On the Consistency of AUC Pairwise Optimization

Wei Gao and **Zhi-Hua Zhou***
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing 210023, China
{gaow, zhouzh}@lamda.nju.edu.cn

Abstract

AUC (Area Under ROC Curve) has been an important criterion widely used in diverse learning tasks. To optimize AUC, many learning approaches have been developed, most working with pairwise surrogate losses. Thus, it is important to study the AUC consistency based on minimizing pairwise surrogate losses. In this paper, we introduce the generalized calibration for AUC optimization, and prove that it is a necessary condition for AUC consistency. We then provide a sufficient condition for AUC consistency, and show its usefulness in studying the consistency of various surrogate losses, as well as the invention of new consistent losses. We further derive regret bounds for exponential and logistic losses, and present regret bounds for more general surrogate losses in the realizable setting. Finally, we prove regret bounds that disclose the equivalence between the pairwise exponential loss of AUC and univariate exponential loss of accuracy.

1 Introduction

AUC (Area Under ROC Curve) has been an important criterion widely used in diverse learning tasks [Freund *et al.*, 2003; Kotlowski *et al.*, 2011; Flach *et al.*, 2011; Zuva and Zuva, 2012]. Owing to its non-convexity and discontinuousness, direct optimization of AUC often leads to NP-hard problems. To make a compromise for avoiding computational difficulties, many pairwise surrogate losses, e.g., exponential loss [Freund *et al.*, 2003; Rudin and Schapire, 2009], hinge loss [Brefeld and Scheffer, 2005; Joachims, 2005; Zhao *et al.*, 2011] and least square loss [Gao *et al.*, 2013], have been widely adopted in practical algorithms.

It is important to study the consistency of these pairwise surrogate losses. In other words, whether the expected risk of learning with surrogate losses converge to the Bayes risk? Here, consistency (also known as Bayes consistency) guarantees the optimization of a surrogate loss will yield an optimal solution with Bayes risk in the limit of infinite sample.

This work presents a theoretical study on the consistency of AUC optimization based on minimizing pairwise surrogate losses. The main contributions include:

i) We introduce the generalized calibration, and prove that it is necessary yet insufficient for AUC consistency (cf. Theorem 1). This is because, for pairwise surrogate losses, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk on each pair of instances from different classes. For example, hinge loss and absolute loss are shown to be calibrated but inconsistent with AUC.

ii) We provide a sufficient condition for the AUC consistency based on minimizing pairwise surrogate losses (cf. Theorem 2). From this finding, we prove that exponential loss, logistic loss and distance-weighted loss are consistent with AUC. In addition, this result suggests the invention of some new consistent surrogate losses such as *q-norm hinge loss* and *general hinge loss*.

iii) We present regret bounds for exponential and logistic losses (cf. Theorem 3 and Corollary 5). For general surrogate losses, we present the regret bounds in the realizable setting (cf. Theorem 4).

iv) We provide regret bounds to disclose the equivalence (cf. Theorems 5 and 6) between the pairwise exponential surrogate loss of AUC and univariate exponential surrogate loss of accuracy. As a result, the univariate exponential loss is consistent AUC, and the pairwise exponential loss is consistent with accuracy by selecting a proper threshold. One direct consequence of this finding is the equivalence between AdaBoost and RankBoost in the limit of infinite sample.

Related Work

The studies on AUC can be traced back to 1970's in signal detection theory [Egan, 1975], and AUC has been an important performance measure for information retrieval and learning to rank, especially in bipartite ranking [Cohen *et al.*, 1999; Freund *et al.*, 2003; Rudin and Schapire, 2009].

Consistency has been an important issue. Zhang [2004b] and Bartlett *et al.* [2006] provided the fundamental analysis for binary classification, and many algorithms such as boosting and SVMs are proven to be consistent. The consistency studies on multi-class and multi-label learnings have been addressed in [Zhang, 2004a; Tewari and Bartlett, 2007] and [Gao and Zhou, 2013], respectively. Much attention has

*Supported by NSFC (61333014, 61321491).

been paid to the consistency of learning to rank [Cossock and Zhang, 2008; Xia *et al.*, 2008; Duchi *et al.*, 2010].

It is noteworthy that previous consistency studies focus on univariate surrogate losses over single instance [Zhang, 2004b; Bartlett *et al.*, 2006], whereas pairwise surrogate losses are defined on pairs of instances from different classes. This difference brings a challenge for studying AUC consistency: for univariate surrogate loss, it is sufficient to study the conditional risk; for pairwise surrogate losses, however, the whole distribution has to be considered (cf. Lemma 1). Because minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk.

Duchi *et al.* [2010] explored the consistency of supervised ranking, which is different from our setting: they considered "instances" consisting of a query, a set of inputs and a weighted graph, and the goal is to order the inputs according to the weighted graph; yet we consider instances with positive or negative labels, and aim to rank positive instances higher than negative ones. Clemencon *et al.* [2008] studied the consistency of ranking, and shown that calibration is a necessary and sufficient condition. We study the consistency of score functions by pairwise surrogate losses, and calibration is necessary but insufficient for AUC consistency (cf. Theorem 1).

Kotlowski *et al.* [2011] studied the AUC consistency based on minimizing univariate exponential and logistic losses, and this study is generalized to proper (composite) surrogate losses in [Agarwal, 2013]. These studies focused on univariate surrogate losses, whereas our work considers pairwise surrogate losses. Almost at the same time of our earlier version [Gao and Zhou, 2012], Uematsu and Lee [2012] provided a sufficient condition similar to Theorem 2 (as to be shown in Section 3.2), but with different proof skills. Later, our Theorem 2 was extended by Menon and Williamson [2014]. [Uematsu and Lee, 2012; Menon and Williamson, 2014] did not provide the other three contributions (i.e., i, iii and iv in the previous section) of our work.

The rest of the paper is organized as follows. Section 2 introduces preliminaries. Section 3 presents consistent conditions. Section 4 gives regret bounds. Section 5 discloses the equivalence of exponential loss. Section 6 concludes.

