JMLR: Workshop and Conference Proceedings vol 35:1–31, 2014

Bayes-Optimal Scorers for Bipartite Ranking

Aditya Krishna Menon ADITYA.MENON@NICTA.COM.AU and **Robert C. Williamson** BOB.WILLIAMSON@NICTA.COM.AU
NICTA and the Australian National University, Canberra, ACT, Australia

Abstract

We address the following seemingly simple question: what is the Bayes-optimal scorer for a bipartite ranking risk? The answer to this question helps elucidate the relationship between bipartite ranking and other established learning problems. We show that the answer is non-trivial in general, but may be easily determined for certain special cases using the theory of proper losses. Our analysis immediately establishes equivalence relationships between several seemingly disparate approaches to bipartite ranking, such as minimising a suitable class-probability estimation risk, and minimising the p-norm push risk proposed in Rudin (2009).

Keywords: Bipartite ranking, p-norm push, class-probability estimation, proper losses

1. Bipartite ranking: an informal introduction

Bipartite ranking problems (Agarwal et al., 2005; Clémençon et al., 2008; Kotlowski et al., 2011) have received considerable attention from the machine learning community. In such problems, we have as input a training set of examples, each of which comprises an *instance* (typically a vector of features describing some entity) with an associated *binary label* (describing whether the instance possesses some attribute, typically denoted "positive" or "negative"). The goal is to learn a *scorer*, which assigns each instance a real number, such that positive instances have a higher score than negative instances. Violations of this condition are penalised according to some loss ℓ, and the *bipartite ranking risk* of a scorer is its expected penalty according to ℓ.

Given a loss ℓ, the *Bayes-optimal* scorers are those that attain the minimal bipartite risk with respect to ℓ. These scorers, by definition, precisely describe what one needs to estimate to achieve good performance. Characterising the Bayes-optimal scorers for a loss ℓ thus delineates the relationship between bipartite ranking and other established problems, clarifies the target scorer sought by existing bipartite ranking algorithms (Cohen et al., 1999; Herbrich et al., 2000; Burges et al., 2005), and suggests new algorithms that minimise alternate risks with the same Bayes-optimal scorers. This characterisation has previously been restricted to the cases of 0-1 loss (Clémençon et al., 2008) and a subset of convex margin losses (Uematsu and Lee, 2012; Gao and Zhou, 2012).

In this paper, we compute the Bayes-optimal scorers for the bipartite ranking risk when ℓ belongs to the family of *proper composite losses* (Reid and Williamson, 2010). This family includes as special cases the 0-1 loss and the margin losses studied in Uematsu and Lee (2012); Gao and Zhou (2012), and consequently we generalise and unify the existing results. We show that in some special cases, the Bayes-optimal scorers have a simple form intimately related to those of other learning problems. Consequently, we find equivalence relationships between several disparate approaches to the bipartite ranking problem, including performing class-probability estimation with a suitable

proper composite loss, and minimising the p-norm push risk, a proposal due to Rudin (2009) which aims to focus accuracy at the head of the ranked list.

We begin the paper with some definitions and notation (§2), and then precisely define the risks of interest to us (§3). We then determine the Bayes-optimal scorers for bipartite ranking (§4) and the p-norm push extension (§5). We then look at the implications of these findings in terms of equivalence relationships between four disparate approaches to bipartite ranking (§6).

2. Preliminary definitions

We define the relevant quantities used in the rest of the paper, and fix some notation.

2.1. Notation

We denote by $\mathbb{R}$ the set of real numbers, and $\mathbb{R}_+ = [0, \infty)$. We use scripted calligraphic fonts, e.g. $\mathcal{X}, \mathcal{Y}$, to denote arbitrary sets. We use $\mathcal{X} \setminus \mathcal{Y}$ to denote set difference, and $\emptyset$ to denote the empty set. We use sans-serif fonts, e.g. X, Y, to denote random variables. The expectation of a random variable is denoted by $\mathbb{E}[\mathsf{X}]$. Given a set $\mathcal{S}$, we denote by $\Delta_{\mathcal{S}}$ by the set of all distributions on $\mathcal{S}$. We denote by $\text{Ber}(\theta)$ the Bernoulli distribution with parameter $\theta \in [0, 1]$.

For any function $f \colon \mathcal{X} \to \mathbb{R}$, we denote by $\operatorname{Argmin}_{x \in \mathcal{X}} f(x)$ the set of all $x \in \mathcal{X}$ such that $f(x) \leq f(x')$ for all $x' \in \mathcal{X}$. When f has a unique minimiser, we denote this by $\operatorname{argmin}_{x \in \mathcal{X}} f(x)$. We denote by $\text{Diff}(f) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the function satisfying $(\text{Diff}(f))(x, x') = f(x) - f(x')$ for every $x, x' \in \mathcal{X}$. For a set $\mathcal{F} = \{f \colon \mathcal{X} \to \mathbb{R}\}$, we define $\text{Diff}(\mathcal{F}) = \{\text{Diff}(f) : f \in \mathcal{F}\}$.

We use the Iverson bracket (Knuth, 1992) $[\![p]\!]$ to denote the indicator function, whose value is 1 if p is true and 0 otherwise. For any $x \in \mathbb{R}$, we define $\text{sign}(x) = [\![x \geq 0]\!] - [\![x = 0]\!]$. The sigmoid function is defined by $\sigma(z) = \frac{1}{1+e^{-z}}$.

2.2. Scorers

We will focus on supervised learning problems involving an instance space $\mathcal{X}$ (often $\mathbb{R}^n$), and a label space $\mathcal{Y} = \{\pm 1\}$. We call an element $x \in \mathcal{X}$ an *instance*, and an element $y \in \{\pm 1\}$ a *label*. A *scorer* s is some function $s \colon \mathcal{X} \to \mathcal{V}$, where $\mathcal{V} \subseteq \mathbb{R}$. A *classifier* is a scorer with $\mathcal{V} = \{\pm 1\}$, and a *class-probability estimator* is a scorer with $\mathcal{V} = [0, 1]$. A *pair-scorer* s_{Pair} for a product space $\mathcal{X} \times \mathcal{X}$ is some function $s_{\text{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathcal{V}$. We call a pair-scorer s_{Pair} *decomposable* if

$$s_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}} := \{\text{Diff}(s) : s \colon \mathcal{X} \to \mathbb{R}\}.$$

2.3. Loss functions

A *loss* ℓ is some measurable function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. We use $\ell_1(v) = \ell(1, v)$ and $\ell_{-1}(v) = \ell(-1, v)$ to denote the individual *partial losses*. We call a loss ℓ *symmetric* if $(\forall v \in \mathbb{R})\, \ell_1(v) = \ell_{-1}(-v)$, or equivalently if it is a *margin loss* i.e. $\ell(y, v) = \phi(yv)$ for some $\phi : \mathbb{R} \to \mathbb{R}$. We define the *conditional ℓ-risk* to be

$$L_\ell(\eta, s) = \mathbb{E}_{\mathsf{Y} \sim \text{Ber}(\eta)} [\ell(\mathsf{Y}, s)] = \eta \ell_1(s) + (1 - \eta)\ell_{-1}(s). \tag{1}$$

A loss of special interest is the *zero-one* or *misclassification loss*, $\ell^{01}(y, v) = [\![yv < 0]\!] + \frac{1}{2}[\![v = 0]\!]$.

A *probability estimation loss* λ is some measurable function $\lambda : \{\pm 1\} \times [0,1] \to \mathbb{R}_+$. We call a probability estimation loss *proper* if (Buja et al., 2005; Reid and Williamson, 2010)

$$(\forall \eta, \eta' \in [0,1])\, L_\lambda(\eta, \eta) \leq L_\lambda(\eta, \eta'). \tag{2}$$

We call a loss *strictly proper* if the inequality is strict. We call a loss ℓ (*strictly*) *proper composite* if there is some invertible *link function* $\Psi : [0,1] \to \mathbb{R}$ such that the probability estimation loss $\lambda(y, v) = \ell(y, \Psi(v))$ is (strictly) proper (Reid and Williamson, 2010). For such losses, we have that for every $\eta \in [0,1], v \in \mathbb{R}$, $L_\ell(\eta, \Psi(\eta)) \leq L_\ell(\eta, v)$. When ℓ is differentiable, its inverse link is (Reid and Williamson, 2010, Corollary 12)

$$(\forall v \in \mathcal{V})\, \Psi^{-1}(v) = \frac{1}{1 - \frac{\ell'_1(v)}{\ell'_{-1}(v)}}. \tag{3}$$

The squared, squared hinge, exponential and logistic loss are all proper composite.

2.4. Conditional distributions

Any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ may be specified exactly by the following three components:

$$(\forall x \in \mathcal{X})\, (P(x), Q(x), \pi) = (\Pr[\mathsf{X} = x | \mathsf{Y} = 1], \Pr[\mathsf{X} = x | \mathsf{Y} = -1], \Pr[\mathsf{Y} = 1]),$$

or alternately by the two components:

$$(\forall x \in \mathcal{X})\, (M(x), \eta(x)) = (\Pr[\mathsf{X} = x], \Pr[\mathsf{Y} = 1 | \mathsf{X} = x]).$$

We refer to P, Q as the *class conditional* densities, and π the *base rate*. We refer to M as the *observation density*, and η the *class-conditional density*. When we wish to refer to these densities, we will explicitly parameterise the distribution $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ as either $D_{P,Q,\pi}$ or $D_{M,\eta}$ as appropriate.

3. Classification, class-probability estimation and bipartite ranking

We describe the problems of interest in this paper by means of their statistical risks.

3.1. Classification and class-probability estimation

Given any $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and loss ℓ, we define the *ℓ-classification risk* for a scorer s to be

$$\mathbb{L}_\ell^D(s) = \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] = \mathbb{E}_{\mathsf{X} \sim M} \left[L_\ell(\eta(\mathsf{X}), s(\mathsf{X}))\right], \tag{4}$$

recalling that $L_\ell(\eta, s)$ is the conditional ℓ-risk (Equation 1). When the infimum is achievable[1], the set of *Bayes-optimal ℓ-scorers* comprises those that minimise the risk:

$$\mathbb{S}_\ell^{D,*} = \operatorname*{Argmin}_{s\colon \mathcal{X} \to \mathbb{R}} \mathbb{L}_\ell^D(s).$$

Under appropriate measurability assumptions, this set may be discerned pointwise, by studying the minimisers of the conditional risk L_ℓ (Steinwart, 2007).

1. The optimal scorer for logistic loss is $s^*(x) = \log \frac{\eta(x)}{1-\eta(x)}$. If the data is separable, i.e. $\eta(x) \in \{0,1\}$ for every x, that would require $s^*(x) \in \{\pm\infty\}$, and so the infimum is not attainable.

In *binary classification* (Devroye et al., 1996), we wish to find a scorer that (approximately) minimises the risk for $\ell = \ell^{01}$, which in a slight abuse of notation we write as $\mathbb{L}^D_{01}$. Directly minimising $\mathbb{L}^D_{01}$ may be computationally challenging due to the non-convexity of ℓ^{01}. A common approach is to instead find a scorer that (approximately) minimises $\mathbb{L}^D_\ell$ for some *surrogate* loss ℓ; via *surrogate regret bounds* (Zhang, 2004; Bartlett et al., 2006), one can quantify how well this scorer performs with respect to ℓ^{01}. When ℓ is proper composite, minimising $\mathbb{L}^D_\ell$ is in fact precisely the goal of the *class-probability estimation* problem (Buja et al., 2005; Reid and Williamson, 2010).

3.2. Bipartite ranking

Given any $D_{P,Q,\pi} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ, we define the *ℓ-bipartite risk* for a pair-scorer s_{Pair} to be

$$\mathbb{L}^D_{\text{Bipart},\ell}(s_{\text{Pair}}) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\frac{\ell_1(s_{\text{Pair}}(\mathsf{X},\mathsf{X}')) + \ell_{-1}(s_{\text{Pair}}(\mathsf{X}',\mathsf{X}))}{2}\right]. \tag{5}$$

When the infimum is achievable, the set of *Bayes-optimal ℓ-bipartite pair-scorers* is

$$\mathcal{S}^{D,*}_{\text{Bipart},\ell} = \operatorname*{Argmin}_{s_{\text{Pair}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}^D_{\text{Bipart},\ell}(s_{\text{Pair}}),$$

and the set of *Bayes-optimal univariate scorers* is

$$\mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},\ell} = \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^D_{\text{Bipart},\ell}(\text{Diff}(s)).$$

In *bipartite ranking* (Agarwal et al., 2005; Clémençon et al., 2008; Uematsu and Lee, 2012), we wish to find a scorer $s\colon \mathcal{X}\to\mathbb{R}$ such that $\mathbb{L}^D_{\text{Bipart},01}(\text{Diff}(s))$ is (approximately) minimised; equivalently, we seek to minimise $\mathbb{L}^D_{\text{Bipart},01}(s_{\text{Pair}})$ over all $s_{\text{Pair}} \in \mathcal{S}_{\text{Decomp}}$. As with classification, for computational reasons one often (approximately) minimises $\mathbb{L}^D_{\text{Bipart},\ell}$ for some surrogate loss ℓ. One minus the risk $\mathbb{L}^D_{\text{Bipart},01}(\text{Diff}(s))$ is more popularly known as the *area under the ROC curve (AUC)* of the scorer s (Agarwal et al., 2005; Clémençon et al., 2008); minimising the risk is thus equivalent to maximising the AUC.

