Journal of Machine Learning Research 15 (2014) 1653-1674 Submitted 7/12; Revised 12/13; Published 5/14

Surrogate Regret Bounds for Bipartite Ranking via Strongly Proper Losses

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

Editor: Nicolas Vayatis

Abstract

The problem of bipartite ranking, 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 (or equivalently, that maximizes the area under the ROC curve), has been widely studied in recent years. A dominant theoretical and algorithmic framework for the problem has been to reduce bipartite ranking to pairwise classification; in particular, it is well known that the bipartite ranking regret can be formulated as a pairwise classification regret, which in turn can be upper bounded using usual regret bounds for classification problems. Recently, Kotlowski et al. (2011) showed regret bounds for bipartite ranking in terms of the regret associated with balanced versions of the standard (non-pairwise) logistic and exponential losses. In this paper, we show that such (non-pairwise) surrogate regret bounds for bipartite ranking can be obtained in terms of a broad class of proper (composite) losses that we term as *strongly proper*. Our proof technique is much simpler than that of Kotlowski et al. (2011), and relies on properties of proper (composite) losses as elucidated recently by Reid and Williamson (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 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 consistent for bipartite ranking; moreover, our results allow us to quantify the bipartite ranking regret in terms of the corresponding surrogate regret. We also obtain tighter surrogate bounds under certain low-noise conditions via a recent result of Clémençon and Robbiano (2011).

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

1. Introduction

Ranking problems arise in a variety of applications ranging from information retrieval to recommendation systems and from computational biology to drug discovery, and have been widely studied in machine learning and statistics in the last several years. Recently, there has been much interest in understanding statistical consistency and regret behavior of algorithms for a variety of ranking problems, including various forms of label/subset ranking as well as instance ranking problems (Clémençon and Vayatis, 2007; Clémençon et al., 2008;

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; Clémençon and Robbiano, 2011; Kotlowski et al., 2011; Uematsu and Lee, 2011; Calauzènes et al., 2012; Lan et al., 2012; Ramaswamy and Agarwal, 2012; Ramaswamy et al., 2013).

In this paper, we study regret bounds for the bipartite instance ranking problem, 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, or equivalently, that maximizes the area under the ROC curve (Freund et al., 2003; Agarwal et al., 2005). A popular algorithmic and theoretical approach to bipartite ranking has been to treat the problem as analogous to pairwise classification (Herbrich et al., 2000; Joachims, 2002; Freund et al., 2003; Rakotomamonjy, 2004; Burges et al., 2005; Clémençon et al., 2008). Indeed, this approach enjoys theoretical support since the bipartite ranking 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 ranking regret (this follows formally from results of 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 bipartite ranking 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 so far on quantifying the ranking regret of a scoring function in terms of the regret associated with such surrogate losses.

Recently, Kotlowski et al. (2011) showed that the bipartite ranking 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). More fundamentally, the balanced losses in their result depend on the underlying distribution and cannot be optimized directly by an algorithm; while it is possible to do so approximately, one then loses the quantitative nature of the bounds.

In this work we obtain quantitative regret bounds for bipartite ranking 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 (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 as special cases. An immediate consequence is that standard algorithms minimizing such losses, such as standard logistic regression and boosting algorithms (assuming a universal function class and appropriate regularization), are in fact consistent for bipartite

ranking. We also obtain tighter surrogate bounds under certain low-noise conditions via a recent result of Clémençon and Robbiano (2011).

The paper is organized as follows. In Section 2 we formally set up the bipartite instance ranking problem and definitions related to loss functions and regret, and provide background on proper (composite) losses. Section 3 summarizes related work that provides the background for our study, 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 a bound on the bipartite ranking regret in terms of the regret associated with any strongly proper loss, together with several examples. Section 6 gives a tighter bound under certain low-noise conditions via a recent result of Clémençon and Robbiano (2011). We conclude with a brief discussion and some open questions in Section 7.

2. Formal Setup, Preliminaries, and Background

This section provides background on the bipartite ranking problem, binary loss functions and regret, and proper (composite) losses.

2.1 Bipartite Ranking

As in binary classification, in bipartite ranking there is an instance space $\mathcal{X}$ and binary labels $\mathcal{Y} = \{\pm 1\}$, with an unknown distribution D 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 $p = \mathbf{P}(Y = 1)$. Given i.i.d. examples $(X_1, Y_1), \ldots, (X_n, Y_n) \sim D$, the goal is to learn a scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$ (where $\bar{\mathbb{R}} = [-\infty, \infty]$) that assigns higher scores to positive instances than to negative ones.[1] Specifically, the goal is to learn a scoring function f with low *ranking error* (or *ranking risk*), defined as[2]

