Noname manuscript No.
(will be inserted by the editor)

The risk of trivial solutions in bipartite top ranking

Aditya Krishna Menon

Received: date / Accepted: date

Abstract Given a sample of instances with binary labels, the bipartite top ranking problem is to produce a ranked list of instances whose *head* is dominated by positives. One popular existing approach to this problem is based on constructing surrogates to a performance measure known as the fraction of positives of the top (PTop). In this paper, we theoretically show that the measure and its surrogates have an undesirable property: for certain noisy distributions, it is optimal to trivially predict *the same score for all instances*. We propose a simple rectification which avoids such trivial solutions, while still focussing on the head of the ranked list and being as easy to optimise.

Keywords Bipartite ranking, top rank optimisation, positives at the top

1 Bipartite (top) ranking problems

Given a sample of instances endowed with binary labels, bipartite ranking is the problem of learning a real-valued scoring function for instances with maximal area under the ROC curve (AUC) (Freund et al, 2003; Agarwal et al, 2005; Agarwal and Niyogi, 2005; Clémençon et al, 2008; Uematsu and Lee, 2012; Gao and Zhou, 2015). By ranking instances according to their scores under a model with high AUC, most positives will be ranked higher than most negatives. This problem has use in many real-world contexts, such as ranking movies according to whether or not a user likes them, and ranking patients according to how likely they are to have a disease.

In many practical settings, interest is usually reserved for the *head* of the ranked list. We shall refer to this as the *bipartite top ranking* problem. For example, when presented with a ranked list of movies, users will typically focus on just the top few results (Järvelin and Kekäläinen, 2000). Similarly, owing to time and monetary costs, doctors can often interact with only those few patients deemed to have highest chance of being ill. Scorers with good AUC will not necessarily have maximal accuracy at the head of the ranked list (Yue et al, 2007; Li et al, 2014b), thus prompting much recent algorithmic effort explicitly targetting

Aditya Krishna Menon
The Australian National University
E-mail: `aditya.menon@anu.edu.au`

the top ranking regime (Rudin, 2009; Agarwal, 2011; Boyd et al, 2012; Narasimhan and Agarwal, 2013; Yun et al, 2014; Li et al, 2014b; Kar et al, 2015; Liu et al, 2015).

One popular measure designed for the top ranking regime is the fraction of positives of the top (PTop) (Agarwal, 2011; Boyd et al, 2012), which computes the fraction of positive instances ranked above *all* negative instances. Intuitively, this penalises scorers which place a negative instance near the head of the list. This measure has inspired convex surrogates which are tractable to optimise, and have demonstrated strong empirical performance (Agarwal, 2011; Rakotomamonjy, 2012; Li et al, 2014b).

Despite these impressive successes, theoretical aspects of the PTop have received less attention. For example, it is natural to ask whether one can establish settings under which PTop optimisation is guaranteed to produce solutions that are superior to standard AUC optimisation. One entry point to studying this issue is the following: assuming we have access to an *infinite* number of samples, and an *arbitrarily powerful* scorer, how does the scorer that optimises the PTop compare to that which optimises the AUC?

In this paper, we theoretically analyse this question, with a surprising conclusion: when our underlying samples have even a tiny amount of noise, the PTop measure and its surrogates asymptotically have a trivial optimal solution that assigns the *same score to every instance* (Proposition 1); thus, optimising a PTop surrogate may be *harmful* in certain top ranking settings. Given the importance of the top ranking regime, and the fact that a non-trivial body of recent work has focussed on the PTop and extensions (Agarwal, 2011; Rakotomamonjy, 2012; Li et al, 2014b), we believe that this result is of interest, and warrants the design of alternate top ranking measures that avoid this issue.

While our result indicates a flaw with the PTop, it is fortunately simple to resolve: we propose a simple rectification of the PTop (§4.2) which provably avoids trivial solutions (Propositions 2, 3), while being as easy to optimise (Proposition 5). Thus, these *rectified PTop* surrogates retain the strengths but eliminate the weakness of standard PTop surrogates. We empirically validate these theoretical findings (§6), demonstrating that both on synthetic and real-world datasets, even when learning with a finite sample, the optimal PTop solution can be trivial or highly sub-optimal, unlike the optimal RectTop solution.