4. Bayes-optimal scorers for the bipartite ranking risk

We now derive the Bayes-optimal scorers for bipartite ranking when $\ell = \ell^{01}$ and when ℓ is any strictly proper composite loss. Knowledge of the optimal scorers gives insight into the problem, and helps relate it to the more familiar tasks of binary classification and class-probability estimation. Before proceeding, we first recall the Bayes-optimal scorers for the latter problems given some $D = D_{P,Q,\pi} = D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$.

4.1. Binary classification and class-probability estimation

For ℓ^{01}, any Bayes-optimal scorer has the same sign as $\eta(x) - 1/2$ when this quantity is nonzero (Devroye et al., 1996, pg. 10), (Bartlett et al., 2006):

$$\mathcal{S}^{D,*}_{01} = \left\{ s\colon \mathcal{X}\to\mathbb{R} : \begin{array}{l} \eta(x) \neq 1/2 \implies \text{sign}(s(x)) = \text{sign}(2\eta(x)-1) \\ \eta(x) = 1/2 \implies s(x) \in \mathbb{R} \end{array} \right\}. \tag{6}$$

Thus, for ℓ^{01}, what is of interest is determining whether or not each instance has a greater than random chance of being labelled positive. When ℓ is a proper composite loss with link Ψ, from the definition of properness (Equation 2) we can specify one minimiser of the conditional risk, which applied pointwise gives:

$$\{\Psi \circ \eta\} \subseteq \mathcal{S}_{\ell}^{D,*}. \tag{7}$$

This is an equality if and only if ℓ is strictly proper composite. Thus, a strictly proper composite loss requires precise information about η, unlike ℓ^{01}. Observe that $\Psi \circ \eta$ may be trivially transformed to give an optimal scorer for ℓ^{01}; thus, exactly solving class-probability estimation also solves binary classification. For an approximate solution, one can bound the excess ℓ^{01} error via a surrogate regret bound (Reid and Williamson, 2009).

4.2. Bipartite ranking with pair-scorers

When looking to establish the Bayes-optimal scorers for bipartite ranking, we immediately face a challenge. Finding the s that minimises $\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s))$ is equivalent to finding the s_{Pair} that minimises $\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}})$ subject to $s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}}$. While the latter constraint seems innocuous, it means we need to reason about a minimiser in a *restricted* function class. Thus, in general, it is no longer possible to simply study the conditional risk and make a pointwise analysis. But if it happens that the optimal *pair*-scorer is in fact decomposable, we can effectively ignore the restricted function class, as the following makes precise. (Proofs not in the main body may be found in the Appendix.)

Proposition 1 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ,*

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset \implies \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}} = \mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}).$$

The result simplifies when *every* Bayes-optimal pair-scorer is decomposable, which is of interest for example when there is a unique optimal pair-scorer.

Corollary 2 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and loss ℓ,*

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \subseteq \mathcal{S}_{\mathrm{Decomp}} \iff \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}).$$

Simply put, the decomposable Bayes-optimal pair-scorers are exactly the Bayes-optimal scorers passed through Diff. Thus, if we can show that $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset$ for a loss ℓ, we automatically deduce the Bayes-optimal scorer. To begin, we must find the Bayes-optimal pair-scorers. To do so, we exploit an equivalence of the bipartite risk to classification on pairs, which is well known for ℓ^{01} (Balcan et al., 2008; Kotlowski et al., 2011; Agarwal, 2013).

Lemma 3 *For any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, loss ℓ and pair-scorer $s_{\mathrm{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,*

$$\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}}) = \mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}}),$$

where $\mathrm{Bipart}(D) \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm 1\}}$ is defined by

$$(\Pr[\mathsf{X},\mathsf{X}'|\mathsf{Z}=1], \Pr[\mathsf{X},\mathsf{X}'|\mathsf{Z}=-1], \Pr[\mathsf{Z}=1]) = (P(\mathsf{X})Q(\mathsf{X}'), P(\mathsf{X}')Q(\mathsf{X}), 1/2).$$

This implies that the Bayes-optimal pair-scorers with respect to D are exactly the Bayes-optimal classifiers with respect to $\mathrm{Bipart}(D)$. Given some $D_{M,\eta}$, let $\mathrm{Bipart}(D)$ have observation-conditional density η_{Pair}. In Appendix C, we show that

$$\eta_{\mathrm{Pair}} = \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta),$$

where $\sigma(\cdot)$ is the sigmoid function. Thus, Equation 6 implies that

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},01} = \left\{ s_{\mathrm{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R} : \begin{array}{rcl} \eta(x) \neq \eta(x') & \implies & \mathrm{sign}(s_{\mathrm{Pair}}(x,x')) = \mathrm{sign}(\eta(x) - \eta(x')) \\ \eta(x) = \eta(x') & \implies & s_{\mathrm{Pair}}(x,x') \in \mathbb{R}. \end{array} \right\}, \tag{8}$$

where we have used the fact that $\mathrm{sign}(2\eta_{\mathrm{Pair}}(x,x') - 1) = \mathrm{sign}(\eta(x) - \eta(x'))$. Similarly, when ℓ is proper composite with link Ψ,

$$\{\Psi \circ \eta_{\mathrm{Pair}}\} = \{\Psi \circ \sigma \circ \mathrm{Diff}(\sigma^{-1} \circ \eta)\} \subseteq \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}. \tag{9}$$

This is an equality if and only if ℓ is strictly proper composite. As with binary classification, the optimal solution for a proper composite loss may be trivially transformed to reside in $\mathcal{S}^{D,*}_{\mathrm{Bipart},01}$.

4.3. Bipartite ranking with univariate scorers: decomposable case

We now return to our goal of determining $\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}$. We first handle the case where there is a decomposable Bayes-optimal pair-scorer, which allows us to easily compute the optimal scorer. Applying Proposition 1 and Equation 8, we have the following for ℓ^{01}.

Proposition 4 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X} \times \{\pm 1\}}$,*

$$\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},01} = \{s \colon \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s\}$$

for some (not necessarily strictly) monotone increasing $\phi : [0,1] \to \mathbb{R}$.

If the ϕ in Proposition 4 is not strictly monotone increasing, then if $\eta(x) = \eta(x')$ for some $x \neq x' \in \mathcal{X}$, it may be that $s(x) \neq s(x')$. Nonetheless, an immediate corollary is that any strictly monotone increasing transform of η is necessarily an optimal univariate scorer.

Corollary 5 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X} \times \{\pm 1\}}$ and any strictly monotone increasing $\phi : [0,1] \to \mathbb{R}$.*

$$\phi \circ \eta \in \mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},01}.$$

We see that like class-probability estimation, bipartite ranking with ℓ^{01} aims to find a transformation of η. Unlike class-probability estimation, one is satisfied with *any* strictly monotone transformation, not necessarily one specified by the loss itself. Loosely, then, bipartite ranking is less "strict" than class-probability estimation.

We now proceed to the case where ℓ is a proper composite loss. To apply Corollary 2, we characterise the subset of proper composite losses for which there exists a decomposable pair-scorer.

Proposition 6 (Decomposability of Bayes-optimal bipartite pair-scorer.) *Given any* $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and a strictly proper composite loss* ℓ *with a differentiable, invertible link function* Ψ*,*

$$\mathcal{S}^{D,*}_{\text{Bipart},\ell} \subseteq \mathcal{S}_{\text{Decomp}} \iff (\exists a \in \mathbb{R} \setminus \{0\})\,(\forall v \in \mathbb{R})\,\Psi^{-1}(v) = \frac{1}{1+e^{-av}}.$$

Proof

($\Longleftarrow$) Let the link function of ℓ have the specified form, so that $\Psi(v) = \frac{1}{a}\log\frac{v}{1-v} = \frac{1}{a}\sigma^{-1}(v)$, and so $(\Psi \circ \sigma)(v) = \frac{v}{a}$. From Equation 9, the Bayes-optimal pair-scorer is

$$s^*_{\text{Pair}} = \frac{1}{a}\cdot \text{Diff}(\sigma^{-1}\circ\eta) = \text{Diff}\left(\left(\frac{1}{a}\cdot\sigma^{-1}\right)\circ\eta\right) \in \mathcal{S}_{\text{Decomp}}.$$

($\Longrightarrow$) If $\mathcal{S}^{D,*}_{\text{Bipart},\ell} \cap \mathcal{S}_{\text{Decomp}} \neq \emptyset$,

$$\Psi \circ \sigma \circ \text{Diff}(\sigma^{-1}\circ\eta) \in \mathcal{S}_{\text{Decomp}}.$$

We wish to determine the Ψ for which this holds. Let $f = \Psi \circ \sigma \circ \log$, so that the above becomes

$$(\forall x, x' \in \mathcal{X})\, f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x'))}}\right) = g(x) - g(x')$$

for some $g : \mathcal{X} \to \mathbb{R}$. Noting that $g(x) - g(x') = g(x) - g(x'') + g(x'') - g(x')$ for any $x'' \in \mathcal{X}$,

$$(\forall x, x', x'' \in \mathcal{X})\, f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x'))}}\right) = f\left(\frac{e^{\sigma^{-1}(\eta(x))}}{e^{\sigma^{-1}(\eta(x''))}}\right) + f\left(\frac{e^{\sigma^{-1}(\eta(x''))}}{e^{\sigma^{-1}(\eta(x'))}}\right).$$

We require this to hold for any D, and thus for any η. Therefore, equivalently, we have

$$(\forall a, b \in \mathbb{R}_+)\, f(a\cdot b) = f(a) + f(b).$$

The function f is continuous by assumed differentiability of Ψ. Thus the only solution to the equation is $f(z) = \frac{1}{a}\cdot\log z$ for some $a \in \mathbb{R}$ (Kannappan, 2009, Corollary 1.43), or equivalently that $\Psi^{-1}(v) = \sigma(a\cdot v) = \frac{1}{1+e^{-av}}$. The case $a = 0$ is ruled out by assumed invertibility of Ψ. ■

While Proposition 6 follows easily from the proper loss machinery, the requirement on the link function is *a priori* non-obvious. We emphasise that the class of proper composite losses satisfying the above condition is "large" in the following sense: one may take *any* strictly proper loss and compose it with any member of the given link family. Some of these compositions result in a non-convex proper composite loss; nonetheless, we are able to easily determine the optimal scorers for all such losses, as below.

Corollary 7 *Given any* $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ *and strictly proper composite loss* ℓ *with inverse link function* $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$ *for some* $a \in \mathbb{R}\setminus\{0\}$*,*

$$\mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},\ell} = \{\Psi\circ\eta + b : b \in \mathbb{R}\} \subseteq \mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},01}.$$

Compared to the case of ℓ^{01}, we see that in this family of surrogate losses, one has only a limited amount of "slack" in the solution, namely, one is allowed to translate the transformed η arbitrarily. Nonetheless, any optimal solution for such a proper composite loss is also optimal for ℓ^{01}, and thus induces the same ranking over instances.

4.4. Bipartite ranking with univariate scorers: non-decomposable case

We now turn to the case where the loss ℓ does *not* have a decomposable Bayes-optimal pair-scorer. As noted earlier, we can no longer resort to using the conditional risk. Fortunately, the simple structure of $\mathcal{S}_{\text{Decomp}}$ means that we can hope to directly compute the risk minimiser via an appropriate derivative. Under some assumptions on the loss, it turns out that the Bayes-optimal scorer is still a strictly monotone transform of η; however, the transform is now *distribution dependent*, rather than simply the fixed link function Ψ.

