JMLR: Workshop and Conference Proceedings vol 35:1–34, 2014

Bayes-Optimal Scorers for Bipartite Ranking

Aditya Krishna Menon ADITYA.MENON@NICTA.COM.AU

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 establish the consistency of the minimisation of surrogate bipartite risks, and elucidates 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 equivalences between several seemingly disparate risks for bipartite ranking, such as minimising a suitable class-probability estimation risk, and minimising the p-norm push risk proposed by Rudin (2009).

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

1. The bipartite ranking problem

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. A canonical choice is for ℓ to be zero-one loss, for which the bipartite risk is one minus the *area under the ROC curve (AUC)* (Agarwal et al., 2005).

The non-convexity of the 0-1 loss hampers direct maximisation of the AUC. A popular strategy is to instead minimise the bipartite risk with respect to some surrogate loss ℓ (Cohen et al., 1999; Herbrich et al., 2000; Burges et al., 2005). While intuitive, it is of interest to establish whether these approaches are consistent for the task of AUC maximisation. A necessary condition for this to hold is for the *Bayes-optimal* scorers under the surrogate loss to match the Bayes-optimal scorers under the 0-1 loss. While the Bayes-optimal scorers for 0-1 loss are well understood (Clémençon et al., 2008), for surrogate losses their study has been restricted to 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 equivalences between the risks for several disparate approaches to bipartite ranking, 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 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 $\mathrm{Ber}(\theta)$ the Bernoulli distribution with parameter $\theta \in [0, 1]$.

For any function $f \colon \mathcal{X} \to \mathbb{R}$, we denote by $\underset{x \in \mathcal{X}}{\mathrm{Argmin}}\, f(x)$ 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 $\underset{x \in \mathcal{X}}{\mathrm{argmin}}\, f(x)$. We denote by $\mathrm{Diff}(f) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the function satisfying $(\mathrm{Diff}(f))(x, x') = f(x) - f(x')$ for every $x, x' \in \mathcal{X}$. For a set of functions $\mathcal{F} = \{f \colon \mathcal{X} \to \mathbb{R}\}$, we define $\mathrm{Diff}(\mathcal{F}) = \{\mathrm{Diff}(f) : f \in \mathcal{F}\}$.

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 $\mathrm{sign}(x) = [\![x \geq 0]\!] - [\![x \leq 0]\!]$. 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 $s_{\mathrm{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathcal{V}$. We call a pair-scorer s_{Pair} *decomposable* if

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

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*. Slightly abusing notation, we sometimes specify a loss via $\ell(v) = (\ell_{-1}(v), \ell_1(v))$. 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

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

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* (Buja et al., 2005; Reid and Williamson, 2010) if the conditional risk $L(\eta, \cdot)$ is minimised by predicting η:

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

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 triplet (P, Q, π), where

$$(\forall x \in \mathcal{X})\; (P(x), Q(x), \pi) = (\Pr[\mathsf{X} = x | \mathsf{Y} = 1], \Pr[\mathsf{X} = x | \mathsf{Y} = -1], \Pr[\mathsf{Y} = 1]), \tag{4}$$

or alternately by the tuple (M, η), where

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

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 now 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

$$\mathbb{L}_\ell^D(s) = \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim D}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] = \mathbb{E}_{\mathsf{X} \sim M}\left[L_\ell(\eta(\mathsf{X}), s(\mathsf{X}))\right], \tag{5}$$

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:

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

[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.

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

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 $\mathbb{L}_{01}^{D}$. Directly minimising $\mathbb{L}_{01}^{D}$ may be computationally challenging due to the non-convexity of ℓ^{01}. A common approach is to instead find a scorer that (approximately) minimises $\mathbb{L}_{\ell}^{D}$ 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 $\mathbb{L}_{\ell}^{D}$ 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

$$\mathbb{L}_{\mathrm{Bipart},\ell}^{D}(s_{\mathrm{Pair}}) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[\frac{\ell_1(s_{\mathrm{Pair}}(\mathsf{X},\mathsf{X}')) + \ell_{-1}(s_{\mathrm{Pair}}(\mathsf{X}',\mathsf{X}))}{2}\right]. \tag{6}$$

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

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

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

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

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 $\mathbb{L}_{\mathrm{Bipart},01}^{D}(\mathrm{Diff}(s))$ is (approximately) minimised; equivalently, we seek to minimise $\mathbb{L}_{\mathrm{Bipart},01}^{D}(s_{\mathrm{Pair}})$ over all $s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}}$. One minus the risk $\mathbb{L}_{\mathrm{Bipart},01}^{D}(\mathrm{Diff}(s))$ equals the *area under the ROC curve (AUC)* of the scorer s (Agarwal et al., 2005; Clémençon et al., 2008), which can be interpreted as the probability of a random positive instance scoring higher than a random negative instance:

$$\mathrm{AUC}^{D}(s) = \mathbb{E}_{\mathsf{X}\sim P,\mathsf{X}'\sim Q}\left[[\![s(\mathsf{X}) > s(\mathsf{X}')]\!] + \frac{1}{2}[\![s(\mathsf{X}) = s(\mathsf{X}')]\!]\right].$$

Minimising the 0-1 bipartite risk is thus equivalent to maximising the AUC.

4. Bayes-optimal scorers for the bipartite ranking risk

There are two approaches to learning a scorer s that approximately minimises $\mathbb{L}_{\mathrm{Bipart},01}^{D}(\mathrm{Diff}(s))$. In the *pointwise approach*, one minimises $\mathbb{L}_{\ell}^{D}(s)$ for some surrogate loss ℓ (Kotlowski et al., 2011). In the *pairwise approach*, one minimises $\mathbb{L}_{\mathrm{Bipart},\ell}^{D}(\mathrm{Diff}(s))$ for some surrogate loss ℓ (Herbrich et al., 2000; Freund et al., 2003; Burges et al., 2005). A key question is whether these approaches are consistent for the task of minimising $\mathbb{L}_{\mathrm{Bipart},01}^{D}(\mathrm{Diff}(s))$. To answer this, it is necessary to establish that the corresponding Bayes-optimal solutions $\mathcal{S}_{\ell}^{D,*}$ and $\mathcal{S}_{\mathrm{Bipart},\ell}^{D,\mathrm{Univ},*}$ fall in the set $\mathcal{S}_{\mathrm{Bipart},01}^{D,\mathrm{Univ},*}$. The nature of the Bayes-optimal solutions is well-understood for the univariate approach, but less so for the pairwise approach. We thus aim to characterise $\mathcal{S}_{\mathrm{Bipart},\ell}^{D,\mathrm{Univ},*}$, for which it will be useful to establish $\mathcal{S}_{\ell}^{D,*}$. In what follows, $D = D_{P,Q,\pi} = D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$.

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):

$$\mathcal{S}_{01}^{D,*} = \{s \colon \mathcal{X} \to \mathbb{R} : \eta(x) \neq 1/2 \implies \operatorname{sign}(s(x)) = \operatorname{sign}(2\eta(x) - 1)\}\,. \tag{7}$$

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:

$$\{\Psi \circ \eta\} \subseteq \mathcal{S}_{\ell}^{D,*}. \tag{8}$$

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

In bipartite ranking, our interest is in determining the Bayes-optimal *univariate* scorers, $\mathcal{S}_{\text{Bipart},\ell}^{D,\text{Univ},*}$. As a warm-up, we first determine the Bayes-optimal *pair*-scorers, $\mathcal{S}_{\text{Bipart},\ell}^{D,*}$. Our first challenge is determining a suitable conditional risk from Equation 6. To do so, we exploit an equivalence of the bipartite risk to a *pairwise classification* risk on a distribution $\text{Bipart}(D)$, which we now define.

Definition 1 *For any $D_{P,Q,\pi} \in \Delta_{\mathcal{X}\times\{\pm1\}}$, let $\text{Bipart}(D) \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm1\}}$ by defined by the triplet $(P_{\text{pair}}, Q_{\text{pair}}, \pi_{\text{pair}})$, where*

$$(\forall x, x' \in \mathcal{X})\,(P_{\text{pair}}(x,x'), Q_{\text{pair}}(x,x'), \pi_{\text{pair}}) = \big(P(x)Q(x'), P(x')Q(x), 1/2\big)\,.$$

The classification risk with respect to $\text{Bipart}(D)$ is equivalent to the bipartite risk with respect to D, as is well known for ℓ^{01} (Balcan et al., 2008; Kotlowski et al., 2011; Agarwal, 2013).

Lemma 2 *For any $D_{P,Q,\pi} \in \Delta_{\mathcal{X}\times\{\pm1\}}$, loss ℓ and pair-scorer $s_{\text{Pair}} \colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$,*

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

Lemma 2 implies that $\mathcal{S}_{\text{Bipart},\ell}^{\text{Bipart}(D),*} = \mathcal{S}_{\text{Bipart},\ell}^{D,*}$. We now just need the following elementary property of the observation-conditional density η_{Pair} of $\text{Bipart}(D)$.