2 Preliminaries

Let $\mathcal{X}$ and $\mathcal{Y} = \{+1, -1\}$ be the input and output spaces, respectively. Suppose that $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}_\mathcal{X}$ corresponds to the instance-marginal distribution over $\mathcal{X}$. Let $\eta(\mathbf{x}) = \Pr[y = +1|\mathbf{x}]$ be the conditional probability over $\mathbf{x}$. For score function $f\colon \mathcal{X} \to \mathbb{R}$, the AUC w.r.t. distribution $\mathcal{D}$ is defined as

$$E[I[(y - y')f(\mathbf{x}) - f(x') > 0] + \tfrac{1}{2} I[f(\mathbf{x}) = f(\mathbf{x}')]|y \neq y']$$

where $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ are drawn i.i.d. from distribution $\mathcal{D}$, and $I[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise. Maximizing the AUC is equivalent to minimizing the expected risk, which can be viewed as a reward formulation as follows.

$$\begin{aligned} R(f) = E[&\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\ell(f, \mathbf{x}, \mathbf{x}') \\ &+ \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\ell(f, \mathbf{x}', \mathbf{x})] \end{aligned} \tag{1}$$

where the expectation takes on $\mathbf{x}$ and $\mathbf{x}'$ drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, and $\ell(f, \mathbf{x}, \mathbf{x}') = I[f(\mathbf{x}) > f(\mathbf{x}')] + \frac{1}{2} I[f(\mathbf{x}) = f(\mathbf{x}')]$ is called *ranking loss*. Write the Bayes risk $R^* = \inf_f [R(f)]$, and we get the set of Bayes optimal functions as

$$\begin{aligned} \mathcal{B} = \{f\colon R(f) = R^*\} = \{f\colon &(f(\mathbf{x}) - f(\mathbf{x}')) \\ &\times (\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0 \text{ if } \eta(\mathbf{x}) \neq \eta(\mathbf{x}')\}. \end{aligned} \tag{2}$$

Ranking loss ℓ is non-convex and discontinuous, and directly optimizing it often leads to NP-hard problems. In practice, we consider pairwise surrogate losses as follows:

$$\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}')),$$

where ϕ is a convex function, e.g., exponential loss $\phi(t) = e^{-t}$ [Freund *et al.*, 2003; Rudin and Schapire, 2009], hinge loss $\phi(t) = \max(0, 1 - t)$ [Brefeld and Scheffer, 2005; Joachims, 2005; Zhao *et al.*, 2011], etc.

We define the expected ϕ-risk as

$$\begin{aligned} R_\phi(f) = E_{\mathbf{x},\mathbf{x}' \sim \mathcal{D}_\mathcal{X}^2}[&\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\phi(f(\mathbf{x}) - f(\mathbf{x}')) \\ &+ \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\phi(f(\mathbf{x}') - f(\mathbf{x}))], \end{aligned} \tag{3}$$

and denote by $R_\phi^* = \inf_f R_\phi(f)$. Given two instances $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, we define the conditional ϕ-risk as

$$\begin{aligned} C(\mathbf{x}, \mathbf{x}', \alpha) = &\eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\phi(\alpha) \\ &+ \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\phi(-\alpha) \end{aligned} \tag{4}$$

where $\alpha = f(\mathbf{x}) - f(\mathbf{x}')$. For simplicity, denote by $\eta = \eta(\mathbf{x})$ and $\eta' = \eta(\mathbf{x}')$ when it is clear from the context.

3 AUC Consistency

We first define the *AUC consistency* as follows:

Definition 1 *The surrogate loss ϕ is said to be consistent with AUC if for every sequence $\{f^{\langle n \rangle}(\mathbf{x})\}_{n \geq 1}$, the following holds over all distributions $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$:*

$$R_\phi(f^{\langle n \rangle}) \to R_\phi^* \text{ then } R(f^{\langle n \rangle}) \to R^*.$$

In binary classification, recall the notion of *classification calibration*, which is a sufficient and necessary condition for consistency of $0/1$ error [Bartlett *et al.*, 2006]. A surrogate loss ϕ is said to be classification-calibrated if, for every $\mathbf{x} \in \mathcal{X}$ with $\eta(\mathbf{x}) \neq 1/2$,

$$\begin{aligned} &\inf_{f(\mathbf{x})(1 - 2\eta(\mathbf{x})) \geq 0} \{\eta(\mathbf{x})\phi(f(\mathbf{x})) + (1 - \eta(\mathbf{x}))\phi(-f(\mathbf{x}))\} \\ &> \inf_{f(\mathbf{x}) \in \mathbb{R}} \{\eta(\mathbf{x})\phi(f(\mathbf{x})) + (1 - \eta(\mathbf{x}))\phi(-f(\mathbf{x}))\}. \end{aligned}$$

We now generalize to *AUC calibration* as follows:

Definition 2 *The surrogate loss ϕ is said to be calibrated if*

$$H^-(\eta, \eta') > H(\eta, \eta') \text{ for any } \eta \neq \eta'$$

where $H^-(\eta, \eta') = \inf_{\alpha\colon \alpha(\eta - \eta') \leq 0} C(\mathbf{x}, \mathbf{x}', \alpha)$, $H(\eta, \eta') = \inf_{\alpha \in \mathbb{R}} C(\mathbf{x}, \mathbf{x}', \alpha)$ and $C(\mathbf{x}, \mathbf{x}', \alpha)$ is defined in Eqn. (4).

We first have

$$R_\phi^* = \inf_f R_\phi(f) \geq E_{\mathbf{x},\mathbf{x}' \sim \mathcal{D}_\mathcal{X}^2} \inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha). \tag{5}$$

It is noteworthy that the equality in the above does not hold for many surrogate losses from the following lemma:

Lemma 1 *For hinge loss $\phi(t) = \max(0, 1-t)$, least square hinge loss $\phi(t) = (\max(0, 1-t))^2$, least square loss $\phi(t) = (1-t)^2$ and absolute loss $\phi(t) = |1-t|$, we have*

$$\inf_f R_\phi(f) > E_{\mathbf{x},\mathbf{x}'\sim\mathcal{D}^2_{\mathcal{X}}} \inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha).$$

Proof We will present detailed proof for hinge loss by contradiction, and similar considerations could be made to other losses. Suppose that there exists a function f^* s.t.

$$R_\phi(f^*) = E_{\mathbf{x},\mathbf{x}'\sim\mathcal{D}^2_{\mathcal{X}}}[\inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha)].$$

For simplicity, we consider three instances $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathcal{X}$ s.t. $\eta(\mathbf{x}_1) < \eta(\mathbf{x}_2) < \eta(\mathbf{x}_3)$. The conditional risk of hinge loss is given by

$$\begin{aligned} C(\mathbf{x}, \mathbf{x}', \alpha) &= \eta(\mathbf{x})(1-\eta(\mathbf{x}'))\max(0, 1-\alpha) \\ &\quad +\eta(\mathbf{x}')(1-\eta(\mathbf{x}))\max(0, 1+\alpha). \end{aligned}$$

Minimizing $C(\mathbf{x}, \mathbf{x}', \alpha)$ gives $\alpha = -1$ if $\eta(\mathbf{x}) < \eta(\mathbf{x}')$. This yields $f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2) = -1$, $f^*(\mathbf{x}_1) - f^*(\mathbf{x}_3) = -1$ and $f^*(\mathbf{x}_2) - f^*(\mathbf{x}_3) = -1$; yet they are contrary each other. ■

From Lemma 1, the study on AUC consistency should focus on the expected ϕ-risk over the whole distribution rather than conditional ϕ-risk on each pair of instances. This is quite different from binary classification where minimizing the expected risk over the whole distribution is equivalent to minimizing the conditional risk on each instance, and thus binary classification focuses on the conditional risk as illustrated in [Zhang, 2004b; Bartlett *et al.*, 2006].

