVM, SIL, and sbMIL. This may be because the these baselines are not explicitly optimizing AUC, indicating that explicitly optimizing instance-level AUC with WSAUC in the context of multi-instance learning is beneficial for achieve better bag-level AUC performance, and the robust training is helpful for the multi-instance learning task. Compared to MIDAM, a dedicated multi-instance method for AUC optimization, WSAUC still demonstrates significant performance improvement. This indicates the effectiveness of introducing rpAUC for multi-instance AUC optimization.

4) Semi-Supervised AUC Optimization With Label Noise: To the best of our knowledge, no existing AUC optimization

---

[1] http://www.uco.es/grupos/kdis/momil/

TABLE II
PERFORMANCE (STD) OF NOISY LABEL AUC OPTIMIZATION

Dataset	Pos. noise	20%			30%			40%		
	Neg. noise	20%	30%	40%	20%	30%	40%	20%	30%	40%
MNIST	DRAUC	$87.2_{\pm4.3}$	$87.1_{\pm4.8}$	$87.1_{\pm4.4}$	$89.7_{\pm4.5}$	$88.9_{\pm4.9}$	$80.7_{\pm6.4}$	$88.4_{\pm4.9}$	$84.5_{\pm6.0}$	$74.1_{\pm8.8}$
	AUC-H	$\mathbf{99.6_{\pm0.1}}$	$\mathbf{99.6_{\pm0.1}}$	$99.4_{\pm0.1}$	$\mathbf{99.6_{\pm0.1}}$	$\mathbf{99.5_{\pm0.1}}$	$99.2_{\pm0.1}$	$99.4_{\pm0.1}$	$99.2_{\pm0.1}$	$98.7_{\pm0.3}$
	AUC-R	$99.3_{\pm0.1}$	$99.2_{\pm0.1}$	$99.0_{\pm0.1}$	$99.2_{\pm0.1}$	$99.0_{\pm0.1}$	$98.8_{\pm0.1}$	$99.0_{\pm0.1}$	$98.8_{\pm0.2}$	$98.2_{\pm0.3}$
	AUC-U	$88.5_{\pm0.8}$	$88.5_{\pm0.8}$	$88.4_{\pm0.8}$	$88.4_{\pm0.8}$	$88.3_{\pm0.8}$	$88.2_{\pm0.7}$	$88.2_{\pm0.8}$	$88.1_{\pm0.7}$	$87.8_{\pm0.6}$
	AUC-B	$99.5_{\pm0.1}$	$99.5_{\pm0.1}$	$\mathbf{99.5_{\pm0.1}}$	$99.5_{\pm0.1}$	$\mathbf{99.5_{\pm0.1}}$	$\mathbf{99.5_{\pm0.1}}$	$\mathbf{99.6_{\pm0.0}}$	$\mathbf{99.5_{\pm0.0}}$	$\mathbf{99.5_{\pm0.1}}$
	WSAUC	$\mathbf{99.6_{\pm0.0}}$	$\mathbf{99.6_{\pm0.0}}$	$\mathbf{99.5_{\pm0.1}}$	$\mathbf{99.6_{\pm0.0}}$	$\mathbf{99.5_{\pm0.0}}$	$99.4_{\pm0.1}$	$99.5_{\pm0.1}$	$99.4_{\pm0.1}$	$99.1_{\pm0.1}$
FMNIST	DRAUC	$93.5_{\pm4.8}$	$89.0_{\pm5.8}$	$78.7_{\pm5.7}$	$91.2_{\pm5.0}$	$89.5_{\pm7.0}$	$80.9_{\pm6.9}$	$90.1_{\pm6.9}$	$88.6_{\pm5.8}$	$76.4_{\pm9.0}$
	AUC-H	$99.3_{\pm0.1}$	$99.2_{\pm0.1}$	$99.1_{\pm0.1}$	$99.2_{\pm0.1}$	$99.2_{\pm0.1}$	$99.0_{\pm0.1}$	$99.1_{\pm0.1}$	$99.0_{\pm0.1}$	$98.8_{\pm0.1}$
	AUC-R	$99.2_{\pm0.1}$	$99.2_{\pm0.1}$	$99.1_{\pm0.1}$	$99.1_{\pm0.1}$	$99.1_{\pm0.1}$	$98.9_{\pm0.1}$	$99.0_{\pm0.1}$	$98.9_{\pm0.1}$	$98.6_{\pm0.2}$
	AUC-U	$96.6_{\pm0.7}$	$96.5_{\pm0.7}$	$96.4_{\pm0.7}$	$96.7_{\pm0.7}$	$96.6_{\pm0.7}$	$96.6_{\pm0.7}$	$96.8_{\pm0.6}$	$96.7_{\pm0.7}$	$96.7_{\pm0.6}$