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

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 ranking error of f is simply the probability that a randomly drawn positive instance receives a lower score under f than a randomly drawn negative instance, with ties broken uniformly at random. The *optimal ranking error* (or *Bayes ranking error* or *Bayes ranking risk*) can be seen to be

$$\begin{aligned} \mathrm{er}_D^{\mathrm{rank},*} &= \inf_{f:\mathcal{X}\to\bar{\mathbb{R}}} \mathrm{er}_D^{\mathrm{rank}}[f] & (2)\\ &= \frac{1}{2p(1-p)} \mathbf{E}_{X,X'}\Big[\min\Big(\eta(X)(1-\eta(X')),\, \eta(X')(1-\eta(X))\Big)\Big]. & (3) \end{aligned}$$

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

$$\mathrm{regret}_D^{\mathrm{rank}}[f] \;=\; \mathrm{er}_D^{\mathrm{rank}}[f] - \mathrm{er}_D^{\mathrm{rank},*}. \quad (4)$$

1. Most algorithms learn real-valued functions; we also allow values $-\infty$ and ∞ for technical reasons.
2. We assume measurability conditions where necessary.

We will be interested in upper bounding the ranking regret of a scoring function f in terms of its regret with respect to certain other (binary) loss functions. In particular, the loss functions we consider will belong to the class of *proper* (composite) loss functions. Below we briefly review some standard notions related to loss functions and regret, and then discuss some properties of proper (composite) losses.

2.2 Loss Functions, Regret, and Conditional Risks and Regret

Assume again a probability distribution D on $\mathcal{X} \times \{\pm 1\}$ as above. Given a prediction space $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$, a binary *loss function* $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ (where $\bar{\mathbb{R}}_+ = [0, \infty]$) 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

$$\mathrm{er}_D^{\ell}[f] = \mathbf{E}_{(X,Y)\sim D}[\ell(Y, f(X))]\,, \tag{5}$$

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

$$\mathrm{er}_D^{\ell,*} = \inf_{f:\mathcal{X}\to\widehat{\mathcal{Y}}} \mathrm{er}_D^{\ell}[f]\,. \tag{6}$$

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:

$$\mathrm{regret}_D^{\ell}[f] = \mathrm{er}_D^{\ell}[f] - \mathrm{er}_D^{\ell,*}\,. \tag{7}$$

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

$$L_\ell(\eta, \widehat{y}) = \mathbf{E}_{Y\sim\eta}[\ell(Y, \widehat{y})] = \eta\,\ell(1, \widehat{y}) + (1-\eta)\,\ell(-1, \widehat{y})\,, \tag{8}$$

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

$$H_\ell(\eta) = \inf_{\widehat{y}\in\widehat{\mathcal{Y}}} L_\ell(\eta, \widehat{y})\,. \tag{9}$$

The *conditional ℓ-regret* $R_\ell : [0,1] \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ is then simply

$$R_\ell(\eta, \widehat{y}) = L_\ell(\eta, \widehat{y}) - H_\ell(\eta)\,. \tag{10}$$

Clearly, we have for $f : \mathcal{X} \to \widehat{\mathcal{Y}}$,

$$\mathrm{er}_D^{\ell}[f] = \mathbf{E}_X[L_\ell(\eta(X), f(X))]\,, \tag{11}$$

and

$$\mathrm{er}_D^{\ell,*} = \mathbf{E}_X[H_\ell(\eta(X))]\,. \tag{12}$$

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.

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.

2.3 Proper and Proper Composite Losses

In this section we review some background material related to proper and proper composite losses, as studied recently in Reid and Williamson (2010, 2011); Gneiting and Raftery (2007); Buja et al. (2005). While the material is meant to be mostly a review, some of the exposition is simplified compared to previous presentations, and we include a new, simple proof of an important fact (Theorem 4).

2.3.1 Proper Losses

We start by considering binary class probability estimation (CPE) loss functions that operate on the prediction space $\widehat{\mathcal{Y}} = [0,1]$. A binary CPE loss function $c : \{\pm 1\} \times [0,1] \to \mathbb{R}_+$ is said to be *proper* if for all $\eta \in [0,1]$,

$$\eta \in \operatorname*{arg\,min}_{\widehat{\eta} \in [0,1]} L_c(\eta, \widehat{\eta}) \,, \tag{13}$$

and *strictly proper* if the minimizer is unique for all $\eta \in [0,1]$. Equivalently, c is proper if for all $\eta \in [0,1]$, $H_c(\eta) = L_c(\eta, \eta)$, and strictly proper if $H_c(\eta) < L_c(\eta, \widehat{\eta})$ for all $\widehat{\eta} \neq \eta$. We have the following basic result:

Lemma 2 (Gneiting and Raftery (2007); Schervish (1989)) *Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a binary CPE loss. If c is proper, then $c(1,\cdot)$ is a decreasing function on $[0,1]$ and $c(-1,\cdot)$ is an increasing function. If c is strictly proper, then $c(1,\cdot)$ is strictly decreasing on $[0,1]$ and $c(-1,\cdot)$ is strictly increasing.*