In summary, we provide two main contributions **C1** and **C2**:

C1: we show that for certain noisy distributions, the PTop measure and its surrogates have a trivial optimal solution that assigns the same score to every instance (Proposition 1).

C2: we propose a simple rectification of the PTop (§4.2) which provably avoids trivial solutions (Propositions 2, 3), while being as easy to optimise (Proposition 5).

2 Background and notation

We begin with some background and notation. See Table 1 for a glossary.

Distributions, losses, and risks. Fix a finite instance space $\mathcal{X}$.[1] Denote by D a distribution over $\mathcal{X} \times \{\pm 1\}$ with support $\mathrm{supp}(D)$, and random variables $(\mathsf{X}, \mathsf{Y}) \sim D$. Any D may be decomposed into $(P, Q, \pi) = (\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = -1), \mathbb{P}(\mathsf{Y} = 1))$ or $(M, \eta) = (\mathbb{P}(\mathsf{X}), \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x))$. We call P, Q the class-conditional distributions, and η the class-probability function. We will write $D = (P, Q, \pi)$ or $D = (M, \eta)$ as appropriate. We call D *separable* if the labels are deterministic i.e., $(\forall x \in \mathcal{X})\, \eta(x) \in \{0, 1\}$.

A *loss* is any $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. A *margin loss* is any $\ell(y, v) = \phi(yv)$ for some non-increasing $\phi \colon \mathbb{R} \to \mathbb{R}_+$; we interchangeably write such an ℓ using its underlying ϕ. Examples

[1] This assumption removes the need for various measure-theoretic details in the proofs and analysis.

Symbol	Meaning	Symbol	Meaning
R_{bal}	Balanced classification risk	D	Joint distribution
R_{rank}	Bipartite ranking risk	P, Q	Class-conditionals
R_{ptop}	PTop risk	η	Class-probability
R_{rtop}	Rectified PTop risk	ϕ, ℓ	(Margin) Loss

Table 1 Glossary of commonly used symbols in the paper.

include the 0-1 loss, which for indicator function $\llbracket \cdot \rrbracket$ is $\ell^{01}(y, v) \doteq \llbracket yv < 0 \rrbracket + \frac{1}{2} \cdot \llbracket v = 0 \rrbracket$, hinge loss $\ell(y, v) = \max(0, 1 - yv)$, and exponential loss $\ell(y, v) = e^{-yv}$.

A *scorer* is any $f \colon \mathcal{X} \to \mathbb{R}$. Given a distribution D and loss ℓ, a *risk* is any $R(\cdot; D, \ell) \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}_+$ which quantifies a scorer's performance. A *Bayes-optimal scorer* for a risk is any minimiser $f^* \in \operatorname{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} R(f; D, \ell)$. For example, the 0-1 loss has as Bayes-optimal scorer any f^* with $\operatorname{sign}(f^*(x)) = \operatorname{sign}(2\eta(x) - 1)$, so that instances which are on average labelled positive are assigned a non-negative score.

To make these ideas concrete, we now define three fundamental learning problems on binary labels. Each seeks a scorer $f \colon \mathcal{X} \to \mathbb{R}$, but with a different risk being minimised.

Binary classification. Binary classification concerns learning a scorer $f \colon \mathcal{X} \to \mathbb{R}$ whose sign agrees with the label of an instance. Formally, the *ℓ-risk* of f under $D = (P, Q, \pi)$ is

$$R_{\mathrm{class}}(f; D, \ell) \doteq \pi \cdot \mathbb{E}_{\mathsf{X} \sim P}\left[\ell(+1, f(\mathsf{X}))\right] + (1 - \pi) \cdot \mathbb{E}_{\mathsf{X} \sim Q}\left[\ell(-1, f(\mathsf{X}))\right]. \tag{1}$$

Then, we seek a scorer with small *misclassification error* $R_{\mathrm{class}}(f; D, \ell^{01})$. On a finite sample $\mathsf{S} \sim D^N$ with positive instances $\{x_i^+\}_{i=1}^{n_+}$ and negative instances $\{x_j^-\}_{j=1}^{n_-}$, as a proxy one can minimise an *empirical surrogate risk* for suitable convex ℓ (Bartlett et al, 2006; Scott, 2012),

$$R_{\mathrm{class}}(f; \mathsf{S}, \ell) = \frac{1}{N} \sum_{i=1}^{n_+} \ell(+1, f(x_i^+)) + \frac{1}{N} \sum_{j=1}^{n_-} \ell(-1, f(x_j^-)). \tag{2}$$