3.1 Calibration is Necessary yet Insufficient for AUC Consistency

We first prove that calibration is a necessary condition for AUC consistency as follows:

Lemma 2 *If the surrogate loss ϕ is consistent with AUC, then ϕ is calibrated, and for convex ϕ, it is differentiable at $t = 0$ with $\phi'(0) < 0$.*

Proof If ϕ is not calibrated, then there exist η_0 and η_0' s.t. $\eta_0 > \eta_0'$ and $H^-(\eta_0, \eta_0') = H(\eta_0, \eta_0')$. This implies the existence of some $\alpha_0 \le 0$ such that

$$\begin{aligned} &\eta_0(1-\eta_0')\phi(\alpha_0) + \eta_0'(1-\eta_0)\phi(-\alpha_0) \\ &\quad = \inf_{\alpha\in\mathbb{R}} \{\eta_0(1-\eta_0')\phi(\alpha) + \eta_0'(1-\eta_0)\phi(-\alpha)\}. \end{aligned}$$

We consider an instance space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ with marginal probability $\Pr[\mathbf{x}_1] = \Pr[\mathbf{x}_2] = 1/2$, $\eta(\mathbf{x}_1) = \eta_0$ and $\eta(\mathbf{x}_2) = \eta_0'$. We construct a sequence $\{f^{\langle n\rangle}\}_{n\neq 1}$ by selecting $f^{\langle n\rangle}(\mathbf{x}_1) = f^{\langle n\rangle}(\mathbf{x}_2) + \alpha_0$, and it is easy to get that

$$R_\phi(f^{\langle n\rangle}) \to R^*_\phi \text{ yet } R(f^{\langle n\rangle}) - R^* = (\eta_0 - \eta_0')/8 \text{ as } n \to \infty,$$

which shows that calibration is a necessary condition.

To prove $\phi'(0) < 0$, we consider the instance space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ with $\Pr[\mathbf{x}_1] = \Pr[\mathbf{x}_2] = 1/2$, $\eta(\mathbf{x}_1) = \eta_1$ and $\eta(\mathbf{x}_2) = \eta_2$. Assume that ϕ is differentiable at $t = 0$ with $\phi'(0) \ge 0$. For convex ϕ, we have $\eta_1(1-\eta_2)\phi(\alpha) + \eta_2(1-\eta_1)\phi(-\alpha) \ge (\eta_1 - \eta_2)\alpha\phi'(0) + (\eta_1(1-\eta_2) + \eta_2(1-\eta_1))\phi(0) \ge (\eta_1(1-\eta_2) + \eta_2(1-\eta_1))\phi(0)$ for $(\eta_1 - \eta_2)\alpha \ge 0$. This follows that

$$\begin{aligned} &\min\Big\{\{\eta_1(1-\eta_2)\phi(0) + \eta_2(1-\eta_1)\phi(0)\}, \\ &\quad \inf_{(\eta_1-\eta_2)\alpha\le 0}\{\eta_1(1-\eta_2)\phi(\alpha) + \eta_2(1-\eta_1)\phi(-\alpha)\}\Big\} \\ &= \inf_{(\eta_1-\eta_2)\alpha\le 0}\{\eta_1(1-\eta_2)\phi(\alpha) + \eta_2(1-\eta_1)\phi(-\alpha)\} \end{aligned}$$

which is contrary to $H^-(\eta_1, \eta_2) > H(\eta_1, \eta_2)$.

Suppose that ϕ is not differentiable at $t = 0$. There exists two subgradients $g_1 > g_2$ such that

$$\phi(t) \ge g_1 t + \phi(0) \text{ and } \phi(t) \ge g_2 t + \phi(0) \text{ for } t \in \mathbb{R}.$$

If $g_1 > g_2 \ge 0$, we select $\eta_1 = g_1/(g_1+g_2)$ and $\eta_2 = g_2/(g_1+g_2)$. It is obvious that $\eta_1 > \eta_2$, and for any $\alpha \ge 0$, we have $\eta_1(1-\eta_2)\phi(\alpha) + \eta_2(1-\eta_1)\phi(-\alpha) \ge \eta_1(1-\eta_2)(g_2\alpha + \phi(0)) + \eta_2(1-\eta_1)(-g_1\alpha + \phi(0)) \ge (\eta_1(1-\eta_2) + \eta_2(1-\eta_1))\phi(0)$.

In a similar manner, we can prove $\eta_1(1-\eta_2)\phi(\alpha) + \eta_2(1-\eta_1)\phi(-\alpha) \ge (\eta_1(1-\eta_2) + \eta_2(1-\eta_1))\phi(0)$ for $g_1 \ge 0 > g_2$, $g_1 > 0 \ge g_2$ and $0 \ge g_1 > g_2$ if $(\eta_1 - \eta_2)\alpha \ge 0$. This follows that $H(\eta_1, \eta_2) = H^-(\eta_1, \eta_2)$, which is contrary to the consistency of ϕ. ■

For the converse direction, we observe that hinge loss and absolute loss are convex with $\phi'(0) < 0$, and thus they are calibrated, yet inconsistent with AUC as follows:

Lemma 3 *For hinge loss $\phi(t) = \max(0, 1-t)$ and absolute loss $\phi(t) = |1-t|$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is inconsistent with AUC.*