We will find it useful to consider *regular* proper losses. As in Gneiting and Raftery (2007), we say a binary CPE 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. The following characterization of regular proper losses is well known (see also Gneiting and Raftery, 2007):

Theorem 3 (Savage (1971)) *A regular binary CPE 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]

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

The following is a characterization of strict properness of a proper loss c in terms of its conditional Bayes risk H_c:

Theorem 4 *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.*

This result can be proved in several ways. A proof in Gneiting and Raftery (2007) is attributed to an argument in Hendrickson and Buehler (1971). If H_c is twice differentiable, an alternative proof follows from a result in Buja et al. (2005); Schervish (1989), which shows that a proper loss c is strictly proper if and only if its 'weight function' $w_c = -H_c''$ satisfies $w_c(\eta) > 0$ for all except at most countably many points $\eta \in [0,1]$; by a very recent

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

result of Stein (2012), this condition is equivalent to strict convexity of the function $-H_c$, or equivalently, strict concavity of H_c. Here we give a third, self-contained proof of the above result that is derived from first principles, and that will be helpful when we study strongly proper losses in Section 4.

Proof [of Theorem 4] Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a proper loss. For the 'if' direction, assume H_c is strictly concave. Let $\eta, \widehat{\eta} \in [0,1]$ such that $\widehat{\eta} \neq \eta$. Then we have

$$\begin{aligned}
L_c(\eta, \widehat{\eta}) - H_c(\eta) &= L_c(\eta, \widehat{\eta}) + H_c(\widehat{\eta}) - H_c(\widehat{\eta}) - H_c(\eta) \\
&= L_c(\eta, \widehat{\eta}) + H_c(\widehat{\eta}) - 2\Big(\tfrac{1}{2}H_c(\eta) + \tfrac{1}{2}H_c(\widehat{\eta})\Big) \\
&> L_c(\eta, \widehat{\eta}) + H_c(\widehat{\eta}) - 2H_c\Big(\frac{\eta + \widehat{\eta}}{2}\Big) \\
&= 2\left(\Big(\frac{\eta + \widehat{\eta}}{2}\Big)c(1, \widehat{\eta}) + \Big(1 - \frac{\eta + \widehat{\eta}}{2}\Big)c(-1, \widehat{\eta})\right) - 2H_c\Big(\frac{\eta + \widehat{\eta}}{2}\Big) \\
&= 2\left(L_c\Big(\frac{\eta + \widehat{\eta}}{2}, \widehat{\eta}\Big) - H_c\Big(\frac{\eta + \widehat{\eta}}{2}\Big)\right) \\
&\geq 0\,.
\end{aligned}$$

Thus c is strictly proper.

Conversely, to prove the 'only if' direction, assume c is strictly 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

$$\begin{aligned}
H_c\big(t\eta_1 + (1-t)\eta_2\big) &= L_c\big(t\eta_1 + (1-t)\eta_2,\, t\eta_1 + (1-t)\eta_2\big) \\
&= t\, L_c\big(\eta_1,\, t\eta_1 + (1-t)\eta_2\big) + (1-t)\, L_c\big(\eta_2,\, t\eta_1 + (1-t)\eta_2\big) \\
&> t\, H_c(\eta_1) + (1-t)\, H_c(\eta_2)\,.
\end{aligned}$$

Thus H_c is strictly concave. ■

2.3.2 Proper Composite Losses

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 $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$ is said to be *proper composite* if it can be written as

$$\ell(y, \widehat{y}) = c(y, \psi^{-1}(\widehat{y})) \tag{14}$$

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 (2010, 2011); Buja et al. (2005), and include several widely used losses such as squared, squared hinge, logistic, and exponential losses.

It is worth noting that for a proper composite loss ℓ formed from a proper loss c, $H_\ell = H_c$. Moreover, any property associated with the underlying proper loss c can also be used to describe the composite loss ℓ; thus 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 bipartite ranking.

3. Related Work

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

3.1 Reduction of Bipartite Ranking to Pairwise Binary Classification

For any 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 $(\widetilde{X}, \widetilde{Y})$.

Then it is easy to see that, under $\widetilde{D}$,

$$\mathbf{P}\big(\widetilde{X} = (x, x')\big) \;=\; \frac{\mathbf{P}(X = x)\,\mathbf{P}(X' = x')\,\big(\eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))\big)}{2p(1-p)} \tag{15}$$

$$\widetilde{\eta}((x, x')) \;=\; \mathbf{P}\big(\widetilde{Y} = 1 \mid \widetilde{X} = (x, x')\big) \;=\; \frac{\eta(x)(1 - \eta(x'))}{\eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))} \tag{16}$$

$$\widetilde{p} \;=\; \mathbf{P}\big(\widetilde{Y} = 1\big) \;=\; \tfrac{1}{2}\,. \tag{17}$$

