JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–16

Surrogate Regret Bounds for the Area Under the ROC Curve via Strongly Proper Losses

Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, India

Abstract

The area under the ROC curve (AUC) is a widely used performance measure in machine learning, and has been widely studied in recent years particularly in the context of bipartite ranking. A dominant theoretical and algorithmic framework for AUC optimization/bipartite ranking has been to reduce the problem to pairwise classification; in particular, it is well known that the AUC regret can be formulated as a pairwise classification regret, which in turn can be upper bounded using usual regret bounds for binary classification. Recently, Kotlowski et al. (2011) showed AUC regret bounds in terms of the regret associated with 'balanced' versions of the standard (non-pairwise) logistic and exponential losses. In this paper, we obtain such (non-pairwise) surrogate regret bounds for the AUC in terms of a broad class of proper (composite) losses that we term *strongly proper*. Our proof technique is considerably simpler than that of Kotlowski et al. (2011), and relies on properties of proper (composite) losses as elucidated recently by Reid and Williamson (2009, 2010, 2011) and others. Our result yields explicit surrogate bounds (with no hidden balancing terms) in terms of a variety of strongly proper losses, including for example logistic, exponential, squared and squared hinge losses. An important consequence is that standard algorithms minimizing a (non-pairwise) strongly proper loss, such as logistic regression and boosting algorithms (assuming a universal function class and appropriate regularization), are in fact AUC-consistent; moreover, our results allow us to quantify the AUC regret in terms of the corresponding surrogate regret. We also obtain tighter surrogate regret bounds under certain low-noise conditions via a recent result of Clémençon and Robbiano (2011).

Keywords: Area under ROC curve (AUC), bipartite ranking, statistical consistency, regret bounds, proper losses, strongly proper losses.

1. Introduction

The area under the ROC curve (AUC) is a widely used performance measure in machine learning, and has been widely studied, particularly in the context of bipartite ranking problems (Freund et al., 2003; Cortes and Mohri, 2004; Agarwal et al., 2005). A variety of algorithms have been developed for optimizing the AUC, again often in the context of bipartite ranking; many of these algorithms effectively reduce the problem to pairwise classification (Herbrich et al., 2000; Joachims, 2002; Freund et al., 2003; Rakotomamonjy, 2004; Burges et al., 2005). In recent years, there has been much interest in understanding statistical consistency and regret behavior of such algorithms (Clémençon et al., 2008; Clémençon and Robbiano, 2011; Kotlowski et al., 2011; Uematsu and Lee, 2011); indeed, there has been much interest in understanding consistency and regret behavior for ranking problems at large, including not only bipartite instance ranking problems under the AUC performance measure, but also other forms of instance ranking problems as well as various types of label/subset ranking problems (Clémençon and Vayatis, 2007; Cossock and Zhang, 2008; Balcan et al., 2008; Ailon and Mohri, 2008; Xia et al., 2008; Duchi et al., 2010; Ravikumar et al., 2011; Buffoni et al., 2011; Calauzènes et al., 2012; Lan et al., 2012).

In this paper, we study regret bounds for the AUC, or equivalently, for bipartite instance ranking problems where instances are labeled positive or negative, and the goal is to learn a scoring function that minimizes the probability of mis-ranking a pair of positive and negative instances, i.e. that maximizes the AUC. As noted above, a popular algorithmic and theoretical approach to AUC optimization has been to reduce the problem to pairwise classification; indeed, this approach enjoys theoretical support, since the AUC regret can be formulated as a pairwise classification regret, and therefore any algorithm minimizing the latter over a suitable class of functions will also minimize the AUC regret (Clémençon et al., 2008, see Section 3.1 for a summary). Nevertheless, it has often been observed that algorithms such as AdaBoost, logistic regression, and in some cases even SVMs, which minimize the exponential, logistic, and hinge losses respectively in the standard (non-pairwise) setting, also yield good AUC performance (Cortes and Mohri, 2004; Rakotomamonjy, 2004; Rudin and Schapire, 2009). For losses such as the exponential or logistic losses, this is not surprising since algorithms minimizing these losses (but not the hinge loss) are known to effectively estimate conditional class probabilities (Zhang, 2004); since the class probability function provides the optimal ranking (Clémençon et al., 2008), it is intuitively clear (and follows formally from results in (Clémençon et al., 2008; Clémençon and Robbiano, 2011)) that any algorithm providing a good approximation to the class probability function should also produce a good ranking. However, there has been very little work on quantifying the AUC regret of a scoring function in terms of the regret associated with such surrogate losses.

Recently, Kotlowski et al. (2011) showed that the AUC regret of a scoring function can be upper bounded in terms of the regret associated with *balanced* versions of the standard (non-pairwise) exponential and logistic losses. However their proof technique builds on analyses involving the reduction of bipartite ranking to pairwise classification, and involves analyses specific to the exponential and logistic losses (see Section 3.2).

In this paper, we obtain quantitative regret bounds for the AUC in terms of a broad class of proper (composite) loss functions that we term *strongly proper*. Our proof technique is considerably simpler than that of Kotlowski et al. (2011), and relies on properties of proper (composite) losses as elucidated recently for example in (Reid and Williamson, 2009, 2010, 2011; Gneiting and Raftery, 2007; Buja et al., 2005). Our result yields explicit surrogate bounds (with no hidden balancing terms) in terms of a variety of strongly proper (composite) losses, including for example logistic, exponential, squared and squared hinge losses. An immediate consequence is that standard algorithms minimizing such losses, such as logistic regression and boosting algorithms (assuming a universal function class and appropriate regularization), are in fact AUC-consistent. We also obtain tighter surrogate regret bounds under certain low-noise conditions via a recent result of Clémençon and Robbiano (2011).

The paper is organized as follows. Section 2 gives preliminaries and background. Section 3 summarizes previous work, namely the reduction of bipartite ranking to pairwise binary classification and the result of Kotlowski et al. (2011). In Section 4 we define and characterize strongly proper losses. Section 5 contains our main result, namely surrogate regret bounds for the AUC in terms of strongly proper losses. Section 6 gives a tighter bound under certain low-noise conditions. We conclude with a brief discussion in Section 7.

2. Preliminaries and Background

Let $\mathcal{X}$ be an instance space and let D be a probability distribution on $\mathcal{X} \times \{\pm 1\}$. For $(X, Y) \sim D$ and $x \in \mathcal{X}$, we denote $\eta(x) = \mathbf{P}(Y = 1 \mid X = x)$ and $p = \mathbf{P}(Y = 1)$. We denote $\bar{\mathbb{R}} = [-\infty, \infty]$ and $\bar{\mathbb{R}}_+ = [0, \infty]$.