Proposition 8 *Given any $D_{M,\eta} = D_{P,Q,\pi} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and a margin-based strictly proper composite loss $\ell(y,v) = \phi(yv)$ with $\phi : \mathbb{R} \to \mathbb{R}_+$ convex. If ϕ' is bounded, or $\mathcal{X}$ is finite,*

$$\mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},\ell} = \{s^* : \mathcal{X} \to \mathbb{R} : \eta = f^D_{s^*} \circ s^*\},$$

where

$$(\forall v \in \mathcal{V})\; f^D_{s^*}(v) := \frac{\pi \mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]}{\pi \mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right] - (1-\pi)\mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell'_{1}(v - s^*(\mathsf{X}'))\right]}.$$

To express any optimal scorer s^* in terms of η, as we have done for the previous cases, it remains to check whether or not the above the function $f^D_{s^*}$ defined above is invertible. The following corollary provides sufficient conditions for this to hold.

Corollary 9 *Suppose $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and $\ell(y,v) = \phi(yv)$ is a margin-based strictly proper composite loss, where ϕ is differentiable, strictly convex, and satisfies*

$$(\forall v \in \mathcal{V})\; \phi'(v) = 0 \iff \phi'(-v) \neq 0.$$

Then if ϕ' is bounded or $\mathcal{X}$ is finite,

$$\mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},\ell} = \{s^* : \mathcal{X} \to \mathbb{R} : s^* = (f^D_{s^*})^{-1} \circ \eta\} \subseteq \mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},01},$$

where $f^D_{s^*}$ is defined as in Proposition 8.

As before, any optimal scorer for such a proper composite loss is also optimal for ℓ^{01}, despite the link function $f^D_{s^*}$ depending on the distribution D. Appendix F provides an empirical illustration that this link is indeed invertible under the specified conditions, albeit distribution dependent. The results of this section established convexity is *sufficient* for the optimal scorer to be a strictly monotone transform of η, while the previous section established convexity is *not necessary*, since one can have a non-convex loss resulting from a suitable link $\Psi = \frac{1}{a}\sigma^{-1}$.

As a final remark, we suspect that the requirement of ϕ' bounded when $\mathcal{X}$ is infinite may be dropped, but defer to future work investigation of minimal conditions for the result to hold.

4.5. The role of the link function in bipartite ranking

In class-probability estimation with a proper composite loss, there is a separation of concerns between the underlying proper loss and the link function Ψ, with the latter primarily chosen for computational convenience, and not affecting statistical properties of the proper loss (Reid and Williamson, 2010). For bipartite ranking, however, such a separation of concerns is guaranteed only when one operates with the family of link functions from Proposition 6. For this family, the Bayes-optimal scorer is any translation of $\Psi \circ \eta$, while Corollary 9 indicates that outside this family, the optimal scorer may be a distribution-dependent transformation of η. Thus, changing the link function in bipartite ranking can change the optimal solutions to the risk in a non-trivial way.

4.6. Relation to existing work

This section generalised and unified several earlier results through the theory of proper losses. For ℓ^{01}, our Corollary 5 is well-known in the context of scorers that maximise the AUC, which is one minus the bipartite ℓ^{01} risk. The result is typically established by the Neyman-Pearson lemma (Torgersen, 1991), whereas we simply use a reduction to binary classification over pairs. For exponential loss with a linear hypothesis class, Ertekin and Rudin (2011) studied the (empirical) Bayes-optimal solutions. For a convex margin loss, Uematsu and Lee (2012) and Gao and Zhou (2012) independently studied conditions for the Bayes-optimal scorers to be transformations of η. Our Proposition 6 is a generalisation of Theorem 7 in Uematsu and Lee (2012) and Lemma 3 of Gao and Zhou (2012), where our result holds for non-symmetric and non-convex proper composite losses; Appendix E has an empirical illustration of this. Our Corollary 9 is essentially equivalent to Theorem 3 of Uematsu and Lee (2012) and Theorem 5 of Gao and Zhou (2012), although we explicitly provide the form of the link function relating η and s^*. (We translate these results in terms of proper losses so that the connection is more apparent in Appendix D.)

5. Bayes-optimal scorers for the p-norm push risk

We now consider the *p-norm push risk*, a family of bipartite risks proposed by Rudin (2009). Their aim is to focus attention at the head of the ranked list; confer Clémençon and Vayatis (2007). We characterise the Bayes-optimal solutions of the p-norm push risk to relate it to those of other learning problems. In the sequel, let $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$.

5.1. The (ℓ, g)-push risk

Rudin (2009); Swamidass et al. (2010) studied a family of risks parameterised by a monotone increasing function designed for the ranking the best problem. Generalising these proposals to the case of an arbitrary loss ℓ and pair-scorer s_{Pair}, we obtain the *(ℓ, g)-push bipartite ranking risk*:

$$\mathbb{L}^D_{\text{push},\ell,g}(s_{\text{Pair}}) = \mathbb{E}_{\mathsf{X}'\sim Q}\left[g\left(\mathbb{E}_{\mathsf{X}\sim P}\left[\frac{\ell_1(s(\mathsf{X},\mathsf{X}')) + \ell_{-1}(s(\mathsf{X}',\mathsf{X}))}{2}\right]\right)\right],$$

where $g(\cdot)$ is a nonnegative, monotone increasing function. When $g(x) = x$, we recover the standard bipartite risk (Equation 5). Rudin (2009) provides a detailed study of the choice $g^p(x) = x^p$ for $p \geq 1$, with margin loss ℓ and decomposable pair-scorer, leading to the *p-norm push* risk:

$$\mathbb{L}^D_{\text{push},\ell,g}(\text{Diff}(s)) = \mathbb{E}_{\mathsf{X}'\sim Q}\left[\left(\mathbb{E}_{\mathsf{X}\sim P}\left[\ell_1(s(\mathsf{X}) - s(\mathsf{X}'))\right]\right)^p\right].$$

For large p, and $\ell = \ell^{01}$, the risk penalises high false negative rates, which is an intuitive explanation for why it is suitable for maximising accuracy at the head of the list. To get a different explanation, we look to compute the Bayes-optimal solutions for this risk, and see how they compare to those for the standard bipartite risk. Following the conventions of the prequel, define:

$$\mathbb{S}^{D,*}_{\text{push},\ell,g} = \operatorname*{Argmin}_{s_{\text{Pair}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}^D_{\text{push},\ell,g}(s_{\text{Pair}})$$
$$\mathbb{S}^{D,\text{Univ},*}_{\text{push},\ell,g} = \operatorname*{Argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{L}^D_{\text{push},\ell,g}(\text{Diff}\circ s).$$

5.2. Bayes-optimal pair-scorers

As with the standard bipartite risk, determining the Bayes-optimal scorer for the (ℓ, g) push is challenging due to the implicit restricted function class $\mathcal{S}_{\text{Decomp}}$. In fact, this is difficult even for the pair-scorer case: the (ℓ, g) push risk is not easily expressible in terms of a conditional risk. Thus, we explicitly compute the derivative of the risk, as in the proof of Proposition 8. (Note that requiring differentiability of the loss means that we cannot compute the optimal solution for ℓ^{01}.)

Proposition 10 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$, a differentiable function $g : \mathcal{X} \to \mathbb{R}$, and a strictly proper composite loss ℓ with link function Ψ, if ℓ'_1, ℓ'_{-1} are bounded or $\mathcal{X}$ is finite,*

$$\mathcal{S}^{D,*}_{\text{push},\ell,g} = \left\{ s^*_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} : s^*_{\text{Pair}} = \Psi \circ \sigma \circ \left(\text{Diff}(\sigma^{-1} \circ \eta) - G^D_{s^*_{\text{Pair}}} \right) \right\}, \tag{10}$$

where

$$G^D_{s_{\text{Pair}}}(x, x') = \log \frac{g'\left(F^D_{s_{\text{Pair}}}(x)\right)}{g'\left(F^D_{s_{\text{Pair}}}(x')\right)}$$
$$F^D_{s_{\text{Pair}}}(x) = \mathbb{E}_{\mathsf{X}\sim P}\left[\frac{\ell_1(s_{\text{Pair}}(\mathsf{X}, x)) + \ell_{-1}(s_{\text{Pair}}(x, \mathsf{X}))}{2}\right].$$

When $g : x \mapsto x$, which yields the standard ℓ-bipartite ranking risk, $G^D \equiv 0$ and so $s^*_{\text{Pair}} = \Psi \circ \eta_{\text{Pair}}$ as in Equation 9. For general (ℓ, g), however, it is unclear how to simplify the term G^D any further, and thus we apparently have to settle for the above implicit equation. Surprisingly, when ℓ is the exponential loss and $g^p(x) = x^p$, we have the following simple characterisation.

Proposition 11 *Pick any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$. Let $\ell^{\exp}(y, v) = e^{-yv}$ be the exponential loss and $g^p(x) = x^p$ for any $p > 0$. Then, if $\mathcal{X}$ is finite,*

$$\mathcal{S}^{D,*}_{\text{push},\ell^{\exp},g^p} = \left\{ \frac{1}{p+1} \cdot \sigma^{-1} \circ \eta_{\text{Pair}} \right\} = \left\{ \frac{1}{p+1} \cdot \text{Diff}(\sigma^{-1} \circ \eta) \right\}.$$

As with Proposition 8, we suspect the finiteness assumption on $\mathcal{X}$ can be dropped, although we have been unsuccessful in establishing this. Nonetheless, for this special case, the optimal scorer can be expressed as $\frac{2}{p+1} \cdot \Psi \circ \eta_{\text{Pair}}$, where Ψ is the link function corresponding to exponential loss; comparing this to the optimal pair-scorer for the standard bipartite risk (Equation 9), we see that the effect of the function $g : x \mapsto x^p$ is equivalent to slightly transforming the loss ℓ; we will explore this more in the next section.

5.3. Bayes-optimal univariate scorers

We now turn attention to computing $\mathcal{S}^{D,\text{Univ},*}_{\text{push},\ell,g}$. For ℓ^{01}, we were unsuccessful in computing the optimal pair-scorer; nonetheless, a different technique lets us establish the optimal univariate scorers.

Proposition 12 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and nonnegative, monotone increasing g,*

$$\phi \circ \eta \in \mathcal{S}^{D,\text{Univ},*}_{\text{push},01,g},$$

for any strictly monotone increasing $\phi : [0, 1] \to \mathbb{R}$.

We see that $\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},01} \cap \mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{push},01,g} \neq \emptyset$, and so the (ℓ^{01}, g)-push maintains the optimal solutions for the standard bipartite risk. For a general proper composite loss, it appears difficult to appeal to the optimal pair-scorer implicitly derived in Proposition 10. Nonetheless, for the special case of exponential loss, we have the following.

Proposition 13 *Pick any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$. Let $\ell^{\exp}(y, v) = e^{-yv}$ be the exponential loss and $g^p(x) = x^p$ for any $p > 0$. Then, if $\mathcal{X}$ is finite,*

$$\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{push},\ell^{\exp},g^p} = \left\{ \frac{1}{p+1}(\sigma^{-1} \circ \eta) + b : b \in \mathbb{R} \right\}.$$

As $p \to \infty$, we see that the Bayes-optimal scorer tends to $\delta_0 + \delta_1$, so that it is only the instances at the absolute bottom ($\eta = 0$) and top ($\eta = 1$) of the ranked list that attention is paid to; no effort is spent in distinguishing between all intermediate instances.

5.4. Existing work

Ertekin and Rudin (2011, Theorem 1) implicitly derived the Bayes-optimal scorer for the p-norm push with exponential loss, by showing that for a linear function class, it coincides with that of the p-classification risk. We shall study a more general equivalence in the next section.

6. Four equivalent approaches to bipartite ranking

Consider the following approaches to outputting a pair-scorer, given a strictly proper composite ℓ:

(1) Minimise the ℓ-classification risk $\mathbb{L}^D_\ell$, and construct the difference pair-scorer.

(2) Minimise the ℓ-bipartite ranking risk $\mathbb{L}^D_{\mathrm{Bipart},\ell}$ over *decomposable* pair-scorers.

(3) Minimise the ℓ-bipartite ranking risk $\mathbb{L}^D_{\mathrm{Bipart},\ell}$ over all pair-scorers.

(4) Minimise the p-norm push risk $\mathbb{L}^D_{\mathrm{push},\ell^{\exp},g^p}$ over *decomposable* pair-scorers.

Superficially, these appear very different: method (4) is the only one that departs from the standard conditional risk framework, method (3) is the only one to use a pair-scorer during minimisation, and method (1) is the only one to operate on single instances rather than pairs. It is thus surprising that our results provide conditions under which all methods have the *same* output; it is further surprising that the condition involves the choice of link function in the loss ℓ, which is typically chosen for computational rather than statistical reasons (Reid and Williamson, 2010).