	AUC-B	$98.9_{\pm0.1}$	$98.9_{\pm0.1}$	$98.8_{\pm0.1}$	$98.9_{\pm0.1}$	$98.9_{\pm0.1}$	$98.9_{\pm0.1}$	$98.8_{\pm0.1}$	$98.8_{\pm0.1}$	$98.9_{\pm0.1}$
	WSAUC	$\mathbf{99.4_{\pm0.1}}$	$\mathbf{99.4_{\pm0.1}}$	$\mathbf{99.3_{\pm0.1}}$	$\mathbf{99.3_{\pm0.0}}$	$\mathbf{99.3_{\pm0.1}}$	$\mathbf{99.2_{\pm0.1}}$	$\mathbf{99.3_{\pm0.0}}$	$\mathbf{99.2_{\pm0.1}}$	$\mathbf{99.1_{\pm0.1}}$
CIFAR10	DRAUC	$38.8_{\pm9.3}$	$40.2_{\pm9.6}$	$37.7_{\pm12.0}$	$39.6_{\pm9.1}$	$35.2_{\pm13.0}$	$40.1_{\pm8.7}$	$40.3_{\pm9.6}$	$30.0_{\pm13.0}$	$40.1_{\pm12.0}$
	AUC-H	$92.8_{\pm0.1}$	$92.0_{\pm0.3}$	$91.7_{\pm0.4}$	$92.1_{\pm0.3}$	$91.3_{\pm0.2}$	$90.1_{\pm0.1}$	$91.5_{\pm0.4}$	$90.3_{\pm0.3}$	$88.2_{\pm0.4}$
	AUC-R	$87.1_{\pm5.5}$	$88.5_{\pm3.1}$	$89.4_{\pm3.2}$	$89.1_{\pm1.3}$	$89.2_{\pm2.5}$	$88.6_{\pm2.4}$	$87.8_{\pm1.9}$	$89.2_{\pm1.4}$	$85.0_{\pm2.2}$
	AUC-U	$73.5_{\pm1.6}$	$73.3_{\pm3.1}$	$75.5_{\pm3.5}$	$72.7_{\pm2.1}$	$74.2_{\pm5.1}$	$73.1_{\pm4.7}$	$71.6_{\pm1.2}$	$74.4_{\pm4.7}$	$71.6_{\pm1.1}$
	AUC-B	$88.5_{\pm0.7}$	$88.4_{\pm1.4}$	$89.0_{\pm1.0}$	$88.0_{\pm1.4}$	$88.7_{\pm0.9}$	$88.4_{\pm0.9}$	$88.1_{\pm0.8}$	$88.4_{\pm1.4}$	$88.2_{\pm0.8}$
	WSAUC	$\mathbf{93.6_{\pm0.2}}$	$\mathbf{92.8_{\pm0.2}}$	$\mathbf{92.3_{\pm0.2}}$	$\mathbf{92.9_{\pm0.2}}$	$\mathbf{92.2_{\pm0.2}}$	$\mathbf{91.1_{\pm0.3}}$	$\mathbf{92.2_{\pm0.3}}$	$\mathbf{90.9_{\pm0.4}}$	$\mathbf{89.5_{\pm0.2}}$
CIFAR100	DRAUC	$47.4_{\pm5.7}$	$39.8_{\pm8.8}$	$37.8_{\pm7.3}$	$38.1_{\pm10.0}$	$44.4_{\pm7.8}$	$41.5_{\pm6.8}$	$44.8_{\pm5.5}$	$41.7_{\pm7.1}$	$43.6_{\pm5.9}$
	AUC-H	$85.4_{\pm0.5}$	$83.5_{\pm1.0}$	$81.7_{\pm1.0}$	$84.0_{\pm0.8}$	$82.0_{\pm0.8}$	$80.1_{\pm1.2}$	$82.7_{\pm0.6}$	$\mathbf{80.4_{\pm0.7}}$	$76.4_{\pm0.6}$
	AUC-R	$86.3_{\pm1.4}$	$82.6_{\pm5.7}$	$80.7_{\pm3.6}$	$83.6_{\pm2.6}$	$\mathbf{83.7_{\pm1.5}}$	$77.0_{\pm5.6}$	$80.0_{\pm3.2}$	$74.6_{\pm5.6}$	$67.4_{\pm2.6}$
	AUC-U	$63.0_{\pm2.4}$	$61.5_{\pm0.6}$	$62.7_{\pm1.5}$	$64.6_{\pm4.0}$	$63.4_{\pm3.1}$	$63.5_{\pm2.0}$	$62.8_{\pm3.4}$	$64.9_{\pm2.4}$	$61.5_{\pm0.9}$
	AUC-B	$78.1_{\pm2.0}$	$78.9_{\pm2.9}$	$78.5_{\pm2.7}$	$79.0_{\pm3.7}$	$79.1_{\pm4.2}$	$78.8_{\pm3.1}$	$79.4_{\pm4.0}$	$78.2_{\pm4.3}$	$\mathbf{78.3_{\pm2.5}}$
	WSAUC	$\mathbf{86.4_{\pm0.4}}$	$\mathbf{85.3_{\pm0.5}}$	$\mathbf{83.3_{\pm0.7}}$	$\mathbf{84.8_{\pm0.5}}$	$82.5_{\pm0.7}$	$\mathbf{80.3_{\pm1.0}}$	$\mathbf{83.1_{\pm0.5}}$	$80.4_{\pm1.1}$	$76.5_{\pm1.3}$

The best performances are bolded.