Moreover, for the 0-1 loss $\ell_{0\text{-}1} : \{\pm 1\} \times \{\pm 1\} \to \{0, 1\}$ given by $\ell_{0\text{-}1}(y, \widehat{y}) = \mathbf{1}(\widehat{y} \neq y)$, we have the following for any pairwise binary classifier $h : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$:

$$\operatorname{er}_{\widetilde{D}}^{0\text{-}1}[h] \;=\; \mathbf{E}_{(\widetilde{X}, \widetilde{Y}) \sim \widetilde{D}}\Big[\mathbf{1}\big(h(\widetilde{X}) \neq \widetilde{Y}\big)\Big] \tag{18}$$

$$\operatorname{er}_{\widetilde{D}}^{0\text{-}1,*} \;=\; \mathbf{E}_{\widetilde{X}}\Big[\min\Big(\widetilde{\eta}(\widetilde{X}), 1 - \widetilde{\eta}(\widetilde{X})\Big)\Big] \tag{19}$$

$$\operatorname{regret}_{\widetilde{D}}^{0\text{-}1}[h] \;=\; \operatorname{er}_{\widetilde{D}}^{0\text{-}1}[h] - \operatorname{er}_{\widetilde{D}}^{0\text{-}1,*}\,. \tag{20}$$

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

$$f_{\text{diff}}(x, x') \;=\; f(x) - f(x')\,. \tag{21}$$

Then it is easy to see that:

$$\operatorname{er}_{D}^{\text{rank}}[f] \;=\; \operatorname{er}_{\widetilde{D}}^{0\text{-}1}[\operatorname{sign} \circ f_{\text{diff}}] \tag{22}$$

$$\operatorname{er}_{D}^{\text{rank},*} \;=\; \operatorname{er}_{\widetilde{D}}^{0\text{-}1,*}\,, \tag{23}$$

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

where $(g \circ f)(u) = g(f(u))$. The equality in Eq. (23) follows from the fact that the classifier $h^*(x, x') = \text{sign}(\eta(x) - \eta(x'))$ achieves the Bayes 0-1 risk, i.e., $\text{er}^{0\text{-}1}_{\widetilde{D}}[h^*] = \text{er}^{0\text{-}1,*}_{\widetilde{D}}$ (Clémençon et al., 2008). Thus

$$\text{regret}^{\text{rank}}_{D}[f] \;\; = \;\; \text{regret}^{0\text{-}1}_{\widetilde{D}}[\text{sign} \circ f_{\text{diff}}]\,, \tag{24}$$

and therefore the ranking regret of a scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$ can be analyzed via upper bounds on the 0-1 regret of the pairwise classifier $(\text{sign} \circ f_{\text{diff}}) : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$.[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 above in terms of the pairwise ℓ_ϕ-regret associated with any classification-calibrated margin loss $\ell_\phi : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$, 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]

$$\widehat{y}^* \in \arg\min_{\widehat{y} \in \bar{\mathbb{R}}} L_\phi(\eta, \widehat{y}) \;\; \Longrightarrow \;\; \widehat{y}^*(\eta - \tfrac{1}{2}) > 0\,. \tag{25}$$

We note in particular that for every proper composite margin loss, the associated link function ψ satisfies $\psi(\frac{1}{2}) = 0$ (Reid and Williamson, 2010), and therefore every strictly proper composite margin loss is classification-calibrated in the sense above.[9]

Theorem 5 (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 $\phi_{\text{exp}}(u) = e^{-u}$ and logistic loss given by $\phi_{\log}(u) = \ln(1 + e^{-u})$, both of which are strictly proper composite losses (see Section 5.2) and are therefore classification-calibrated, one has

$$g_{\text{exp}}(z) \;\leq\; \sqrt{2z} \tag{26}$$

$$g_{\log}(z) \;\leq\; \sqrt{2z}\,. \tag{27}$$

As we describe below, 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.

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).
8. We abbreviate $L_\phi = L_{\ell_\phi}$, $\text{er}^{\phi}_{D} = \text{er}^{\ell_\phi}_{D}$, etc.
9. We note that in general, every strictly proper (composite) loss is classification-calibrated with respect to any cost-sensitive zero-one loss, using a more general definition of classification calibration with an appropriate threshold (e.g., see (Reid and Williamson, 2010)).

