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
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. Clemenccon et al. [2008] studied the consistency of ranking, and showed 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 \( \eta(x) = \text{Pr}[y = +1|x] \) be the conditional probability over \( x \). For score function \( f : \mathcal{X} \rightarrow \mathbb{R} \), the AUC w.r.t. distribution \( \mathcal{D} \) is defined as

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

where \((x, y)\) and \((x', y')\) 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.

\[
R(f) = E[\eta(x)(1 - \eta(x'))\ell(f, x, x') + \eta(x')(1 - \eta(x))\ell(f, x', x)]
\]

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

\[
\mathcal{B} = \{f : R(f) = R^*\} = \{f : (f(x) - f(x')) \times (\eta(x) - \eta(x')) > 0 \text{ if } \eta(x) \neq \eta(x')\}.
\]

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

\[
\Psi(f, x, x') = \phi(f(x) - f(x'))
\]

where \( \phi \) 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 \( \phi \)-risk as

\[
R_\phi(f) = E_{x, x' \sim \mathcal{D}_\mathcal{X}}[\eta(x)(1 - \eta(x'))\phi(f(x) - f(x')) + \eta(x')(1 - \eta(x))\phi(f(x') - f(x))],
\]

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

\[
C(x, x', \alpha) = \eta(x)(1 - \eta(x'))\phi(\alpha) + \eta(x')(1 - \eta(x))\phi(-\alpha)
\]

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

3 AUC Consistency

We first define the AUC consistency as follows:

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

\[
R_\phi(f^{(n)}) \rightarrow R^*_\phi \text{ then } R(f^{(n)}) \rightarrow R^*.
\]

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 \( \phi \) is said to be classification-calibrated if, for every \( x \in \mathcal{X} \) with \( \eta(x) \neq 1/2 \),

\[
\inf_{f(x) \in \mathbb{R}} \left\{ \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)) \right\} > \inf_{f(x) \in \mathbb{R}} \left\{ \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)) \right\}.
\]

We now generalize to AUC calibration as follows:

Definition 2 The surrogate loss \( \phi \) is said to be calibrated if

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

where \( H^- (\eta, \eta') = \inf_{\alpha} \{\alpha \eta - \alpha \eta'\} \leq C(x, x', \alpha) \), \( H(\eta, \eta') = \inf_{\alpha} \{C(x, x', \alpha) + C(x, x', \alpha)\} \) is defined in Eqn. (4).

We first have

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

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
\[
\inf \int f^*(f) > E_{x,x' \sim D^*} \inf \alpha (C(\eta(x), \eta(x'), \alpha)).
\]

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 $f^* \text{ s.t.}$
\[
R(f^*) = E_{x,x' \sim D^*} \inf \alpha (C(\eta(x), \eta(x'), \alpha)).
\]
For simplicity, we consider three instances $x_1, x_2, x_3 \in \mathcal{X}$ s.t. $\eta(x_1) < \eta(x_2) < \eta(x_3)$. The conditional risk of hinge loss is given by
\[
C(x, x', \alpha) = \eta(x)(1 - \eta(x')) \max(0, 1 - \alpha) + \eta(x')(1 - \eta(x)) \max(0, 1 + \alpha).
\]
Minimizing $C(x, x', \alpha)$ gives $\alpha = -1$ if $\eta(x) < \eta(x')$. This yields $f^*(x_1) = f^*(x_2) = -1$, $f^*(x_1) - f^*(x_3) = -1$ and $f^*(x_2) - f^*(x_3) = -1$; yet they are contrary each other. ■

From Lemma 1, the study on AUC consistency should focus on the expected $\phi$-risk over the whole distribution rather than conditional $\phi$-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 $\phi$ is consistent with AUC, then $\phi$ is calibrated, and for convex $\phi$, it is differentiable at $t = 0$ with $\phi'(0) < 0$.