AUC and Bipartite Ranking. The AUC of a scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$ w.r.t. D, used as a performance measure in bipartite ranking problems, can be written as follows (Cortes and Mohri, 2004; Agarwal et al., 2005; Clémençon et al., 2008):[1,2]

$$\mathrm{AUC}_D[f] \;=\; \mathbf{E}\Big[\mathbf{1}\big((Y - Y')(f(X) - f(X')) > 0\big) + \tfrac{1}{2}\,\mathbf{1}\big(f(X) = f(X')\big) \;\Big|\; Y \neq Y'\Big]\,,$$

where $(X, Y), (X', Y')$ are assumed to be drawn i.i.d. from D, and $\mathbf{1}(\cdot)$ is 1 if its argument is true and 0 otherwise; thus the AUC of f is simply the probability that a randomly drawn positive instance is ranked higher by f (receives a higher score under f) than a randomly drawn negative instance, with ties broken uniformly at random. The optimal AUC is

$$\mathrm{AUC}^*_D \;=\; \sup_{f:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{AUC}_D[f] \;=\; 1 - \frac{1}{2p(1-p)} \mathbf{E}_{X,X'}\Big[\min\Big(\eta(X)(1-\eta(X')),\, \eta(X')(1-\eta(X))\Big)\Big]\,.$$

The AUC *regret* of a scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$ is then simply

$$\mathrm{regret}^{\mathrm{AUC}}_D[f] \;=\; \mathrm{AUC}^*_D - \mathrm{AUC}_D[f]\,.$$

We will be interested in upper bounding the AUC regret of a scoring function f in terms of its regret with respect to various (binary) loss functions.

Loss Functions and Regret. A binary loss function on a prediction space $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ is a function $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ that assigns a penalty $\ell(y, \widehat{y})$ for predicting $\widehat{y} \in \widehat{\mathcal{Y}}$ when the true label is $y \in \{\pm 1\}$.[3] For any such loss ℓ, the *ℓ-error* (or *ℓ-risk*) of a function $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ is defined as

$$\mathrm{er}^\ell_D[f] = \mathbf{E}_{(X,Y)\sim D}[\ell(Y, f(X))]\,,$$

and the *optimal ℓ-error* (or *optimal ℓ-risk* or *Bayes ℓ-risk*) is defined as

$$\mathrm{er}^{\ell,*}_D = \inf_{f:\mathcal{X}\to\widehat{\mathcal{Y}}} \mathrm{er}^\ell_D[f]\,.$$

The *ℓ-regret* of a function $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ is the difference of its ℓ-error from the optimal ℓ-error:

$$\mathrm{regret}^\ell_D[f] = \mathrm{er}^\ell_D[f] - \mathrm{er}^{\ell,*}_D\,.$$

The *conditional ℓ-risk* $L_\ell : [0, 1] \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ is defined as[4]

$$L_\ell(\eta, \widehat{y}) = \mathbf{E}_{Y\sim\eta}[\ell(Y, \widehat{y})] = \eta\,\ell(1, \widehat{y}) + (1 - \eta)\,\ell(-1, \widehat{y})\,,$$

[1] One typically works with real-valued functions; we also allow values $-\infty$ and ∞ for technical reasons.

[2] We assume measurability conditions wherever necessary.

[3] Most loss functions take values in $\mathbb{R}_+$, but some loss functions (such as the logistic loss, described later) can assign a loss of ∞ to certain label-prediction pairs.

[4] Note that we overload notation by using η here to refer to a number in $[0, 1]$; the usage should be clear from context.

where $Y \sim \eta$ denotes a $\{\pm 1\}$-valued random variable taking value $+1$ with probability η. The *conditional Bayes ℓ-risk* $H_\ell : [0,1]\to\bar{\mathbb{R}}_+$ is defined as

$$H_\ell(\eta) = \inf_{\widehat{y}\in\widehat{\mathcal{Y}}} L_\ell(\eta, \widehat{y}) \,.$$

Clearly, $\mathrm{er}_D^\ell[f] = \mathbf{E}_X[L_\ell(\eta(X), f(X))]$ and $\mathrm{er}_D^{\ell,*} = \mathbf{E}_X[H_\ell(\eta(X))]$. We note the following:

Lemma 1 *For any $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and binary loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}}\to\bar{\mathbb{R}}_+$, the conditional Bayes ℓ-risk H_ℓ is a concave function on* $[0,1]$.

The proof follows simply by observing that H_ℓ is defined as the pointwise infimum of a family of linear (and therefore concave) functions, and therefore is itself concave.

Proper and Proper Composite Losses. Proper losses in their basic form are defined on the prediction space $\widehat{\mathcal{Y}} = [0,1]$ and facilitate class probability estimation. A loss function $c : \{\pm 1\} \times [0,1]\to\bar{\mathbb{R}}_+$ is said to be *proper* if for all $\eta \in [0,1]$,

$$\eta \in \operatorname*{arg\,min}_{\widehat{\eta}\in[0,1]} L_c(\eta, \widehat{\eta}) \,,$$

and *strictly proper* if the minimizer is unique for all $\eta \in [0,1]$. Equivalently, c is proper if $\forall\eta \in [0,1]$, $H_c(\eta) = L_c(\eta, \eta)$, and strictly proper if $H_c(\eta) < L_c(\eta, \widehat{\eta})\ \forall\widehat{\eta} \neq \eta$. As in (Gneiting and Raftery, 2007), we say a loss $c : \{\pm 1\} \times [0,1]\to\bar{\mathbb{R}}_+$ is *regular* if $c(1, \widehat{\eta}) \in \mathbb{R}_+\ \forall\widehat{\eta} \in (0,1]$ and $c(-1, \widehat{\eta}) \in \mathbb{R}_+\ \forall\widehat{\eta} \in [0,1)$, i.e. if $c(y, \widehat{\eta})$ is finite for all $y, \widehat{\eta}$ except possibly for $c(1,0)$ and $c(-1,1)$, which are allowed to be infinite. We recall the following well known results:

Theorem 2 (Savage (1971)) *A regular loss $c : \{\pm 1\} \times [0,1]\to\bar{\mathbb{R}}_+$ is proper if and only if for all $\eta, \widehat{\eta} \in [0,1]$ there exists a superderivative $H_c'(\widehat{\eta})$ of H_c at $\widehat{\eta}$ such that*[5]

$$L_c(\eta, \widehat{\eta}) \;\; = \;\; H_c(\widehat{\eta}) + (\eta - \widehat{\eta}) \cdot H_c'(\widehat{\eta}) \,.$$

Theorem 3 (Hendrickson and Buehler (1971); Schervish (1989)) *A proper loss $c : \{\pm 1\} \times [0,1]\to\bar{\mathbb{R}}_+$ is strictly proper if and only if H_c is strictly concave.*

The notion of properness can be extended to binary loss functions operating on prediction spaces $\widehat{\mathcal{Y}}$ other than $[0,1]$ via composition with a *link* function $\psi : [0,1]\to\widehat{\mathcal{Y}}$. Specifically, for any $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$, a loss function $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}}\to\bar{\mathbb{R}}_+$ is said to be *proper composite* if it can be written as

$$\ell(y, \widehat{y}) = c(y, \psi^{-1}(\widehat{y}))$$

for some proper loss $c : \{\pm 1\} \times [0,1]\to\bar{\mathbb{R}}_+$ and strictly increasing (and therefore invertible) link function $\psi : [0,1]\to\widehat{\mathcal{Y}}$. Proper composite losses have been studied recently in (Reid and Williamson, 2009, 2010, 2011; Buja et al., 2005), and include several widely used losses such as squared, squared hinge, logistic, and exponential losses.