Lemma 3 *For any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm1\}}$, $\text{Bipart}(D)$ has observation-conditional density given by*

$$\eta_{\text{Pair}} = \sigma \circ \text{Diff}(\sigma^{-1} \circ \eta). \tag{9}$$

Thus, combining Equations 7 and 9, we have

$$\mathcal{S}_{\text{Bipart},01}^{D,*} = \big\{s_{\text{Pair}} \colon \mathcal{X}\times\mathcal{X} \to \mathbb{R} : \eta(x) \neq \eta(x') \implies \operatorname{sign}(s_{\text{Pair}}(x,x')) = \operatorname{sign}(\eta(x) - \eta(x'))\big\}\,, \tag{10}$$

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,

$$\{\Psi\circ\eta_{\mathrm{Pair}}\}=\{\Psi\circ\sigma\circ\mathrm{Diff}(\sigma^{-1}\circ\eta)\}\subseteq\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell}. \tag{11}$$

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

When looking to establish the Bayes-optimal univariate scorers for bipartite ranking, we immediately face a challenge. Finding the set of scorers s that minimise $\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s))$ is equivalent to finding the set of pair-scorers s_{Pair} that minimise $\mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}})$ 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 make a pointwise analysis via the conditional risk. 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.

Proposition 4 *Given any $D\in\Delta_{\mathcal{X}\times\{\pm1\}}$ and loss ℓ,*

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

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 5 *Given any $D\in\Delta_{\mathcal{X}\times\{\pm1\}}$ 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 univariate 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.

4.3.1. DECOMPOSABLE CASE

We first handle the case where there is a decomposable Bayes-optimal pair-scorer, which allows us to easily compute the optimal scorer. Observing from Equation 10 that $\{\mathrm{Diff}(\eta)\}\subseteq\mathcal{S}^{D,*}_{\mathrm{Bipart},01}\cap\mathcal{S}_{\mathrm{Decomp}}$, we have the following characterisation of the optimal univariate scorers for ℓ^{01}.

Proposition 6 *Given any $D_{M,\eta}\in\Delta_{\mathcal{X}\times\{\pm1\}}$,*

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

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

The fact that ϕ in Proposition 6 need not be *strictly* monotone increasing means that for some $x\neq x'\in\mathcal{X}$, we may have $\eta(x)=\eta(x')$ but $s(x)\neq s(x')$. Nonetheless, an immediate corollary is that any strictly monotone increasing transform of η is necessarily an optimal univariate scorer.

Corollary 7 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and any strictly monotone increasing $\phi : [0,1] \to \mathbb{R}$,*

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

Combined with Equation 8, Corollary 7 implies that $\mathcal{S}^{D,*}_{\ell} \subseteq \mathcal{S}^{D,\text{Univ},*}_{\text{Bipart},01}$ for a strictly proper composite loss. Surrogate regret bounds (Agarwal, 2013) may further be developed to establish consistency of the pointwise approach for this family of losses.

We now proceed to the case where ℓ is a proper composite loss. To apply Corollary 5, we characterise the subset of proper composite losses for which there exists a decomposable pair-scorer.

Proposition 8 (Decomposability of Bayes-optimal bipartite pair-scorer.) *Given any strictly proper composite loss ℓ with a differentiable, invertible link function Ψ,*

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

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 11, the Bayes-optimal pair-scorer is

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

($\Longrightarrow$) See Appendix A. ■

What is special about the particular family of links in Proposition 8, which are scaled versions of the sigmoid? The answer is simply that the scorer η_{Pair} involves a sigmoid link function (Equation 9). This form of η_{Pair} in turn can be understood in terms of utility representations for binary relations on sets (Roberts, 1984, pg. 273 – 280); for details, see (Menon and Williamson, 2014).

Observe that the class of proper composite losses satisfying the conditions of Proposition 8 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 9 *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 $a \in \mathbb{R} \setminus \{0\}$,*

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

Further, we may transfer surrogate regret bounds from binary classification to relate the excess pairwise ℓ risk of a scorer $s : \mathcal{X} \to \mathbb{R}$ to its the excess pairwise ℓ^{01} risk. Such a bound implies that minimising certain pairwise surrogate risks is consistent for AUC maximisation.

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

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

where

$$\text{regret}^{D,\text{Univ}}_{\text{Bipart},\ell}(s) = \mathbb{L}^{D}_{\text{Bipart},\ell}(\text{Diff}(s)) - \inf_{t:\mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\text{Bipart},\ell}(\text{Diff}(t)).$$

The function $F_\ell : [0,1] \to \mathbb{R}_+$ is exactly that which appears in bounds relating 0-1 to ℓ classification regret for proper composite losses and may be specified in terms of the Bayes-risk of the proper composite loss ℓ Reid and Williamson (2009, Theorem 3).

Proposition 10 is a simple consequence of the reduction of bipartite ranking to classification over pairs, and in particular the fact that $\mathrm{regret}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s) = \mathrm{regret}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s))$ when the Bayes-optimal pair-scorer is decomposable. When the optimal pair-scorer is *not* decomposable, the two regrets will no longer coincide, and more effort is needed to derive a surrogate regret bound. This further illustrates the value of the decomposability of the Bayes-optimal pair-scorer.

4.3.2. 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 $\mathcal{S}_{\mathrm{Decomp}}$ 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 11 *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 D has finite support*

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

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 12 *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 D has finite support

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

where $f^D_{s^*}$ is defined as in Proposition 11.

We make some observations on the result in Proposition 11. First, while the results of this section rely on convexity of the loss ℓ, the previous section established that convexity is *not necessary*, since one can have a non-convex loss resulting from a suitable link $\Psi = \frac{1}{a}\sigma^{-1}$. Second, we suspect that the requirement of ϕ' bounded when D does not have finite support may be dropped, but defer to future work investigation of minimal conditions for the result to hold. Third, 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, changing the link function is seen to change the Bayes-optimal solutions in a non-trivial way. It is unclear for example whether the use of non-sigmoidal links, which induce non-decomposable Bayes-optimal pair-scorers, still allow for the derivation of surrogate regret bounds, and hence consistency.

4.4. Relation to existing work

This section generalised and unified several earlier results through the theory of proper losses. For ℓ^{01}, our Corollary 7 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 8 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 D has an empirical illustration of this. Our Corollary 12 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^*; Appendix E empirically illustrates the distribution-dependent nature of the link function. (We translate these results in terms of proper losses so that the connection is more apparent in Appendix C.)

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) and 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 6). 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:

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

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 11. (Note that requiring differentiability of the loss means that we cannot compute the optimal solution for ℓ^{01}.)

Proposition 13 *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 D has finite support*

$$\mathcal{S}^{D,*}_{\text{push},\ell,g} = \left\{ s^*_{\text{Pair}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R} : s^*_{\text{Pair}} = \Psi\circ\sigma\circ\left(\text{Diff}(\sigma^{-1}\circ\eta) - G^{D}_{s^*_{\text{Pair}}}\right)\right\}, \tag{12}$$

where

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

When $g : x \mapsto x$, we obtain the standard ℓ-bipartite ranking risk, $G^D \equiv 0$ and so $s^*_{\text{Pair}} = \Psi\circ\eta_{\text{Pair}}$ as in Equation 11. 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. Interestingly, when ℓ is the exponential loss and $g^p(x) = x^p$, we have the following simple characterisation.

Proposition 14 *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 D has finite support*

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

As with Proposition 11, we suspect the finiteness assumption on the support of D 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 11), 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 $\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{push},\ell,g}$. For ℓ^{01}, we were unsuccessful in computing the optimal pair-scorer; nonetheless, a different technique lets us establish the optimal univariate scorers.

Proposition 15 *Given any $D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm 1\}}$ and nonnegative, monotone increasing g,*

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

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. This is not surprising: if one can exactly recover η (or a strictly monotone transform thereof), then one can perfectly rank elements at the top of the ranked list.

For a general proper composite loss, it appears difficult to appeal to the optimal pair-scorer implicitly derived in Proposition 13. For the special case of exponential loss, the optimal pair-scorer immediately implies the form of the optimal univariate scorer.

Proposition 16 *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 D has finite support,*

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

We see that changing p results in a scaling of the link function $\Psi = \sigma^{-1}$. One might then hope to understand the p-norm push by considering a family of proper composite risks, where each member of the family comprises some fixed proper loss composed with an appropriately scaled sigmoidal link. However, for a proper composite risk, scaling of the link function simply corresponds to a scaling of the prediction space of the scorer. Thus, even on a finite sample, and a restricted function class, the family of proper composite risks have optimal solutions that are scalings of one another.