TABLE III
PERFORMANCE (STD) OF POSITIVE-UNLABELED AUC OPTIMIZATION

Dataset	Pos. ratio	20%		30%		40%	
	Label ratio	5%	10%	5%	10%	5%	10%
MNIST	PU-AUC	$86.8_{\pm3.1}$	$96.8_{\pm0.5}$	$77.8_{\pm7.4}$	$94.8_{\pm1.1}$	$67.2_{\pm0.7}$	$88.8_{\pm2.9}$
	SAMULT^{P+U}	$86.7_{\pm3.7}$	$97.5_{\pm0.4}$	$74.5_{\pm5.7}$	$95.3_{\pm1.1}$	$65.8_{\pm8.4}$	$89.1_{\pm2.9}$
	WSAUC	$\mathbf{93.8_{\pm3.2}}$	$\mathbf{98.6_{\pm0.1}}$	$\mathbf{88.5_{\pm3.5}}$	$\mathbf{98.2_{\pm0.2}}$	$\mathbf{84.2_{\pm4.6}}$	$\mathbf{95.7_{\pm1.3}}$
FMNIST	PU-AUC	$90.7_{\pm2.4}$	$98.5_{\pm0.3}$	$81.1_{\pm6.1}$	$98.2_{\pm0.3}$	$67.6_{\pm10.4}$	$97.2_{\pm0.6}$
	SAMULT^{P+U}	$92.4_{\pm2.1}$	$98.6_{\pm0.3}$	$83.0_{\pm6.3}$	$98.4_{\pm0.3}$	$67.9_{\pm10.6}$	$97.3_{\pm0.6}$
	WSAUC	$\mathbf{98.1_{\pm1.1}}$	$\mathbf{99.0_{\pm0.1}}$	$\mathbf{89.0_{\pm1.2}}$	$\mathbf{98.9_{\pm0.1}}$	$\mathbf{88.8_{\pm1.7}}$	$\mathbf{98.6_{\pm0.1}}$
CIFAR10	PU-AUC	$62.6_{\pm7.1}$	$81.7_{\pm1.2}$	$57.6_{\pm6.7}$	$77.2_{\pm3.1}$	$51.8_{\pm6.3}$	$70.1_{\pm2.2}$
	SAMULT^{P+U}	$56.9_{\pm7.1}$	$81.4_{\pm1.8}$	$58.9_{\pm7.7}$	$78.0_{\pm2.2}$	$51.9_{\pm6.6}$	$72.5_{\pm3.7}$
	WSAUC	$\mathbf{76.6_{\pm7.9}}$	$\mathbf{86.7_{\pm0.7}}$	$\mathbf{73.5_{\pm8.7}}$	$\mathbf{82.7_{\pm1.9}}$	$\mathbf{65.4_{\pm9.3}}$	$\mathbf{76.9_{\pm5.1}}$
CIFAR100	PU-AUC	$56.0_{\pm3.6}$	$68.9_{\pm0.9}$	$54.4_{\pm3.2}$	$65.8_{\pm2.1}$	$52.6_{\pm3.1}$	$60.9_{\pm2.8}$
	SAMULT^{P+U}	$54.2_{\pm3.8}$	$69.6_{\pm1.3}$	$52.7_{\pm3.5}$	$67.1_{\pm1.0}$	$51.1_{\pm3.2}$	$62.4_{\pm1.8}$
	WSAUC	$\mathbf{64.4_{\pm5.4}}$	$\mathbf{73.7_{\pm1.0}}$	$\mathbf{62.4_{\pm5.5}}$	$\mathbf{70.5_{\pm1.1}}$	$\mathbf{59.6_{\pm5.6}}$	$\mathbf{66.6_{\pm2.9}}$

The best performances are bolded.

method learns with both unlabeled and noisily labeled data simultaneously. Therefore, we compare the following baselines in the task of semi-supervised AUC optimization: SAUC-LS [27], a semi-supervised extension of AUC optimization with a novel loss function that utilizes the unlabeled data. OptAG [27], a semi-supervised AUC optimization method based on generative models. PNU-AUC [28], a semi-supervised

AUC optimization approach that combines unbiased PU and NU AUC optimization, where the unbiased PU and NU estimation is achieved by compensating the bias with the risk estimated on the labeled data. SAMULT [29], a semi-supervised AUC optimization method that utilizes the unbiased risk estimation. In addition to showing the performance in the normal semi-supervised setup with 0% noise ratio, we also evaluate the performance in

TABLE IV
PERFORMANCE (STD) OF MULTI-INSTANCE AUC OPTIMIZATION

Dataset	Musk1	Musk2	fox	tiger	elephant
miSVM	$78.8_{\pm 10.5}$	$75.7_{\pm 12.2}$	$52.8_{\pm 9.8}$	$78.6_{\pm 9.1}$	$76.7_{\pm 10.1}$
MISVM	$83.3_{\pm 12.0}$	$83.9_{\pm 14.6}$	$55.5_{\pm 10.8}$	$83.6_{\pm 10.3}$	$88.4_{\pm 6.0}$
MissSVM	$78.9_{\pm 10.1}$	$75.8_{\pm 12.5}$	$50.1_{\pm 9.4}$	$78.2_{\pm 8.9}$	$77.4_{\pm 9.2}$
SIL	$90.7_{\pm 9.1}$	$75.9_{\pm 13.4}$	$59.3_{\pm 11.2}$	$85.9_{\pm 10.6}$	$86.1_{\pm 6.4}$
sbMIL	$74.5_{\pm 18.9}$	$73.9_{\pm 16.0}$	$63.0_{\pm 8.4}$	$78.9_{\pm 7.9}$	$83.4_{\pm 7.5}$
MIDAM (smx)	$83.4_{\pm 12.0}$	$90.5_{\pm 6.8}$	$62.2_{\pm 18.8}$	$86.1_{\pm 7.1}$	$87.3_{\pm 10.4}$
MIDAM (att)	$82.6_{\pm 10.7}$	$84.3_{\pm 10.7}$	$73.3_{\pm 9.7}$	$86.7_{\pm 6.6}$	$90.6_{\pm 6.9}$
WSAUC	$\mathbf{96.0_{\pm 6.0}}$	$\mathbf{98.6_{\pm 4.5}}$	$\mathbf{90.4_{\pm 5.5}}$	$\mathbf{95.9_{\pm 3.7}}$	$\mathbf{96.6_{\pm 7.1}}$

The best performances are bolded.