Note that for a proper composite loss ℓ formed from a proper loss c, $H_\ell = H_c$. We will refer to a proper composite loss ℓ formed from a regular proper loss c as *regular proper composite*, a composite loss formed from a strictly proper loss as *strictly proper composite*, etc. In Section 4, we will define and characterize *strongly proper* (composite) losses, which we will use to obtain regret bounds for the AUC.

[5] Here $u \in \mathbb{R}$ is a superderivative of H_c at $\widehat{\eta}$ if for all $\eta \in [0,1]$, $H_c(\widehat{\eta}) - H_c(\eta) \geq u(\widehat{\eta} - \eta)$.

3. Previous Work

As noted above, a popular theoretical and algorithmic framework for bipartite ranking/AUC optimization has been to reduce the problem to pairwise classification. Below we review this reduction in the context of our setting and notation, and then summarize the result of Kotlowski et al. (2011) which builds on this pairwise reduction.

3.1. Reduction of Bipartite Ranking to Pairwise Binary Classification

Given a distribution D on $\mathcal{X} \times \{\pm 1\}$, consider the distribution $\widetilde{D}$ on $(\mathcal{X} \times \mathcal{X}) \times \{\pm 1\}$ defined as follows:

1. Sample (X, Y) and (X', Y') i.i.d. from D;
2. If $Y = Y'$, then go to step 1; else set[6]

$$\widetilde{X} = (X, X')\,, \quad \widetilde{Y} = \operatorname{sign}(Y - Y')$$

and return $(\widetilde{X}, \widetilde{Y})$.

Now for any scoring function $f : \mathcal{X} {\rightarrow} \bar{\mathbb{R}}$, define $f_{\text{diff}} : \mathcal{X} \times \mathcal{X} {\rightarrow} \bar{\mathbb{R}}$ as

$$f_{\text{diff}}(x, x') \;=\; f(x) - f(x')\,. \tag{1}$$

Also define the 0-1 loss $\ell_{\text{0-1}} : \{\pm 1\} \times \{\pm 1\} {\rightarrow} \{0, 1\}$ as $\ell_{\text{0-1}}(y, \widehat{y}) = \mathbf{1}(\widehat{y} \neq y)$. Then it can be shown that

$$\text{AUC}_D[f] \;=\; 1 - \text{er}_{\widetilde{D}}^{\text{0-1}}[\operatorname{sign} \circ f_{\text{diff}}]\,; \qquad \text{AUC}_D^* \;=\; 1 - \text{er}_{\widetilde{D}}^{\text{0-1},*}\,,$$

where $(g \circ f)(u) = g(f(u))$; the second equality follows from the fact that the classifier $h^*(x, x') = \operatorname{sign}(\eta(x) - \eta(x'))$ (which is of the form in Eq. (1) for $f = \eta$) achieves the Bayes 0-1 risk, i.e. $\text{er}_{\widetilde{D}}^{\text{0-1}}[h^*] = \text{er}_{\widetilde{D}}^{\text{0-1},*}$ (Clémençon et al., 2008). Thus

$$\text{regret}_D^{\text{AUC}}[f] \;=\; \text{regret}_{\widetilde{D}}^{\text{0-1}}[\operatorname{sign} \circ f_{\text{diff}}]\,, \tag{2}$$

and therefore the AUC regret of a scoring function $f : \mathcal{X} {\rightarrow} \bar{\mathbb{R}}$ can be analyzed via upper bounds on the 0-1 regret of the pairwise classifier $(\operatorname{sign} \circ f_{\text{diff}}) : \mathcal{X} \times \mathcal{X} {\rightarrow} \{\pm 1\}$.[7]

In particular, as noted in (Clémençon et al., 2008), applying a result of Bartlett et al. (2006), we can upper bound the pairwise 0-1 regret of any pairwise classifier $h : \mathcal{X} \times \mathcal{X} {\rightarrow} \{\pm 1\}$ in terms of the pairwise ℓ_ϕ-regret associated with any classification-calibrated margin loss

[6] Throughout the paper, $\operatorname{sign}(u) = +1$ if $u > 0$ and -1 otherwise.

[7] Note that the setting here is somewhat different from that of Balcan et al. (2008) and Ailon and Mohri (2008), who consider a *subset* version of bipartite ranking where each instance consists of some finite subset of objects to be ranked; there also the problem is reduced to a (subset) pairwise classification problem, and it is shown that given any (subset) pairwise classifier h, a subset ranking function f can be constructed such that the resulting subset ranking regret is at most twice the subset pairwise classification regret of h (Balcan et al., 2008), or in expectation at most equal to the pairwise classification regret of h (Ailon and Mohri, 2008).

$\ell_\phi : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$, i.e. any loss of the form $\ell_\phi(y, \widehat{y}) = \phi(y\widehat{y})$ for some function $\phi : \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ satisfying $\forall\, \eta \in [0,1], \eta \neq \frac{1}{2}$,[8]

$$\widehat{y}^* \in \arg\min_{\widehat{y} \in \bar{\mathbb{R}}} L_\phi(\eta, \widehat{y}) \quad \Longrightarrow \quad \widehat{y}^*(\eta - \tfrac{1}{2}) > 0\,.$$

Theorem 4 (Bartlett et al. (2006); see also Clémençon et al. (2008)) *Let $\phi : \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ be such that the margin loss $\ell_\phi : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ defined as $\ell_\phi(y, \widehat{y}) = \phi(y\widehat{y})$ is classification-calibrated as above. Then $\exists$ strictly increasing function $g_\phi : \bar{\mathbb{R}}_+ \to [0,1]$ with $g_\phi(0) = 0$ such that for any $\widetilde{f} : \mathcal{X} \times \mathcal{X} \to \bar{\mathbb{R}}$,*

$$\text{regret}^{0\text{-}1}_{\widetilde{D}}[\text{sign} \circ \widetilde{f}] \;\leq\; g_\phi\Big(\text{regret}^{\phi}_{\widetilde{D}}[\widetilde{f}]\Big)\,.$$

Bartlett et al. (2006) give a construction for g_ϕ; in particular, for the exponential loss given by $\phi_{\exp}(u) = e^{-u}$ and logistic loss given by $\phi_{\log}(u) = \ln(1 + e^{-u})$, both of which are known to be classification-calibrated, one has

$$g_{\exp}(z) \;\leq\; \sqrt{2z}\,; \qquad g_{\log}(z) \;\leq\; \sqrt{2z}\,. \tag{3}$$

Kotlowski et al. (2011) build on these observations to bound the ranking regret in terms of the regret associated with balanced versions of the exponential and logistic losses.