Proof  If $\phi$ is not calibrated, then there exist $\eta_0$ and $\eta_0'$ s.t. $\eta_0 > \eta_0'$ and $H^- (\eta_0, \eta_0') = H (\eta_0, \eta_0')$. This implies the existence of some $\alpha_0 \leq 0$ such that
\[
\eta_0(1 - \eta_0')\phi(\alpha_0) + \eta_0'(1 - \eta_0)\phi(-\alpha_0) = \inf_{\alpha \in \mathbb{R}} \{ \eta_0(1 - \eta_0')\phi(\alpha) + \eta_0'(1 - \eta_0)\phi(-\alpha) \}.
\]
We consider an instance space $\mathcal{X} = \{x_1, x_2\}$ with marginal probability $Pr[x_1] = Pr[x_2] = 1/2$, $\eta(x_1) = \eta_0$ and $\eta(x_2) = \eta_0'$. We construct a sequence $\{f^{(n)}\}_{n \geq 1}$ by selecting $f^{(n)}(x_1) = f^{(n)}(x_2) + \alpha_0$, and it is easy to get that
\[
R_\phi(f^{(n)}) \to R_\phi^* \text{ yet } R(f^{(n)}) - R^* = (\eta_0 - \eta_0')/8 \text{ as } n \to \infty,
\]
which shows that calibration is a necessary condition.

To prove $\phi'(0) < 0$, we consider the instance space $\mathcal{X} = \{x_1, x_2\}$ with $Pr[x_1] = Pr[x_2] = 1/2$, $\eta(x_1) = \eta_1$ and $\eta(x_2) = \eta_2$. Assume that $\phi$ is differentiable at $t = 0$ with $\phi'(0) \geq 0$. For convex $\phi$, we have $\eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \geq (\eta_1 - \eta_2)\alpha\phi'(0) + (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0)$ for $(\eta_1 - \eta_2)\alpha \geq 0$.

This follows that
\[
\min \{ \{\eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha)\} \}
\]
which is contrary to $H^-(\eta_1, \eta_2) > H(\eta_1, \eta_2)$.

Suppose that $\phi$ is not differentiable at $t = 0$. There exists two subgradients $g_1 > g_2$ such that
\[
\phi(t) \geq g_1t + \phi(0) \text{ and } \phi(t) \geq g_2t + \phi(0) \text{ for } t \in \mathbb{R}.
\]
If $g_1 > g_2 \geq 0$, 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 $\alpha \geq 0$, we have $\eta_1(1 - \eta_2)\phi(\alpha) + \eta_2(1 - \eta_1)\phi(-\alpha) \geq \eta_1(1 - \eta_2)(g_2\alpha + \phi(0)) + \eta_2(1 - \eta_1)(-g_1\alpha + \phi(0)) \geq (\eta_1(1 - \eta_2) + \eta_2(1 - \eta_1))\phi(0)$.

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

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 $\hat{\psi}(f, x, x') = \phi(f(x) - f(x'))$ 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 $\mathcal{X} = \{x_1, x_2, x_3\}$, and assume that, for $1 \leq i \leq 3$, the marginal probability $Pr[x_i] = 1/3$ and conditional probability $\eta_i = \eta_i(x_i)$ 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 $f_i = f(x_i)$ for $1 \leq i \leq 3$. Eqn. (3) gives
\[
R_\phi(f) = \kappa_0 + \kappa_1 \sum_{i=1}^{3} \sum_{j \neq i}^{3} \eta_i(1 - \eta_j) \max(0, 1 + f_j - f_i)
\]
where $\kappa_0 > 0$ and $\kappa_1 > 0$ are constants and independent to $f$. Minimizing $R_\phi(f)$ yields the optimal expected $\phi$-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
\[
R_\phi(f^*) = \kappa_0 + \kappa_1 (5\eta_1 + 2\eta_2 - 2\eta_1\eta_2 - 3\eta_1\eta_3 - 2\eta_2\eta_3) = R^*_\phi + \kappa_1 (2\eta_1 - \eta_2 - \eta_1\eta_3) = R^*_\phi + \kappa_1 (2\eta_1 - \eta_2 - \eta_1\eta_3)/2.
\]
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 \( \Psi(f, x, x') = \phi(f(x) - f(x')) \) is consistent with AUC if \( \phi : \mathbb{R} \to \mathbb{R} \) is a convex, differentiable and non-increasing function s.t. \( \phi'(0) < 0 \).

**Proof** It suffices to prove \( \inf_{f \in \Phi} R_\phi(f) > \inf_{f \in \Phi} R_\phi(f) \) for convex, differentiable and non-increasing function \( \phi \) s.t. \( \phi'(0) < 0 \). Assume that \( \inf_{f \in \Phi} R_\phi(f) = \inf_{f \in \Phi} R_\phi(f) \), i.e., there is an optimal function \( f^* \) s.t. \( R_\phi(f^*) = \inf_{f \in \Phi} R_\phi(f) \). This holds that, for \( \phi(x) = e^{x} \) 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 \( \Psi(f, x, x') = \phi(f(x) - f(x')) \) 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

\[
\phi(t) = \frac{1}{t} \quad \text{for } t \geq c \quad \text{and } \quad \phi(t) = \frac{1}{t} (2 - \frac{t}{\epsilon}) \quad \text{otherwise.}
\]

**Corollary 2** For distance-weighted loss, the surrogate loss \( \Psi(f, x, x') = \phi(f(x) - f(x')) \) 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.