TABLE V
PERFORMANCE (STD) OF SEMI-SUPERVISED AUC OPTIMIZATION WITH/WITHOUT NOISY LABELS

Dataset	Noise ratio	0%		20%		30%	
	Label ratio	5%	10%	5%	10%	5%	10%
MNIST	SAUC-LS	$93.9_{\pm 6.5}$	$97.5_{\pm 6.6}$	$78.9_{\pm 1.7}$	$87.3_{\pm 4.7}$	$73.6_{\pm 0.2}$	$73.7_{\pm 0.1}$
	OptAG	$97.1_{\pm 5.2}$	$98.4_{\pm 4.4}$	$90.3_{\pm 5.7}$	$96.4_{\pm 8.0}$	$76.7_{\pm 1.0}$	$80.6_{\pm 2.3}$
	PNU-AUC	$95.9_{\pm 2.3}$	$98.5_{\pm 0.1}$	$91.2_{\pm 2.7}$	$98.5_{\pm 0.2}$	$83.2_{\pm 4.2}$	$97.5_{\pm 0.5}$
	SAMULT	$\mathbf{98.4_{\pm 2.8}}$	$\mathbf{99.3_{\pm 0.1}}$	$93.2_{\pm 2.5}$	$\mathbf{98.9_{\pm 0.1}}$	$83.9_{\pm 3.9}$	$98.2_{\pm 0.2}$
	WSAUC	$98.3_{\pm 2.7}$	$\mathbf{99.3_{\pm 0.1}}$	$\mathbf{96.6_{\pm 2.4}}$	$98.8_{\pm 0.1}$	$\mathbf{95.2_{\pm 2.0}}$	$\mathbf{98.3_{\pm 0.1}}$
FMNIST	SAUC-LS	$98.3_{\pm 4.7}$	$98.7_{\pm 2.3}$	$54.0_{\pm 2.9}$	$60.0_{\pm 1.7}$	$47.0_{\pm 3.2}$	$49.6_{\pm 2.2}$
	OptAG	$\mathbf{98.7_{\pm 2.2}}$	$98.9_{\pm 1.0}$	$95.8_{\pm 6.6}$	$98.3_{\pm 4.3}$	$52.7_{\pm 3.2}$	$57.3_{\pm 2.0}$
	PNU-AUC	$97.8_{\pm 0.8}$	$98.9_{\pm 0.1}$	$93.5_{\pm 1.5}$	$98.9_{\pm 0.1}$	$89.8_{\pm 2.9}$	$98.7_{\pm 0.1}$
	SAMULT	$98.4_{\pm 0.9}$	$\mathbf{99.3_{\pm 0.0}}$	$96.1_{\pm 1.1}$	$\mathbf{99.0_{\pm 0.1}}$	$90.3_{\pm 2.6}$	$\mathbf{98.8_{\pm 0.1}}$
	WSAUC	$98.4_{\pm 0.9}$	$\mathbf{99.3_{\pm 0.1}}$	$\mathbf{97.7_{\pm 0.9}}$	$\mathbf{99.0_{\pm 0.1}}$	$\mathbf{97.4_{\pm 0.9}}$	$\mathbf{98.8_{\pm 0.1}}$
CIFAR10	SAUC-LS	$63.7_{\pm 0.7}$	$63.9_{\pm 0.6}$	$63.4_{\pm 0.7}$	$63.0_{\pm 0.5}$	$63.3_{\pm 0.6}$	$62.5_{\pm 0.3}$
	OptAG	$63.3_{\pm 0.8}$	$58.2_{\pm 2.3}$	$62.9_{\pm 0.8}$	$58.4_{\pm 2.0}$	$63.2_{\pm 0.7}$	$58.0_{\pm 1.7}$
	PNU-AUC	$\mathbf{78.5_{\pm 5.8}}$	$88.9_{\pm 0.1}$	$60.8_{\pm 7.8}$	$85.4_{\pm 1.1}$	$57.9_{\pm 8.1}$	$82.5_{\pm 0.9}$
	SAMULT	$70.3_{\pm 7.2}$	$89.2_{\pm 0.4}$	$65.8_{\pm 8.4}$	$\mathbf{86.9_{\pm 0.8}}$	$60.8_{\pm 8.7}$	$\mathbf{84.5_{\pm 0.6}}$
	WSAUC	$70.9_{\pm 7.1}$	$\mathbf{89.4_{\pm 0.4}}$	$\mathbf{66.5_{\pm 8.0}}$	$86.7_{\pm 0.7}$	$\mathbf{65.8_{\pm 0.7}}$	$83.9_{\pm 0.7}$
CIFAR100	SAUC-LS	$42.5_{\pm 1.2}$	$43.5_{\pm 0.5}$	$41.8_{\pm 1.1}$	$42.7_{\pm 0.5}$	$41.8_{\pm 1.0}$	$42.5_{\pm 0.4}$
	OptAG	$42.2_{\pm 1.0}$	$45.7_{\pm 1.5}$	$41.6_{\pm 0.8}$	$45.8_{\pm 2.0}$	$41.7_{\pm 1.0}$	$46.3_{\pm 2.0}$
	PNU-AUC	$61.0_{\pm 4.9}$	$76.8_{\pm 0.5}$	$56.6_{\pm 4.1}$	$73.8_{\pm 0.4}$	$55.4_{\pm 4.3}$	$69.6_{\pm 1.3}$
	SAMULT	$\mathbf{64.0_{\pm 4.5}}$	$\mathbf{78.1_{\pm 0.7}}$	$60.1_{\pm 4.7}$	$\mathbf{75.1_{\pm 0.3}}$	$57.4_{\pm 4.7}$	$71.2_{\pm 1.2}$
	WSAUC	$\mathbf{64.0_{\pm 4.5}}$	$77.9_{\pm 0.6}$	$\mathbf{61.4_{\pm 5.1}}$	$\mathbf{75.1_{\pm 0.5}}$	$\mathbf{60.3_{\pm 4.0}}$	$\mathbf{71.7_{\pm 1.0}}$