Proof We present detailed proof for hinge loss and similar proof could be made to absolute loss. We consider the instance space $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, and assume that, for $1 \le i \le 3$, the marginal probability $\Pr[\mathbf{x}_i] = 1/3$ and conditional probability $\eta_i = \eta(\mathbf{x}_i)$ s.t. $\eta_1 < \eta_2 < \eta_3$, $2\eta_2 < \eta_1 + \eta_3$ and $2\eta_1 > \eta_2 + \eta_1\eta_3$. We write $f_i = f(\mathbf{x}_i)$ for $1 \le i \le 3$. Eqn. (3) gives

$$R_\phi(f) = \kappa_0 + \kappa_1 \sum_{i=1}^{3}\sum_{j\neq i} \eta_i(1-\eta_j)\max(0, 1+f_j-f_i)$$

where $\kappa_0 > 0$ and $\kappa_1 > 0$ are constants and independent to f. Minimizing $R_\phi(f)$ yields the optimal expected ϕ-risk

$$R^*_\phi = \kappa_0 + \kappa_1(3\eta_1 + 3\eta_2 - 2\eta_1\eta_2 - 2\eta_1\eta_3 - 2\eta_2\eta_3)$$

when $f^* = (f_1^*, f_2^*, f_3^*)$ s.t. $f_1^* = f_2^* = f_3^* - 1$. Note that $f' = (f_1', f_2', f_3')$ s.t. $f_1' + 1 = f_2' = f_3' - 1$ is not the optimal solution w.r.t. hinge loss since

$$\begin{aligned} R_\phi(f') &= \kappa_0 + \kappa_1(5\eta_1 + 2\eta_2 - 2\eta_1\eta_2 - 3\eta_1\eta_3 - 2\eta_2\eta_3) \\ &= R^*_\phi + \kappa_1(2\eta_1 - \eta_2 - \eta_1\eta_3) = R^*_\phi + \kappa_1(\eta_2 - \eta_1)/2. \end{aligned}$$

This completes the proof. ■

Together with Lemma 2 and Lemma 3, we have

Theorem 1 *Calibration is necessary yet insufficient for AUC consistency.*

The study on AUC consistency is not parallel to that of binary classification where the classification calibration is necessary and sufficient for the consistency of $0/1$ error in [Bartlett *et al.*, 2006]. The main difference is that, for AUC consistency, minimizing the expected risk over the whole distribution is not equivalent to minimizing the conditional risk on each pair of instances as shown by Lemma 1.

3.2 Sufficient Condition for AUC Consistency

We now present a sufficient condition for AUC consistency.

Theorem 2 *The surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC if $\phi\colon \mathbb{R} \to \mathbb{R}$ is a convex, differentiable and non-increasing function s.t. $\phi'(0) < 0$.*

Proof It suffices to prove $\inf_{f \notin \mathcal{B}} R_\phi(f) > \inf_f R_\phi(f)$ for convex, differentiable and non-increasing function ϕ s.t. $\phi'(0) < 0$. Assume that $\inf_{f \notin \mathcal{B}} R_\phi(f) = \inf_f R_\phi(f)$, i.e., there is an optimal function f^* s.t. $R_\phi(f^*) = \inf_f R_\phi(f)$ and $f^* \notin \mathcal{B}$, i.e., for some $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have $f^*(\mathbf{x}_1) \leq f^*(\mathbf{x}_2)$ yet $\eta(\mathbf{x}_1) > \eta(\mathbf{x}_2)$. Recall the ϕ-risk's definition in Eqn. (3)

$$R_\phi(f) = \int_{\mathcal{X}} \int_{\mathcal{X}} \eta(\mathbf{x})(1 - \eta(\mathbf{x}'))\phi(f(\mathbf{x}) - f(\mathbf{x}')) + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))\phi(f(\mathbf{x}') - f(\mathbf{x}))d\Pr(\mathbf{x})d\Pr(\mathbf{x}').$$

We introduce function h_1 s.t. $h_1(\mathbf{x}) = 0$ if $\mathbf{x} \neq \mathbf{x}_1$ and $h_1(\mathbf{x}_1) = 1$ otherwise, and write $g(\gamma) = R_\phi(f^* + \gamma h_1)$ for any $\gamma \in \mathbb{R}$, and thus g is convex. For optimal function f^*, we have $g'(0) = 0$ which implies that

$$\int_{\mathcal{X} \setminus \mathbf{x}_1} \eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}))\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x})) - \eta(\mathbf{x})(1 - \eta(\mathbf{x}_1))\phi'(f^*(\mathbf{x}) - f^*(\mathbf{x}_1))d\Pr(\mathbf{x}) = 0. \quad (6)$$

Similarly, we have

$$\int_{\mathcal{X} \setminus \mathbf{x}_2} \eta(\mathbf{x}_2)(1 - \eta(\mathbf{x}))\phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x})) - \eta(\mathbf{x})(1 - \eta(\mathbf{x}_2))\phi'(f^*(\mathbf{x}) - f^*(\mathbf{x}_2))d\Pr(\mathbf{x}) = 0. \quad (7)$$

For convex differentiable and non-increasing function ϕ, we have $\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x})) \leq \phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x})) \leq 0$ if $f^*(\mathbf{x}_1) \leq f^*(\mathbf{x}_2)$. This follows

$$\eta(\mathbf{x}_1)\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x})) \leq \eta(\mathbf{x}_2)\phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x})) \quad (8)$$

for $\eta(\mathbf{x}_1) > \eta(\mathbf{x}_2)$. In a similar manner, we have

$$(1 - \eta(\mathbf{x}_2))\phi'(f^*(\mathbf{x}) - f^*(\mathbf{x}_2)) \leq (1 - \eta(\mathbf{x}_1))\phi'(f^*(\mathbf{x}) - f^*(\mathbf{x}_1)). \quad (9)$$

If $f^*(\mathbf{x}_1) = f^*(\mathbf{x}_2)$, then we have

$$\eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}_2))\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)) - \eta(\mathbf{x}_2)(1 - \eta(\mathbf{x}_1))\phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x}_1)) < 0$$

from $\phi'(0) < 0$ and $\eta(\mathbf{x}_1) > \eta(\mathbf{x}_2)$, which is contrary to Eqns. (6) and (7) by combining Eqns. (8) and (9).