3.2. Result of Kotlowski et al. (2011)

For any binary loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$, define the *balanced* ℓ-loss $\ell_{\text{bal}} : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ as

$$\ell_{\text{bal}}(y, \widehat{y}) \;=\; \frac{1}{2p}\ell(1, \widehat{y}) \cdot \mathbf{1}(y = 1) + \frac{1}{2(1-p)}\ell(-1, \widehat{y}) \cdot \mathbf{1}(y = -1)\,. \tag{4}$$

Such a balanced loss depends on the underlying distribution D via $p = \mathbf{P}(Y = 1)$. Kotlowski et al. (2011) show the following, via analyses specific to the exponential and logistic losses:

Theorem 5 (Kotlowski et al. (2011)) *For any $f : \mathcal{X} \to \bar{\mathbb{R}}$,*

$$\text{regret}^{\exp}_{\widetilde{D}}[f_{\text{diff}}] \;\leq\; \frac{9}{4}\text{regret}^{\exp,\text{bal}}_{D}[f]\,; \qquad \text{regret}^{\log}_{\widetilde{D}}[f_{\text{diff}}] \;\leq\; 2\,\text{regret}^{\log,\text{bal}}_{D}[f]\,.$$

Combining this with Eq. (2), Theorem 4, and Eq. (3) then gives the following bounds on the AUC regret in terms of the (non-pairwise) balanced exponential and logistic regrets:

$$\text{regret}^{\text{AUC}}_{D}[f] \;\leq\; \frac{3}{\sqrt{2}}\sqrt{\text{regret}^{\exp,\text{bal}}_{D}[f]}\,; \qquad \text{regret}^{\text{AUC}}_{D}[f] \;\leq\; 2\sqrt{\text{regret}^{\log,\text{bal}}_{D}[f]}\,.$$

This suggests that an algorithm that produces a function $f : \mathcal{X} \to \bar{\mathbb{R}}$ with low balanced exponential or logistic regret will also have low AUC regret. Unfortunately, since the balanced losses depend on the unknown distribution D, they cannot be optimized by an algorithm directly.[9] Below we obtain upper bounds on the AUC regret of a function f directly in terms of its loss-based regret (with no balancing terms) for a wide range of proper (composite) loss functions that we term *strongly proper*, including the exponential and logistic losses as special cases.

[8] We abbreviate $L_\phi = L_{\ell_\phi}$, $\text{er}^\phi_D = \text{er}^{\ell_\phi}_D$, etc.

[9] We note it is possible to optimize approximately balanced losses, e.g. by estimating p from the data.

4. Strongly Proper Losses

We define strongly proper losses as follows:

Definition 6 *Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a binary loss and let $\lambda > 0$. We say c is* λ-strongly proper *if for all $\eta, \widehat{\eta} \in [0,1]$,*

$$L_c(\eta, \widehat{\eta}) - H_c(\eta) \ \geq \ \frac{\lambda}{2}(\eta - \widehat{\eta})^2 \,.$$

We have the following necessary and sufficient conditions for strong properness:

Lemma 7 *Let $\lambda > 0$. If $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ is λ-strongly proper, then H_c is λ-strongly concave.*

Proof Let c be λ-strongly proper. Let $\eta_1, \eta_2 \in [0,1]$ such that $\eta_1 \neq \eta_2$, and let $t \in (0,1)$. Then we have

$$\begin{aligned} H_c\big(t\eta_1 + (1-t)\eta_2\big) &= L_c\big(t\eta_1 + (1-t)\eta_2,\, t\eta_1 + (1-t)\eta_2\big) \\ &= t\, L_c\big(\eta_1,\, t\eta_1 + (1-t)\eta_2\big) + (1-t)\, L_c\big(\eta_2,\, t\eta_1 + (1-t)\eta_2\big) \\ &\geq t\left(H_c(\eta_1) + \frac{\lambda}{2}(1-t)^2(\eta_1 - \eta_2)^2\right) + (1-t)\left(H_c(\eta_2) + \frac{\lambda}{2}t^2(\eta_1 - \eta_2)^2\right) \\ &= t\, H_c(\eta_1) + (1-t)\, H_c(\eta_2) + \frac{\lambda}{2}t(1-t)(\eta_1 - \eta_2)^2 \,. \end{aligned}$$

Thus H_c is λ-strongly concave. ■

Lemma 8 *Let $\lambda > 0$ and let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a regular proper loss. If H_c is λ-strongly concave, then c is λ-strongly proper.*

Proof Let $\eta, \widehat{\eta} \in [0,1]$. By Theorem 2, there exists a superderivative $H'_c(\widehat{\eta})$ of H_c at $\widehat{\eta}$ such that

$$L_c(\eta, \widehat{\eta}) \ = \ H_c(\widehat{\eta}) + (\eta - \widehat{\eta}) \cdot H'_c(\widehat{\eta}) \,.$$

This gives

$$\begin{aligned} L_c(\eta, \widehat{\eta}) - H_c(\eta) &= H_c(\widehat{\eta}) - H_c(\eta) + (\eta - \widehat{\eta}) \cdot H'_c(\widehat{\eta}) \\ &\geq \frac{\lambda}{2}(\widehat{\eta} - \eta)^2 \,, \qquad \text{since } H_c \text{ is } \lambda\text{-strongly concave.} \end{aligned}$$

Thus c is λ-strongly proper. ■

This gives us the following characterization of strong properness for regular proper losses:

Theorem 9 *Let $\lambda > 0$ and let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a regular proper loss. Then c is λ-strongly proper if and only if H_c is λ-strongly concave.*

It is interesting to compare this result with Theorem 3, which gives a similar characterization of strict properness of a proper loss c in terms of strict concavity of H_c. Section 5.2 contains examples of strongly proper (composite) losses. Theorem 9 will form our main tool in establishing strong properness of many of these loss functions.

5. Regret Bounds via Strongly Proper Losses

We start by recalling the following result of Clémençon et al. (2008) (adapted to account for ties, and for the conditioning on $Y \neq Y'$):

Theorem 10 (Clémençon et al. (2008)) *For any $f : \mathcal{X} \rightarrow \bar{\mathbb{R}}$,*

$$\begin{aligned}\text{regret}_D^{\text{AUC}}[f] \;=\; \frac{1}{2p(1-p)}\mathbf{E}_{X,X'}\Big[\big|\eta(X)-\eta(X')\big| \cdot \Big(\mathbf{1}\big((f(X)-f(X'))(\eta(X)-\eta(X'))<0\big) \\ +\; \tfrac{1}{2}\mathbf{1}\big(f(X)=f(X')\big)\Big)\Big]\,.\end{aligned}$$

As noted by Clémençon and Robbiano (2011), this leads to the following corollary on the AUC regret of any plug-in ranking (scoring) function based on an estimate $\widehat{\eta}$:

Corollary 11 *For any $\widehat{\eta} : \mathcal{X} \rightarrow [0,1]$,*

$$\text{regret}_D^{\text{AUC}}\big[\,\widehat{\eta}\,\big] \;\leq\; \frac{1}{p(1-p)}\mathbf{E}_X\big[\big|\widehat{\eta}(X)-\eta(X)\big|\big]\,.$$

For completeness, a proof is given in Appendix A. We are now ready for our main result.