The best performances are bolded.

situations where the labeled data is affected by noise, with noise ratios of 20% or 30%. We test the baselines with label ratios of $\{5\%, 10\%\}$ to show the impact of the amount of labeled data.

The results are shown in Table V. We observe that when the data is noisy-free, WSAUC achieves similar performance to PNU-AUC and SAMULT on all four datasets. However, when the labeled data is both limited (5% label ratio) and noisy (20% or 30% noise ratio), WSAUC outperforms the baselines by a large margin. This indicates that WSAUC is particularly useful when dealing with scarce and contaminated labeled data. In other scenarios, WSAUC performs just as well as the prior state-of-the-art methods. Such results indicate that WSAUC is more robust to the label incomplete and inaccurate problems.

C. rpAUC Versus AUC as Training Objective

To support our claim that rpAUC is a robust training objective for weakly supervised AUC optimization, we empirically

demonstrate how much the rpAUC can outperform AUC as the training objective when the noise ratio varies. Optimizing normal AUC can be regarded as a degenerate version of optimizing rpAUC by ablating the removal step of the noisy instance pairs. We follow the problem setup in Section IV-A with varying θ_A and θ_B to show the performance increments of optimizing rpAUC over AUC.

We adopt the CIFAR100 dataset and vary the class proportions θ_A and $(1 - \theta_B)$ in the range of $[0.65, 1.0]$ with step size of 0.05. The smaller the values of θ_A and $(1 - \theta_B)$, the higher the noise ratio in the corresponding dataset. 5% of the data is used for training, to investigate the algorithm performance when data is relatively scarce. The experiments are repeated 10 times for each class proportion combination. The results are shown in Fig. 2.

It can be observed in the plot that when the dataset is clean, maximizing AUC or rpAUC yields similar result. As the noise ratio increases, maximizing rpAUC shows greater improvement over maximizing AUC as the optimization objective during
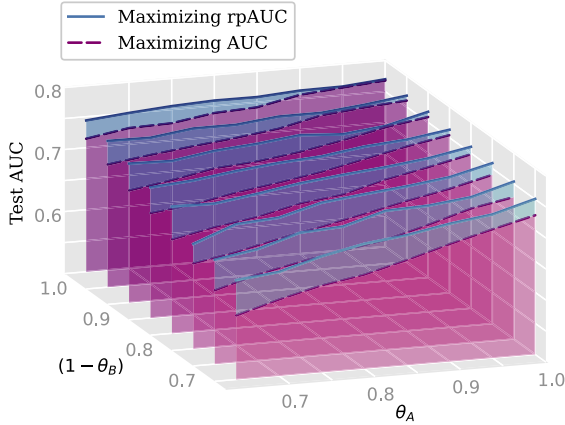
Fig. 2. Performance comparison of using AUC and rpAUC as training objective, under different noise ratios. Blue shading indicates the improvement in test AUC achieved by maximizing rpAUC during training.
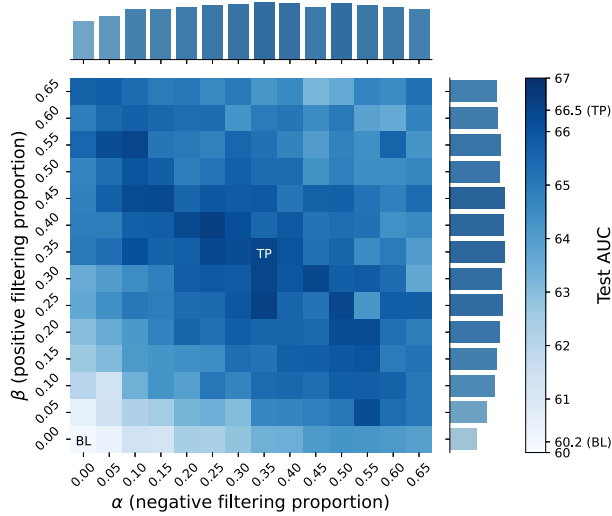

Fig. 3. Hyperparameter sensitivity. BL (baseline): rpAUC degenerates to full AUC when both α and β are set to zero. TP (true proportion): the filtering proportions are identical to the true proportions, i.e., $\alpha = 1 - \theta_A$ and $\beta = \theta_B$.

training. This result validates that the rpAUC is a more robust training objective for weakly supervised AUC optimization.