As with the ℓ^{01} case, this is not surprising, and indicates that the p-norm push risk must be understood in terms of its behaviour under a restricted function class or finite sample. Along these lines, Ertekin and Rudin (2011, Theorem 1) showed that for a linear function class, the p-norm push risk with exponential loss is equivalent to the proper composite risk corresponding to the p-*classification loss*, given by $\ell(v) = \left(e^{-v}, \frac{1}{p}e^{vp}\right)$. As we demonstrate in Appendix F, varying p in the p-classification loss no longer results in a simple scaling, because the parameter p is embedded in the underlying proper loss itself. This observation gives a means of designing alternate proper composite losses for focussing at the head of the ranked list; we give some examples of such losses in Table 1.

6. Four equivalent risks for 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.

Name	$\ell_{-1}(v)$	$\ell_1(v)$	$\Psi(\eta)$
p-classification	$\frac{1}{p} e^{vp}$	e^{-v}	$\frac{1}{p+1} \log \frac{\eta}{1-\eta}$
Asymmetric A	$\frac{2}{\sqrt{\sigma(-v)}}$	$2 \tanh^{-1}(\sqrt{\sigma(-v)})$	$\log \frac{\eta}{1-\eta}$
Asymmetric B	$\begin{cases} \frac{1}{2} e^{\frac{v}{2}} & v > 0 \\ \log\left(\frac{1+e^v}{2}\right) + \frac{1}{2} & v \leq 0 \end{cases}$	$\begin{cases} \frac{1}{2} e^{-\frac{v}{2}} & v > 0 \\ \log\left(\frac{1+e^{-v}}{2}\right) & v \leq 0 \end{cases}$	$\log \frac{\eta}{1-\eta}$

Table 1: Candidate proper composite alternatives to the p-norm push.

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

Table 2: 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.

(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. However, our results provide conditions under which all methods have the *same* output, meaning that the corresponding risks have equivalent minimisers.

Proposition 17 *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 the support of D is finite and $p = a - 1$ for $a > 1$, method (4) also produces the same pair-scorer.*

Proof By Equation 8 and Corollary 9, methods (1) and (2) produce the same scorer $\Psi \circ \eta$, up to a translation which is nullified by the Diff operator. By Equation 11, this pair-scorer is equivalent to that produced by method (3). Further, if $p = a - 1$ for $a > 1$, then by Proposition 16, method (4) returns $\Psi \circ \eta$ up to a translation which is nullified by the Diff operator. ■

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 2 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 the risk equivalences

Our definition of "equivalent" is that two risks have the same optimal scorer. This does not imply that the corresponding 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).

With these caveats in mind, the results 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; the quality of this reduction can be quantified with surrogate regret bounds (Agarwal, 2013; Narasimhan and Agarwal, 2013).

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, the equivalence between methods (1) and (2) was made by Gao and Zhou (2012, Lemma 4). For the special case of convex margin losses, the equivalence between methods (2) and (3) was shown by Uematsu and Lee (2012). Rudin and Schapire (2009, Theorem 10), Ertekin and Rudin (2011, Theorem 3) showed the equivalence between methods (1) and (2), and Ertekin and Rudin (2011, Theorem 1) the equivalence between methods (1) and (4), holds when the minimisation is over a linear hypothesis class and D has finite support.

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. This characterisation helps establish consistency of minimisation of certain pairwise surrogate risks for the task of minimising the 0-1 bipartite risk. 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 the risks for four seemingly disparate approaches to bipartite ranking. We believe our results illustrate the value of the proper loss machinery in studying ranking problems. One can 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); we refer the reader to (Menon and Williamson, 2014) for details.

Acknowledgments

This work was supported by the Australian Research Council (RCW) and NICTA (AKM and RCW). NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. Thanks to Cynthia Rudin and the COLT referees for their helpful comments.

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 *Advances 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.

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.

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, December 2003.

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.

Aditya Krishna Menon and Robert C. Williamson. Bipartite ranking: risk, optimality, and equivalences. Submitted to JMLR, 2014.

Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances 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.

Fred S. Roberts. *Measurement theory with Applications to Decision Making, Utility, and the Social Sciences*, volume 7 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, Reading, MA, 1984.

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.

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.

Emir H. Shuford Jr., Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. doi: 10.1007/BF02289503.

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 [**Lemma 2**] By Equation 6,

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

where in the penultimate equation we have simply renamed the random variables in the second expression.

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

Proof [**Lemma 3**]

We show the result of Lemma 3, and additionally collect some identities about the distribution $\mathrm{Bipart}(D)$. Suppose we have a distribution $D_{P,Q,\pi} = D_{M,\eta} \in \Delta_{\mathcal{X}\times\{\pm1\}}$. Let the random variable triplet $(\mathsf{X},\mathsf{X}',\mathsf{Z})$ be such that, for any $x, x' \in \mathcal{X}$ and $z \in \{\pm 1\}$,

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

Further suppose that X, X' are conditionally independent given Z. Then, the above summarise a distribution $\mathrm{Bipart}(D) \in \Delta_{\mathcal{X}\times\mathcal{X}\times\{\pm1\}}$, from which a sample (x, x', z) may be drawn via the following process:

- Draw $z \sim \mathrm{Ber}(1/2)$
- Draw $x \sim [\![z = 1]\!] P + [\![z = -1]\!] Q$
- Draw $x' \sim [\![z = -1]\!] P + [\![z = 1]\!] Q$.

From these, we may derive other marginals and conditionals:

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

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

The last two identities follows because

$$\sigma^{-1}(\eta(x)) = \sigma^{-1}(\pi) + \log \frac{P(x)}{Q(x)}.$$

■

Proof [Proposition 4] The ($\Longleftarrow$) direction is immediate, since $\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell} \neq \emptyset$ and thus $\mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},\ell}) \neq \emptyset$. We show the ($\Longrightarrow$) direction.

($\subseteq$). 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 of s^*_{Pair},

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

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, by definition,

$$s_{\mathrm{Pair}} \in \underset{t_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}}}{\mathrm{Argmin}}\; \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(t_{\mathrm{Pair}}).$$

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 *un*constrained optimisation that lies in $\mathcal{S}_{\mathrm{Decomp}}$, call it t_{Pair}. Clearly t_{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 5] ($\Longrightarrow$) follows by Proposition 4, and ($\Longleftarrow$) follows by definition of decomposability. ■

Proof [Proposition 6] Let $\mathcal{A} = \mathcal{S}^{D,*}_{\mathrm{Bipart},01} \cap \mathcal{S}_{\mathrm{Decomp}}$. By Equation 10,

$$\begin{aligned}
\mathcal{A} &= \left\{ s_{\mathrm{Pair}} \in \mathcal{S}_{\mathrm{Decomp}} : \eta(x) \neq \eta(x') \implies \mathrm{sign}(s_{\mathrm{Pair}}(x, x')) = \mathrm{sign}(\eta(x) - \eta(x')) \right\} \\
&= \mathrm{Diff}\left(\left\{ s \colon \mathcal{X} \to \mathbb{R} : \eta(x) \neq \eta(x') \implies \mathrm{sign}(s(x) - s(x')) = \mathrm{sign}(\eta(x) - \eta(x')) \right\} \right) \\
&= \mathrm{Diff}\left(\{ s \colon \mathcal{X} \to \mathbb{R} : \eta = \phi \circ s \text{ for } \phi \text{ monotone increasing } \} \right) \text{ by Lemma 18.}
\end{aligned}$$

Since $\mathcal{A}$ is nonempty, $\mathcal{A} = \mathrm{Diff}(\mathcal{S}^{D,\mathrm{Univ},*}_{\mathrm{Bipart},01})$ by Proposition 4. For any sets of scorers $\mathcal{S}_1, \mathcal{S}_2$, $\mathrm{Diff}(\mathcal{S}_1) = \mathrm{Diff}(\mathcal{S}_2) \implies (\forall s_1 \in \mathcal{S}_1)\, (\exists s_2 \in \mathcal{S}_2, c \in \mathbb{R})\, s_1 = s_2 + c$, i.e. the scorers in the two sets must be related by a linear translation. But if for a scorer s we have $\eta = \phi \circ s$ for some monotone ϕ, then it must also be true that $\eta = \tilde{\phi} \circ (s + c)$ where $\tilde{\phi} : x \mapsto \phi(x - c)$ is also monotone. Thus,

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

■

Proof [Implication direction of Proposition 8] We follow the general strategy of Uematsu and Lee (2012, Theorem 7). If $\mathcal{S}^{D,*}_{\mathrm{Bipart},\ell} \cap \mathcal{S}_{\mathrm{Decomp}} \neq \emptyset$,