3.2 Result of Kotlowski et al. (2011)

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

$$\ell_{\text{bal}}(y, \widehat{y}) \;=\; \frac{1}{2p}\ell(1, \widehat{y}) \cdot \mathbf{1}(y = 1) + \frac{1}{2(1-p)}\ell(-1, \widehat{y}) \cdot \mathbf{1}(y = -1)\,. \tag{28}$$

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

Theorem 6 (Kotlowski et al. (2011)) *For any $f : \mathcal{X} \to \bar{\mathbb{R}}$,*

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

Combining this with the results of Eq. (24), Theorem 5, and Eqs. (26-27) then gives the following bounds on the ranking regret of any scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$ in terms of the (non-pairwise) balanced exponential and logistic regrets of f:

$$\text{regret}^{\text{rank}}_{D}[f] \;\leq\; \frac{3}{\sqrt{2}}\sqrt{\text{regret}^{\text{exp,bal}}_{D}[f]} \tag{29}$$

$$\text{regret}^{\text{rank}}_{D}[f] \;\leq\; 2\sqrt{\text{regret}^{\text{log,bal}}_{D}[f]}\,. \tag{30}$$

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 ranking regret. Unfortunately, since the balanced losses depend on the unknown distribution D, they cannot be optimized by an algorithm directly.[10] Kotlowski et al. (2011) provide some justification for why in certain situations, minimizing the usual exponential or logistic loss may also minimize the balanced versions of these losses; however, by doing so, one loses the quantitative nature of the above bounds. Below we obtain upper bounds on the ranking 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.

4. Strongly Proper Losses

We define strongly proper losses as follows:

Definition 7 *Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a binary CPE 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:

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

Lemma 8 *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 The proof is similar to the 'only if' direction in the proof of Theorem 4. 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

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

Thus H_c is λ-strongly concave. ■

Lemma 9 *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 3, 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 10 *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.*

We note that from Lemma 8, another way to think about strongly proper losses is that the weight function $w(\eta) = -H''(\eta)$, used to express the proper loss as a weighted mixture of cost-sensitive misclassification losses (e.g., Buja et al., 2005; Reid and Williamson, 2010), is bounded below by a positive constant.

Several examples of strongly proper (composite) losses will be provided in Section 5.2 and Section 5.3. Theorem 10 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 11 (Clémençon et al. (2008)) *For any $f : \mathcal{X} \to \bar{\mathbb{R}}$,*

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

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

Corollary 12 *For any $\widehat{\eta} : \mathcal{X} \to [0,1]$,*

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

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

5.1 Main Result

Theorem 13 *Let $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and let $\lambda > 0$. Let $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ be a λ-strongly proper composite loss. Then for any $f : \mathcal{X} \to \widehat{\mathcal{Y}}$,*

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

Proof Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a λ-strongly proper loss and $\psi : [0,1] \to \widehat{\mathcal{Y}}$ 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 $f : \mathcal{X} \to \widehat{\mathcal{Y}}$. Then we have

$$\begin{aligned}
\mathrm{regret}_D^{\mathrm{rank}}[f] &= \mathrm{regret}_D^{\mathrm{rank}}[\psi^{-1} \circ f]\,, \quad \text{since } \psi \text{ is strictly increasing} \\
&\leq \frac{1}{p(1-p)}\mathbf{E}_X\big[\big|\psi^{-1}(f(X)) - \eta(X)\big|\big]\,, \quad \text{by Corollary 12} \\
&= \frac{1}{p(1-p)}\sqrt{\Big(\mathbf{E}_X\big[\big|\psi^{-1}(f(X)) - \eta(X)\big|\big]\Big)^2} \\
&\leq \frac{1}{p(1-p)}\sqrt{\mathbf{E}_X\Big[\big(\psi^{-1}(f(X)) - \eta(X)\big)^2\Big]}\,, \\
&\qquad\qquad \text{by convexity of } \phi(u) = u^2 \text{ and Jensen's inequality} \\
&\leq \frac{1}{p(1-p)}\sqrt{\frac{2}{\lambda}\,\mathbf{E}_X\big[R_c(\eta(X), \psi^{-1}(f(X)))\big]}\,, \quad \text{since } c \text{ is } \lambda\text{-strongly proper} \\
&= \frac{1}{p(1-p)}\sqrt{\frac{2}{\lambda}\,\mathbf{E}_X\big[R_\ell(\eta(X), f(X))\big]} \\
&= \frac{\sqrt{2}}{p(1-p)\sqrt{\lambda}}\sqrt{\mathrm{regret}_D^{\ell}[f]}.
\end{aligned}$$

■

Theorem 13 shows that for any strongly proper composite loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$, a function $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ with low ℓ-regret will also have low ranking regret. Below we give several examples of such strongly proper (composite) loss functions; properties of some of these losses are summarized in Table 1.