D. Hyperparameter Sensitivity

Using rpAUC requires two hyperparameters: the FPR threshold α and the TPR threshold β. These thresholds dictate the filtering proportions in the negative data (or $\mathcal{X}_B$) and the positive data (or $\mathcal{X}_A$), respectively. To investigate how they impact the model performance, we alter their values and show the resulting variations in the model performance.

We again adopt the CIFAR100 dataset for the experiment, and the proportions θ_A and $(1 - \theta_B)$ are both set to 0.65. The data size is shrink to 5% to enlarge the impact of the contamination. The hyperparameter α and β are chosen in the range of $[0.65, 1.0]$ with step size of 0.05. The results are shown in Fig. 3.

In Fig. 3, the abbreviation "BL" denotes the baseline scenario, where both α and β are set to zero. In this case, rpAUC degenerates to the full AUC. The results reveal that as α and β increase,

the model performance consistently surpasses that of the baseline case, suggesting that a relatively arbitrary assignment of hyperparameter values can lead to performance enhancement. The test AUC reaches its peak when the hyperparameters closely approximate the true proportions (annotated as "TP"), aligning with our understanding of maximizing rpAUC. When the true proportions are unknown, it is advisable to make larger guesses of α and β. This is illustrated in the bar plot, where the average performance gradually declines as the hyperparameters exceed the true proportions.

VIII. CONCLUSION AND FUTURE WORK

In this work, we study the weakly supervised AUC optimization problem. We present two main results:

1) We show that the AUC optimization problems in various weakly supervised scenarios can be uniformly framed as minimizing the AUC risk on contaminated sets, and the empirical risk minimization is consistent with the maximization of true AUC.
2) We propose the reversed partial AUC (rpAUC), which serves as a robust training objective for maximizing AUC under contaminated supervision.

Based on these results, we present the WSAUC framework, which provides a unified solution for multiple weakly supervised AUC optimization problems.

The unified WSAUC framework may be helpful for adapting specialize weakly supervised AUC optimization methods to new scenarios. It may also enable the development of new methods that can be applied to multiple scenarios, e.g., multiclass AUC optimization. These directions are left for future work.

REFERENCES

[1] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[2] S. J. Mason and N. E. Graham, "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation," *Quart. J. Roy. Meteorol. Soc.*, vol. 128, no. 584, pp. 2145–2166, 2002.
[3] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
[4] T. Yang and Y. Ying, "AUC maximization in the era of Big Data and AI: A survey," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–37, 2022.
[5] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 313–320.
[6] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, 2003.
[7] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 377–384.
[8] A. Herschtal and B. Raskutti, "Optimising area under the ROC curve using gradient descent," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 49–56.
[9] P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang, "Online AUC maximization," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 233–240.

[10] Y. Ying, L. Wen, and S. Lyu, "Stochastic online AUC maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 451–459.

[11] W. Gao and Z.-H. Zhou, "On the consistency of AUC pairwise optimization," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 939–945.

[12] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the ROC curve," *J. Mach. Learn. Res.*, vol. 6, pp. 393–425, 2005.

[13] M. Liu, Z. Yuan, Y. Ying, and T. Yang, "Stochastic AUC maximization with deep neural networks," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[14] Z. Yuan, Z. Guo, N. Chawla, and T. Yang, "Compositional training for end-to-end deep AUC maximization," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–14.

[15] H. Narasimhan and S. Agarwal, "SVM pAUC tight: A new support vector method for optimizing partial AUC based on a tight convex upper bound," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2013, pp. 167–175.

[16] H. Yang, K. Lu, X. Lyu, and F. Hu, "Two-way partial AUC and its properties," *Statist. Methods Med. Res.*, vol. 28, pp. 184–195, 2019.

[17] F. Cheng, G. Fu, X. Zhang, and J. Qiu, "Multi-objective evolutionary algorithm for optimizing the partial area under the ROC curve," *Knowl.-Based Syst.*, vol. 170, pp. 61–69, 2019.

[18] T. Iwata, A. Fujino, and N. Ueda, "Semi-supervised learning for maximizing the partial AUC," in *Proc. 34nd AAAI Conf. Artif. Intell.*, 2020, pp. 4239–4246.

[19] Z. Yang, Q. Xu, S. Bao, Y. He, X. Cao, and Q. Huang, "When all we need is a piece of the pie: A generic framework for optimizing two-way partial AUC," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 11820–11829.

[20] Z. Liu and T. Hyslop, "Partial AUC for differentiated gene detection," in *Proc. IEEE Int. Conf. Bioinf. Bioeng.*, 2010, pp. 310–311.

[21] Z. Bai, X.-L. Zhang, and J. Chen, "Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6819–6823.

[22] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2017.

[23] X. Zhu, A. B. Goldberg, R. Brachman, and T. Dietterich, *Introduction to Semi-Supervised Learning.* San Rafael, CA, USA: Morgan and Claypool Publishers, 2009.