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

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. ■

Proof [Corollary 9] By Proposition 8 and Corollary 5,

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

Further, by Equation 11,

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

The result follows because

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

■

Proof [Proposition 10] For any $s \colon \mathcal{X} \to \mathbb{R}$, $s_{\mathrm{Pair}} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define

$$\mathrm{regret}^D_\ell(s) = \mathbb{L}^D_\ell(s) - \inf_{t : \mathcal{X} \to \mathbb{R}} \mathbb{L}^D_\ell(t)$$

$$\mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}}) = \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(s_{\mathrm{Pair}}) - \inf_{t_{\mathrm{Pair}} \colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(t_{\mathrm{Pair}})$$
$$\mathrm{regret}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s) = \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)) - \inf_{t:\mathcal{X}\to\mathbb{R}} \mathbb{L}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(t)).$$

By the reduction of bipartite ranking to classification over pairs (Lemma 2),

$$\mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)) = \mathrm{regret}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s)).$$

Existing surrogate regret bounds for proper composite losses (Reid and Williamson, 2009) imply that there exists some convex $F_\ell : [0,1] \to \mathbb{R}_+$ such that

$$F_\ell\left(\mathrm{regret}^{\mathrm{Bipart}(D)}_{01}(\mathrm{Diff}(s))\right) \leq \mathrm{regret}^{\mathrm{Bipart}(D)}_{\ell}(\mathrm{Diff}(s)),$$

or equivalently,

$$F_\ell\left(\mathrm{regret}^{D}_{\mathrm{Bipart},01}(\mathrm{Diff}(s))\right) \leq \mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)).$$

Since ℓ^{01} and ℓ induce a decomposable pair-scorer, the minimisations in the two regrets above are effectively over $\mathcal{S}_{\mathrm{Decomp}}$, i.e.,

$$\mathrm{regret}^{D}_{\mathrm{Bipart},\ell}(\mathrm{Diff}(s)) = \mathrm{regret}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s).$$

Thus, we can write the regret bound as

$$F_\ell\left(\mathrm{regret}^{D,\mathrm{Univ}}_{\mathrm{Bipart},01}(s)\right) \leq \mathrm{regret}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s).$$

■

Proof [**Proposition 11**] We follow the general strategy of Uematsu and Lee (2012, Theorem 3). 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 the support of D 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)

$$\delta\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s;t) = \lim_{\epsilon\to 0} \frac{\mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s+\epsilon t) - \mathbb{L}^{D,\mathrm{Univ}}_{\mathrm{Bipart},\ell}(s)}{\epsilon}$$

$$= F'_{s,t}(0),$$

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,

$$\left| \frac{\phi((\mathrm{Diff}(s+\epsilon t))(x,x')) - \phi((\mathrm{Diff}(s))(x,x'))}{\epsilon} \right| \leq \left|\phi((\mathrm{Diff}(s+t))(x,x')) - \phi((\mathrm{Diff}(s))(x,x'))\right| \tag{13}$$
$$\leq \phi((\mathrm{Diff}(s+t))(x,x')) + \phi((\mathrm{Diff}(s))(x,x')),$$

where Equation 13 is because $\phi(a+\epsilon b) \leq \epsilon\phi(a+b) + (1-\epsilon)\phi(a)$ 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

$$F'_{s,t}(0) = \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[(t(\mathsf{X}) - t(\mathsf{X}')) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] \tag{14}$$
$$= \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}) - s(\mathsf{X}'))\right] - \mathbb{E}_{\mathsf{X} \sim Q, \mathsf{X}' \sim P}\left[t(\mathsf{X}) \cdot \phi'(s(\mathsf{X}') - s(\mathsf{X}))\right] \tag{15}$$
$$= \int_{\mathcal{X}} t(x) \cdot r(x)\, dx,$$

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 14 and 15, we relied on the fact that the individual terms are finite:

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

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

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

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

$$(\forall x, x' \in \mathcal{X})\, |t(x) - t(x')| \leq \frac{|\phi(t(x) - t(x')) - \phi(0)|}{|\phi'(0)|}$$

$$\leq \frac{1}{|\phi'(0)|} \cdot (\phi(t(x) - t(x')) + \phi(0)) \text{ by the triangle inequality.}$$

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

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

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,

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

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 $\mathbb{L}^{D,\text{Univ}}_{\text{Bipart},\ell}$. By convexity of $\mathbb{L}^{D,\text{Univ}}_{\text{Bipart},\ell}$, 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,

$$(\forall t \in \mathcal{L}(D)) \int_{\mathcal{X}} t(x) \cdot r(x)\, dx = 0. \tag{16}$$

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

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

where the last line is since $\phi(z) < \infty$ for every $z \in \mathbb{R}$. By assumption, Equation 16 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 $x_0 \in \mathcal{X}$,

$$\begin{aligned}
\frac{\eta(x_0)}{1-\eta(x_0)} \cdot \frac{1-\pi}{\pi} &= \frac{P(x_0)}{Q(x_0)} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\phi'(s^*(\mathsf{X}) - s^*(x_0))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[\phi'(s^*(x_0) - s^*(\mathsf{X}'))\right]} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\ell'_1(s^*(\mathsf{X}) - s^*(x_0)) - \ell'_{-1}(s^*(x_0) - s^*(\mathsf{X}))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[-\ell'_1(s^*(x_0) - s^*(\mathsf{X}')) + \ell'_{-1}(s^*(\mathsf{X}) - s^*(x_0))\right]} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\ell'_{-1}(s^*(x_0) - s^*(\mathsf{X})) - \ell'_1(s^*(\mathsf{X}) - s^*(x_0))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[\ell'_1(s^*(x_0) - s^*(\mathsf{X}')) - \ell'_{-1}(s^*(\mathsf{X}) - s^*(x_0))\right]} \\
&= \frac{\mathbb{E}_{\mathsf{X} \sim P}\left[\ell'_{-1}(s^*(x_0) - s^*(\mathsf{X}))\right]}{\mathbb{E}_{\mathsf{X}' \sim Q}\left[\ell'_1(s^*(x_0) - s^*(\mathsf{X}'))\right]} \text{ since } \ell \text{ is symmetric,}
\end{aligned}$$

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 12**] 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 13**] First, in the notation of the theorem statement,

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

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

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

Now consider

$$\begin{aligned} F'_{s,t}(0) &= \mathbb{E}_{\mathsf{X}' \sim Q}\left[g'\left(F^D_{s_{\mathrm{Pair}}}(\mathsf{X}') \right) \cdot \mathbb{E}_{\mathsf{X} \sim P}\left[t_{\mathrm{Pair}}(\mathsf{X}, \mathsf{X}') \cdot \frac{\ell'_1(s_{\mathrm{Pair}}(\mathsf{X}, \mathsf{X}'))}{2} + \right.\right. \\ &\qquad\qquad \left.\left. t_{\mathrm{Pair}}(\mathsf{X}', \mathsf{X}) \cdot \frac{\ell'_{-1}(s_{\mathrm{Pair}}(\mathsf{X}', \mathsf{X}))}{2} \right]\right] \\ &= \frac{1}{2} \int_{\mathcal{X} \times \mathcal{X}} t_{\mathrm{Pair}}(x, x') \cdot \big(P(x) Q(x') \cdot g'(F^D_{s_{\mathrm{Pair}}}(x')) \cdot \ell'_1(s_{\mathrm{Pair}}(x, x')) + \\ &\qquad\qquad P(x') Q(x) \cdot g'(F^D_{s_{\mathrm{Pair}}}(x)) \cdot \ell'_1(s_{\mathrm{Pair}}(x', x)) \big) \, dx \, dx', \end{aligned}$$

where as in the proof of Proposition 11, the interchange of derivative and expectation is justified when the support of D 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 14] We establish this by verifying that $s_{\text{Pair}} = \frac{1}{p+1}\text{Diff}(\sigma^{-1} \circ \eta)$ satisfies the implicit equation in Proposition 13. We begin with the term $A^D(x)$ as defined in Proposition 13. Plugging in $g^p(x) = x^p$ and

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

we get

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

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

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

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

with the result now a simple function of η, and

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

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 12, we get

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

Therefore $s_{\text{Pair}} = \frac{1}{p+1}\text{Diff}(\sigma^{-1} \circ \eta)$ satisfies the implicit equation of Proposition 13, 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