Proposition 14 *Given any $D \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and strictly proper composite loss ℓ with inverse link function $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$ for some $a \in \mathbb{R} \setminus \{0\}$, methods (1), (2) and (3) produce the same pair-scorer; if $\mathcal{X}$ is finite and $p = a - 1$ for $a > 1$, method (4) also produces the same pair-scorer.*

Proof By Equation 7 and Corollary 7, methods (1) and (2) produce the same scorer $\Psi \circ \eta$, up to a translation which is nullified by the Diff operator. By Equation 9, this pair-scorer is equivalent to that produced by method (3). Further, if $p = a - 1$ for $a > 1$, then by Proposition 13, method (4) returns $\Psi \circ \eta$ up to a translation which is nullified by the Diff operator. ■

(1) Diff $\left(\operatorname*{argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[e^{-\mathsf{Y}s(\mathsf{X})}\right]\right)$	(2) Diff $\left(\operatorname*{argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[e^{-(s(\mathsf{X})-s(\mathsf{X}'))}\right]\right)$
(3) $\operatorname*{argmin}_{s_{\text{Pair}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[e^{-s_{\text{Pair}}(\mathsf{X},\mathsf{X}')}\right]$	(4) Diff $\left(\operatorname*{argmin}_{s\colon \mathcal{X}\to\mathbb{R}} \mathbb{E}_{\mathsf{X}'\sim Q}\left[\left(\mathbb{E}_{\mathsf{X}\sim P}\left[e^{-(s(\mathsf{X})-s(\mathsf{X}'))}\right]\right)^p\right]\right)$

Table 1: Four methods for obtaining a pair-scorer in a bipartite ranking problem, using exponential loss. Our results show that the all methods produce the same output.

Name	$\ell_1(v)$	$\ell_{-1}(v)$	$\Psi(\eta)$
Generalised Logistic	$\frac{1}{p}\log(1+e^{-vp})$	$\frac{1}{p}\log(1+e^{vp})$	$\frac{1}{p}\log\frac{\eta}{1-\eta}$
Generalised Exponential	$\frac{1}{p}e^{-vp}$	$\frac{1}{p}e^{vp}$	$\frac{1}{2p}\log\frac{\eta}{1-\eta}$
p-classification	e^{-v}	$\frac{1}{p}e^{vp}$	$\frac{1}{p+1}\log\frac{\eta}{1-\eta}$

Table 2: Proper composite losses with same Bayes-optimal scorers as the p-norm push.

In hindsight, these equivalences are not surprising by virtue of the Bayes-optimal scorer for each type of risk depending on the observation-conditional density η. They are not however *a priori* obvious, given how ostensibly different the risks appear. To illustrate these superficial differences, Table 1 provides a concrete example of the four methods when $\ell = \ell^{\exp}$ is the exponential loss, whose link $\Psi = \frac{1}{2}\sigma^{-1}$ satisfies the required condition.

6.1. Implications of equivalences

Our definition of "equivalent" is that two methods have the same optimal scorer. This does not imply that the methods are interchangeable in practice. A statistical caveat to these equivalences is that they ignore the issue of finite samples and a restricted function class. When one or both of these issues is relevant, it may be that one of these methods is more preferable. A computational caveat is that methods (2) – (4) rely on minimisation over pairs of examples. On a finite training set, this requires roughly quadratic complexity, compared to the linear complexity of method (1). These practical issues deserve investigation, but are beyond the scope of this paper.

This caveat in mind, we believe the results at least illuminate similarities between seemingly disparate approaches. For the problem of minimising the ℓ-bipartite risk for an appropriate surrogate ℓ, the above provides evidence that minimising the ℓ-classification risk is a suitable proxy. That is, performing class-probability estimation is a suitable proxy for ranking; this can be formalised with surrogate regret bounds (Agarwal, 2013; Narasimhan and Agarwal, 2013).

Similarly, for the problem of minimising the p-norm push objective, we have evidence that minimising the ℓ-classification or bipartite risk is a suitable proxy. Table 2 summarises several losses with the same Bayes-optimal scorer as the p-norm push, including the p-classification loss discussed in Ertekin and Rudin (2011). Appendix G reports some experiments that illustrate that such losses do indeed give comparable performance to the p-norm push.

6.2. Relation to existing work

Subsets of the above equivalences have been observed earlier under special cases. For the specific case of exponential loss and a linear hypothesis class, the equivalence between methods (1) and (2) was made by Ertekin and Rudin (2011, Theorem 3), Gao and Zhou (2012, Lemma 4), while the equivalence between method (1) and (4) was shown in Ertekin and Rudin (2011, Theorem 1); here, method (1) represents AdaBoost, and method (2) RankBoost. For the special case of convex margin losses, the equivalence between methods (2) and (3) was shown by Uematsu and Lee (2012).

7. Conclusion

We derived the Bayes-optimal scorers for bipartite ranking under the proper composite family of losses, including as special cases the 0-1, logistic and exponential losses. The theory of proper composite losses illuminated certain special cases where this optimal scorer has an especially simple form, related to that of the optimal scorer for the class-probability estimation risk. We further studied Bayes-optimal scorers for a generalised family of bipartite risks, namely the p-norm push risk (Rudin, 2009). Consequently, we established equivalences between four seemingly disparate approaches to bipartite ranking. We believe our results illustrate the value of the proper loss machinery in studying ranking problems. In future work we aim to use this machinery to yield further insight into bipartite ranking, for example by relating the optimal bipartite risk to an f-divergence (Reid and Williamson, 2011), and studying integral representations analogous to those for proper composite risks (Reid and Williamson, 2010).

References

Shivani Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SIAM International Conference on Data Mining (SDM)*, pages 839–850, 2011.

Shivani Agarwal. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *Conference on Learning Theory (COLT)*, pages 338–353, 2013.

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, December 2005.

Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B . Sorkin. Robust reductions from ranking to classification. *Machine Learning*, 72 (1-2):139–153, 2008. doi: 10.1007/s10994-008-5058-6.

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

Stephen P. Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In *Neural Information Processing Systems (NIPS)*, pages 962–970, 2012.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. www-stat.wharton.upenn.edu/~buja, 2005. Unpublished manuscript.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML)*, pages 89–96, New York, NY, USA, 2005. ACM. doi: 10.1145/1102351.1102363.

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

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, April 2008. doi: 10.1214/009052607000000910.

William W. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10(1):243–270, May 1999.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Şeyda Ertekin and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, Oct 2011.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley Interscience, New York, 1999.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC optimization. *CoRR*, abs/1208.0645, 2012.

Izrail M. Gelfand and Sergei V. Fomin. *Calculus of Variations*. Dover, 2000.

Mariano Giaquinta and Stefan Hildebrandt. *Calculus of Variations I: The Lagrangian formalism*. Springer-Verlag, Berlin, 2nd edition, 2004.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132, Cambridge, MA, 2000. MIT Press.

Palaniappan Kannappan. *Functional equations and inequalities with applications.* New York, NY: Springer, 2009. doi: 10.1007/978-0-387-89492-8.

Donald E. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, May 1992. doi: 10.2307/2325085.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Hüllermeier. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning (ICML)*, pages 1113–1120, 2011.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Neural Information Processing Systems (NIPS)*, pages 2913–2921, 2013.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning (ICML)*, pages 897–904, New York, NY, USA, 2009. ACM. doi: 10.1145/1553374.1553489.

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

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

Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, December 2009.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007. doi: 10.1007/s00365-006-0662-3.

S. Joshua Swamidass, Chloé-Agathe Azencott, Kenny Daily, and Pierre Baldi. A CROC stronger than ROC. *Bioinformatics*, 26(10):1348–1356, May 2010. doi: 10.1093/bioinformatics/btq140.

Erik N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

John L. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*. Undergraduate Texts in Mathematics. Springer, 1996.

Kazuki Uematsu and Yoonkyung Lee. On theoretically optimal ranking functions in bipartite ranking. http://www.stat.osu.edu/~yklee/mss/bipartrank.rev.pdf, 2012. Unpublished manuscript.

Elodie Vernet, Mark D. Reid, and Robert C. Williamson. Composite multiclass losses. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 24*, pages 1224–1232, 2011.

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

Appendix A. Proofs

Proof [Proposition 1] ($\subseteq$). Pick any $s^*_{\mathrm{Pair}} \in \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}}$. Then $s^*_{\mathrm{Pair}} = \mathrm{Diff}(s^*)$ for some $s^* \colon \mathcal{X} \to \mathbb{R}$. By optimality,

$$(\forall s \colon \mathcal{X} \to \mathbb{R})\, \mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s^*) = \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s^*_{\mathrm{Pair}}) \leq \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)) = \mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s).$$

Thus $s^* \in \mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}$, and so $s^*_{\mathrm{Pair}} \in \mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell})$.

($\supseteq$). Pick any $s^* \in \mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}$, and let $s^*_{\mathrm{Pair}} = \mathrm{Diff}(s^*)$. Then

$$s^*_{\mathrm{Pair}} \in \operatorname*{Argmin}_{s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}}).$$

This is a constrained optimisation problem. When $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset$, there is at least one solution to the unconstrained optimisation that lies in $\mathcal{S}_{\mathrm{Decomp}}$, call it s_{Pair}. Clearly s_{Pair} is a feasible solution for the constrained problem above. Thus, it must have an identical risk to s^*_{Pair}. But then s^*_{Pair} is a solution to the unconstrained problem as well, and so $s^*_{\mathrm{Pair}} \in \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}}$. ■

Proof [Corollary 2] ($\implies$) follows by Proposition 1, and ($\impliedby$) follows by definition. ■

Proof [Lemma 3] By Equation 5,

$$\begin{aligned}
\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}}) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\frac{\ell_1(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}')) + \ell_{-1}(s_{\mathrm{Pair}}(\mathsf{X}',\mathsf{X}))}{2}\right] \\
&= \frac{1}{2}\cdot \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_1(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right] + \frac{1}{2}\cdot \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_{-1}(s_{\mathrm{Pair}}(\mathsf{X}',\mathsf{X}))\right] \\
&= \frac{1}{2}\cdot \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell_1(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right] + \frac{1}{2}\cdot \mathbb{E}_{\mathsf{X}\sim Q,\mathsf{X}'\sim P}\left[\ell_{-1}(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right] \\
&= \frac{1}{2}\cdot \mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim (P\times Q)}\left[\ell_1(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right] + \frac{1}{2}\cdot \mathbb{E}_{(\mathsf{X},\mathsf{X}')\sim (Q\times P)}\left[\ell_{-1}(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}'))\right].
\end{aligned}$$

By definition of $\mathrm{Bipart}(D)$, this is exactly $\mathbb{L}^{\mathrm{Bipart}(D)}_{\ell}(s_{\mathrm{Pair}})$. As noted in the body, this result is well-known for the case of ℓ^{01} (Balcan et al., 2008; Kotlowski et al., 2011; Agarwal, 2013). ■

Proof [Proposition 4] Let $\mathcal{A} = \mathcal{S}^{D,*}_{\mathrm{Bipart},01} \cap \mathcal{S}_{\mathrm{Decomp}}$. By definition,

$$\begin{aligned}
\mathcal{A} &= \left\{ s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}} : \begin{array}{rcl} \eta(x) \neq \eta(x') & \implies & \mathrm{sign}(s_{\mathrm{Pair}}(x,x')) = \mathrm{sign}(\eta(x) - \eta(x')) \\ \eta(x) = \eta(x') & \implies & s_{\mathrm{Pair}}(x,x') \in \mathbb{R}. \end{array} \right\} \\
&= \mathrm{Diff}\left(\left\{ s \colon \mathcal{X} \to \mathbb{R} : \begin{array}{rcl} \eta(x) \neq \eta(x') & \implies & \mathrm{sign}(s(x) - s(x')) = \mathrm{sign}(\eta(x) - \eta(x')) \\ \eta(x) = \eta(x') & \implies & s(x) - s(x') \in \mathbb{R}. \end{array} \right\} \right) \\
&= \mathrm{Diff}\left(\{ s \colon \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s \} \right) \text{ by Lemma 15 .}
\end{aligned}$$

Since $\mathcal{A}$ is nonempty, $\mathcal{A} = \mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},01})$ by Proposition 1. Thus the result follows. ■

Proof [**Corollary 7**] By Proposition 6 and Corollary 2,

$$\mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}) = \mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}.$$

Further, by Equation 9,

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} = \mathrm{Diff}\left(\frac{1}{a}\cdot\sigma^{-1}\circ\eta\right) = \mathrm{Diff}(\Psi\circ\eta).$$

The result follows because

$$\mathrm{Diff}(f) = \mathrm{Diff}(g) \iff (\exists b\in\mathbb{R})\, f = g + b.$$

■

Proof [**Proposition 8**] For fixed D, let $\mathcal{L}(D)$ denote the space of all Lebesgue-measurable scorers $s\colon \mathcal{X}\to\mathbb{R}$, with addition and scalar multiplication defined pointwise, such that

$$\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\phi(s(\mathsf{X}) - s(\mathsf{X}'))\right] < \infty.$$

Then $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell} : \mathcal{L}(D)\to\mathbb{R}$ is a functional whose minimisers may be determined by considering an appropriate notion of functional derivative. We shall employ the Gâteaux variation. This coincides with the standard directional derivative when $\mathcal{X}$ is finite, where the minimisation is effectively over finite dimensional vectors.