5.2 Examples

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

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

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

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

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

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

with

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

Thus $H_{\exp}$ is 4-strongly concave, and so by Theorem 10, we have $\ell_{\exp}$ is 4-strongly proper composite. Therefore applying Theorem 13 we have for any $f : \mathcal{X} \to \bar{\mathbb{R}}$,

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

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

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

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

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

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

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

with

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

Thus $H_{\log}$ is 4-strongly concave, and so by Theorem 10, we have $\ell_{\log}$ is 4-strongly proper composite. Therefore applying Theorem 13 we have for any $f : \mathcal{X} \to \bar{\mathbb{R}}$,

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

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 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 is 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*

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

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 13 we have for any* $f : \mathcal{X} \to [-1, 1]$*,*

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

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 ranking 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*:*

$$\bar{f}(x) \;=\; \begin{cases} -1 & \text{if } f(x) < -1 \\ f(x) & \text{if } f(x) \in [-1, 1] \\ 1 & \text{if } f(x) > 1. \end{cases}$$

5.3 Constructing 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:

$$c(1, \widehat{\eta}) \;=\; H(\widehat{\eta}) + (1 - \widehat{\eta})H'(\widehat{\eta}) \tag{31}$$

$$c(-1, \widehat{\eta}) \;=\; H(\widehat{\eta}) - \widehat{\eta}H'(\widehat{\eta}) \,, \tag{32}$$

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. (31-32) (and any link function ψ) is λ-strongly proper composite.

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
Canonical 'exponential'	$\bar{\mathbb{R}}$	$\sqrt{1+\left(\frac{\widehat{y}}{2}\right)^2}-\frac{y\widehat{y}}{2}$	$\sqrt{\frac{1-\widehat{\eta}}{\widehat{\eta}}}$	$\sqrt{\frac{\widehat{\eta}}{1-\widehat{\eta}}}$	$\frac{2\widehat{\eta}-1}{\sqrt{\widehat{\eta}(1-\widehat{\eta})}}$	4
Canonical squared	$[-1,1]$	$\frac{1}{4}(1-y\widehat{y})^2$	$(1-\widehat{\eta})^2$	$\widehat{\eta}^2$	$2\widehat{\eta}-1$	2
Canonical spherical	$[-1,1]$	$1-\frac{1}{2}\left(\sqrt{2-\widehat{y}^2}+y\widehat{y}\right)$	$1-\frac{\widehat{\eta}}{\sqrt{\widehat{\eta}^2+(1-\widehat{\eta})^2}}$	$1-\frac{1-\widehat{\eta}}{\sqrt{\widehat{\eta}^2+(1-\widehat{\eta})^2}}$	$\frac{2\widehat{\eta}-1}{\sqrt{\widehat{\eta}^2+(1-\widehat{\eta})^2}}$	1

Table 1: Examples of strongly proper composite losses $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ satisfying the conditions of Theorem 13, together with prediction space $\widehat{\mathcal{Y}}$, proper loss $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$, link function $\psi : [0,1] \to \widehat{\mathcal{Y}}$, and strong properness parameter λ.

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. (31-32) 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 13, we have for any $f : \mathcal{X} \to [0,1]$,

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

The loss c_{spher} above corresponds to the spherical scoring rule *described in Gneiting and Raftery (2007).*

We also note that, for every strictly proper loss $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$, there is an associated 'canonical' link function $\psi : [0,1] \to \widehat{\mathcal{Y}}$ defined as

$$\psi(\widehat{\eta}) \;=\; c(-1, \widehat{\eta}) - c(1, \widehat{\eta})\,, \tag{33}$$

where $\widehat{\mathcal{Y}} = \{\psi(\widehat{\eta}) : \widehat{\eta} \in [0,1]\}$. We refer to composite losses comprised of such a strictly proper loss c with the corresponding canonical link ψ as *canonical* proper composite losses. Clearly, multiplying c by a factor $\alpha > 0$ results in the corresponding canonical link ψ also being multiplied by α; adding a constant (or a function $\theta(y, \widehat{\eta}) = \theta(\widehat{\eta})$ that is independent of y) to c has no effect on ψ. Conversely, given any $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and any (strictly increasing) link function $\psi : [0,1] \to \widehat{\mathcal{Y}}$, there is a unique strictly proper loss $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ (up to addition of constants or functions of the form $\theta(y, \widehat{\eta}) = \theta(\widehat{\eta})$) for which ψ is canonical; this is obtained using Eqs. (31-32) with H satisfying $H'(\widehat{\eta}) = -\psi(\widehat{\eta})$ (with possible addition of a term $\theta(\widehat{\eta})$ to both $c(1, \widehat{\eta})$ and $c(-1, \widehat{\eta})$ thus constructed). Canonical proper composite losses $\ell(y, \widehat{y})$ have some desirable properties, including for example convexity in their second argument $\widehat{y}$ for each $y \in \{\pm 1\}$; we refer the reader to Buja et al. (2005); Reid and Williamson (2010) for further discussion of such properties.

We note that the logistic loss in Example 2 is a canonical proper composite loss. On the other hand, as noted in Buja et al. (2005), the link ψ_{exp} associated with the exponential loss in Example 1 is not the canonical link for the proper loss c_{exp} (see Example 5). The squared loss in Example 3 is almost canonical, modulo a scaling factor; one needs to scale either the link function or the loss appropriately (Example 6).