If $f^*(\mathbf{x}_1) < f^*(\mathbf{x}_2)$, then $\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)) \leq \phi'(0) < 0$, $\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)) \leq \phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x}_1)) \leq 0$, and

$$\eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}_2))\phi'(f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)) \leq \eta(\mathbf{x}_2)(1 - \eta(\mathbf{x}_1))\phi'(f^*(\mathbf{x}_2) - f^*(\mathbf{x}_1))$$

which is also contrary to Eqns. (6) and (7) by combining Eqns. (8) and (9).This theorem follows as desired. ■

From Theorem 2, we have

Corollary 1 *For exponential loss $\phi(t) = e^{-t}$ and logistic loss $\phi(t) = \ln(1 + e^{-t})$, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Marron *et al*. (2007) introduced the *distance-weighted loss* method for high-dimensional yet small-size sample, which was reformulated by Bartlett *et al*. (2006), for any $\epsilon > 0$, as

$$\phi(t) = \tfrac{1}{t} \text{ for } t \geq \epsilon; \text{ and } \phi(t) = \tfrac{1}{\epsilon}\left(2 - \tfrac{t}{\epsilon}\right) \text{ otherwise.}$$

Corollary 2 *For distance-weighted loss, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Lemma 3 proves the inconsistency of hinge loss, and also shows the difficulty for consistency without differentiability. We now derive some variants of hinge loss that are consistent. For example, the *q-norm hinge loss*: $\phi(t) = (\max(0, 1 - t))^q$ for $q > 1$ is consistent as follows:

Corollary 3 *For q-norm hinge loss, the surrogate loss $\phi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

From this corollary, it is immediate to get the consistency of *least-square hinge loss* $\phi(t) = (\max(0, 1 - t))^2$.

For $\epsilon > 0$, define the *general hinge loss* as $\phi(t) = 1 - t$ for $t \leq 1 - \epsilon$; $\phi(t) = 0$ for $t \geq 1 + \epsilon$; and $\phi(t) = (t - 1 - \epsilon)^2/4\epsilon$ otherwise.

Corollary 4 *For general hinge loss, the surrogate loss $\Psi(f, \mathbf{x}, \mathbf{x}') = \phi(f(\mathbf{x}) - f(\mathbf{x}'))$ is consistent with AUC.*

Hinge loss is inconsistent with AUC, but we can use the general hinge loss to approach hinge loss when $\epsilon \to 0$. In addition, it is also interesting to derive other consistent surrogate losses under the guidance of Theorem 2.

4 Regret Bounds

We will present the regret bounds for exponential and logistic losses, and for general losses under the realizable setting.

4.1 Regret Bounds for Exponential and Logistic Losses

We begin with a special property as follows:

Proposition 1 *For exponential loss and logistic loss, we have*

$$\inf_f R_\phi(f) = E_{\mathbf{x}, \mathbf{x}' \sim \mathcal{D}^2_{\mathcal{X}}} \inf_\alpha C(\eta(\mathbf{x}), \eta(\mathbf{x}'), \alpha).$$

Proof We provide the detailed proof for exponential loss. For a fixed instance $\mathbf{x}_0 \in \mathcal{X}$ and $f(\mathbf{x}_0)$, we set

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{2} \ln \frac{\eta(\mathbf{x})(1 - \eta(\mathbf{x}_0))}{\eta(\mathbf{x}_0)(1 - \eta(\mathbf{x}))} \quad \text{for } \mathbf{x} \neq \mathbf{x}_0.$$

This holds that, for $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) = \frac{1}{2} \ln \frac{\eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}_2))}{\eta(\mathbf{x}_2)(1 - \eta(\mathbf{x}_1))},$$

which minimizes $C(\eta(\mathbf{x}_1), \eta(\mathbf{x}_2), \alpha)$ by $\alpha = f(\mathbf{x}_1) - f(\mathbf{x}_2)$. We complete the proof as desired. ■

Proposition 1 is specific to the exponential and logistic loss, and does not hold for hinge loss, absolute loss, etc. Based on this proposition, we study the regret bounds for exponential and logistic loss by focusing on conditional risk as follows:

Theorem 3 *For constants $\kappa_0 > 0$ and $0 < \kappa_1 \leq 1$, we have*

$$R(f) - R^* \leq \kappa_0 (R_\phi(f) - R_\phi^*)^{\kappa_1},$$

if $f^* \in \arg\inf_f R_\phi(f)$ satisfies that, for $\eta(\mathbf{x}) \neq \eta(\mathbf{x}')$, $(f^*(\mathbf{x}) - f^*(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0$ and $|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq \kappa_0 \big(C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}'))\big)^{\kappa_1}$.

Proof This proof is partly motivated from [Zhang, 2004b]. From Eqns. (1) and (2), we have

$$\begin{aligned} R(f) - R^* &= E_{(\eta(\mathbf{x})-\eta(\mathbf{x}'))(f(\mathbf{x})-f(\mathbf{x}'))<0}[|\eta(\mathbf{x}) - \eta(\mathbf{x}')|] \\ &\quad + E_{f(\mathbf{x})=f(\mathbf{x}')}[|\eta(\mathbf{x}') - \eta(\mathbf{x})|]/2 \\ &\leq E_{(\eta(\mathbf{x})-\eta(\mathbf{x}'))(f(\mathbf{x})-f(\mathbf{x}'))\leq 0}[|\eta(\mathbf{x}) - \eta(\mathbf{x}')|] \end{aligned}$$

which yields that, by our assumption and Jensen's inequality,

$$\begin{aligned} R(f) - R^* \leq \kappa_0 \big(&E_{(\eta(\mathbf{x})-\eta(\mathbf{x}'))(f(\mathbf{x})-f(\mathbf{x}'))\leq 0}[C(\eta(\mathbf{x}), \\ &\eta(\mathbf{x}'), 0) - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}'))]\big)^{\kappa_1} \end{aligned}$$

for $0 < \kappa_1 < 1$. This remains to prove that

$$\begin{aligned} &E[C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}'))] \\ &\leq E[C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}')) \\ &\qquad - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}'))] \end{aligned}$$

where the expectations take over $(\eta(\mathbf{x}) - \eta(\mathbf{x}'))(f(\mathbf{x}) - f(\mathbf{x}')) \leq 0$. To see it, we consider the following cases:
1) For $\eta(\mathbf{x}) = \eta(\mathbf{x}')$ and convex ϕ, we have

$$C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) \leq C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}'));$$

2) For $f(\mathbf{x}) = f(\mathbf{x}')$, we have

$$C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) = C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}'));$$

3) For $(\eta(\mathbf{x}) - \eta(\mathbf{x}'))(f(\mathbf{x}) - f(\mathbf{x}')) < 0$, we derive that 0 is between $f(\mathbf{x}) - f(\mathbf{x}')$ and $f^*(\mathbf{x}) - f^*(\mathbf{x}')$ from assumption $(f^*(\mathbf{x}) - f^*(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0$. For convex ϕ, we have $C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) \leq \max(C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}'))$ and $C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}'))) = C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}'))$. The theorem follows as desired. ■