When $\min(\pi, 1 - \pi) \ll 1/2$, it is common to instead use the *balanced ℓ-risk*,

$$R_{\mathrm{bal}}(f; D, \ell) \doteq \mathbb{E}_{\mathsf{X} \sim P}\left[\ell(+1, f(\mathsf{X}))\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\ell(-1, f(\mathsf{X}))\right], \tag{3}$$

which for ℓ^{01} is known as the *balanced error* (Chan and Stolfo, 1998; Brodersen et al, 2010).

Class-probability estimation. Class-probability estimation concerns learning a scorer $f \colon \mathcal{X} \to \mathbb{R}$ that is an invertible transformation of η. Formally, we seek a scorer with small risk $R_{\mathrm{class}}(f; D, \ell)$ for ℓ whose Bayes-optimal scorer is $f^* = \Psi \circ \eta$ for some invertible $\Psi \colon (0, 1) \to \mathbb{R}$. Such ℓ are called *strictly proper composite* with *link function* Ψ (Reid and Williamson, 2010). Examples include the logistic loss $\ell(y, v) = \log(1 + e^{-yv})$ with $\Psi(u) = \log \frac{u}{1-u}$, and exponential loss $\ell(y, v) = e^{-yv}$ with $\Psi(u) = {}^1\!/_2 \cdot \log \frac{u}{1-u}$. On a finite sample, one can minimise the empirical ℓ-risk per Equation 2.

Bipartite ranking: from AUC to PTop. Bipartite ranking concerns learning a scorer $f \colon \mathcal{X} \to \mathbb{R}$ that ranks the positives above the negatives (Freund et al, 2003; Agarwal et al, 2005). Intuitively, this involves ordering instances according to the values of the underlying class-probability η. Formally, for margin loss ϕ, define

$$R_{\mathrm{rank}}(f; D, \phi) \doteq \mathbb{E}_{\mathsf{X} \sim P, \mathsf{X}' \sim Q}\left[\phi(f(\mathsf{X}) - f(\mathsf{X}'))\right]. \tag{4}$$

Then, we seek a scorer with small *pairwise disagreement*, $R_{\mathrm{rank}}(f; D, \ell^{01})$, which is also one minus the *area under the ROC curve* (*AUC*) of f. We refer to $R_{\mathrm{rank}}(\cdot; D, \phi)$ as the *AUC ϕ-risk*.

In practice, performance at the *head* of the ranking induced by f is crucial (Clémençon and Vayatis, 2007), which we call the *top ranking* regime. Intuitively, in contrast to bipartite ranking, we seek to only accurately order those instances with large values of η. One popular measure for this regime is the *fraction of positives at the top* (*PTop*) (Agarwal, 2011),

$$R_{\text{ptop}}(f; \mathsf{S}, \phi) \doteq \max_{1 \leq j \leq n_-} \frac{1}{n_+} \sum_{i=1}^{n_+} \phi(f(x_i^+) - f(x_j^-)). \tag{5}$$

Then, the PTop for the scorer f is one minus $R_{\text{ptop}}(f; \mathsf{S}, \ell^{01})$, *viz.* the fraction of positives that are ranked above the highest negative; this strongly penalises errors at the head of the ranking induced by f (Li et al, 2014b). We will refer to $R_{\text{ptop}}(f; \mathsf{S}, \phi)$ as the empirical PTop ϕ-risk. Compared to other top ranking measures such as the average precision and discounted cumulative gain, one appeal of the PTop is that it is simple to optimise (Li et al, 2014b).

3 The risk of trivial PTop optimisers

Our first contribution is to show that for non-separable distributions, the PTop is asymptotically optimised by trivially assigning the same score to all instances. To begin, we provide the distributional version to the PTop.