Pick any $s, t\in\mathcal{L}(D)$. For any $\epsilon > 0$, define

$$\begin{aligned} F_{s,t}(\epsilon) &= \mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s+\epsilon t) \\ &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\phi(s(\mathsf{X}) - s(\mathsf{X}') + \epsilon(t(\mathsf{X}) - t(\mathsf{X}')))\right]. \end{aligned}$$

The Gâteaux variation of $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}$ at s in the direction of t is (Troutman, 1996, pg. 45), (Giaquinta and Hildebrandt, 2004, pg. 10)

$$\begin{aligned} \delta\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s;t) &= \lim_{\epsilon\to 0}\frac{\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s+\epsilon t) - \mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s)}{\epsilon} \\ &= F'_{s,t}(0), \end{aligned}$$

assuming the latter exists. To show that $F'_{s,t}(0)$ exists, we will justify interchange of the derivative and expectation. For any $\epsilon\in(0,1]$ and $x, x'\in\mathcal{X}$, by convexity and nonnegativity of ϕ,

$$\begin{aligned} \left|\frac{\phi((\mathrm{Diff}(s+\epsilon t))(x,x')) - \phi((\mathrm{Diff}(s))(x,x'))}{\epsilon}\right| &\le \left|\phi((\mathrm{Diff}(s+t))(x,x')) - \phi((\mathrm{Diff}(s))(x,x'))\right| \\ &\le \phi((\mathrm{Diff}(s+t))(x,x')) + \phi((\mathrm{Diff}(s))(x,x')), \end{aligned} \tag{11}$$

where Equation 11 is because $\phi(a+\epsilon b)\le\epsilon\phi(a+b) + (1-\epsilon)\phi(a)$ for any $a, b\in\mathbb{R}$.

By assumption, $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s+t)$ and $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s)$ are both finite. Further,

$$\lim_{\epsilon\to 0}\frac{\phi(s(x) - s(x') + \epsilon(t(x) - t(x'))) - \phi(s(x) - s(x'))}{\epsilon} = (t(x) - t(x'))\cdot\phi'(s(x) - s(x')).$$

Thus, by the dominated convergence theorem (Folland, 1999, pg. 56), we have

$$\begin{aligned} F'_{s,t}(0) &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[(t(\mathsf{X}) - t(\mathsf{X}')) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] && (12)\\ &= \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] - \mathbb{E}_{\mathsf{X}\sim Q,\mathsf{X}'\sim P}\left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}') - s(\mathsf{X}))\right] && (13)\\ &= \int_{\mathcal{X}} t(x) \cdot r(x)\, dx, \end{aligned}$$

where

$$(\forall x \in \mathcal{X})\, r(x) := P(x) \cdot \mathbb{E}_{\mathsf{X}'\sim Q}\left[\phi'(s(x) - s(\mathsf{X}'))\right] - Q(x) \cdot \mathbb{E}_{\mathsf{X}\sim P}\left[\phi'(s(\mathsf{X}) - s(x))\right].$$

In the splitting the expectation in Equations 12 and 13, we relied on the fact that the individual terms are finite:

$$\begin{aligned} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\left|t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right|\right] &< +\infty\\ \mathbb{E}_{\mathsf{X}\sim Q,\mathsf{X}'\sim P}\left[\left|t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}') - s(\mathsf{X}))\right|\right] &< +\infty. \end{aligned}$$

When $\mathcal{X}$ is finite, the expectations are summations, and this is immediate by finiteness of each of the terms in the sum. When $\mathcal{X}$ is infinite, we assumed that ϕ' is bounded. Consequently,

$$\mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\left|t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right|\right] < \sup_{z\in\mathbb{R}} |\phi'(z)| \cdot \mathbb{E}_{\mathsf{X}\sim P}\left[|t(\mathsf{X})|\right],$$

and similarly for the second term. Therefore we simply need to show that

$$\begin{aligned} \mathbb{E}_{\mathsf{X}\sim P}\left[|t(\mathsf{X})|\right] &< +\infty\\ \mathbb{E}_{\mathsf{X}'\sim Q}\left[\left|t(\mathsf{X}')\right|\right] &< +\infty. \end{aligned}$$

To show this, we lower bound the nonnegative convex function ϕ with its Taylor expansion at 0:

$$\begin{aligned} (\forall x, x' \in \mathcal{X})\, |t(x) - t(x')| &\leq \frac{|\phi(t(x) - t(x')) - \phi(0)|}{|\phi'(0)|}\\ &\leq \frac{1}{|\phi'(0)|} \cdot (\phi(t(x) - t(x')) + \phi(0)) \text{ by the triangle inequality.} \end{aligned}$$

We can then bound the expectation of $\mathrm{Diff}(t)$:

$$\begin{aligned} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\left|t(\mathsf{X}) - t(\mathsf{X}')\right|\right] &\leq \frac{1}{|\phi'(0)|} \cdot \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\phi(t(\mathsf{X}) - t(\mathsf{X}')) + \phi(0)\right]\\ &< +\infty, \end{aligned}$$

where we use the fact that $t \in \mathcal{L}(D)$, and $\phi'(0) \neq 0$ since ℓ is strictly proper composite (Vernet et al., 2011, Proposition 14). Unrolling the expectation,

$$\begin{aligned} \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\left|t(\mathsf{X}) - t(\mathsf{X}')\right|\right] &= \int_{\mathcal{X}\times\mathcal{X}} p(x)q(x')|t(x) - t(x')|\, d((x, x'))\\ &= \int_{\mathcal{X}} p(x) \cdot \left(\int_{\mathcal{X}} q(x')|t(x) - t(x')|\, dx'\right) dx \text{ by Tonelli's theorem} \end{aligned}$$

$$
\begin{aligned}
&\geq \int_{\mathcal{X}} p(x) \cdot \left(\int_{\mathcal{X}} q(x')(|t(x')| - |t(x)|)\, dx' \right) dx \text{ by the reverse triangle inequality} \\
&= \int_{\mathcal{X}} p(x) \cdot \left(\int_{\mathcal{X}} q(x')|t(x')|\, dx' - |t(x)| \right) dx \\
&=: \int_{\mathcal{X}} p(x) \cdot u(x)\, dx.
\end{aligned}
$$

Since the left hand side is finite, the function u must be finite almost everywhere. But $u(x) = \mathbb{E}_{\mathsf{X}' \sim Q}\left[|t(\mathsf{X}')|\right] - |t(x)|$, where the first term does not depend on x. Thus, since $t(x)$ is finite for every $x \in \mathcal{X}$, we must have $\mathbb{E}_{\mathsf{X}' \sim Q}\left[|t(\mathsf{X}')|\right] < +\infty$. A similar argument, where the order of the double integration is reversed, shows that $\mathbb{E}_{\mathsf{X} \sim P}\left[|t(\mathsf{X})|\right] < +\infty$.

Now suppose $s^* \colon \mathcal{X} \to \mathbb{R}$ minimises the functional $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}$. By convexity of $\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}$, it is necessary and sufficient that the Gâteaux variation is zero for every $t \in \mathcal{L}(D)$ (Gelfand and Fomin, 2000, Theorem 2), (Troutman, 1996, Proposition 3.3). That is,

$$
(\forall t \in \mathcal{L}(D)) \int_{\mathcal{X}} t(x) \cdot r(x)\, dx = 0. \tag{14}
$$

It is then necessary and sufficient that r is zero (almost) everywhere. Sufficiency is immediate; to see necessity, let $\mathcal{A} \subseteq \mathcal{X}$ be the set of points where r is nonzero. If $\mathcal{A} = \emptyset$ we are done, so suppose that $\mathcal{A} \neq \emptyset$. For any $\mathcal{A}' \subseteq \mathcal{A}$, let $t_{\mathcal{A}'} : x \mapsto [\![x \in \mathcal{A}']\!]$ be the indicator function on the set. Then $t_{\mathcal{A}'} \in \mathcal{L}(D)$ because

$$
\begin{aligned}
\mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[\phi((\mathrm{Diff}(t))(\mathsf{X}, \mathsf{X}'))\right] &= \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[[\![\mathsf{X} \in \mathcal{A}', \mathsf{X}' \notin \mathcal{A}']\!]\phi(1) + [\![\mathsf{X} \notin \mathcal{A}', \mathsf{X}' \in \mathcal{A}']\!]\phi(-1)\right] \\
&= P(\mathcal{A}')Q(\mathcal{X} \setminus \mathcal{A}')\phi(1) + P(\mathcal{X} \setminus \mathcal{A}')Q(\mathcal{A}')\phi(-1) \\
&< \infty,
\end{aligned}
$$

where the last line is since $\phi(z) < \infty$ for every $z \in \mathbb{R}$. By assumption, Equation 14 holds for every $t_{\mathcal{A}'}$. But that implies

$$
(\forall \mathcal{A}' \subseteq \mathcal{A}) \int_{\mathcal{A}'} r(x)\, dx = 0,
$$

which in turn implies that $r(x) \equiv 0$ on $\mathcal{A}$, which is a contradiction.

Thus, for s^* to minimise the risk, it is necessary and sufficient that for (almost) every $x_0 \in \mathcal{X}$,

$$
P(x_0) \cdot \mathbb{E}_{\mathsf{X}' \sim Q}\left[\phi'(s^*(x_0) - s^*(\mathsf{X}'))\right] = Q(x_0) \cdot \mathbb{E}_{\mathsf{X} \sim P}\left[\phi'(s^*(\mathsf{X}) - s^*(x_0))\right],
$$

which means for (almost) every $x_0 \in \mathcal{X}$,

$$
\begin{aligned}
\frac{\eta(x_0)}{1 - \eta(x_0)} \cdot \frac{1 - \pi}{\pi} &= \frac{P(x_0)}{Q(x_0)} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\phi'(s^*(\mathsf{X}) - s^*(x_0))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[\phi'(s^*(x_0) - s^*(\mathsf{X}'))\right]} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\ell_1'(s^*(\mathsf{X}) - s^*(x_0)) - \ell_{-1}'(s^*(x_0) - s^*(\mathsf{X}))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[-\ell_1'(s^*(x_0) - s^*(\mathsf{X}')) + \ell_{-1}'(s^*(\mathsf{X}) - s^*(x_0))\right]} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\ell_{-1}'(s^*(x_0) - s^*(\mathsf{X})) - \ell_1'(s^*(\mathsf{X}) - s^*(x_0))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[\ell_1'(s^*(x_0) - s^*(\mathsf{X}')) - \ell_{-1}'(s^*(\mathsf{X}) - s^*(x_0))\right]}
\end{aligned}
$$

$$= \frac{\mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(s^*(x_0) - s^*(\mathsf{X}))\right]}{\mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell'_1(s^*(x_0) - s^*(\mathsf{X}'))\right]} \text{ since } \ell \text{ is symmetric,}$$

which means

$$\eta = f^D_{s^*} \circ s^*,$$

where $f^D_{s^*}$ is given by

$$(f^D_{s^*})(v) := \frac{\pi\mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]}{\pi\mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right] - (1-\pi)\mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell'_1(v - s^*(\mathsf{X}'))\right]}.$$

■

Proof [**Corollary 9**] We show that $f^D_{s^*}$ strictly monotone, by establishing the strict monotonicity of

$$g(v) := \frac{\mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell'_1(v - s^*(\mathsf{X}'))\right]}{\mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]}.$$

The derivative of this function is

$$g'(v) = \frac{\mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\ell''_1(v - s^*(\mathsf{X}')) - \ell''_{-1}(v - s^*(\mathsf{X}))\ell'_1(v - s^*(\mathsf{X}'))\right]}{\left(\mathbb{E}_{\mathsf{X}\sim P}\left[\ell'_{-1}(v - s^*(\mathsf{X}))\right]\right)^2}.$$