**Corollary 3** For \( q \)-norm hinge loss, the surrogate loss \( \phi(f(x) - f(x')) = \phi(f(x) - f(x')) \) 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 \( t \leq 1 - c \); \( \phi(t) = 0 \) for \( t \geq 1 + \epsilon \); and \( \phi(t) = (t - 1 - c)^2/4\epsilon \) otherwise.

**Corollary 4** For general hinge loss, the surrogate loss \( \Psi(f, x, x') = \phi(f(x) - f(x')) \) 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

\[
\inf_{f} R_\phi(f) = E_{x, x' \sim \mathcal{D}_{n}} \inf_{\alpha} C(\eta(x), \eta(x'), \alpha).
\]

**Proof** We provide the detailed proof for exponential loss. For a fixed instance \( x_0 \in \mathcal{X} \) and \( f(x_0) \), we set

\[
f(x) = f(x_0) + \frac{1}{2} \ln \frac{\eta(x)(1 - \eta(x))}{\eta(x_0)(1 - \eta(x_0))} \quad \text{for } x \neq x_0.
\]

This holds that, for \( x_1, x_2 \in \mathcal{X} \),

\[
f(x_1) - f(x_2) = \frac{1}{2} \ln \frac{\eta(x_1)(1 - \eta(x_2))}{\eta(x_2)(1 - \eta(x_1))},
\]

which minimizes \( C(\eta(x_1), \eta(x_2), \alpha) \) by \( \alpha = f(x_1) - f(x_2) \). 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 \( \eta(x) \neq \eta(x') \),
\[
(f^*(x) - f^*(x'))(\eta(x) - \eta(x')) > 0
\]
and similarly consider logistic loss. The optimal function \( \phi \) satisfies
\[
I\left[ f(x) - f^*(x') \right] = 0
\]
for \( \eta(x) \neq \eta(x') \). The regret bounds for exponential and logistic loss are given, respectively, by
\[
R(f) - R^* \leq \sqrt{R_\phi(f) - R_\phi^*},
\]
\[
R(f) - R^* \leq 2\sqrt{R_\phi(f) - R_\phi^*}.
\]

**Proof** This proof is partly motivated from [Zhang, 2004b]. From Eqs. (1) and (2), we have
\[
R(f) - R^* = E_{\eta(x)}, \eta(x') \left[ f(x) - f^*(x') \right] < 0 \]
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 \( \text{sgn}(f(x)) \). The goal is to improve the accuracy by minimizing

\[
R_{\text{acc}}(f) = E_{(x,y) \sim D}[I[yf(x) < 0]] \\
= E_x [\eta(x)I[f(x) < 0] + (1 - \eta(x))I[f(x) > 0]].
\]

Denote by \( R_{\text{acc}}^* = \inf f \ R_{\text{acc}}(f) \), and we set the goal of optimal solutions for accuracy as follows:

\[
B_{\text{acc}} = \{ f : f(x)(\eta(x) - 1/2) > 0 \text{ for } \eta(x) \neq 1/2 \}.
\]

The popular formulation, called surrogate losses, is given by

\[
\phi_{\text{acc}}(f(x), y) = \phi(yf(x)),
\]

where \( \phi \) is a convex function such as exponential loss [Freund and Schapire, 1997], logistic loss [Friedman et al., 2000], etc. We define the expected \( \phi_{\text{acc}} \)-risk as

\[
R_{\phi_{\text{acc}}}(f) = E_{(x,y) \sim D}[\phi(yf(x))] = E_x[C_{\text{acc}}(\eta(x), f(x))]
= E_x[(1 - \eta(x))\phi(-f(x)) + \eta(x)\phi(f(x))],
\]

and denote by \( R_{\phi_{\text{acc}}}^* = \inf f \ R_{\phi_{\text{acc}}}(f) \).