5.1. Main Result

Theorem 12 *Let $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and let $\lambda > 0$. Let $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \rightarrow \bar{\mathbb{R}}_+$ be a λ-strongly proper composite loss. Then for any $f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$,*

$$\text{regret}_D^{\text{AUC}}[f] \;\leq\; \frac{\sqrt{2}}{p(1-p)\sqrt{\lambda}}\sqrt{\text{regret}_D^{\ell}[f]}\,.$$

Proof Let $c : \{\pm 1\} \times [0,1] \rightarrow \bar{\mathbb{R}}_+$ be a λ-strongly proper loss and $\psi : [0,1] \rightarrow \widehat{\mathcal{Y}}$ be a (strictly increasing) link function such that $\ell(y,\widehat{y}) = c(y,\psi^{-1}(\widehat{y}))$ for all $y \in \{\pm 1\}, \widehat{y} \in \widehat{\mathcal{Y}}$. Let $f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$. Then we have,

$$\begin{aligned}
\text{regret}_D^{\text{AUC}}[f] \;&=\; \text{regret}_D^{\text{AUC}}[\psi^{-1} \circ f]\,, \quad \text{since } \psi \text{ is strictly increasing} \\
&\leq\; \frac{1}{p(1-p)}\mathbf{E}_X\big[\big|\psi^{-1}(f(X))-\eta(X)\big|\big]\,, \quad \text{by Corollary 11} \\
&=\; \frac{1}{p(1-p)}\sqrt{\Big(\mathbf{E}_X\big[\big|\psi^{-1}(f(X))-\eta(X)\big|\big]\Big)^2} \\
&\leq\; \frac{1}{p(1-p)}\sqrt{\mathbf{E}_X\Big[\big(\psi^{-1}(f(X))-\eta(X)\big)^2\Big]}\,, \\
&\qquad\qquad\qquad \text{by convexity of } \phi(u)=u^2 \text{ and Jensen's inequality} \\
&\leq\; \frac{1}{p(1-p)}\sqrt{\frac{2}{\lambda}\,\mathbf{E}_X\big[L_c(\eta(X),\psi^{-1}(f(X)))-H_c(\eta(X))\big]}\,, \quad \text{since } c \text{ is } \lambda\text{-strongly proper} \\
&=\; \frac{1}{p(1-p)}\sqrt{\frac{2}{\lambda}\,\mathbf{E}_X\big[L_\ell(\eta(X),(f(X))-H_\ell(\eta(X))\big]} \\
&=\; \frac{\sqrt{2}}{p(1-p)\sqrt{\lambda}}\sqrt{\text{regret}_D^{\ell}[f]}\,.
\end{aligned}$$

This proves the result. ■

Table 1: Examples of strongly proper composite losses $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \rightarrow \bar{\mathbb{R}}_+$ together with prediction space $\widehat{\mathcal{Y}}$, underlying proper loss $c : \{\pm 1\} \times [0,1] \rightarrow \bar{\mathbb{R}}_+$, link function $\psi : [0,1] \rightarrow \widehat{\mathcal{Y}}$, and strong properness parameter λ. The spherical loss is described in Appendix B.

Loss	$\widehat{\mathcal{Y}}$	$\ell(y,\widehat{y})$	$c(y,\widehat{\eta})$		$\psi(\widehat{\eta})$	λ
			$y=1$	$y=-1$		
Exponential	$\bar{\mathbb{R}}$	$e^{-y\widehat{y}}$	$\sqrt{\frac{1-\widehat{\eta}}{\widehat{\eta}}}$	$\sqrt{\frac{\widehat{\eta}}{1-\widehat{\eta}}}$	$\frac{1}{2}\ln\left(\frac{\widehat{\eta}}{1-\widehat{\eta}}\right)$	4
Logistic	$\bar{\mathbb{R}}$	$\ln(1+e^{-y\widehat{y}})$	$-\ln\widehat{\eta}$	$-\ln(1-\widehat{\eta})$	$\ln\left(\frac{\widehat{\eta}}{1-\widehat{\eta}}\right)$	4
Squared	$[-1,1]$	$(1-y\widehat{y})^2$	$4(1-\widehat{\eta})^2$	$4\widehat{\eta}^2$	$2\widehat{\eta}-1$	8
Spherical	$[0,1]$	$c(y,\widehat{y})$	$1-\frac{\widehat{\eta}}{\sqrt{\widehat{\eta}^2+(1-\widehat{\eta})^2}}$	$1-\frac{1-\widehat{\eta}}{\sqrt{\widehat{\eta}^2+(1-\widehat{\eta})^2}}$	$\widehat{\eta}$	1

Theorem 12 shows that for any strongly proper composite loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \rightarrow \bar{\mathbb{R}}_+$, a function $f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$ with low ℓ-regret will also have low AUC regret. Below we give examples of such strongly proper (composite) loss functions. Properties of these losses are summarized in Table 1; the spherical loss is described in Appendix B.

5.2. Examples

Example 1 (Exponential loss) *The exponential loss* $\ell_{\text{exp}} : \{\pm 1\} \times \bar{\mathbb{R}} \rightarrow \bar{\mathbb{R}}_+$ *defined as*

$$\ell_{\text{exp}}(y,\widehat{y}) \;=\; e^{-y\widehat{y}}$$

is a proper composite loss with associated proper loss $c_{\text{exp}} : \{\pm 1\} \times [0,1] \rightarrow \bar{\mathbb{R}}_+$ *and link function* $\psi_{\text{exp}} : [0,1] \rightarrow \bar{\mathbb{R}}$ *given by*

$$c_{\text{exp}}(y,\widehat{\eta}) \;=\; \left(\frac{1-\widehat{\eta}}{\widehat{\eta}}\right)^{y/2} \quad ; \quad \psi_{\text{exp}}(\widehat{\eta}) \;=\; \frac{1}{2}\ln\left(\frac{\widehat{\eta}}{1-\widehat{\eta}}\right).$$

It is easily verified that c_{exp} *is regular. Moreover, it can be seen that*

$$H_{\text{exp}}(\eta) \;=\; 2\sqrt{\eta(1-\eta)}\,,$$

with

$$-H''_{\text{exp}}(\eta) \;=\; \frac{1}{2(\eta(1-\eta))^{3/2}} \;\geq\; 4 \quad \forall \eta \in [0,1]\,.$$

Thus H_{exp} *is 4-strongly concave, and so by Theorem 9, we have* ℓ_{exp} *is 4-strongly proper composite. Therefore applying Theorem 12 we have for any* $f : \mathcal{X} \rightarrow \bar{\mathbb{R}}$,

$$\text{regret}_D^{\text{AUC}}[f] \;\leq\; \frac{1}{\sqrt{2}\,p(1-p)}\sqrt{\text{regret}_D^{\text{exp}}[f]}\,.$$