By strict convexity of ℓ, the terms $\ell''_1(v - s^*(\mathsf{X}'))$ and $\ell''_{-1}(v - s^*(\mathsf{X}))$ are positive. Further, by (Vernet et al., 2011, Proposition 15), ℓ_1 and ℓ_{-1} are respectively increasing and decreasing, or vice-versa. By assumption their derivatives cannot simultaneously be zero. Therefore the expectand is always positive or negative for every v, and hence $g'(v)$ is always strictly positive or negative. Thus g is strictly monotone, which means $f^D_{s^*}$ is as well. Therefore, $s^* = (f^D_{s^*})^{-1} \circ \eta$. ■

Proof [**Proposition 10**] First, in the notation above,

$$\mathbb{L}^D_{\text{push},\ell,g}(s_{\text{Pair}}) = \mathbb{E}_{\mathsf{X}'\sim Q}\left[g\left(F^D_{s_{\text{Pair}}}(\mathsf{X}')\right)\right].$$

For fixed D, let $\mathcal{L}(D)$ denote the space of all Lebesgue-measurable pair-scorers $s_{\text{Pair}} \colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, with addition and scalar multiplication defined pointwise, such that $\mathbb{L}^D_{\text{push},\ell,g}(s_{\text{Pair}}) < \infty$. As before, we consider the Gâteaux variation of the functional. Pick any $s_{\text{Pair}}, t_{\text{Pair}} \in \mathcal{L}(D)$. For any $\epsilon > 0$, define

$$\begin{aligned} F_{s,t}(\epsilon) &= \mathbb{L}^D_{\text{push},\ell,g}(s_{\text{Pair}} + \epsilon t_{\text{Pair}}) \\ &= \mathbb{E}_{\mathsf{X}'\sim Q}\left[g\left(F^D_{s_{\text{Pair}}+\epsilon t_{\text{Pair}}}(\mathsf{X}')\right)\right]. \end{aligned}$$

Now consider

$$F'_{s,t}(0) = \mathbb{E}_{\mathsf{X}'\sim Q}\left[g'\left(F^D_{s_{\text{Pair}}}(\mathsf{X}')\right)\cdot\mathbb{E}_{\mathsf{X}\sim P}\left[t_{\text{Pair}}(\mathsf{X},\mathsf{X}')\cdot\frac{\ell'_1(s_{\text{Pair}}(\mathsf{X},\mathsf{X}'))}{2} + t_{\text{Pair}}(\mathsf{X}',\mathsf{X})\cdot\frac{\ell'_{-1}(s_{\text{Pair}}(\mathsf{X}',\mathsf{X}))}{2}\right]\right]$$

$$= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{X}} t_{\mathrm{Pair}}(x,x') \cdot \big(P(x)Q(x') \cdot g'(F^D_{s_{\mathrm{Pair}}}(x')) \cdot \ell_1'(s_{\mathrm{Pair}}(x,x')) + P(x')Q(x) \cdot g'(F^D_{s_{\mathrm{Pair}}}(x)) \cdot \ell_1'(s_{\mathrm{Pair}}(x',x))\big)\, dx\, dx',$$

where as in the proof of Proposition 8, the interchange of derivative and expectation is justified when $\mathcal{X}$ is finite, or when the derivatives ℓ_1', ℓ_{-1}' are bounded.

For the optimal pair-scorer s^*_{Pair}, the derivative must be zero for every t_{Pair}. A sufficient condition for this to hold is that the second term in the integrand is zero for (almost) every $x, x' \in \mathcal{X}$.

Now, since ℓ is strictly proper composite, for any $\eta \in [0,1]$, the solution to

$$\eta\ell_1'(s) + (1-\eta)\ell_{-1}'(s) = 0$$

is $s = \Psi(\eta)$, by virtue of the above being the derivative of the conditional risk. Thus, the solution to

$$\frac{a}{a+b}\ell_1'(s) + \frac{b}{a+b}\ell_{-1}'(s) = 0$$

for $a, b > 0$ is $s = \Psi(a/(a+b)) = \Psi(\sigma(\log(a/b)))$. Letting $a := g'\left(F^D_{s_{\mathrm{Pair}}}(x')\right) \cdot P(x)Q(x')$ and $b := g'\left(F^D_{s_{\mathrm{Pair}}}(x)\right) \cdot Q(x)P(x')$, the optimal pair-scorer is, for every $x, x' \in \mathcal{X}$,

$$\begin{aligned} s^*_{\mathrm{Pair}}(x,x') &= \Psi \circ \sigma \circ \log \frac{P(x)Q(x')g'(F^D_{s_{\mathrm{Pair}}}(x'))}{P(x')Q(x)g'(F^D_{s_{\mathrm{Pair}}}(x))} \\ &= \Psi \circ \sigma \circ \left(\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x')) - G^D_{s^*_{\mathrm{Pair}}}\right) \text{ since } \frac{P(x)}{Q(x)} = \frac{\eta(x)}{1-\eta(x')} \cdot \frac{1-\pi}{\pi}, \end{aligned}$$

i.e.

$$s^*_{\mathrm{Pair}} = \Psi \circ \sigma \circ \left(\mathrm{Diff}(\sigma^{-1} \circ \eta) - G^D_{s^*_{\mathrm{Pair}}}\right).$$

The result follows by dividing through by the numerator. ■

Proof [Proposition 11] We establish this by verifying that $s_{\mathrm{Pair}} = \frac{1}{p+1}\mathrm{Diff}(\sigma^{-1} \circ \eta)$ satisfies the implicit equation in Proposition 10. We begin with the term $A^D(x)$ as defined in Proposition 10. Plugging in $g^p(x) = x^p$ and

$$s_{\mathrm{Pair}} = \frac{1}{p+1} \cdot \sigma^{-1} \circ \eta_{\mathrm{Pair}} = \frac{1}{p+1} \cdot \mathrm{Diff}(\sigma^{-1} \circ \eta),$$

we get

$$\begin{aligned} (\forall x \in \mathcal{X})\, A^D_{s_{\mathrm{Pair}}}(x) &= \mathbb{E}_{\mathsf{X}\sim P}\left[\frac{\ell_1(s_{\mathrm{Pair}}(\mathsf{X},x)) + \ell_{-1}(s_{\mathrm{Pair}}(x,\mathsf{X}))}{2}\right] \\ &= \mathbb{E}_{\mathsf{X}\sim P}\left[\frac{e^{-s_{\mathrm{Pair}}(\mathsf{X},x)} + e^{s_{\mathrm{Pair}}(x,\mathsf{X})}}{2}\right] \\ &= \frac{1}{2}\mathbb{E}_{\mathsf{X}\sim P}\left[\left(\frac{\eta_{\mathrm{Pair}}(\mathsf{X},x)}{1-\eta_{\mathrm{Pair}}(\mathsf{X},x)}\right)^{-1/(p+1)} + \left(\frac{\eta_{\mathrm{Pair}}(x,\mathsf{X})}{1-\eta_{\mathrm{Pair}}(x,\mathsf{X})}\right)^{1/(p+1)}\right] \\ &= \mathbb{E}_{\mathsf{X}\sim P}\left[\exp((\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(\mathsf{X})))/(p+1))\right] \end{aligned}$$

$$= \exp(\sigma^{-1}(\eta(x))/(p+1)) \cdot \mathbb{E}_{\mathsf{X}\sim P}\left[\exp(-\sigma^{-1}(\eta(\mathsf{X})))/(p+1))\right],$$

where crucially the dependence on η is separated from the dependence on the rest of the distribution.

Thus, for $g^p(x) = x^p$,

$$(\forall x, x' \in \mathcal{X})\ \frac{g'\left(A^D_{s_{\mathrm{Pair}}}(x)\right)}{g'\left(A^D_{s_{\mathrm{Pair}}}(x')\right)} = \frac{\exp(\sigma^{-1}(\eta(x)) \cdot (p-1)/(p+1))}{\exp(\sigma^{-1}(\eta(x')) \cdot (p-1)/(p+1))}$$

with the result now a simple function of η, and

$$(\forall x, x' \in \mathcal{X})\ \log \frac{g'\left(A^D_{s_{\mathrm{Pair}}}(x)\right)}{g'\left(A^D_{s_{\mathrm{Pair}}}(x')\right)} = \frac{(p-1)}{(p+1)} \cdot (\sigma^{-1}(\eta(x)) - \sigma^{-1}(\eta(x'))).$$

Now recall that the link function for exponential loss is $\Psi = \frac{1}{2}\sigma^{-1}$. Plugging the above into the right hand side of Equation 10, we get

$$\begin{aligned}\Psi \circ \sigma \circ (\mathrm{Diff}(\sigma^{-1} \circ \eta) - B^D_{s^*_{\mathrm{Pair}}}) &= \left(\frac{1}{2} - \frac{p-1}{2(p+1)}\right)\mathrm{Diff}(\sigma^{-1} \circ \eta) \\ &= \frac{1}{p+1}\mathrm{Diff}(\sigma^{-1} \circ \eta) \\ &= s_{\mathrm{Pair}}.\end{aligned}$$

Therefore $s_{\mathrm{Pair}} = \frac{1}{p+1}\mathrm{Diff}(\sigma^{-1} \circ \eta)$ satisfies the implicit equation of Proposition 10, and hence must be an optimal pair-scorer for exponential loss.

To see why exponential loss simplifies matters, we note that the risk can be decomposed into

$$\mathbb{L}^D_{\mathrm{push},\exp,g}(\mathrm{Diff}(s)) = \left(\mathbb{E}_{\mathsf{X}\sim P}\left[e^{-s(\mathsf{X})}\right]\right)^p \cdot \left(\mathbb{E}_{\mathsf{X}'\sim Q}\left[e^{ps(\mathsf{X}')}\right]\right).$$

This decomposition into the product of two expectations simplifies the derivatives considerably. In fact, an alternate strategy to determine the minimisers of the risk is to consider the scorer that maximises $\mathbb{E}_{\mathsf{X}'\sim Q}\left[e^{ps(\mathsf{X}')}\right]$ subject to $\left(\mathbb{E}_{\mathsf{X}\sim P}\left[e^{-s(\mathsf{X})}\right]\right)^p$ being a constant; this is reminiscent of the Neyman-Pearson approach to arguing for the optimal scorers for the AUC, which incidentally is the strategy we shall employ for proving Proposition 12. ■

Proof [Proposition 12] For any $s\colon \mathcal{X} \to \mathbb{R}$ and $t \in \mathbb{R}$, let

$$\begin{aligned}\mathrm{FNR}^D_s(t) &= \mathbb{E}_{\mathsf{X}\sim P}\left[\ell^{01}(1, s(\mathsf{X}) - t)\right] = \Pr_{\mathsf{X}\sim P}\left[s(\mathsf{X}) < t\right] + \frac{1}{2}\Pr_{\mathsf{X}\sim P}\left[s(\mathsf{X}) = t\right] \\ \mathrm{FPR}^D_s(t) &= \mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell^{01}(-1, s(\mathsf{X}') - t)\right] = \Pr_{\mathsf{X}'\sim Q}\left[s(\mathsf{X}') > t\right] + \frac{1}{2}\Pr_{\mathsf{X}'\sim Q}\left[s(\mathsf{X}') = t\right]\end{aligned}$$

denote the false-negative and false-positive rates respectively of s using a threshold t. Observe that we can write:

$$\begin{aligned}\mathbb{L}^D_{\mathrm{push},01,g}(\mathrm{Diff}(s)) &= \mathbb{E}_{\mathsf{X}'\sim Q}\left[g\left(\mathbb{E}_{\mathsf{X}\sim P}\left[\ell^{01}(1, s(\mathsf{X}) - s(\mathsf{X}'))\right]\right)\right] \\ &= \mathbb{E}_{\mathsf{X}'\sim Q}\left[g\left(\mathrm{FNR}^D_s(s(\mathsf{X}'))\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathsf{X}' \sim Q} \left[\int_{-\infty}^{\infty} \delta_{s(\mathsf{X}')}(t) \cdot g\left(\mathrm{FNR}_s^D(t)\right) \, dt \right] \\
&= \int_{-\infty}^{\infty} \mathbb{E}_{\mathsf{X}' \sim Q} \left[\delta_{s(\mathsf{X}')}(t) \cdot g\left(\mathrm{FNR}_s^D(t)\right) \right] \, dt \\
&= \int_{-\infty}^{\infty} \Pr_{\mathsf{X}' \sim Q} \left[s(\mathsf{X}') = t \right] \cdot g\left(\mathrm{FNR}_s^D(t)\right) \, dt \\
&= \int_{-\infty}^{\infty} -(\mathrm{FPR}_s^D)'(t) \cdot g\left(\mathrm{FNR}_s^D(t)\right) \, dt \\
&= \int_0^1 g\left(\mathrm{FNR}_s^D((\mathrm{FPR}_s^D)^{-1}(\alpha))\right) \, d\alpha,
\end{aligned}
$$

where δ_{x_0} denotes the Dirac delta function centred at x_0, i.e. the generalised function satisfying $\int_{\mathbb{R}} f(x)\delta_{x_0}(x)\,dx = f(x_0)$ for any f continuous at x_0, and the interchange of expectation and integration is valid by nonnegativity of the integrand. That is, the (ℓ, g)-push risk can be interpreted as the area under the parametric curve

$$\{(\mathrm{FPR}_s^D(t), g(\mathrm{FNR}_s^D(t))) : t \in \mathbb{R}\}.$$

Following the Neyman-Pearson approach to ROC maximisation (Clémençon et al., 2008), we equivalently wish to solve for each $\alpha \in [0, 1]$

$$\operatorname*{Argmin}_{s\colon \mathcal{X} \to \mathbb{R}, t \in \mathbb{R}} g(\mathrm{FNR}_s^D(t)) : \mathrm{FPR}_s^D(t) = \alpha.$$