3.1 Distributional version of PTop

The definition of the PTop in Equation 5 was only on a finite sample S. Assuming $\mathsf{S} \sim D^N$, we may view Equation 5 as the empirical version of the risk (Agarwal, 2011, Equation 5.21)

$$R_{\text{ptop}}(f; D, \phi) \doteq \max_{x' \in \text{supp}(Q)} \mathbb{E}_{\mathsf{X} \sim P} \left[\phi(f(\mathsf{X}) - f(x'))\right], \tag{6}$$

which we refer to as the *PTop ϕ-risk*.

Explicating this risk has two advantages. First, it makes transparent the distinction to the AUC. For a scorer f and instance x', let

$$\begin{aligned} \text{rank}^+(f, x'; D) &\doteq \mathbb{E}_{\mathsf{X} \sim P} \left[\ell^{01}(+1, f(\mathsf{X}) - f(x'))\right] \\ &= \mathbb{P}_{\mathsf{X} \sim P} \left(f(\mathsf{X}) < f(x')\right) + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X} \sim P} \left(f(\mathsf{X}) = f(x')\right), \end{aligned} \tag{7}$$

viz. the fraction of positives scoring lower than x', plus a penalty for ties per Agarwal (2011, Footnote 3). It is desirable to ensure that $\text{rank}^+(f; x', D)$ is small for negative instances x'. Observe now that

$$\begin{aligned} R_{\text{rank}}(f; D, \ell^{01}) &= \mathbb{E}_{\mathsf{X}' \sim Q} \left[\text{rank}^+(f, \mathsf{X}'; D)\right] \\ R_{\text{ptop}}(f; D, \ell^{01}) &= \max_{x' \in \text{supp}(Q)} \text{rank}^+(f, x'; D), \end{aligned} \tag{8}$$

i.e. the AUC seeks *most* negatives to have low rank, but the PTop seeks *every* negative to have low rank; the latter thus more directly targets the top ranking regime.

Second, with Equation 6 in place, we can study the Bayes-optimal scorers for the PTop with a general ϕ, which provides some insight into the risk.

3.2 Bayes-optimal scorers for the PTop risk

The Bayes-optimal scorers for a risk are those scorers one theoretically converges to, given infinite data and an arbitrarily flexible scorer class. While neither assumption is practically relevant, these scorers are nonetheless a useful theoretical device, and examining their form can offer insight as to the (un)suitability of a risk for a particular problem. In particular, if the Bayes-optimal scorer does not have a sensible form, it is an indication that the finite-sample and restricted scorer risk minimiser may perform sub-optimally.

For example, in binary classification, a basic restriction imposed on a surrogate loss ℓ is that it is *classification-calibrated* (Bartlett et al, 2006), i.e., that the Bayes-optimal scorer agrees with that of the 0-1 loss. This can be seen as a necessary condition to establishing statistical consistency of surrogate loss minimisation; if the Bayes-optimal solution of the loss does not agree with that of the 0-1 loss, then one cannot possibly establish convergence of a finite sample solution to the optimal one for 0-1 loss.

In bipartite ranking, the set of Bayes-optimal scorers for the AUC comprise all strictly increasing transformations of the underlying class-probability η (Clémençon et al, 2008). This indicates why optimising the AUC may be sub-optimal for top ranking problems: there is no preference for accurately modelling larger values of η.

We now show that the Bayes-optimal scorers for the PTop ϕ-risk ostensibly fare better.

Proposition 1 *Pick any distribution $D = (M, \eta)$ and non-increasing margin loss $\phi \colon \mathbb{R} \to \mathbb{R}_+$ with attainable minimum and $\phi(0) < \phi(0^-)$. Pick $f^* \colon \mathcal{X} \to \mathbb{R}$ with*

$$(\forall x \in \mathcal{X})\, f^*(x) \in \begin{cases} \underset{v \in \mathbb{R}}{\operatorname{argmin}}\, \phi(v) & \textit{if } \eta(x) = 1 \\ \{0\} & \textit{if } \eta(x) \in (0,1) \\ (-\infty, 0] & \textit{if } \eta(x) = 0. \end{cases} \tag{9}$$

Then, $\underset{f \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}}\, R_{\text{ptop}}(f; D, \phi) = \{f^* + C \mid C \in \mathbb{R}\}$.