Example 2 (Logistic loss) *The logistic loss* $\ell_{\text{exp}} : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ *defined as*

$$\ell_{\text{log}}(y, \widehat{y}) \;=\; \ln(1 + e^{-y\widehat{y}})$$

is a proper composite loss with associated proper loss $c_{\text{log}} : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ *and link function* $\psi_{\text{log}} : [0,1] \to \bar{\mathbb{R}}$ *given by*

$$c_{\text{log}}(1, \widehat{\eta}) \;=\; -\ln \widehat{\eta}\,; \qquad c_{\text{log}}(-1, \widehat{\eta}) \;=\; -\ln(1 - \widehat{\eta})\,; \qquad \psi_{\text{log}}(\widehat{\eta}) \;=\; \ln\left(\frac{\widehat{\eta}}{1 - \widehat{\eta}}\right).$$

Again, it is easily verified that c_{log} *is regular. Moreover, it can be seen that*

$$H_{\text{log}}(\eta) \;=\; -\eta \ln \eta - (1-\eta)\ln(1-\eta)\,,$$

with

$$-H''_{\text{log}}(\eta) \;=\; \frac{1}{\eta(1-\eta)} \;\geq\; 4 \qquad \forall \eta \in [0,1]\,.$$

Thus H_{log} *is 4-strongly concave, and so by Theorem 9, we have* ℓ_{log} *is 4-strongly proper composite. Therefore applying Theorem 12 we have for any* $f : \mathcal{X} \to \bar{\mathbb{R}}$*,*

$$\text{regret}^{\text{AUC}}_D[f] \;\leq\; \frac{1}{\sqrt{2}\,p(1-p)}\sqrt{\text{regret}^{\text{log}}_D[f]}\,.$$

Example 3 (Squared and squared hinge losses) *The (binary) squared loss* $(1 - y\widehat{y})^2$ *and squared hinge loss* $((1 - y\widehat{y})_+)^2$ *(where* $u_+ = \max(u, 0)$*) are generally defined for* $\widehat{y} \in \mathbb{R}$*. To obtain class probability estimates from a predicted value* $\widehat{y} \in \mathbb{R}$*, one then truncates* $\widehat{y}$ *to* $[-1,1]$*, and uses* $\widehat{\eta} = \frac{\widehat{y}+1}{2}$ *(Zhang, 2004). To obtain a proper composite loss, we can take* $\widehat{\mathcal{Y}} = [-1,1]$*; in this range, both losses coincide, and we can define* $\ell_{\text{sq}} : \{\pm 1\} \times [-1,1] \to [0,4]$ *as*

$$\ell_{\text{sq}}(y, \widehat{y}) \;=\; (1 - y\widehat{y})^2\,.$$

This forms a proper composite loss with associated proper loss $c_{\text{sq}} : \{\pm 1\} \times [-1,1] \to [0,4]$ *and link function* $\psi_{\text{sq}} : [0,1] \to [-1,1]$ *given by*

$$c_{\text{sq}}(1, \widehat{\eta}) \;=\; 4(1 - \widehat{\eta})^2\,; \qquad c_{\text{sq}}(-1, \widehat{\eta}) \;=\; 4\widehat{\eta}^2\,; \qquad \psi_{\text{sq}}(\widehat{\eta}) \;=\; 2\widehat{\eta} - 1\,.$$

It can be seen that

$$L_{\text{sq}}(\eta, \widehat{\eta}) \;=\; 4\eta(1 - \widehat{\eta})^2 + 4(1-\eta)\widehat{\eta}^2$$

and

$$H_{\text{sq}}(\eta) \;=\; 4\eta(1-\eta)\,,$$

so that

$$L_{\text{sq}}(\eta, \widehat{\eta}) - H_{\text{sq}}(\eta) \;=\; 4(\eta - \widehat{\eta})^2\,.$$

Thus ℓ_{sq} *is 8-strongly proper composite, and so applying Theorem 12 we have for any* $f : \mathcal{X} \to [-1,1]$*,*

$$\text{regret}^{\text{AUC}}_D[f] \;\leq\; \frac{1}{2\,p(1-p)}\sqrt{\text{regret}^{\text{sq}}_D[f]}\,.$$

Note that, if a function $f : \mathcal{X} \to \mathbb{R}$ *is learned, then our bound in terms of* ℓ_{sq}*-regret applies to the AUC regret of an appropriately transformed function* $\bar{f} : \mathcal{X} \to [-1,1]$*, such as that obtained by truncating values* $f(x) \notin [-1,1]$ *to the appropriate endpoint* -1 *or* 1*.*

6. Tighter Bounds under Low-Noise Conditions

In essence, our results exploit the fact that for a λ-strongly proper composite loss ℓ formed from a λ-strongly proper loss c and link function ψ, given any scoring function f, the $L_2(\mu)$ distance (where μ denotes the marginal density of D on $\mathcal{X}$) between $\psi^{-1}(f(X))$ and $\eta(X)$ (and therefore the $L_1(\mu)$ distance between $\psi^{-1}(f(X))$ and $\eta(X)$, which gives an upper bound on the AUC regret of f) can be upper bounded precisely in terms of the ℓ-regret of f. From this perspective, $\widehat{\eta} = \psi^{-1} \circ f$ can be treated as a 'plug-in' scoring function, which we analyzed via Corollary 11.

Recently, Clémençon and Robbiano (2011) showed that, under certain low-noise assumptions, one can obtain tighter bounds on the bipartite ranking/AUC regret of a plug-in scoring function $\widehat{\eta} : \mathcal{X} \rightarrow [0,1]$ than that offered by Corollary 11. Specifically, Clémençon and Robbiano (2011) consider the following noise assumption for bipartite ranking (inspired by the noise condition studied in (Tsybakov, 2004) for binary classification):

Noise Assumption NA(α) $(\alpha \in [0,1])$: *A distribution D on $\mathcal{X} \times \{\pm 1\}$ satisfies assumption* NA(α) *if $\exists$ a constant $C > 0$ such that for all $x \in \mathcal{X}$ and $t \in [0,1]$,*

$$\mathbf{P}_X\big(\big|\eta(X) - \eta(x)\big| \leq t\big) \;\leq\; C \cdot t^{\alpha}\,.$$

Note that $\alpha = 0$ imposes no restriction on D, while larger values of α impose greater restrictions. Clémençon and Robbiano (2011) showed the following result (adapted slightly to our setting, where the AUC is conditioned on $Y \neq Y'$):