[24] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Mach. Learn.*, vol. 109, no. 4, pp. 719–760, 2020.

[25] B. Han et al., "A survey of label-noise representation learning: Past, present and future," 2021, *arXiv: 2011.04406*.

[26] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, 2018.

[27] A. Fujino and N. Ueda, "A semi-supervised AUC optimization method with generative models," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 883–888.

[28] T. Sakai, G. Niu, and M. Sugiyama, "Semi-supervised AUC optimization based on positive-unlabeled learning," *Mach. Learn.*, vol. 107, pp. 767–794, 2018.

[29] Z. Xie and M. Li, "Semi-supervised AUC optimization without guessing labels of unlabeled data," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4310–4317.

[30] N. Charoenphakdee, J. Lee, and M. Sugiyama, "On symmetric losses for learning from corrupted labels," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 961–970.

[31] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 919–926.

[32] N. Lu, G. Niu, A. K. Menon, and M. Sugiyama, "On the minimal supervision for training any binary classifier from only unlabeled data," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–13.

[33] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1196–1204.

[34] A. K. Menon, B. Van Rooyen, C. S. Ong, and R. C. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 125–134.

[35] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1919–1925.

[36] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

[37] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "Mentor-Net: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.

[38] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.

[39] H. Yang et al., "Learning from noisy labels via dynamic loss thresholding," 2021, *arXiv:2104.02570*.

[40] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.

[41] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.

[42] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 322–330.

[43] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2691–2699.

[44] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–9.

[45] B. Han et al., "Masking: A new perspective of noisy supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5841–5851.

[46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–13.

[47] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11244–11253.

[48] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 6448–6458.

[49] X. Xia et al., "Are anchor points really indispensable in label-noise learning?," in *Proc. Ad. Neural Inf. Process. Syst.*, 2019, pp. 6838–6849.

[50] S. Wu et al., "Class2Simi: A noise reduction perspective on learning with noisy labels," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 11285–11295.

[51] C. Scott, "A rate of convergence for mixture proportion estimation, with application to learning from noisy labels," in *Proc. 18th Int. Conf. Artif. Intell. Stat.*, 2015, pp. 838–846.

[52] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.

[53] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, no. 2, pp. 103–134, 2000.

[54] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[55] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. 16th Int. Conf. Mach. Learn.*, 1999, pp. 200–209.

[56] O. Chapelle, M. Chi, and A. Zien, "A continuation method for semi-supervised SVMs," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 185–192.

[57] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou, "Cost-sensitive semi-supervised support vector machine," in *Proc. 24th AAAI Conf. Artif. Intell.*, 2010, pp. 500–505.

[58] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 19–26.

[59] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 912–919.

[60] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, 1998, pp. 92–100.

[61] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[62] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3546–3554.

[63] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[64] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 529–536.

[65] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[67] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–16.

[68] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 387–394.

[69] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *Proc. 18th Int. Joint Conf. Artif. Intell.*, 2003, pp. 587–592.

[70] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 448–455.

[71] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick, "Presence-only data and the EM algorithm," *Biometrics*, vol. 65, no. 2, pp. 554–563, 2008.

[72] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2008, pp. 213–220.

[73] M. C. du Plessis, G. Niu, and M. Sugiyama, "Analysis of learning from positive and unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 719–760.

[74] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1674–1684.

[75] C. Gong, H. Shi, T. Liu, C. Zhang, J. Yang, and D. Tao, "Loss decomposition and centroid estimation for positive and unlabeled learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 918–932, Mar. 2021.

[76] M. C. du Plessis, G. Niu, and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1386–1394.

[77] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1, pp. 31–71, 1997.

[78] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 577–584.

[79] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon, "Robust multiple-instance learning ensembles using random subspace instance selection," *Pattern Recognit.*, vol. 58, pp. 83–99, 2016.

[80] Y. Xiao, B. Liu, and Z. Hao, "A sphere-description-based approach for multiple-instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 242–257, Feb. 2017.

[81] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 179–186.

[82] X.-S. Wei, J. Wu, and Z.-H. Zhou, "Scalable algorithms for multi-instance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 975–987, Apr. 2017.

[83] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[84] X. Wang, Y. Yan, P. Tang, W. Liu, and X. Guo, "Bag similarity network for deep multi-instance learning," *Inf. Sci.*, vol. 504, pp. 578–588, 2019.

[85] Y.-F. Li, L.-Z. Guo, and Z.-H. Zhou, "Towards safe weakly supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 334–346, Jan. 2021.

[86] Z.-Y. Zhang, P. Zhao, Y. Jiang, and Z.-H. Zhou, "Learning from incomplete and inaccurate supervision," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5854–5868, Dec. 2022.

[87] C. Gong, J. Yang, J. J. You, and M. Sugiyama, "Centroid estimation with guaranteed efficiency: A general framework for weakly supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2841–2855, Jun. 2022.

[88] J. Li, R. Socher, and S. C. Hoi, "DivedeMix: Learning with noisy labels as semi-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–11.

[89] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 906–914.

[90] Y. Lei and Y. Ying, "Stochastic proximal AUC maximization," *J. Mach. Learn. Res.*, vol. 22, no. 61, pp. 2832–2876, 2021.