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

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 15. ■

Proof [**Proposition 15**] 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:

$$\begin{aligned} \mathbb{L}^D_{\mathrm{push},01,g}(\mathrm{Diff}(s)) &= \mathbb{E}_{\mathsf{X}' \sim Q}\left[g\left(\mathbb{E}_{\mathsf{X} \sim P}\left[\ell^{01}(1, s(\mathsf{X}) - s(\mathsf{X}'))\right]\right)\right] \\ &= \mathbb{E}_{\mathsf{X}' \sim Q}\left[g\left(\mathrm{FNR}^D_s(s(\mathsf{X}'))\right)\right] \\ &= \mathbb{E}_{\mathsf{X}' \sim Q}\left[\int_{-\infty}^{\infty} \delta_{s(\mathsf{X}')}(t) \cdot g\left(\mathrm{FNR}^D_s(t)\right)\, dt\right] \\ &= \int_{-\infty}^{\infty} \mathbb{E}_{\mathsf{X}' \sim Q}\left[\delta_{s(\mathsf{X}')}(t) \cdot g\left(\mathrm{FNR}^D_s(t)\right)\right]\, dt \\ &= \int_{-\infty}^{\infty} \Pr_{\mathsf{X}' \sim Q}\left[s(\mathsf{X}') = t\right] \cdot g\left(\mathrm{FNR}^D_s(t)\right)\, dt \\ &= \int_{-\infty}^{\infty} -(\mathrm{FPR}^D_s)'(t) \cdot g\left(\mathrm{FNR}^D_s(t)\right)\, dt \\ &= \int_0^1 g\left(\mathrm{FNR}^D_s((\mathrm{FPR}^D_s)^{-1}(\alpha))\right)\, d\alpha, \end{aligned}$$

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

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

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

$$\underset{s \colon \mathcal{X} \to \mathbb{R}, t \in \mathbb{R}}{\mathrm{Argmin}}\ g(\mathrm{FNR}^D_s(t)) : \mathrm{FPR}^D_s(t) = \alpha.$$

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 16**] By Proposition 14, the unique optimal pair-scorer is

$$s^*_{\text{Pair}} = \frac{1}{p+1}\text{Diff}(\sigma^{-1} \circ \eta) = \text{Diff}\left(\frac{1}{p+1}(\sigma^{-1} \circ \eta)\right),$$

which is decomposable. Corollary 5 may be adapted here to argue that any optimal univariate scorer s^* must satisfy $s^*_{\text{Pair}} = \text{Diff}(s^*)$, 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 18 *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 ($\impliedby$). This is easily verified by the definition of monotonicity.

($\implies$). We will construct such a monotone increasing ϕ. For any $y \in \text{Im}(g)$, 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 19 *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 19, $f = \phi_1 \circ g$ 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. Interpretation of (Uematsu and Lee, 2012) in terms of proper losses

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

Proposition 20 ((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 21 ((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 22 *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

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

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 23 *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 19, 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 12, 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 8 where ℓ is a margin loss.

Lemma 24 *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 D. Empirical illustration of Corollary 9

We present an empirical illustration of Corollary 9 for an *asymmetric* proper composite loss. We work with an instance space $\mathcal{X}$ comprising N isolated points $\{x_1, \ldots, x_n\}$. We assume a distribution

D such that the instance $M_i = \Pr[\mathsf{X} = x_i]$, and $\eta_i = \Pr[\mathsf{Y} = 1 | \mathsf{X} = x_i]$. A scorer $s \colon \mathcal{X} \to \mathbb{R}$ is then some vector in $\mathbb{R}^n$. Given a loss ℓ, the bipartite risk of the scorer s is

$$\begin{aligned}\mathbb{L}^{D}_{\ell,\mathrm{Bipart}}(s) &= \sum_{i=1}^{N} \sum_{j=1}^{N} [\eta_i(1-\eta_j) \cdot (\ell_1(s_i - s_j) + \ell_{-1}(s_j - s_i)] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} [\eta_i(1-\eta_j) \cdot (\ell_1(\langle s, e_i - e_j \rangle) + \ell_{-1}(\langle s, e_j - e_i \rangle)] ,\end{aligned}$$

where e_i is the ith standard basis vector in $\mathbb{R}^n$. The Bayes-optimal risk is simply the minimiser of the above objective, and may be computed by numerical optimisation.

We perform 10 repetitions of the following experiment: for $N = 10$ instances, we draw $\eta_i \sim \mathrm{Beta}(4, 3)$, $Z_i \sim \mathrm{Beta}(6, 2)$, and set $M_i = Z_i / \sum_j Z_j$. We then scale the η's to lie in $[0.01, 0.99]$, which is necessary to ensure the attainability of the risk minimiser. Given this distribution, we minimised the bipartite risk using L-BFGS, obtaining the Bayes-optimal scorer s^*. As the risk is invariant to translations, we transform the solution so that its minimum value equals $\Psi(0.01)$ (thus agreeing with that of the expected optimal solution). We collect the corresponding pairs of (η_i, s^*_i) values for all 10 repetitions. We then plot the graph of the resulting η values versus the s^* values. If s^* is a strictly monotone transform of η, then the plot will reflect this (as the different η values from the trials represent different sampling points of the domain of this function).

Figure 1 shows the results where ℓ is the p-classification loss for $p = 2$,

$$\ell(v) = \left(e^{-v}, \frac{1}{2} e^{2v} \right).$$

We see that the relationship between η and s^* is strictly monotone. Also shown on the graph is the plot of η versus $\Psi \circ \eta$, where $\Psi = \frac{1}{2}\sigma^{-1}$; this perfectly agrees with the observed s^* values, as predicted by the theory.

Appendix E. Empirical illustration of Corollary 12

We present an empirical illustration of Corollary 12, showing that for a proper composite loss whose Bayes-optimal pair-scorer is non-decomposable, the optimal univariate scorer is a strictly monotone transform of η, but that the transformation is distribution dependent. We repeat the setup of Appendix D, except that we now work with ℓ being the squared loss, $\ell(y, v) = (1 - yv)^2$, and the canonical boosting loss (Buja et al., 2005),

$$\ell(y, v) = \frac{yv}{2} + \sqrt{1 + \frac{v^2}{4}}.$$

Squared loss employs the identity link, while the canonical boosting loss uses the link $\Psi(\eta) = \frac{2\eta - 1}{\sqrt{\eta(1-\eta)}}$, and thus do not induce a decomposable pair-scorer.

Figure 2 shows that the relationship between η and s^* for these losses across multiple trials is *not* monotone, and significantly deviates from the optimal solution in the class-probability estimation setting, viz. $s^* = \Psi(\eta)$ for Ψ the identity mapping. This indicates that in general, the relationship between η and s^* is distribution dependent. Figures 3 and 4 further studies the relationship between the two quantities for each individual trial. We see that, for a given trial (or equivalently for a given distribution), the relationship between η and s^* is strictly monotone, as expected.

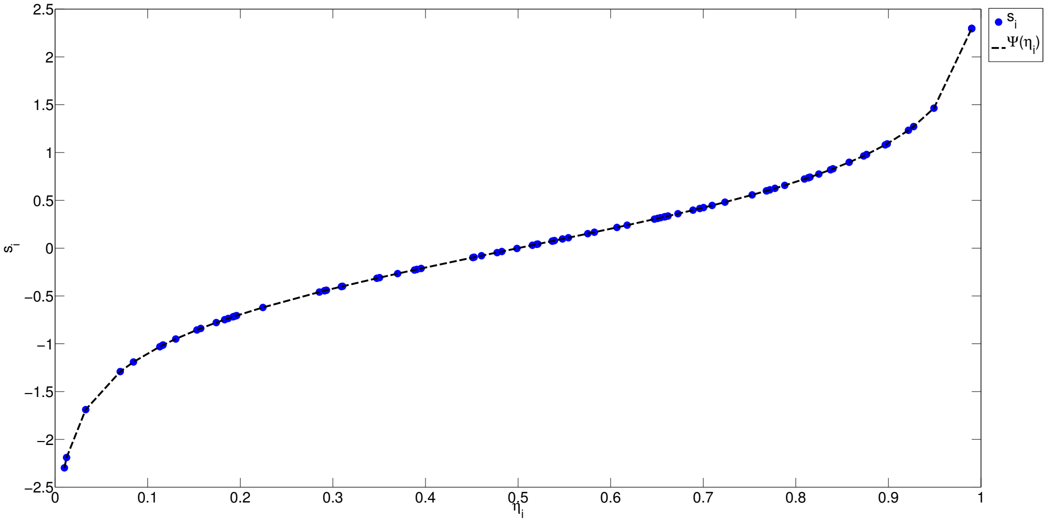

Figure 1: Results of 10 simulation trials to illustrate Corollary 9 for the case of an asymmetric loss. Here, the η_i and M_i values are varied across each trial, and the relationship between the (η, s^*) pairs across all trials is plotted. The relationship exactly matches that of $s^* = \Psi(\eta)$.