Based on this theorem, we have

Corollary 5 *The regret bounds for exponential and logistic loss are given, respectively, by*

$$\begin{aligned} R(f) - R^* &\leq \sqrt{R_\phi(f) - R_\phi^*}\,, \\ R(f) - R^* &\leq 2\sqrt{R_\phi(f) - R_\phi^*}\,. \end{aligned}$$

Proof We will present detailed proof for exponential loss and similarly consider logistic loss. The optimal function f^* satisfies $f^*(\mathbf{x}) - f^*(\mathbf{x}') = \frac{1}{2}\ln\frac{\eta(\mathbf{x})(1-\eta(\mathbf{x}'))}{\eta(\mathbf{x}')(1-\eta(\mathbf{x}))}$ by minimizing $C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f(\mathbf{x}) - f(\mathbf{x}'))$. This follows $(f^*(\mathbf{x}) - f^*(\mathbf{x}'))(\eta(\mathbf{x}) - \eta(\mathbf{x}')) > 0$ for $\eta(\mathbf{x}) \neq \eta(\mathbf{x}')$, $C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) = \eta(\mathbf{x})(1 - \eta(\mathbf{x}')) + \eta(\mathbf{x}')(1 - \eta(\mathbf{x}))$, and $C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}')) = \sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}))}$ $\times\sqrt{\eta(\mathbf{x}')(1-\eta(\mathbf{x}'))}$. Therefore, we have

$$\begin{aligned} &C(\eta(\mathbf{x}), \eta(\mathbf{x}'), 0) - C(\eta(\mathbf{x}), \eta(\mathbf{x}'), f^*(\mathbf{x}) - f^*(\mathbf{x}')) \\ &= \big(\sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}'))} - \sqrt{\eta(\mathbf{x}')(1-\eta(\mathbf{x}))}\big)^2 \\ &= \frac{|\eta(\mathbf{x}) - \eta(\mathbf{x}')|^2}{(\sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}'))} + \sqrt{\eta(\mathbf{x}')(1-\eta(\mathbf{x}))})^2} \\ &\geq |\eta(\mathbf{x}) - \eta(\mathbf{x}')|^2, \end{aligned}$$

where we use the fact $\eta(x), \eta(x') \in [0, 1]$. We complete the proof by applying Theorem 3. ■

4.2 Regret Bounds for Realizable Setting

We now define the realizable setting as:

Definition 3 *A distribution $\mathcal{D}$ is said to be realizable if $\eta(\mathbf{x})(1 - \eta(\mathbf{x})) = 0$ for each $\mathbf{x} \in \mathcal{X}$.*

This setting has been studied for bipartite ranking [Rudin and Schapire, 2009] and multi-class classification [Long and Servedio, 2013]. Under this setting, we have

Theorem 4 *For some $\kappa > 0$, if $R_\phi^* = 0$, then we have*

$$R(f) - R^* \leq \kappa(R_\phi(f) - R_\phi^*)$$

when $\phi(t) \geq 1/\kappa$ for $t \leq 0$ and $\phi(t) \geq 0$ for $t > 0$.

Proof Let $\mathcal{D}_+$ and $\mathcal{D}_-$ denote the positive and negative instance distributions, respectively. Eqn. (1) gives that $R(f)$ equals to

$$E_{\mathbf{x}\sim\mathcal{D}_+, \mathbf{x}'\sim\mathcal{D}_-}[I[f(\mathbf{x}) < f(\mathbf{x}')] + \tfrac{1}{2} I[f(\mathbf{x}) = f(\mathbf{x}')]]$$

and $R^* = \inf_f[R(f)] = 0$ for $f(\mathbf{x}) > f(\mathbf{x}')$. From Eqn. (3), we get the ϕ-risk $R_\phi(f) = E_{\mathbf{x}\sim\mathcal{D}_+, \mathbf{x}'\sim\mathcal{D}_-}[\phi(f(\mathbf{x}) - f(\mathbf{x}'))]$. Then, $R(f) - R^* = E_{\mathbf{x}\sim\mathcal{D}_+, \mathbf{x}'\sim\mathcal{D}_-}[I[f(\mathbf{x}) < f(\mathbf{x}')] + I[f(\mathbf{x}) = f(\mathbf{x}')]/2] \leq E_{\mathbf{x}\sim\mathcal{D}_+, \mathbf{x}'\sim\mathcal{D}_-}[\kappa\phi(f(\mathbf{x}) - f(\mathbf{x}'))] = \kappa(R_\phi(f) - R_\phi^*)$, which completes the proof. ■

Based on this theorem, we have

Corollary 6 *For exponential loss, hinge loss, general hinge loss, q-norm hinge loss, and least square loss, we have*

$$R(f) - R^* \leq R_\phi(f) - R_\phi^*,$$

and for logistic loss, we have

$$R(f) - R^* \leq \tfrac{1}{\ln 2}(R_\phi(f) - R_\phi^*).$$

Hinge loss is consistent with AUC under the realizable setting yet inconsistent for the general case as shown in Lemma 3. Corollary 5 shows regret bounds for exponential and logistic loss in the general case, whereas the above corollary provides tighter regret bounds under the realizable setting.

5 Equivalence Between AUC and Accuracy Optimization with Exponential Losses

In binary classification, we try to learn a score function $f \in \mathcal{X} \to \mathbb{R}$, and make predictions based on $\mathrm{sgn}[f(\mathbf{x})]$. The goal is to improve the accuracy by minimizing

$$\begin{aligned} R_{\mathrm{acc}}(f) &= E_{(\mathbf{x},y)\sim\mathcal{D}}\left[I\left[yf(\mathbf{x})<0\right]\right] \\ &= E_{\mathbf{x}}\left[\eta\left(\mathbf{x}\right)I\left[f\left(\mathbf{x}\right)<0\right]+\left(1-\eta\left(\mathbf{x}\right)\right)I\left[f\left(\mathbf{x}\right)>0\right]\right]. \end{aligned}$$

Denote by $R^*_{\mathrm{acc}} = \inf_f R_{\mathrm{acc}}(f)$, and we get the set of optimal solutions for accuracy as follows:

$$\mathcal{B}_{\mathrm{acc}} = \{f\colon f(\mathbf{x})(\eta(\mathbf{x})-1/2)>0 \text{ for } \eta(\mathbf{x})\neq 1/2\}.$$