Example 5 (Canonical proper composite loss associated with c_{exp}) *Let $c_{\text{exp}} : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be as in Example 1. The corresponding canonical link $\psi_{\text{exp,can}} : [0,1] \to \bar{\mathbb{R}}$ is given by*

$$\psi_{\text{exp,can}}(\widehat{\eta}) \;=\; \sqrt{\frac{\widehat{\eta}}{1-\widehat{\eta}}} - \sqrt{\frac{1-\widehat{\eta}}{\widehat{\eta}}} \;=\; \frac{2\widehat{\eta} - 1}{\sqrt{\widehat{\eta}(1-\widehat{\eta})}}\,.$$

With a little algebra, it can be seen that the resulting canonical proper composite loss $\ell_{\text{exp,can}} : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ is given by

$$\ell_{\text{exp,can}}(y, \widehat{y}) \;=\; \sqrt{1 + \left(\frac{\widehat{y}}{2}\right)^2} - \frac{y\widehat{y}}{2}\,.$$

Since we saw c_{exp} is 4-strongly proper, we have $\ell_{\text{exp,can}}$ is 4-strongly proper composite, and therefore we have from Theorem 13 that for any $f : \mathcal{X} \to \bar{\mathbb{R}}$,

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

Example 6 (Canonical squared loss) *For $c_{\text{sq}} : \{\pm 1\} \times [0,1] \to [0,4]$ defined as in Example 3, the canonical link $\psi_{\text{sq,can}} : [0,1] \to \widehat{\mathcal{Y}}$ is given by*

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

with $\widehat{\mathcal{Y}} = [-4, 4]$*, and the resulting canonical squared loss* $\ell_{\text{sq,can}} : \{\pm 1\} \times [-4, 4] \to [0, 4]$ *is given by*

$$\ell_{\text{sq,can}}(y, \widehat{y}) \;=\; \Big(1 - \frac{y\widehat{y}}{4}\Big)^2 .$$

Since we saw c_{sq} *is 4-strongly proper, we have* $\ell_{\text{sq,can}}$ *is 4-strongly proper composite, giving for any* $f : \mathcal{X} \to [-4, 4]$*,*

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

For practical purposes, this is equivalent to using the loss $\ell_{\text{sq}} : \{\pm 1\} \times [-1, 1] \to [0, 4]$ *defined in Example 3. Alternatively, we can start with a scaled version of the squared proper loss* $c_{\text{sq}'} : \{\pm 1\} \times [0, 1] \to [0, 1]$ *defined as*

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

for which the associated canonical link $\psi_{\text{sq}',\text{can}} : [0, 1] \to \widehat{\mathcal{Y}}$ *is given by*

$$\psi_{\text{sq}',\text{can}}(\widehat{\eta}) \;=\; \widehat{\eta}^2 - (1 - \widehat{\eta})^2 \;=\; 2\widehat{\eta} - 1 \,,$$

with $\widehat{\mathcal{Y}} = [-1, 1]$*, and the resulting canonical squared loss* $\ell_{\text{sq}',\text{can}} : \{\pm 1\} \times [-1, 1] \to [0, 1]$ *is given by*

$$\ell_{\text{sq}',\text{can}}(y, \widehat{y}) \;=\; \frac{(1 - y\widehat{y})^2}{4} \,.$$

Again, it can be verified that $c_{\text{sq}'}$ *is regular; in this case* $H_{\text{sq}'}(\eta) = \eta(1-\eta)$ *which is 2-strongly concave, giving that* $H_{\text{sq}',\text{can}}$ *is 4-strongly proper composite. Therefore applying Theorem 13 we have for any* $f : \mathcal{X} \to [-1, 1]$*,*

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

Again, for practical purposes, this is equivalent to using the loss ℓ_{sq} *defined in Example 3.*

Example 7 (Canonical spherical loss) *For* $c_{\text{spher}} : \{\pm 1\} \times [0, 1] \to \mathbb{R}$ *defined as in Example 4, the canonical link* $\psi_{\text{spher,can}} : [0, 1] \to \widehat{\mathcal{Y}}$ *is given by*

$$\psi_{\text{spher,can}}(\widehat{\eta}) \;=\; \frac{2\widehat{\eta} - 1}{\sqrt{\widehat{\eta}^2 + (1 - \widehat{\eta})^2}} \,,$$

with $\widehat{\mathcal{Y}} = [-1, 1]$*. The resulting canonical spherical loss* $\ell_{\text{spher,can}} : \{\pm 1\} \times [-1, 1] \to \mathbb{R}$ *is given by*

$$\ell_{\text{spher,can}}(y, \widehat{y}) \;=\; 1 - \frac{1}{2}\Big(\sqrt{2 - \widehat{y}^2} + y\widehat{y}\Big) \,.$$

Since we saw c_{spher} *is 1-strongly proper, we have* $\ell_{\text{spher,can}}$ *is 1-strongly proper composite, and therefore we have from Theorem 13 that for any* $f : \mathcal{X} \to [-1, 1]$*,*

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

6. Tighter Bounds under Low-Noise Conditions

In essence, our results exploit the fact that for any scoring function $f : \mathcal{X} \to \bar{\mathbb{R}}$, given a strongly proper composite loss ℓ, one can construct a class probability estimator from f whose distance from the true class probability function η can be upper bounded in terms of the ℓ-regret of f. Specifically, if ℓ is a strongly proper composite loss with underlying strongly proper loss c and link function ψ, then the L_2 (and therefore L_1) distance between $\psi^{-1}(f(X))$ and $\eta(X)$ (with respect to μ, the marginal density of D on $\mathcal{X}$) 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 class probability estimator and therefore a 'plug-in' scoring function, which we analyzed via Corollary 12.