Theorem 13 (Clémençon and Robbiano (2011)) *Let $\alpha \in [0,1)$ and $q \in [1,\infty)$. Then $\exists$ a constant $C_{\alpha,q} > 0$ such that for any distribution D on $\mathcal{X} \times \{\pm 1\}$ satisfying noise assumption* NA(α) *and any $\widehat{\eta} : \mathcal{X} \rightarrow [0,1]$,*

$$\text{regret}_D^{\text{AUC}}[\widehat{\eta}] \;\leq\; \frac{C_{\alpha,q}}{p(1-p)} \Big(\mathbf{E}_X\big[\big|\widehat{\eta}(X) - \eta(X)\big|^q\big]\Big)^{\frac{1+\alpha}{q+\alpha}}\,.$$

This allows us to obtain the following tighter version of our regret bound in terms of strongly proper losses under the same noise assumption:

Theorem 14 *Let $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and $\lambda > 0$, and let $\alpha \in [0,1)$. Let $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \rightarrow \bar{\mathbb{R}}_+$ be a λ-strongly proper composite loss. Then $\exists$ a constant $C_\alpha > 0$ such that for any distribution D on $\mathcal{X} \times \{\pm 1\}$ satisfying noise assumption* NA(α) *and any $f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$,*

$$\text{regret}_D^{\text{AUC}}[f] \;\leq\; \frac{C_\alpha}{p(1-p)} \left(\frac{2}{\lambda}\right)^{\frac{1+\alpha}{2+\alpha}} \left(\text{regret}_D^{\ell}[f]\right)^{\frac{1+\alpha}{2+\alpha}}\,.$$

Proof Let $c : \{\pm 1\} \times [0,1] \rightarrow \bar{\mathbb{R}}_+$ be a λ-strongly proper loss and $\psi : [0,1] \rightarrow \widehat{\mathcal{Y}}$ be a (strictly increasing) link function such that $\ell(y, \widehat{y}) = c(y, \psi^{-1}(\widehat{y}))$ for all $y \in \{\pm 1\}, \widehat{y} \in \widehat{\mathcal{Y}}$. Let D be a distribution on $\mathcal{X} \times \{\pm 1\}$ satisfying noise assumption NA(α) and let $f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$. Then we have,

$$\begin{aligned}
\mathrm{regret}_D^{\mathrm{AUC}}[f] &= \mathrm{regret}_D^{\mathrm{AUC}}[\psi^{-1}\circ f]\,, \quad \text{since } \psi \text{ is strictly increasing} \\
&\leq \frac{C_{\alpha,2}}{p(1-p)}\Big(\mathbf{E}_X\Big[\big(\psi^{-1}(f(X)) - \eta(X)\big)^2\Big]\Big)^{\frac{1+\alpha}{2+\alpha}}\,, \\
&\qquad\qquad\qquad\qquad \text{by Theorem 13, taking } q=2 \\
&\leq \frac{C_{\alpha,2}}{p(1-p)}\left(\frac{2}{\lambda}\,\mathbf{E}_X\big[L_c(\eta(X),\psi^{-1}(f(X))) - H_c(\eta(X))\big]\right)^{\frac{1+\alpha}{2+\alpha}}\,, \\
&\qquad\qquad\qquad\qquad \text{since } c \text{ is } \lambda\text{-strongly proper} \\
&= \frac{C_{\alpha,2}}{p(1-p)}\left(\frac{2}{\lambda}\,\mathbf{E}_X\big[L_\ell(\eta(X),f(X)) - H_\ell(\eta(X))\big]\right)^{\frac{1+\alpha}{2+\alpha}} \\
&= \frac{C_{\alpha,2}}{p(1-p)}\left(\frac{2}{\lambda}\right)^{\frac{1+\alpha}{2+\alpha}}\Big(\mathrm{regret}_D^{\ell}[f]\Big)^{\frac{1+\alpha}{2+\alpha}}\,.
\end{aligned}$$

The result follows by setting $C_\alpha = C_{\alpha,2}$. ■

For $\alpha = 0$, as noted above, there is no restriction on D, and so the above result gives the same dependence on $\mathrm{regret}_D^{\ell}[f]$ as that obtained from Theorem 12. On the other hand, as α approaches 1, the exponent of the $\mathrm{regret}_D^{\ell}[f]$ term in the above bound approaches $\frac{2}{3}$, which improves over the exponent of $\frac{1}{2}$ in Theorem 12.

7. Conclusion and Open Questions

We have obtained upper bounds on the AUC regret of a scoring function in terms of the (non-pairwise) regret associated with a broad class of proper (composite) losses that we have termed *strongly proper*. This class includes several widely used losses such as exponential, logistic, squared and squared hinge losses as special cases. An important consequence is that standard algorithms minimizing a (non-pairwise) strongly proper loss, such as logistic regression and boosting algorithms (assuming a universal function class and appropriate regularization), are in fact AUC-consistent; this explains previous empirical observations of good AUC performance of these algorithms. While our main contribution is in deriving quantitative regret bounds for the AUC in terms of such commonly used surrogate losses, the definition and characterization of strongly proper losses may also be of interest in its own right, and may find applications elsewhere.

The strongly proper composite losses that we have considered, such as the exponential, logistic, squared and spherical losses, are also margin-based classification-calibrated losses, which means the AUC regret can also be upper bounded in terms of the regret associated with pairwise versions of these losses via the reduction to pairwise classification (Section 3.1). A natural question that arises is whether it is possible to characterize conditions on the distribution under which algorithms based on one of the two approaches (minimizing a pairwise form of the loss as in RankBoost/pairwise logistic regression, or minimizing the standard loss as in AdaBoost/standard logistic regression) lead to faster convergence than those based on the other. We hope the tools and results established here may help in studying such questions in the future.

Acknowledgments

Thanks to Harish G. Ramaswamy and Arun Rajkumar for helpful discussions. Thanks also to Yoonkyung Lee for inviting me to give a talk at the ASA Conference on Statistical Learning and Data Mining held in Ann Arbor, Michigan, in June 2012; part of this work was done while preparing for that talk. This research was supported in part by a Ramanujan Fellowship from the Department of Science and Technology, Government of India.

References

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B. Sorkin. Robust reductions from ranking to classification. *Machine Learning*, 72:139–153, 2008.

Peter Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

David Buffoni, Clément Calauzenes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation: Structure and applications. Technical report, University of Pennsylvania, November 2005.

C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems 25*, pages 197–205. 2012.

Stéphan Clémençon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36:844–874, 2008.

Stéphan Clémençon and Sylvain Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Stéphan Clémençon and Nicolas Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

David Cossock and Tong Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

John Duchi, Lester Mackey, and Michael I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:1916–1921, 1971.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM Conference on Knowledge Discovery and Data Mining*, 2002.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Tie-Yan Liu. Statistical consistency of ranking methods in a rank-differentiable probability space. In *Advances in Neural Information Processing Systems 25*, pages 1241–1249. 2012.