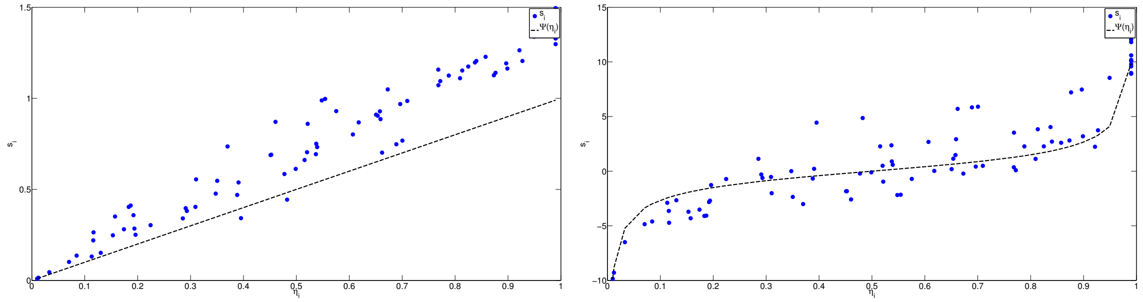

Figure 2: Results of 10 simulation trials to illustrate Proposition 11 for the case of squared and canonical boosting losses. Here, the η_i and M_i values are varied across each trial, and the relationship between the (η, s^*) pairs across all trials is plotted.

Appendix F. The p-classification loss and beyond

For any $p > 0$, the p-classification loss (Ertekin and Rudin, 2011)

$$\ell_{\exp,p}(v) = \left(\frac{1}{p} e^{vp}, e^{-v}\right)$$

is proper composite, with inverse link function

$$\Psi^{-1}(v) = \frac{1}{1 - \frac{\ell'_{\exp,p,1}(v)}{\ell'_{\exp,p,-1}(v)}} = \frac{1}{1 + e^{-(p+1)v}} = \sigma((p+1)v),$$

so that

$$\Psi(q) = \frac{1}{p+1}\sigma^{-1}(q) = \log\left(\frac{q}{1-q}\right)^{\frac{1}{p+1}}.$$

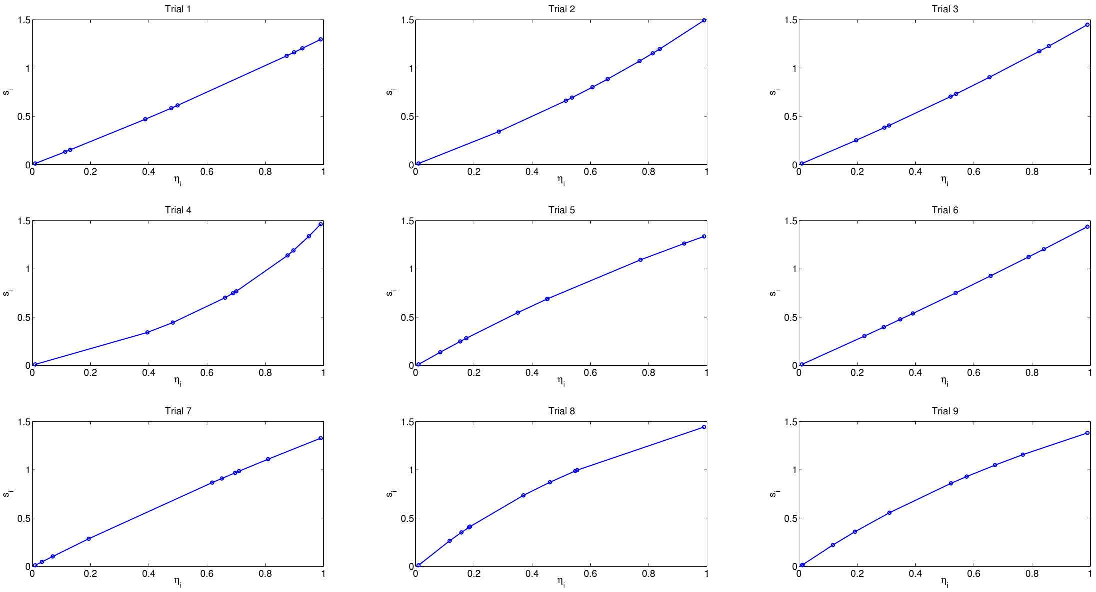

Figure 3: Results of 9 simulation trials to illustrate Proposition 11 for the case of squared loss. Here, the η_i and M_i values are varied across each trial, and each panel represents the relationship between η and s^* for a specific trial.

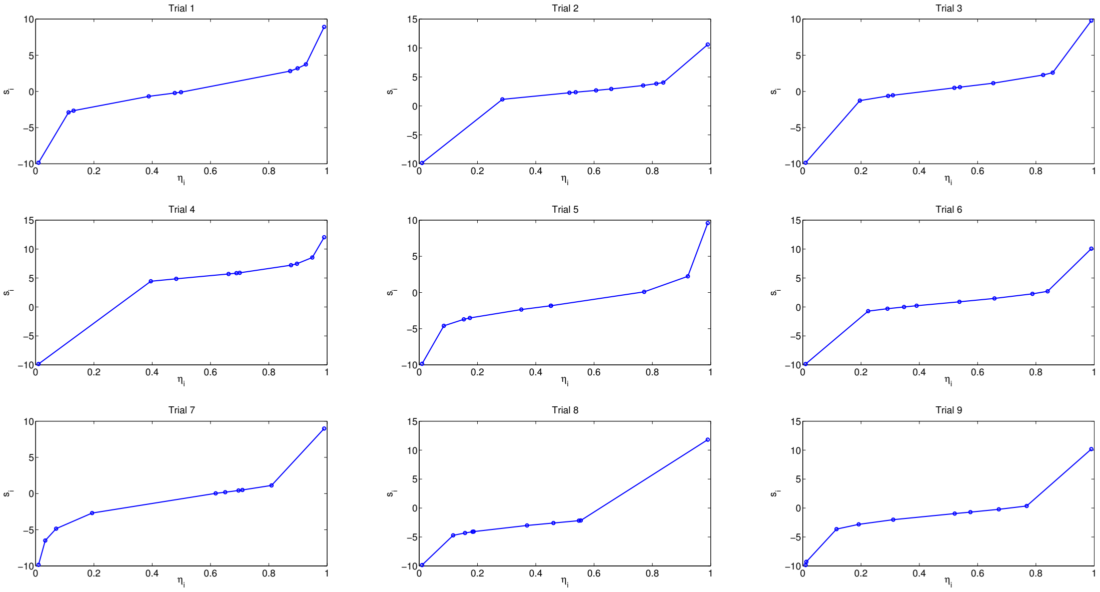

Figure 4: Results of 9 simulation trials to illustrate Proposition 11 for the case of canonical boosting loss. Here, the η_i and M_i values are varied across each trial, and each panel represents the relationship between η and s^* for a specific trial.

The underlying proper loss is

$$\lambda_p(q) = \ell_{\exp,p}(\Psi(q))$$

$$= \left(\left(\frac{1-q}{q} \right)^{\frac{1}{p+1}}, \frac{1}{p} \cdot \left(\frac{q}{1-q} \right)^{1-\frac{1}{p+1}} \right).$$

This is a generalised version of the boosting loss (the case $p = 1$).

The above proper loss may be understood via its *weight function* (Shuford Jr. et al., 1966; Reid and Williamson, 2010). Given a proper loss λ, its weight function $w : [0,1] \to \mathbb{R}_+$ lets one express λ as a weighted combination of cost-sensitive losses:

$$\lambda(y, v) = \int_0^1 w(c) \ell^{\mathrm{CS}(c)}(y, v) \, dc, \tag{17}$$

where $\ell^{\mathrm{CS}(c)}$ is the *cost-sensitive loss* with cost ratio c,

$$\ell^{\mathrm{CS}(c)}(y, v) = (1-c)[\![y = 1 \wedge v \leq c]\!] + c[\![y = -1 \wedge v > c]\!].$$

The weight function describes the relative importance paid to various cost ratios. When w places more weight on larger values of c, the loss intuitively favours accurate prediction of large values of η.

The proper loss corresponding to p-classification may be understood via its weight function,

$$\begin{aligned} w_p(c) &= -\frac{\lambda'_{p,1}(c)}{1-c} \text{ (by (Reid and Williamson, 2010, Theorem 1))} \\ &= \frac{1}{p+1} \cdot \frac{1}{c^{1+\frac{1}{p+1}}(1-c)^{2-\frac{1}{p+1}}}. \end{aligned}$$

As p increases[2], the loss is seen to place relatively more weight on larger values of c. Given the equivalence to the p-norm push risk, we thus have some insight as to how the risk encourages solutions to maximise accuracy at the head of the ranked list.