The popular formulation, called surrogate losses, is given by

$$\phi_{\mathrm{acc}}(f(\mathbf{x}),y) = \phi(yf(\mathbf{x})),$$

where ϕ is a convex function such as exponential loss [Freund and Schapire, 1997], logistic loss [Friedman *et al.*, 2000], etc. We define the expected ϕ_{acc}-risk as

$$\begin{aligned} R_{\phi_{\mathrm{acc}}}(f) &= E_{(\mathbf{x},y)\sim\mathcal{D}}[\phi(yf(\mathbf{x}))] = E_{\mathbf{x}}[C_{\mathrm{acc}}(\eta(\mathbf{x}),f(\mathbf{x}))] \\ &= E_{\mathbf{x}}[(1-\eta(\mathbf{x}))\phi(-f(\mathbf{x}))+\eta(\mathbf{x})\phi(f(\mathbf{x}))], \end{aligned}$$

and denote by $R^*_{\phi_{\mathrm{acc}}} = \inf_f R_{\phi_{\mathrm{acc}}}(f)$.

Theorem 5 *For exponential loss and classifier f, we have*

$$R_\phi(f) - R^*_\phi \le R_{\phi_{acc}}(f)(R_{\phi_{acc}}(f) - R^*_{\phi_{acc}}).$$

Proof For accuracy's exponential surrogate loss, we have

$$R_{\phi_{\mathrm{acc}}}(f) - R^*_{\phi_{\mathrm{acc}}} = E_{\mathbf{x}}\Big(\sqrt{\eta(\mathbf{x})e^{-f(\mathbf{x})}} - \sqrt{(1-\eta(\mathbf{x}))e^{f(\mathbf{x})}}\Big)^2$$

and for AUC's exponential surrogate loss, we have

$$\begin{aligned} R_\phi(f) - R^*_\phi = E_{\mathbf{x},\mathbf{x}'}\Big(&\sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}'))e^{-f(\mathbf{x})+f(\mathbf{x}')}} \\ &- \sqrt{\eta(\mathbf{x}')(1-\eta(\mathbf{x}))e^{f(\mathbf{x})-f(\mathbf{x}')}}\Big)^2. \end{aligned}$$

By using $(ab-cd)^2 \le a^2(b-d)^2 + d^2(a-c)^2$, we have

$$R_\phi(f) - R^*_\phi \le 4E_{\mathbf{x}}[(1-\eta(\mathbf{x}))e^{f(\mathbf{x})}](R_{\phi_{\mathrm{acc}}}(f) - R^*_{\phi_{\mathrm{acc}}})$$

and in a similar manner, we have

$$R_\phi(f) - R^*_\phi \le 4E_{\mathbf{x}}[\eta(\mathbf{x})e^{-f(\mathbf{x})}](R_{\phi_{\mathrm{acc}}}(f) - R^*_{\phi_{\mathrm{acc}}}).$$

This follows $R_\phi(f) - R^*_\phi \le R_{\phi_{\mathrm{acc}}}(f)(R_{\phi_{\mathrm{acc}}}(f) - R^*_{\phi_{\mathrm{acc}}})$. ■

For ranking function f, we select a proper threshold to construct classifier by

$$\begin{aligned} t^*_f &= \operatorname*{arg\,min}_{t\in(-\infty,+\infty)} E_{\mathbf{x}}\left[\eta(\mathbf{x})e^{-f(\mathbf{x})+t} + (1-\eta(\mathbf{x}))e^{f(\mathbf{x})-t}\right] \\ &= \tfrac{1}{2}\ln(E_{\mathbf{x}}[\eta(\mathbf{x})e^{-f(\mathbf{x})}]/\ln E_{\mathbf{x}}[(1-\eta(\mathbf{x}))e^{f(\mathbf{x})}]). \end{aligned}$$

Based on such threshold, we have

Theorem 6 *For ranking function f and exponential loss,*

$$R_{\phi_{acc}}(f - t^*_f) - R^*_{\phi_{acc}} \le 2\sqrt{R_\phi(f) - R^*_\phi}$$

by selecting the threshold t^*_f defined above.

Proof For score function $f(\mathbf{x})$, we have

$$\begin{aligned} R_{\phi_{\mathrm{acc}}}(f - t^*_f) - R^*_{\phi_{\mathrm{acc}}} = &-2E_{\mathbf{x}}\sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}))} \\ &+2\sqrt{E_{\mathbf{x}}[\eta(\mathbf{x})e^{-f(\mathbf{x})}]E_{\mathbf{x}}[(1-\eta(\mathbf{x}))e^{f(\mathbf{x})}]}. \end{aligned}$$

For pairwise exponential loss of AUC, we have

$$\begin{aligned} &R_\phi(f) - R^*_\phi = 2E_{\mathbf{x}}[\eta(\mathbf{x})e^{-f(\mathbf{x})}]E_{\mathbf{x}}[1-\eta(\mathbf{x})e^{f(\mathbf{x})}] - \\ &2(E_{\mathbf{x}}[\sqrt{\eta(\mathbf{x})(1-\eta(\mathbf{x}))}])^2 \ge (R_{\phi_{\mathrm{acc}}}(f - t^*_f) - R^*_{\phi_{\mathrm{acc}}})^2/2 \end{aligned}$$

which completes the proof. ■

Together with Corollary 5, Theorems 5 and 6, and [Zhang, 2004b, Theorem 2.1], we have

Theorem 7 *For exponential loss and classifier f, we have*

$$\begin{aligned} R(f) - R^* &\le \Big(R_{\phi_{acc}}(f)(R_{\phi_{acc}}(f) - R^*_{\phi_{acc}})\Big)^{1/2} \\ R_{acc}(f) - R^*_{acc} &\le \sqrt{2}(R_{\phi_{acc}}(f) - R^*_{\phi_{acc}})^{1/2}. \end{aligned}$$

For exponential loss and ranking function f, we have

$$\begin{aligned} R(f) - R^* &\le (R_\phi(f) - R^*_\phi)^{1/2} \\ R_{acc}(f - t^*_f) - R^*_{acc} &\le 2(R_\phi(f) - R^*_\phi))^{1/4}. \end{aligned}$$

This theorem discloses the asymptotic equivalence between univariate exponential loss of accuracy and the pairwise exponential loss of AUC. As a result, AdaBoost and RankBoost are equivalent, i.e., both of them optimize AUC and accuracy simultaneously, because AdaBoost and RankBoost essentially optimize $\phi_{\mathrm{acc}}(f(\boldsymbol{x}),y) = e^{-yf(\boldsymbol{x})}$ and $\phi(f,\boldsymbol{x},\boldsymbol{x}') = e^{-(f(\boldsymbol{x})-f(\boldsymbol{x}'))}$, respectively.