A. Rakotomamonjy. Optimizing area under ROC curves with SVMs. In *Proceedings of the ECAI-2004 Workshop on ROC Analysis in AI*, 2004.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2010*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 618–626, 2011.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

Cynthia Rudin and Robert E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, 2009.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

M. J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17:1856–1879, 1989.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Kazuki Uematsu and Yoonkyung Lee. On theoretically optimal ranking functions in bipartite ranking. Technical Report 863, Department of Statistics, The Ohio State University, December 2011.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

Appendix A. Proof of Corollary 11

Proof Let $\widehat{\eta} : \mathcal{X} \rightarrow [0,1]$. By Theorem 10, we have

$$\text{regret}_D^{\text{AUC}}[\widehat{\eta}] \quad \leq \quad \frac{1}{2p(1-p)} \mathbf{E}_{X,X'} \Big[\big| \eta(X) - \eta(X') \big| \cdot \mathbf{1}\big((\widehat{\eta}(X) - \widehat{\eta}(X'))(\eta(X) - \eta(X')) \leq 0 \big) \Big] \,.$$

The result follows by observing that for any $x, x' \in \mathcal{X}$,

$$(\widehat{\eta}(x) - \widehat{\eta}(x'))(\eta(x) - \eta(x')) \leq 0 \quad \Longrightarrow \quad |\eta(x) - \eta(x')| \leq |\widehat{\eta}(x) - \eta(x)| + |\widehat{\eta}(x') - \eta(x')| \,.$$

To see this, note that the statement is trivially true if $\eta(x) = \eta(x')$. If $\eta(x) > \eta(x')$, then we have

$$\begin{aligned}
(\widehat{\eta}(x) - \widehat{\eta}(x'))(\eta(x) - \eta(x')) \leq 0 \quad & \Longrightarrow \quad \widehat{\eta}(x) \leq \widehat{\eta}(x') \\
& \Longrightarrow \quad \eta(x) - \eta(x') \leq (\eta(x) - \widehat{\eta}(x)) + (\widehat{\eta}(x') - \eta(x')) \\
& \Longrightarrow \quad \eta(x) - \eta(x') \leq |\eta(x) - \widehat{\eta}(x)| + |\widehat{\eta}(x') - \eta(x')| \\
& \Longrightarrow \quad |\eta(x) - \eta(x')| \leq |\widehat{\eta}(x) - \eta(x)| + |\widehat{\eta}(x') - \eta(x')| \,.
\end{aligned}$$

The case $\eta(x) < \eta(x')$ can be proved similarly. Thus we have

$$\begin{aligned}
\text{regret}_D^{\text{AUC}}[\widehat{\eta}] \quad & \leq \quad \frac{1}{2p(1-p)} \mathbf{E}_{X,X'} \Big[\big| \widehat{\eta}(X) - \eta(X) \big| + \big| \widehat{\eta}(X') - \eta(X') \big| \Big] \\
& = \quad \frac{1}{p(1-p)} \mathbf{E}_X \Big[\big| \widehat{\eta}(X) - \eta(X) \big| \Big] \,.
\end{aligned}$$

■

Appendix B. Additional Examples of Strongly Proper Losses

In general, given any concave function $H : [0,1] \to \mathbb{R}_+$, one can construct a proper loss $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ with $H_c = H$ as follows:

$$\begin{aligned} c(1,\widehat{\eta}) &= H(\widehat{\eta}) + (1-\widehat{\eta})H'(\widehat{\eta}) && (5) \\ c(-1,\widehat{\eta}) &= H(\widehat{\eta}) - \widehat{\eta}H'(\widehat{\eta})\,, && (6) \end{aligned}$$

where $H'(\widehat{\eta})$ denotes any superderivative of H at $\widehat{\eta}$. It can be verified that this gives $L_c(\eta,\widehat{\eta}) = H(\widehat{\eta}) + (\eta-\widehat{\eta})H'(\widehat{\eta})$ for all $\eta,\widehat{\eta} \in [0,1]$, and therefore $H_c(\eta) = H(\eta)$ for all $\eta \in [0,1]$. Moreover, if H is such that $H(\widehat{\eta}) + (1-\widehat{\eta})H'(\widehat{\eta}) \in \mathbb{R}_+ \ \forall \widehat{\eta} \in (0,1]$ and $H(\widehat{\eta}) - \widehat{\eta}H'(\widehat{\eta}) \in \mathbb{R}_+ \ \forall \widehat{\eta} \in [0,1)$, then the loss c constructed above is also regular. Thus, starting with any λ-strongly concave function $H : [0,1] \to \mathbb{R}_+$ satisfying these regularity conditions, any proper composite loss ℓ formed from the loss function c constructed according to Eqs. (5-6) (and any link function ψ) is λ-strongly proper composite.

Example 4 (Spherical loss) *Consider starting with the function $H_{\text{spher}} : [0,1] \to \mathbb{R}$ defined as*

$$H_{\text{spher}}(\eta) \;=\; 1 - \sqrt{\eta^2 + (1-\eta)^2}\,.$$

Then

$$H'_{\text{spher}}(\eta) \;=\; \frac{-(2\eta-1)}{\sqrt{\eta^2 + (1-\eta)^2}}$$

and

$$-H''_{\text{spher}}(\eta) \;=\; \frac{1}{(\eta^2 + (1-\eta)^2)^{3/2}} \;\geq\; 1 \qquad \forall \eta \in [0,1]\,,$$

and therefore H_{spher} is 1-strongly concave. Moreover, since H_{spher} and H'_{spher} are both bounded, the conditions for regularity are also satisfied. Thus we can use Eqs. (5-6) to construct a 1-strongly proper loss $c_{\text{spher}} : \{\pm 1\} \times [0,1] \to \mathbb{R}$ as follows:

$$c_{\text{spher}}(1,\widehat{\eta}) \;=\; H_{\text{spher}}(\widehat{\eta}) + (1-\widehat{\eta})H'_{\text{spher}}(\widehat{\eta}) \;=\; 1 - \frac{\widehat{\eta}}{\sqrt{\widehat{\eta}^2 + (1-\widehat{\eta})^2}}$$

$$c_{\text{spher}}(-1,\widehat{\eta}) \;=\; H_{\text{spher}}(\widehat{\eta}) - \widehat{\eta}H'_{\text{spher}}(\widehat{\eta}) \;=\; 1 - \frac{1-\widehat{\eta}}{\sqrt{\widehat{\eta}^2 + (1-\widehat{\eta})^2}}\,.$$

Therefore by Theorem 12, we have for any $f : \mathcal{X} \to [0,1]$,

$$\text{regret}_D^{\text{AUC}}[f] \;\leq\; \frac{\sqrt{2}}{p(1-p)}\sqrt{\text{regret}_D^{\text{spher}}[f]}\,.$$

The loss c_{spher} above corresponds to the spherical scoring rule *described in (Gneiting and Raftery, 2007). Clearly, any (strictly increasing) link function $\psi : [0,1] \to \widehat{\mathcal{Y}}$ for $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ applied to c_{spher} will then yield a* 1*-strongly proper composite loss $\ell_{\text{spher},\psi} : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \mathbb{R}$ for which the same regret bound as above holds; see (Buja et al., 2005; Reid and Williamson, 2010) for a discussion on* canonical *link functions, which ensure the resulting composite loss is convex.*