The assumption that $\phi(0) < \phi(0^-)$ implies that for the 0-1 loss, we must have $\phi(0) < 1$. This is guaranteed by our definition of the loss, for which $\phi(0) = 1/2$.

Proposition 1 is best illustrated through some examples.

Example 1 For $\phi = \ell^{01}$, we have optimal scorer

$$f^*(x) \in \begin{cases} (0, \infty) & \text{if } \eta(x) = 1 \\ \{0\} & \text{if } \eta(x) \in (0,1) \\ (-\infty, 0] & \text{if } \eta(x) = 0. \end{cases}$$

This makes concrete the intuition that the PTop focusses on the head of the ranked list: it seeks to discriminate only those instances that are *deterministically positive* from the rest. One does not expend effort trying to order other instances, in contrast to the AUC.

Example 2 For $\phi(v) = e^{-v}$, with unattainable minimum, a limiting optimal scorer is

$$f^*(x) \in \begin{cases} +\infty & \text{if } \eta(x) = 1 \\ \{0\} & \text{if } \eta(x) \in (0,1) \\ (-\infty, 0] & \text{if } \eta(x) = 0. \end{cases}$$

By contrast, when using the exponential loss for the classification risk $R_{\text{class}}(f; D, \phi)$, as in AdaBoost, $f^*(x) = \frac{1}{2} \log \frac{\eta(x)}{1-\eta(x)}$ (Buja et al, 2005, Equation 8).

3.3 The risk of trivial Bayes-optimal scorers

Our argument for the sensibility of the PTop in Examples 1 and 2 assumed that *there exist* deterministically positive instances, i.e., $1 \in \mathrm{Im}(\eta)$. However, Proposition 1 equally applies when this assumption does *not* hold, so that $\mathrm{Im}(\eta) \subseteq [0, 1)$. Here, we see a different picture: Equation 9 implies that one optimal solution is the trivial constant scorer

$$(\forall x \in \mathcal{X})\, f^*(x) = 0,$$

which is is in fact the *only* optimal scorer if further $\mathrm{Im}(\eta) \subseteq (0, 1)$. This trivial optimal scorer will not distinguish between the actual highly ranked elements at all!

Distributions with $\mathrm{Im}(\eta) \subseteq (0, 1)$ may arise as a result of the common scenario where labels are subject to noise; the following explicates one such standard learning setup, where the true-class probability comprises a linear score passed through a nonlinear link.

Corollary 1 *Pick any finite $\mathcal{X} \subset \mathbb{R}^d$ and distribution $D = (M, \eta)$ with $\eta(x) = u(\langle w^*, x\rangle)$ for some $w^* \in \mathbb{R}^d$ and strictly monotone $u \colon \mathbb{R} \to (0, 1)$. Then, for any non-increasing $\phi \colon \mathbb{R} \to \mathbb{R}_+$ with $\phi(0) < \phi(0^-)$, $\mathbf{0} \in \underset{w}{\operatorname{argmin}}\, R_{\mathrm{ptop}}(w; D, \phi)$ for the all-zeros vector $\mathbf{0} \in \mathbb{R}^d$.*

3.4 Discussion of results

Both Proposition 1 and Corollary 1 are to our knowledge novel, and show that minimising the (surrogate) PTop risk can produce trivial solutions. In hindsight, this property is clear by the mere definition of the PTop in Equation 6: when P and Q have overlapping support, our maximum will compare every pair of points, in which case it is optimal to make all scores the same. This suggests a potential issue with both *measuring* top ranking performance with the standard ℓ^{01} PTop risk, and *attaining* good top ranking performance by optimising a surrogate PTop risk.

Nonetheless, a natural concern is that these results are of no practical significance. After all, they are statements about the *distributional minimiser* over *all possible scorers*. Both these qualifiers deserve comment. First, are the results relevant if we optimise over a restricted class of scorers (e.g. linear scorers)? In fact, if our scorer class contains constant scorers (which is trivially true for linear scorers), then the optimal PTop scorer within our class will match the Bayes-optimal scorer, i.e. $f^* \equiv 0$. Importantly, this holds even if the scorer class is incapable of modelling the true η. We will see an example of this in §6.1.