Rudin and Schapire [2009] established the equivalence between AdaBoost and RankBoost for finite training sample based on the assumption of equal contribution between negative and positive classes. Our work does not make any assumption, and regret bounds show the equivalence between pairwise and univariate exponential loss, providing a new explanation between AdaBoost and RankBoost.

In [Menon and Williamson, 2014], there is a proposition:

Proposition 10 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm1\}}$, strictly proper composite loss ℓ with inverse link function $\Psi^{-1}(v) = 1/(1+e^{-av})$ for some $a \in \mathbb{R}\setminus\{0\}$, and scorer $s\colon \mathcal{X}\to\mathbb{R}$, there exists a convex function $F_\ell\colon [0,1]\to\mathbb{R}_+$ such that*

$$F_\ell\left(\mathit{regret}^{\mathcal{D},Univ}_{Bipart,01}(s)\right) \le \mathit{regret}^{\mathcal{D},Univ}_{Bipart,\ell}(s)$$

where $\mathit{regret}^{\mathcal{D},Univ}_{Bipart,\ell}(s)$ equals to

$$\mathbb{L}^D_{Bipart,\ell}(\mathit{Diff}(s)) - \inf_{t\colon \mathcal{X}\to\mathbb{R}}[\mathbb{L}^D_{Bipart,\ell}(\mathit{Diff}(t))]$$

where "Univ" means univariate loss, and the other notations please refer [Menon and Williamson, 2014]. This proposition shows that the univariate exponential loss is consistent with AUC optimization. In our Theorems 5 and 6, we show that the univariate exponential loss is equivalent to pairwise exponential loss, for the consistency of optimizing all performance measures such as AUC, rankloss, precision-recall,

etc. Note that the cited proposition does not involve pairwise loss, needless to say the equivalence between pairwise and univariate losses; moreover, the cited proposition considers only AUC for performance measure, whereas we consider all performance measures. [1]

6 Conclusion

This work studies the consistency of AUC optimization by minimizing pairwise surrogate losses. We first showed that calibration is necessary yet insufficient for AUC consistency. We then provide a new sufficient condition, and show the consistency of exponential loss, logistic loss, least-square hinge loss, etc. Further, we derive regret bounds for exponential and logistic losses, and obtain the regret bounds for many surrogate losses under the realizable setting. Finally, we provide regret bounds to show the equivalence between the pairwise exponential loss of AUC and univariate exponential loss of accuracy, with a direct consequence that AdaBoost and RankBoost are equivalent in the limit of infinite sample.

References

[Agarwal, 2013] S. Agarwal. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *COLT*, pages 338–353, 2013.

[Bartlett *et al.*, 2006] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Am. Stat. Assoc.*, 101(473):138–156, 2006.

[Brefeld and Scheffer, 2005] U. Brefeld and T. Scheffer. AUC maximizing support vector learning. In *ICML Workshop*, 2005.

[Clemencon *et al.*, 2008] S. Clemencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Ann. Stat.*, 36(2):844–874, 2008.

[Cohen *et al.*, 1999] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *Neural Comput.*, 10:243–270, 1999.

[Cossock and Zhang, 2008] D. Cossock and T. Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE T. Inform. Theory*, 54(11):5140–5154, 2008.

[Duchi *et al.*, 2010] J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *ICML*, pages 327–334, 2010.

[Egan, 1975] J. Egan. *Signal detection theory and ROC curve, Series in Cognition and Perception*. Academic Press, New York, 1975.

[Flach *et al.*, 2011] P. A. Flach, J. Hernández-Orallo, and C. F. Ramirez. A coherent interpretation of AUC as a measure of aggregated classification performance. In *ICML*, pages 657–664, 2011.

[Freund and Schapire, 1997] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *JCSS*, 55(1):119–139, 1997.

[Freund *et al.*, 2003] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

[Friedman *et al.*, 2000] J. Friedman, T. Hastie, and R. Tibshirani. Addtive logistic regression: A statistical view of boosting (with discussions). *Ann. Stat.*, 28(2):337–407, 2000.

[Gao and Zhou, 2012] W. Gao and Z.-H. Zhou. On the consistency of AUC optimization. *CORR abs/1208.0645*, 2012.

[Gao and Zhou, 2013] W. Gao and Z.-H. Zhou. On the consistency of multi-label learning. *AIJ*, 199:22–44, 2013.

[Gao *et al.*, 2013] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou. One-pass auc optimization. In *ICML*, pages 906–914, 2013.

[Joachims, 2005] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, pages 377–384, 2005.

[Kotlowski *et al.*, 2011] W. Kotlowski, K. Dembczynski, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In *ICML*, pages 1113–1120, 2011.

[Long and Servedio, 2013] P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *ICML*, pages 801–809, 2013.

[Marron *et al.*, 2007] J. Marron, M. Todd, and J. Ahn. Distance-weighted discrimination. *J. Am. Stat. Assoc.*, 102(480):1267–1271, 2007.

[Menon and Williamson, 2014] A. K. Menon and R. C. Williamson. Bayes-optimal scorers for bipartite ranking. In *COLT*, pages 68–106, 2014.

[Rudin and Schapire, 2009] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *JMLR*, 10:2193–2232, 2009.

[Tewari and Bartlett, 2007] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *JMLR*, 8:1007–1025, 2007.

[Uematsu and Lee, 2012] K. Uematsu and Y. Lee. On theoretically optimal ranking functions in bipartite ranking. Technical report, 2012.

[Xia *et al.*, 2008] F. Xia, T. Y. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: Theory and algorithm. In *ICML*, pages 1192–1199, 2008.

[Zhang, 2004a] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *JMLR*, 5:1225–1251, 2004.

[Zhang, 2004b] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.*, 32(1):56–85, 2004.

[Zhao *et al.*, 2011] P. Zhao, S. Hoi, R. Jin, and T. Yang. Online AUC maximization. In *ICML*, pages 233–240, 2011.

[Zuva and Zuva, 2012] K. Zuva and T. Zuva. Evaluation of information retrieval systems. *Int. J. Comput. Sci. Inform. Tech.*, 4:35–43, 2012.

[1] This explanation is added on request by a reviewer.