Since g is a monotone increasing function, it preserves the optimal solution of the case of $g(x) = x$ (although potentially introducing new ones), which is the standard Neyman-Pearson problem. This means that for monotone increasing g, one family of optimal solutions is given by $s^* = \phi \circ \eta$, where ϕ is strictly monotone increasing. ■

Proof [**Proposition 13**] By Proposition 11, the unique optimal pair-scorer is $s^*_{\mathrm{Pair}} = \frac{1}{p+1}\mathrm{Diff}(\sigma^{-1} \circ \eta) = \mathrm{Diff}\left(\frac{1}{p+1}(\sigma^{-1} \circ \eta)\right)$, which is decomposable. Corollary 2 may be adapted here to argue that any optimal univariate scorer s^* must satisfy $s^*_{\mathrm{Pair}} = \mathrm{Diff}(s^*)$, and so $s^* = \frac{1}{p+1}(\sigma^{-1} \circ \eta) + b$ for some $b \in \mathbb{R}$. ■

Appendix B. Assorted lemmas

We collect some assorted lemmas that are employed in the above proofs.

Lemma 15 *Let $f, g : \mathcal{X} \to \mathbb{R}$. Then,*

$$(\forall x, x' \in \mathcal{X})\, f(x) < f(x') \implies g(x) < g(x')$$

if and only if $f = \phi \circ g$ for some monotone increasing $\phi : \mathbb{R} \to \mathbb{R}$.

Proof ($\Longleftarrow$). This is easily verified by the definition of monotonicity.

($\Longrightarrow$). We will construct such a monotone increasing ϕ. For any $y \in \mathrm{Im}(g)$, let

$$\mathcal{I}(y) = \{x \in \mathcal{X} : g(x) = y\}$$

be the preimage of y under g. For any $y \in \mathbb{R}$, let

$$\phi(y) := \min\{f(x) : x \in \mathcal{I}(y)\}.$$

We will check that $f = \phi \circ g$, and that ϕ is monotone increasing.

First, note that for any $x, x' \in \mathcal{I}(y)$, by definition $g(x) = g(x')$. By assumption,

$$g(x) \geq g(x') \implies f(x) \geq f(x')$$

and by symmetry

$$g(x) \leq g(x') \implies f(x) \leq f(x')$$

so that

$$g(x) = g(x') \implies f(x) = f(x').$$

Thus for any $x, x' \in \mathcal{I}(y)$, $f(x) = f(x')$. Thus, for any $x \in \mathcal{I}(y)$,

$$\phi(y) = f(x).$$

Now, for any $x_0 \in \mathcal{X}$,

$$\begin{aligned}\phi(g(x_0)) &= \min\{f(x) : x \in \mathcal{I}(g(x_0))\} \\ &= f(x_0).\end{aligned}$$

Thus, $f = \phi \circ g$. To see that ϕ is monotone increasing, pick $y < y'$, and $x \in \mathcal{I}(y), x' \in \mathcal{I}(y')$. Then $y = g(x) < g(x') = y'$. Since $g(x) < g(x')$ implies $f(x) = \phi(y) < \phi(y') = f(x')$, we see that $y < y' \implies \phi(y) < \phi(y')$. ■

Lemma 16 *Let $f, g : \mathcal{X} \to \mathbb{R}$. Then,*

$$(\forall x, x' \in \mathcal{X})\ \mathrm{sign}(f(x) - f(x')) = \mathrm{sign}(g(x) - g(x'))$$

if and only if $f = \phi \circ g$ for some strictly monotone increasing $\phi : \mathbb{R} \to \mathbb{R}$.

Proof We can equivalently write the condition as

$$(\forall x, x' \in \mathcal{X})\, f(x) < f(x') \iff g(x) < g(x').$$

Thus, by Lemma 16, $f = \phi_1 \circ g$ for some monotone increasing ϕ_1, and $g = \phi_2 \circ f$ for some monotone increasing ϕ_2. Thus $f = \phi_1 \circ \phi_2 \circ f$, and so $\phi_1 = \phi_2^{-1}$. This implies that ϕ_1 and ϕ_2 are invertible, or equivalently, that they both correspond to strictly monotone increasing transforms. ■

Appendix C. Properties of the derived ranking distribution $\mathrm{Bipart}(D)$

Suppose we have a distribution $D_{P,Q,\pi} = D_{M,\eta}$. We will use $(\mathsf{X}, \mathsf{X}', \mathsf{Z})$ to denote the random variables over $\mathcal{X} \times \mathcal{X}$ and ± 1 respectively. By definition, for any $x, x' \in \mathcal{X}$ and $z \in \{\pm 1\}$,

$$\begin{aligned}
\Pr[\mathsf{Z} = z] &= \frac{1}{2} \\
\Pr[\mathsf{X} = x | \mathsf{Z} = z] &= [\![z = 1]\!] P(x) + [\![z = -1]\!] Q(x) \\
\Pr[\mathsf{X}' = x' | \mathsf{Z} = z] &= [\![z = 1]\!] Q(x') + [\![z = -1]\!] P(x').
\end{aligned}$$

From these, we may derive other marginals and conditionals:

$$\begin{aligned}
\Pr[\mathsf{X} = x, \mathsf{X}' = x' | \mathsf{Z} = z] &= \Pr[\mathsf{X} = x | \mathsf{Z} = z] \cdot \Pr[\mathsf{X}' = x' | \mathsf{Z} = z] \\
&= [\![z = 1]\!] P(x) Q(x') + [\![z = -1]\!] P(x') Q(x) \\
\Pr[\mathsf{X} = x, \mathsf{X}' = x'] &= \frac{P(x)Q(x') + P(x')Q(x)}{2} \\
&= \frac{1}{2\pi(1-\pi)} \cdot \Pr[x] \Pr[x'] \cdot (\eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))) \\
\Pr[\mathsf{X} = x] &= \frac{P(x) + Q(x)}{2} \\
\Pr[\mathsf{X} = x | \mathsf{X}' = x'] &= \frac{P(x)Q(x') + P(x')Q(x)}{P(x) + Q(x)} \\
\Pr[\mathsf{Z} = 1 | \mathsf{X} = x] &= \frac{P(x)}{P(x) + Q(x)} \\
&= \sigma(\sigma^{-1}(\Pr[\mathsf{Y} = 1 | \mathsf{X} = x]) - \sigma^{-1}(\pi)) \\
\Pr[\mathsf{Z} = 1 | \mathsf{X} = x, \mathsf{X}' = x'] &= \frac{P(x)Q(x')}{P(x)Q(x') + P(x')Q(x)} \\
&= \frac{1}{1 + \frac{Q(x)}{P(x)} \cdot \frac{P(x')}{Q(x')}} \\
&= \sigma(\sigma^{-1}(\Pr[\mathsf{Z} = 1 | \mathsf{X} = x]) - \sigma^{-1}(\Pr[\mathsf{Z} = 1 | \mathsf{X}' = x'])) \\
&= \sigma(\sigma^{-1}(\Pr[\mathsf{Y} = 1 | \mathsf{X} = x]) - \sigma^{-1}(\Pr[\mathsf{Y} = 1 | \mathsf{X}' = x'])) \\
&= \sigma((\mathrm{Diff}(\sigma^{-1} \circ \eta))(x, x')).
\end{aligned}$$

The last two identities follows because

$$\sigma^{-1}(\Pr[\mathsf{Y} = 1 | \mathsf{X} = x]) = \log \frac{\pi}{1 - \pi} + \log \frac{P(x)}{Q(x)},$$

where $\Pr[\mathsf{Y} = 1 | \mathsf{X} = x]$ is understood to mean the observation conditional density in the original space $D \in \Delta_{\mathcal{X} \times \{\pm 1\}}$.

Appendix D. Interpretation of (Uematsu and Lee, 2012) in terms of proper losses

The following are the results shown in (Uematsu and Lee, 2012).

Proposition 17 ((Uematsu and Lee, 2012, Theorem 3)) *Suppose* $\ell(y,v) = \phi(yv)$ *for some* $\phi : \mathbb{R} \to \mathbb{R}_+$*, where* ϕ *is differentiable, monotone decreasing, convex, and* $\phi'(0) < 0$*. For a given distribution* $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$*, let*

$$s^* \in \mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}.$$

Then,

$$(\forall x, x' \in \mathcal{X})\, \eta(x) \neq \eta(x') \implies \mathrm{sign}(\mathrm{Diff}(s^*)(x,x')) = \mathrm{sign}(\eta(x) - \eta(x')).$$

If ϕ *is strictly convex, then the above also holds when* $\eta(x) = \eta(x')$*.*

Proposition 18 ((Uematsu and Lee, 2012, Theorem 7)) *Suppose* $\ell(y,v) = \phi(yv)$ *for some* $\phi : \mathbb{R} \to \mathbb{R}_+$*, where* ϕ *is differentiable, strictly monotone decreasing, convex, and* $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ *is strictly increasing. Given any* $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$*,*

$$\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \subseteq \mathcal{S}_{\mathrm{Decomp}}$$

if and only if $\phi'(-s)/\phi'(s) = e^{as}$ *for some* $a > 0$*.*

We show how to interpret these in terms of proper composite losses. First, we show that the conditions of their theorems imply that ℓ is a proper composite margin loss.

Proposition 19 *Let* ϕ *be differentiable, monotone decreasing, strictly convex, and* $\phi'(0) < 0$*. Then,* $\ell(y,v) = \phi(yv)$ *is proper composite.*

Proof Let ϕ meet the stated conditions. Since ϕ is convex and monotone decreasing with $\phi'(0) < 0$, then it must be true that

$$(\forall v \in \mathbb{R})(\phi'(v) \neq 0 \vee \phi'(-v) \neq 0).$$

Further, the function

$$f(v) := \frac{\phi'(v)}{\phi'(-v)}$$

is continuous by differentiability of ϕ, and monotone by monotonicity and convexity of ϕ, since

$$f'(v) = \frac{1}{(\phi'(v))^2} \cdot \big(\phi'(-v)\phi''(v) + \phi''(-v)\phi'(v)\big) \leq 0.$$

When ϕ is *strictly* convex, f is strictly monotone because the numerator above cannot be 0. Thus, the conditions of Corollary 16 in (Vernet et al., 2011) hold, and so ℓ is proper composite. ■

Proposition 20 *Let* ϕ *be differentiable, strictly monotone decreasing, convex, and* $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ *is strictly increasing. Then,* $\ell(y,v) = \phi(yv)$ *is proper composite.*

Proof The proof follows by the conditions of Corollary 16 in (Vernet et al., 2011), as before; with invertibility $f : s \mapsto \frac{\phi'(-s)}{\phi'(s)}$ directly assumed rather than derived as a consequence of strict convexity. ■

By Lemma 16, the statement of their Theorem 3 is equivalent to saying that s^* is a strictly monotone increasing transform of η. Thus, this result is equivalent to Corollary 9, except that the latter explicitly provides the form of the link function relating η and s^*.

The following shows that the conditions in their Theorem 7 imply that the inverse link function is of the form $\Psi^{-1}(v) = \frac{1}{1+e^{-av}}$, which means the result is a special case of Proposition 6 where ℓ is a margin loss.