The weight function perspective suggests the construction of other proper composite losses suitable for maximising accuracy at the head of the ranked list. For example, consider the asymmetric weight function

$$w(c) = \frac{1}{c(1-c)^{3/2}}$$

which, when composed with the sigmoid link, yields the loss

$$\ell_A(v) = \left(\frac{2}{\sqrt{\sigma(-v)}}, 2\tanh^{-1}(\sqrt{\sigma(-v)}) \right).$$

As another example, consider the weight function

$$w(c) = \begin{cases} \frac{1}{2c^{3/2}(1-c)^{3/2}} & \text{if } c > \frac{1}{2} \\ \frac{1}{c(1-c)} & \text{else,} \end{cases}$$

[2] Note that as $p \to \infty$, we have the limiting weight function $w_\infty(c) = \frac{1}{c(1-c)^2}$, which results in unbounded partial losses. Nonetheless, the weight results in valid losses for every finite $p \geq 1$.

which, when composed with the sigmoid link, yields the loss

$$\ell_B(v) = \left(\begin{cases} \frac{1}{2} e^{v/2} & v > 0 \\ \log\left(\frac{1+e^{-v}}{2}\right) + \frac{1}{2} & v \leq 0 \end{cases}, \begin{cases} \frac{1}{2} e^{-v/2} & v > 0 \\ \log\left(\frac{1+e^{-v}}{2}\right) + \frac{1}{2} & v \leq 0 \end{cases} \right).$$

As with the p-classification loss, one can consider parameterised weight functions where the skew is varied. One may also consider weight functions that place bounded mass for values of c less than some parameterised threshold; for details, see (Menon and Williamson, 2014).

Appendix G. Experiments with the p-norm push

We present experiments that assess the efficacy of the proper composite losses in Table 1 (which we call "Asymmetric A" and "Asymmetric B") for the problem of maximising accuracy at the head of the ranked list. We consider all combinations of the three risk types considered in this paper – proper, bipartite, and p-norm push – and a selection of proper composite losses losses – logistic, exponential, p-classification, Asymmetric A, and Asymmetric B. The aim of our experiments is *not* to position the new losses as 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 four UCI datasets: `ionosphere`, `housing`, `german` and `car`. 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 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 $\lambda \in \{10^{-6}, 10^{-5}, \dots, 10^{1}\}$, and where appropriate the constant $p \in \{1, 2, 4, 8, 16, 32\}$. 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 – 6. No single method clearly outperforms all others in all metrics. However, we observe that the candidate proper composite losses are very competitive with p-classification and the p-norm push, as evidenced from the average ranks across all datasets reported in Table 7.

Method	AUC	MRR	DCG	AP	PTop
Proper Logistic	0.9113 ± 0.0208 (14)	0.0422 ± 0.0115 (11)	0.1966 ± 0.0108 (13)	0.9243 ± 0.0339 (14)	13.0000 ± 17.0880 (6)
Proper Exponential	0.9128 ± 0.0166 (13)	0.0482 ± 0.0078 (2)	0.2034 ± 0.0070 (2)	0.9262 ± 0.0318 (12)	12.8000 ± 12.9499 (7)
Proper P-Classification	0.9152 ± 0.0160 (5)	0.0426 ± 0.0106 (9)	0.1973 ± 0.0100 (11)	0.9349 ± 0.0232 (2)	11.6000 ± 8.8487 (9)
Proper Asymmetric A	0.9133 ± 0.0250 (11)	0.0470 ± 0.0089 (5)	0.2020 ± 0.0081 (4)	0.9336 ± 0.0360 (4)	**16.8000** ± **10.8720 (1)**
Proper Asymmetric B	0.9135 ± 0.0161 (10)	0.0470 ± 0.0064 (5)	0.2019 ± 0.0062 (5)	0.9249 ± 0.0289 (13)	9.6000 ± 12.4016 (11)
Bipartite Logistic	0.9157 ± 0.0195 (2)	**0.0492** ± **0.0047 (1)**	**0.2039** ± **0.0036 (1)**	0.9316 ± 0.0315 (7)	14.8000 ± 15.1228 (3)
Bipartite Exponential	0.9149 ± 0.0149 (7)	0.0424 ± 0.0116 (10)	0.1970 ± 0.0109 (12)	0.9292 ± 0.0292 (11)	13.0000 ± 12.7475 (6)
Bipartite P-Classification	0.9151 ± 0.0287 (6)	0.0473 ± 0.0111 (4)	0.2017 ± 0.0104 (6)	0.9294 ± 0.0361 (10)	15.6000 ± 14.6731 (2)
Bipartite Asymmetric A	**0.9176** ± **0.0187 (1)**	0.0475 ± 0.0094 (3)	0.2022 ± 0.0087 (3)	0.9343 ± 0.0298 (3)	13.6000 ± 13.0499 (5)
Bipartite Asymmetric B	0.9147 ± 0.0160 (8)	0.0448 ± 0.0078 (7)	0.1998 ± 0.0073 (9)	0.9308 ± 0.0277 (9)	13.6000 ± 14.2934 (5)
P-Norm Logistic	0.9129 ± 0.0182 (12)	0.0436 ± 0.0094 (8)	0.1988 ± 0.0084 (10)	0.9314 ± 0.0292 (8)	14.0000 ± 11.6833 (4)
P-Norm Exponential	0.9154 ± 0.0147 (4)	0.0460 ± 0.0081 (6)	0.2006 ± 0.0076 (7)	**0.9354** ± **0.0222 (1)**	12.0000 ± 9.5131 (8)
P-Norm Asymmetric A	0.9142 ± 0.0169 (9)	0.0395 ± 0.0105 (12)	0.1947 ± 0.0092 (14)	0.9330 ± 0.0275 (5)	13.0000 ± 10.9087 (6)
P-Norm Asymmetric B	0.9155 ± 0.0169 (3)	0.0448 ± 0.0083 (7)	0.2001 ± 0.0076 (8)	0.9317 ± 0.0263 (6)	11.4000 ± 11.8870 (10)

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

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.7597 ± 0.0415 (2)	0.0435 ± 0.0441 (6)	0.1924 ± 0.0364 (6)	0.1490 ± 0.0623 (8)	0.0000 ± 0.0000 (2)
Proper Exponential	0.7563 ± 0.0824 (3)	0.0540 ± 0.0436 (4)	0.2008 ± 0.0455 (3)	**0.1762 ± 0.0752 (1)**	**0.4000 ± 0.8944 (1)**
Proper P-Classification	0.7344 ± 0.0964 (9)	0.0324 ± 0.0149 (12)	0.1841 ± 0.0168 (11)	0.1404 ± 0.0628 (11)	0.0000 ± 0.0000 (2)
Proper Asymmetric A	0.7560 ± 0.0462 (4)	0.0398 ± 0.0301 (8)	0.1889 ± 0.0319 (7)	0.1475 ± 0.0630 (9)	0.0000 ± 0.0000 (2)
Proper Asymmetric B	**0.7633 ± 0.0362 (1)**	0.0495 ± 0.0627 (5)	0.1953 ± 0.0538 (5)	0.1496 ± 0.0612 (7)	0.0000 ± 0.0000 (2)
Bipartite Logistic	0.7280 ± 0.1085 (12)	0.0364 ± 0.0229 (10)	0.1828 ± 0.0209 (12)	0.1707 ± 0.0839 (5)	**0.4000 ± 0.8944 (1)**
Bipartite Exponential	0.7306 ± 0.0882 (10)	0.0635 ± 0.0536 (2)	0.2090 ± 0.0508 (2)	0.1740 ± 0.0837 (3)	**0.4000 ± 0.8944 (1)**
Bipartite P-Classification	0.7282 ± 0.0889 (11)	0.0372 ± 0.0212 (9)	0.1856 ± 0.0217 (9)	0.1704 ± 0.0841 (6)	**0.4000 ± 0.8944 (1)**
Bipartite Asymmetric A	0.7534 ± 0.0823 (7)	0.0356 ± 0.0189 (11)	0.1847 ± 0.0178 (10)	0.1754 ± 0.0760 (2)	**0.4000 ± 0.8944 (1)**
Bipartite Asymmetric B	0.7252 ± 0.1077 (13)	0.0418 ± 0.0375 (7)	0.1863 ± 0.0360 (8)	0.1709 ± 0.0840 (4)	**0.4000 ± 0.8944 (1)**
P-Norm Logistic	0.6987 ± 0.1159 (14)	0.0286 ± 0.0134 (14)	0.1802 ± 0.0202 (14)	0.1190 ± 0.0501 (14)	0.0000 ± 0.0000 (2)
P-Norm Exponential	0.7377 ± 0.0691 (8)	0.0320 ± 0.0197 (13)	0.1803 ± 0.0196 (13)	0.1442 ± 0.0621 (10)	0.0000 ± 0.0000 (2)
P-Norm Asymmetric A	0.7548 ± 0.0650 (6)	0.0559 ± 0.0716 (3)	0.1990 ± 0.0597 (4)	0.1403 ± 0.0592 (12)	0.0000 ± 0.0000 (2)
P-Norm Asymmetric B	0.7549 ± 0.0510 (5)	**0.0694 ± 0.0622 (1)**	**0.2127 ± 0.0568 (1)**	0.1390 ± 0.0543 (13)	0.0000 ± 0.0000 (2)

Table 4: Results of various “ranking the best” methods on `housing` dataset.