Recently, Clémençon and Robbiano (2011) showed that, under certain low-noise assumptions, one can obtain tighter bounds on the ranking risk of a plug-in scoring function $\widehat{\eta} : \mathcal{X} \to [0,1]$ than that offered by Corollary 12. 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 ranking risk is conditioned on $Y \neq Y'$):

Theorem 14 (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 $\widehat{\eta} : \mathcal{X} \to [0,1]$,*

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

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 15 *Let $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ and $\lambda > 0$, and let $\alpha \in [0,1)$. Let $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ 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 $f : \mathcal{X} \to \widehat{\mathcal{Y}}$,*

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

Proof Let $c : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a λ-strongly proper loss and $\psi : [0,1] \to \widehat{\mathcal{Y}}$ 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 $f : \mathcal{X} \to \widehat{\mathcal{Y}}$. Then

$$\begin{aligned}
\text{regret}_D^{\text{rank}}[f] &= \text{regret}_D^{\text{rank}}[\psi^{-1} \circ f], \quad \text{since } \psi \text{ is strictly increasing} \\
&\leq \frac{C_{\alpha,2}}{p(1-p)} \left(\mathbf{E}_X \left[\left(\psi^{-1}(f(X)) - \eta(X) \right)^2 \right] \right)^{\frac{1+\alpha}{2+\alpha}}, \\
&\qquad \text{by Theorem 14, taking } q = 2 \\
&\leq \frac{C_{\alpha,2}}{p(1-p)} \left(\frac{2}{\lambda} \mathbf{E}_X \left[R_c(\eta(X), \psi^{-1}(f(X))) \right] \right)^{\frac{1+\alpha}{2+\alpha}}, \\
&\qquad \text{since } c \text{ is } \lambda\text{-strongly proper} \\
&= \frac{C_{\alpha,2}}{p(1-p)} \left(\frac{2}{\lambda} \mathbf{E}_X \left[R_\ell(\eta(X), f(X)) \right] \right)^{\frac{1+\alpha}{2+\alpha}} \\
&= \frac{C_{\alpha,2}}{p(1-p)} \left(\frac{2}{\lambda} \right)^{\frac{1+\alpha}{2+\alpha}} \left(\text{regret}_D^\ell[f] \right)^{\frac{1+\alpha}{2+\alpha}}.
\end{aligned}$$

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 $\text{regret}_D^\ell[f]$ as that obtained from Theorem 13. On the other hand, as α approaches 1, the exponent of the $\text{regret}_D^\ell[f]$ term in the above bound approaches $\frac{2}{3}$, which improves over the exponent of $\frac{1}{2}$ in Theorem 13.

7. Conclusion and Open Questions

We have obtained upper bounds on the bipartite ranking 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* (composite) losses. This class includes several widely used losses such as exponential, logistic, squared and squared hinge losses as special cases.

The definition and characterization of strongly proper losses may be of interest in its own right and may find applications elsewhere. As one example, we recently found strongly proper losses to be useful in analyzing consistency of algorithms for binary classification in class imbalance settings (Menon et al., 2013). An open question concerns the necessity of the regularity condition in the characterization of strong properness of a proper loss in terms of strong concavity of the conditional Bayes risk (Theorem 10). The characterization of strict properness of a proper loss in terms of strict concavity of the conditional Bayes risk (Theorem 4) does not require such an assumption, and one wonders whether it may be possible to remove the regularity assumption in the case of strong properness as well.

Many of the strongly proper composite losses that we have considered, such as the exponential, logistic, squared and spherical losses, are margin-based losses, which means the bipartite ranking 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, and to the anonymous reviewers of this work for helpful comments. 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.

Appendix A. Proof of Corollary 12

Proof Let $\widehat{\eta} : \mathcal{X} \to [0,1]$. By Theorem 11, we have

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

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

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

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

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

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

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

■

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.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. 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 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.

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

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.

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

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

Thorsten 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.

Aditya K. Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

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

Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems 25*, pages 2087–2095. 2012.

Harish G. Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems 26*, pages 1475–1483. 2013.

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. 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.

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

Oliver Stein. Twice differentiable characterizations of convexity notions for functions on full dimensional convex sets. *Schedae Informaticae*, 21:55–63, 2012.

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.