Theorem 5 For exponential loss and classifier \( f \), we have

\[
R_{\phi}(f) - R_{\phi}^* \leq R_{\phi_{\text{acc}}}(f)(R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*).
\]

Proof For accuracy’s exponential surrogate loss, we have

\[
R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^* = E_x\left(\sqrt{\eta(x)e^{-f(x)}} - \sqrt{(1 - \eta(x))e^{f(x)}}\right)^2
\]

and for AUC’s exponential surrogate loss, we have

\[
R_{\phi}(f) - R_{\phi}^* = E_x, x'\left(\sqrt{\eta(x)(1 - \eta(x'))e^{-f(x) + f(x')}}
- \sqrt{\eta(x')(1 - \eta(x))e^{f(x) - f(x')}}\right)^2.
\]

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

\[
R_{\phi}(f) - R_{\phi}^* \leq 4E_x[(1 - \eta(x))e^{f(x)}](R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*)
\]

and in a similar manner, we have

\[
R_{\phi}(f) - R_{\phi}^* \leq 4E_x[\eta(x)e^{-f(x)}](R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*).
\]

By using the above inequality, this follows

\[
R_{\phi}(f) - R_{\phi}^* \leq R_{\phi_{\text{acc}}}(f)(R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*).
\]

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

\[
t^*_f = \arg\min_{t \in (-\infty, +\infty)} E_x[\eta(x)e^{-f(x) + t} + (1 - \eta(x))e^{f(x) - t}]
= \frac{1}{2} \ln(E_x[\eta(x)e^{-f(x)}]/E_x[(1 - \eta(x))e^{f(x)}]).
\]

Based on such threshold, we have

Theorem 6 For ranking function \( f \) and exponential loss,

\[
R_{\phi_{\text{acc}}}(f - t^*_f) - R_{\phi_{\text{acc}}}^* \leq 2\sqrt{R_{\phi}(f) - R_{\phi}^*},
\]

by selecting the threshold \( t^*_f \) defined above.

Proof For score function \( f(x) \), we have

\[
R_{\phi_{\text{acc}}}(f - t^*_f) - R_{\phi_{\text{acc}}}^* = -2E_x\sqrt{\eta(x)(1 - \eta(x))}
+ 2\sqrt{E_x[\eta(x)e^{-f(x)}]E_x[(1 - \eta(x))e^{f(x)}]}.
\]

For pairwise exponential loss of AUC, we have

\[
R_{\phi}(f) - R_{\phi}^* = 2E_x[\eta(x)e^{-f(x)}]E_x[(1 - \eta(x))e^{f(x)}]
- 2(E_x[\sqrt{\eta(x)(1 - \eta(x))}]^2 \geq (R_{\phi_{\text{acc}}}(f - t^*_f) - R_{\phi_{\text{acc}}}^*)^2/2
\]

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

\[
R(f) - R^* \leq \left(R_{\phi_{\text{acc}}}(f)(R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*)\right)^{1/2}
\]

\[
R_{\text{acc}}(f) - R_{\text{acc}}^* \leq \sqrt{2}(R_{\phi_{\text{acc}}}(f) - R_{\phi_{\text{acc}}}^*)^{1/2}.
\]

For exponential loss and ranking function \( f \), we have

\[
R(f) - R^* \leq (R_{\phi}(f) - R_{\phi}^*)^{1/2}
\]

\[
R_{\text{acc}}(f - t^*_f) - R_{\text{acc}}^* \leq 2(R_{\phi}(f) - R_{\phi}^*)^{1/4}.
\]

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 \( \phi_{\text{acc}}(f(x), y) = e^{-yf(x)} \) and \( \phi(f, x, x') = e^{-(f(x) - f(x'))} \), 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 \( D_{\text{st}, n} \in \Delta_{\mathcal{X} \times \{\pm 1\}} \), strictly proper composite loss \( \ell \) with inverse link function \( \Psi^{-1}(v) = 1/(1 + e^{-\alpha v}) \) for some \( \alpha \in \mathbb{R} \setminus \{0\} \), and scorers \( s : \mathcal{X} \to \mathbb{R} \), there exists a convex function \( F_{\ell} : [0, 1] \to \mathbb{R}_+ \) such that

\[
F_{\ell}\left(\text{regret}_{\text{Bipart}, 01}^D(s)\right) \leq \text{regret}_{\text{Bipart}, \ell}^D(s)
\]

where \( \text{regret}_{\text{Bipart}, \ell}^D(s) \) equals to

\[
\inf_{t : \mathcal{X} \to \mathbb{R}} \| D_{\text{Bipart}, \ell}(\text{Diff}(s)) - \inf_{t : \mathcal{X} \to \mathbb{R}} [D_{\text{Bipart}, \ell}(\text{Diff}(t))] \|
\]

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.\footnote{This explanation is added on request by a reviewer.}

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