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Proper Logistic	0.8125 ± 0.0281 (6)	0.0156 ± 0.0036 (13)	0.1513 ± 0.0043 (12)	0.6245 ± 0.0629 (5)	2.4000 ± 1.9494 (4)
Proper Exponential	0.8131 ± 0.0311 (3)	0.0196 ± 0.0064 (4)	0.1547 ± 0.0069 (6)	0.6218 ± 0.0676 (10)	1.8000 ± 2.0494 (7)
Proper P-Classification	0.8115 ± 0.0282 (9)	0.0193 ± 0.0049 (6)	0.1552 ± 0.0043 (5)	0.6226 ± 0.0621 (8)	2.4000 ± 2.3022 (4)
Proper Asymmetric A	0.8130 ± 0.0284 (4)	0.0155 ± 0.0075 (14)	0.1497 ± 0.0080 (14)	0.6261 ± 0.0604 (2)	**3.2000** ± **2.2804 (1)**
Proper Asymmetric B	0.8129 ± 0.0274 (5)	0.0177 ± 0.0096 (8)	0.1533 ± 0.0096 (7)	0.6246 ± 0.0627 (4)	2.0000 ± 2.0000 (6)
Bipartite Logistic	0.8140 ± 0.0295 (2)	0.0166 ± 0.0034 (10)	0.1528 ± 0.0041 (10)	0.6241 ± 0.0676 (6)	2.6000 ± 2.7019 (3)
Bipartite Exponential	0.8118 ± 0.0268 (8)	**0.0219** ± **0.0058 (1)**	0.1563 ± 0.0044 (2)	0.6216 ± 0.0631 (11)	2.4000 ± 1.9494 (4)
Bipartite P-Classification	0.8131 ± 0.0298 (3)	0.0216 ± 0.0037 (2)	**0.1569** ± **0.0034 (1)**	0.6245 ± 0.0657 (5)	2.2000 ± 1.6432 (5)
Bipartite Asymmetric A	0.8131 ± 0.0306 (3)	0.0200 ± 0.0045 (3)	0.1554 ± 0.0049 (4)	0.6224 ± 0.0709 (9)	2.6000 ± 2.7019 (3)
Bipartite Asymmetric B	0.8115 ± 0.0255 (9)	0.0191 ± 0.0048 (7)	0.1532 ± 0.0054 (8)	0.6209 ± 0.0601 (12)	2.2000 ± 1.7889 (5)
P-Norm Logistic	0.8140 ± 0.0286 (2)	0.0157 ± 0.0036 (12)	0.1510 ± 0.0051 (13)	0.6258 ± 0.0625 (3)	2.8000 ± 2.5884 (2)
P-Norm Exponential	0.8105 ± 0.0277 (10)	0.0195 ± 0.0047 (5)	0.1556 ± 0.0055 (3)	0.6206 ± 0.0614 (13)	2.0000 ± 2.0000 (6)
P-Norm Asymmetric A	0.8121 ± 0.0291 (7)	0.0168 ± 0.0044 (9)	0.1531 ± 0.0053 (9)	0.6236 ± 0.0634 (7)	2.2000 ± 2.1679 (5)
P-Norm Asymmetric B	**0.8148** ± **0.0282 (1)**	0.0159 ± 0.0019 (11)	0.1521 ± 0.0018 (11)	**0.6276** ± **0.0628 (1)**	2.2000 ± 2.1679 (5)

Table 5: Results of various “ranking the best” methods on `german` dataset.

Method	AUC	MRR	DCG	AP	PTop
Proper Logistic	0.9976 ± 0.0012 (2)	**0.0237** ± **0.0261 (1)**	**0.1506** ± **0.0226 (1)**	0.9391 ± 0.0370 (4)	**13.2000** ± **3.9623 (1)**
Proper Exponential	0.9976 ± 0.0012 (2)	0.0080 ± 0.0017 (12)	0.1327 ± 0.0046 (12)	0.9376 ± 0.0339 (6)	12.8000 ± 3.9623 (3)
Proper P-Classification	0.9968 ± 0.0022 (5)	0.0099 ± 0.0048 (8)	0.1366 ± 0.0073 (7)	0.9316 ± 0.0394 (10)	12.8000 ± 3.9623 (3)
Proper Asymmetric A	0.9976 ± 0.0012 (2)	0.0157 ± 0.0180 (3)	0.1416 ± 0.0187 (5)	0.9385 ± 0.0367 (5)	13.0000 ± 3.9370 (2)
Proper Asymmetric B	0.9974 ± 0.0009 (3)	0.0075 ± 0.0034 (13)	0.1313 ± 0.0081 (13)	0.9297 ± 0.0309 (12)	12.0000 ± 4.4721 (6)
Bipartite Logistic	0.9976 ± 0.0013 (2)	0.0106 ± 0.0098 (7)	0.1347 ± 0.0117 (8)	0.9371 ± 0.0375 (7)	12.8000 ± 3.9623 (3)
Bipartite Exponential	0.9976 ± 0.0012 (2)	0.0089 ± 0.0055 (9)	0.1345 ± 0.0103 (9)	0.9364 ± 0.0348 (8)	12.2000 ± 4.1473 (5)
Bipartite P-Classification	**0.9977** ± **0.0012 (1)**	0.0063 ± 0.0027 (14)	0.1293 ± 0.0062 (14)	0.9394 ± 0.0340 (2)	12.4000 ± 4.1593 (4)
Bipartite Asymmetric A	**0.9977** ± **0.0012 (1)**	0.0081 ± 0.0028 (11)	0.1337 ± 0.0047 (10)	**0.9401** ± **0.0350 (1)**	**13.2000** ± **3.9623 (1)**
Bipartite Asymmetric B	0.9976 ± 0.0011 (2)	0.0085 ± 0.0036 (10)	0.1335 ± 0.0083 (11)	0.9392 ± 0.0337 (3)	13.0000 ± 4.0000 (2)
P-Norm Logistic	0.9976 ± 0.0013 (2)	0.0170 ± 0.0094 (2)	0.1437 ± 0.0109 (3)	0.9392 ± 0.0384 (3)	**13.2000** ± **3.9623 (1)**
P-Norm Exponential	0.9968 ± 0.0021 (5)	0.0119 ± 0.0060 (6)	0.1397 ± 0.0102 (6)	0.9307 ± 0.0370 (11)	12.8000 ± 3.9623 (3)
P-Norm Asymmetric A	0.9966 ± 0.0022 (6)	0.0150 ± 0.0097 (5)	0.1418 ± 0.0109 (4)	0.9289 ± 0.0382 (13)	13.0000 ± 3.9370 (2)
P-Norm Asymmetric B	0.9972 ± 0.0014 (4)	0.0151 ± 0.0088 (4)	0.1444 ± 0.0140 (2)	0.9333 ± 0.0340 (9)	12.4000 ± 3.9115 (4)

Table 6: Results of various “ranking the best” methods on `car` dataset.

Method	**AUC**	**MRR**	**DCG**	**AP**	**PTop**
Proper Logistic	6.0000	7.7500	8.0000	7.7500	3.2500
Proper Exponential	5.2500	**5.5000**	5.7500	7.2500	4.5000
Proper P-Classification	7.0000	8.7500	8.5000	7.7500	4.5000
Proper Asymmetric A	5.2500	7.5000	7.5000	5.0000	**1.5000**
Proper Asymmetric B	4.7500	7.7500	7.5000	9.0000	6.2500
Bipartite Logistic	4.5000	7.0000	7.7500	6.2500	2.5000
Bipartite Exponential	6.7500	**5.5000**	6.2500	8.2500	4.0000
Bipartite P-Classification	5.2500	7.2500	7.5000	5.7500	3.0000
Bipartite Asymmetric A	**3.0000**	7.0000	6.7500	**3.7500**	2.5000
Bipartite Asymmetric B	8.0000	7.7500	9.0000	7.0000	3.2500
P-Norm Logistic	7.5000	9.0000	10.0000	7.0000	2.2500
P-Norm Exponential	6.7500	7.5000	7.2500	8.7500	4.7500
P-Norm Asymmetric A	7.0000	7.2500	7.7500	9.2500	3.7500
P-Norm Asymmetric B	3.2500	5.7500	**5.5000**	7.2500	5.2500

Table 7: Average ranks of various “ranking the best” methods for each performance measure across all datasets.