Lemma 21 *Let ℓ be a differentiable proper composite margin loss with link function Ψ, so that $\ell(y, v) = \phi(yv)$ for some differentiable $\phi : \mathbb{R} \to \mathbb{R}_+$. Then, for any $a \in \mathbb{R} \setminus \{0\}$,*

$$\Psi^{-1}(s) = \frac{1}{1 + e^{-as}} \iff \phi'(-s)/\phi'(s) = e^{as}.$$

Proof The link function for a differentiable proper composite loss satisfies

$$\Psi^{-1}(s) = \frac{1}{1 - \frac{\ell_1'(s)}{\ell_{-1}'(s)}} = \frac{1}{1 + \frac{\phi'(s)}{\phi'(-s)}} = \frac{1}{1 + e^{-as}}.$$

■

Appendix E. Empirical illustration of Corollary 7

We present an empirical illustration of the assertion that Corollary 7 holds for an *asymmetric* proper composite loss. We work with a one dimensional discrete distribution over instances drawn from $[N]$, where the instance i has probability M_i of being drawn. Each instance i has an associated probability η_i of having a positive label. We compute the Bayes-optimal univariate scorer with the p-classification loss for $p = 2$,

$$\ell_1(v) = e^{-v}, \ell_{-1}(v) = \frac{1}{2}e^{2v}.$$

We minimised the risk using L-BFGS. We performed this operation 9 times for $N = 100$. In each trial, we drew $\eta_i \sim \text{Uniform}[0, 1]$, $Z_i \sim \text{Uniform}[0, 1]$, and set $M_i = Z_i / \sum_j Z_j$. We plot for each trial the graph of η versus s in Figure 1. We see in each case that the optimal scorer is identical. It can be verified that it is further $\Psi \circ \eta$, as expected.

Ideally one would like to empirically illustrate the claim for non-convex losses that use the specified link. However, determining the optimal scorer in such cases is of course computationally challenging.

Appendix F. Empirical illustration of Corollary 9

We present an empirical illustration of the assertions that for a proper composite loss whose Bayes-optimal pair-scorer is non-decomposable, (a) the optimal univariate scorer is a strictly monotone transform of η, and (b) the transformation is distribution dependent. As in the previous section, we work with a one dimensional discrete distribution over instances drawn from $[N]$, where the

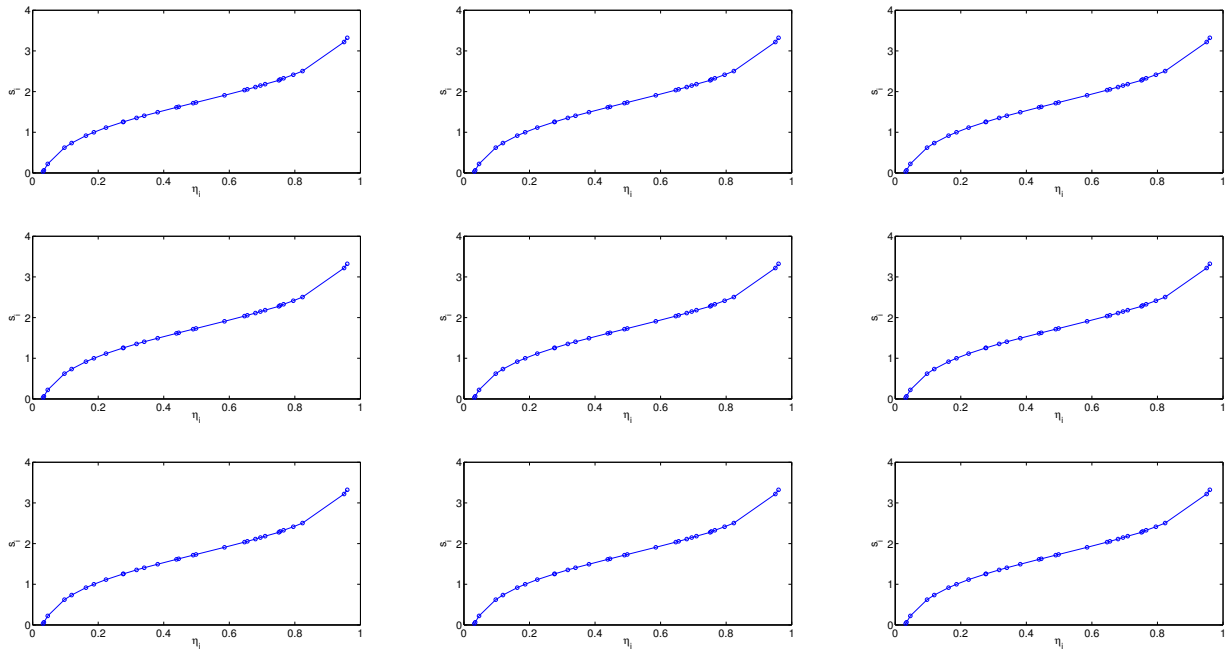

Figure 1: Results of 9 simulation trials to illustrate Corollary 7 for the case of an asymmetric loss. Here, the η_i and M_i values are varied across each trial, and each panel represents the relationship between η and s for a specific trial. We see that the relationship in each trial is identical.

instance i has probability M_i of being drawn, and probability η_i of having a positive label. We compute the Bayes-optimal univariate scorer with squared loss, i.e. we find

$$\min_{s_1,\ldots,s_N} \mathbb{E}_{\mathsf{X},\mathsf{X}'\sim M}\left[\eta(\mathsf{X})(1-\eta(\mathsf{X}'))\ell_1(s(\mathsf{X})-s(\mathsf{X}'))\right] = \min_{s_1,\ldots,s_N} \sum_{i,j=1}^{N} M_i M_j \eta_i (1-\eta_j)(1-(s_i-s_j))^2.$$

This is a weighted least squares problem, and so admits a closed form solution. We performed this operation 5 times for $N = 100$. In each trial, we drew $\eta_i \sim \text{Uniform}[0,1]$, $Z_i \sim \text{Uniform}[0,1]$, and set $M_i = Z_i / \sum_j Z_j$. We plot for each trial the graph of η versus s in Figure 2. We see in each case that (a) the relationship between the two quantities is strictly monotone increasing, and (b) that the relationship differs across the trials.

Appendix G. Experiments with the p-norm push

We conducted experiments comparing the alternatives to the p-norm push in Table 2. As baselines, we used class-probability estimation with logistic and exponential loss, and the p-norm push with exponential loss. The aim of our experiments is *not* to put forth a superior alternative to the existing p-classification and p-norm push approaches. Rather, we wish to demonstrate that the proper composite interpretation gives one way of generating a family of losses for this problem.

We compare these methods on a three UCI datasets: `ionosphere`, `german`, and `magic`. Each method was trained with a regularised linear model, where the training objective was minimised using L-BFGS (Nocedal and Wright, 2006, pg. 177). For each dataset, we created 5 random

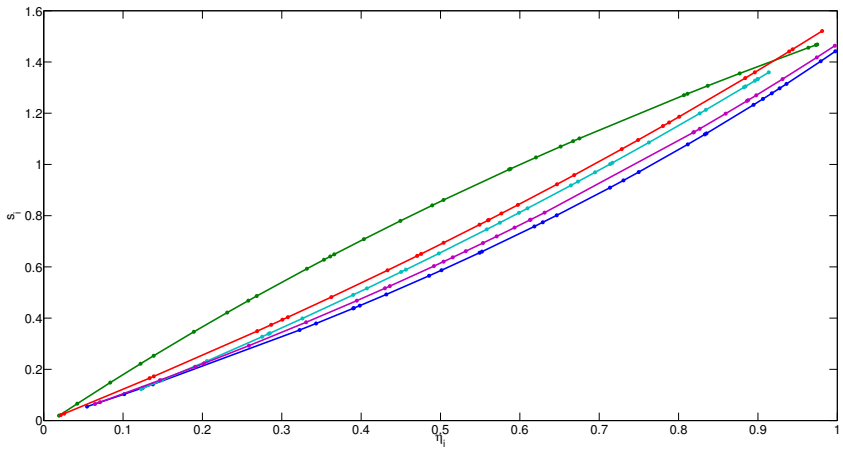

Figure 2: Results of 5 simulation trials to illustrate Proposition 8 for the case of squared loss. Here, the η_i and M_i values are varied across each trial, and each coloured plot represents the relationship between η and s for a specific trial.

train-test splits in the ratio $2:1$. For each split, we performed 5-fold cross-validation on the training set to tune the strength of regularisation $\lambda \in \{10^{-6}, 10^{-5}, \ldots, 10^{1}\}$, and where appropriate the constant $p \in \{1, 2, 4, 8, 16\}$. We then evaluated performance on the test set, and report the average across all splits. As performance measures, we used the AUC, MRR, DCG, AP, and PTop (Agarwal, 2011; Boyd et al., 2012). For all measures, a higher score is better. Parameter tuning was done based on the AP on the test folds.

The results are summarised in Tables 3 – 5. (We were unable to optimise the p-norm push on the `magic` dataset within a reasonable amount of time.) No single method clearly outperforms all others in all metrics. However, we observe that, as expected, the candidate proper composite losses are very competitive with p-classification and the p-norm push.

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Logistic	0.9113 ± 0.0208	0.0422 ± 0.0115	0.1966 ± 0.0108	0.9243 ± 0.0339	13.0000 ± 17.0880
Exponential	0.9128 ± 0.0166	$\mathbf{0.0482 \pm 0.0078}$	$\mathbf{0.2034 \pm 0.0070}$	0.9262 ± 0.0318	12.8000 ± 12.9499
Generalised Logistic	0.9115 ± 0.0189	0.0409 ± 0.0119	0.1963 ± 0.0115	0.9266 ± 0.0324	13.6000 ± 16.4408
Generalised Exponential	$\mathbf{0.9220 \pm 0.0184}$	0.0462 ± 0.0098	0.2009 ± 0.0092	$\mathbf{0.9416 \pm 0.0272}$	$\mathbf{16.0000 \pm 11.5109}$
P-Classification	0.9166 ± 0.0183	0.0447 ± 0.0094	0.1991 ± 0.0088	0.9364 ± 0.0246	12.4000 ± 9.5289
P-Norm Push	0.9180 ± 0.0202	0.0463 ± 0.0098	0.2010 ± 0.0095	0.9366 ± 0.0289	14.0000 ± 13.2476

Table 3: Results of various "ranking the best" methods on `ionosphere` dataset.

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Logistic	0.8121 ± 0.0285	0.0157 ± 0.0036	0.1514 ± 0.0043	0.6236 ± 0.0637	2.4000 ± 1.9494
Exponential	0.8131 ± 0.0311	0.0197 ± 0.0065	0.1549 ± 0.0071	0.6217 ± 0.0677	1.8000 ± 2.0494
Generalised Logistic	0.8123 ± 0.0301	0.0187 ± 0.0038	0.1539 ± 0.0035	0.6225 ± 0.0682	2.2000 ± 2.1679
Generalised Exponential	$\mathbf{0.8133 \pm 0.0304}$	$\mathbf{0.0242 \pm 0.0068}$	$\mathbf{0.1599 \pm 0.0067}$	0.6239 ± 0.0637	2.2000 ± 1.7889
P-Classification	0.8117 ± 0.0279	0.0190 ± 0.0051	0.1549 ± 0.0041	0.6230 ± 0.0617	2.4000 ± 2.3022
P-Norm Push	0.8121 ± 0.0282	0.0183 ± 0.0029	0.1551 ± 0.0046	0.6236 ± 0.0604	2.2000 ± 1.9235

Table 4: Results of various "ranking the best" methods on `german` dataset.

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Logistic	0.8372 ± 0.0032	0.0014 ± 0.0001	0.0917 ± 0.0002	0.8781 ± 0.0049	6.0000 ± 2.4495
Exponential	0.8377 ± 0.0026	0.0015 ± 0.0002	$\mathbf{0.0918 \pm 0.0002}$	0.8806 ± 0.0044	2.6000 ± 1.6733
Generalised Logistic	0.8374 ± 0.0031	0.0015 ± 0.0001	0.0917 ± 0.0001	0.8781 ± 0.0051	6.2000 ± 2.3875
Generalised Exponential	0.8379 ± 0.0025	$\mathbf{0.0016 \pm 0.0001}$	$\mathbf{0.0918 \pm 0.0002}$	0.8808 ± 0.0043	4.2000 ± 2.6833
P-Classification	$\mathbf{0.8406 \pm 0.0028}$	0.0015 ± 0.0001	$\mathbf{0.0918 \pm 0.0002}$	$\mathbf{0.8937 \pm 0.0043}$	$\mathbf{8.4000 \pm 13.3903}$

Table 5: Results of various "ranking the best" methods on `magic` dataset.