[91] Z. Xie and M. Li, "Cutting the software building efforts in continuous integration by semi-supervised online AUC optimization," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2875–2881.

[92] S. Azizi et al., "Big self-supervised models advance medical image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3458–3468.

[93] A. Feizi, "Hierarchical detection of abnormal behaviors in video surveillance through modeling normal behaviors based on AUC maximization," *Soft Comput.*, vol. 24, pp. 10401–10413, 2020.

[94] N. Usunier, M.-R. Amini, and P. Gallinari, "A data-dependent generalisation error bound for the AUC," in *Proc. Workshop ROC Anal Mach. Learn.*, 2005, pp. 1–9.

[95] D. Arpit et al., "A closer look at memorization in deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 233–242.

[96] D. Zhu, G. Li, B. Wang, X. Wu, and T. Yang, "When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 27548–27573.

[97] L. Feng et al., "Pointwise binary classification with pairwise confidence comparisons," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 3252–3262.

[98] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 22911–22924.

[99] C. Zhang, W. Shi, L. Luo, and B. Gu, "Doubly robust AUC optimization against noisy and adversarial samples," in *Proc. 29th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2023, pp. 3195–3205.

[100] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1167–1174.

[101] S. Ray and M. Craven, "Supervised versus multiple instance learning: An empirical comparison," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 697–704.

[102] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 105–112.

[103] D. Zhu et al., "Provable multi-instance deep AUC maximization with stochastic pooling," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 43205–43227.

Zheng Xie received the BEng degree in computer science from Xi'an Jiaotong University, in 2016. Currently he is working toward the PhD degree with the Department of Computer Science and Technology, Nanjing University. He is a member of the LAMDA Group. His research interests mainly include machine learning and data mining, especially in learning from weak supervision or imbalanced supervision. He has received a number of awards including Outstanding Graduate of Xi'an Jiaotong University, in 2016, AAAI Scholarship, in 2018, selected in the Program for Outstanding PhD Candidate of Nanjing University, in 2021, and so forth.

Yu Liu received the BSc degree in mathematics from Nanjing University, in 2021. Currently he is working toward the master's degree with the School of Artificial Intelligence, Nanjing University. He is a member of the LAMDA Group. His research interests mainly include machine learning and data mining.

Hao-Yuan He received the BSc degree in computer science from Nanjing Tech University, in 2021. Currently, he is working toward the master's degree with the School of Artificial Intelligence, Nanjing University. He is a member of the LAMDA Group. His research interests mainly include machine learning and data mining, especially in weakly supervised learning and neuro-symbolic learning.

Ming Li (Member, IEEE) is currently a professor with the LAMDA group, National Key Laboratory for Novel Software Technology, Nanjing University. His major research interests include machine learning and data mining, especially on software mining. He has served as the area chair of IJCAI, IEEE ICDM, etc, senior PC member of the premium conferences in artificial intelligence such as AAAI, and PC members for other premium conferences, such as KDD, NeurIPS, ICML, etc. He is the founding chair of the International Workshop on Software Mining. He has been granted various awards including the PAKDD Early Career Award, the NSFC Excellent Youth Award, the New Century Excellent Talents program of the Education Ministry of China, etc.

Zhi-Hua Zhou (Fellow, IEEE) received the BSc, MSc, and PhD degrees in computer science from Nanjing University, China, in 1996, 1998, and 2000, respectively. He joined the Department of Computer Science and Technology, Nanjing University as an assistant professor in 2001 and is currently a professor, the head of the Department of Computer Science and Technology, and the dean of the School of Artificial Intelligence. He is also the founding director of the LAMDA Group. He has authored the books Ensemble Methods: Foundations and Algorithms, Evolutionary Learning: Advances in Theories and Algorithms, Machine Learning (in Chinese) and authored or coauthored more than 200 papers in top-tier international journals or conference proceedings. He also holds 24 patents. His research interests include artificial intelligence, machine learning, and data mining. He was the recipient of various awards or honors including the National Natural Science Award of China, the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, the ACML Distinguished Contribution Award, the PAKDD Distinguished Contribution Award, IEEE ICDM Outstanding Service Award, and the Microsoft Professorship Award. He is the editor-in-chief of the *Frontiers of Computer Science*, the associate editor-in-chief of the *Science China Information Sciences*, action or associate editor of *Machine Learning*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *ACM Transactions on Knowledge Discovery from Data*. He was the associate editor-in-chief for *Chinese Science Bulletin*, an associate editor for *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Neural Networks and Learning Systems*, *ACM Transactions on Intelligent Systems and Technology*, and *Neural Networks*. He founded Asian Conference on Machine Learning, was an advisory committee member for IJCAI from 2015 to 2016, steering committee member for ICDM, ACML, PAKDD and PRICAI, and the chair of various conferences, including the program chair of IJCAI 2021, AAAI 2019, the general chair of ICDM 2016, the senior area chair of NeurIPS, the area chair of ICML, AAAI, IJCAI, and KDD. He was the chair of the IEEE CIS Data Mining Technical Committee from 2015 to 2016, the chair of the CCF-AI from 2012 to 2019, and the chair of the CAAI Machine Learning Technical Committee from 2006 to 2015. He is a foreign member of the Academy of Europe, and a fellow of the ACM, AAAI, AAAS, IAPR, IET/IEE, CCF, and CAAI.