Second, are the results relevant if we optimise on on an empirical sample S, as is always true in practice? Here, the optimal solution may indeed be non-trivial, because the *empirical* distribution is often separable (as in a finite sample the various instances are typically distinct). However, our results imply that for non-separable distributions, *as we increase the number of samples*[2] *we will converge to a trivial scorer*. Such "anti-consistency" is clearly undesirable for any learning method. We will see an example of this in §6.2.

Fundamentally, a modification of PTop that avoids its behaviour for non-separable distributions while preserving its behaviour for separable distributions would be *strictly preferable* to work with. We now provide one such simple modification to the risk.

[2] Our result assumes iid draws of instances *and* labels. If however one fixes the set of instances and only draws labels randomly, the optimal scorers may depart from the trivial form implied by Proposition 1.

Risk	Expression	Risk	Expression
PTop	$\mathop{\mathbb{E}}_{P} e^{-f(\mathsf{X})} + \mathop{\mathbb{E}}_{Q} \begin{cases} +\infty & \text{if } f(\mathsf{X}) > 0 \\ 0 & \text{else} \end{cases}$	RectTop	$\mathop{\mathbb{E}}_{P} e^{-f(\mathsf{X})} + \mathop{\mathbb{E}}_{Q} \begin{cases} +\infty & \text{if } f(\mathsf{X}) > 0 \\ e^{f(\mathsf{X})} & \text{else} \end{cases}$
PTop_{App}	$\mathop{\mathbb{E}}_{P} e^{-f(\mathsf{X})} + \mathop{\mathbb{E}}_{Q} \begin{cases} \frac{1}{p} e^{p \cdot f(\mathsf{X})} & \text{if } f(\mathsf{X}) > 0 \\ \frac{1}{p} e^{p \cdot f(\mathsf{X})} & \text{else} \end{cases}$	$\text{RectTop}_{\text{App}}$	$\mathop{\mathbb{E}}_{P} e^{-f(\mathsf{X})} + \mathop{\mathbb{E}}_{Q} \begin{cases} \frac{1}{p}(e^{p \cdot f(\mathsf{X})} - 1) & \text{if } f(\mathsf{X}) > 0 \\ e^{f(\mathsf{X})} & \text{else} \end{cases}$

Table 2 Comparison of (approximate) PTop risk and proposed rectification (RectTop) for exponential loss.

4 RectTop: a rectification of the PTop

Our second contribution is a simple rectification of PTop that dispels trivial Bayes-optimal scorers. Our proposal relies on relating the PTop and a balanced classification risk (Equation 3). This is used to further design a family of differentiable approximations to the risk, akin to how the p-norm push (Rudin, 2009) approximates the PTop. For concreteness, Table 2 contrasts the existing and proposed risks for the case of exponential loss.

4.1 Relating PTop and balanced classification risk

There are several ways one might reasonably seek to modify the PTop risk so as to avoid trivial solutions. Our approach is to avail of the rich set of tools available to study loss functions for learning with binary labels. To do so, we observe that we may re-interpret the optimisation of the PTop risk as a special kind of balanced loss minimisation.

Lemma 1 *For any distribution D, non-increasing margin loss $\phi \colon \mathbb{R} \to \mathbb{R}_+$, and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ that is closed under translation,*

$$\min_{f \in \mathcal{F}} R_{\text{ptop}}(f; D, \phi) = \min_{f \in \mathcal{F}} R_{\text{bal}}(f; D, \ell) \tag{10}$$

$$\textit{where } \ell(+1, v) = \phi(v) \qquad \ell(-1, v) = \begin{cases} +\infty & \textit{if } v > 0 \\ 0 & \textit{if } v \leq 0. \end{cases} \tag{11}$$

The peculiar form of Equation 11 is a result of the second minimisation being implicitly constrained. As an illustrative example, the optimisation of the PTop (Equation 5) with a linear scorer $f \colon x \mapsto \langle w, x \rangle + b$ can be re-written as the constrained optimisation

$$\min_{w,b} \frac{1}{n_+} \sum_{i=1}^{n_+} \phi(\langle w, x_i^+ \rangle + b) \colon (\forall j) \langle w, x_j^- \rangle + b \leq 0, \tag{12}$$

with the bias term b playing the role of the (negated) maximum negative score. Interpreting the constraint as a loss per Equation 11 explicates why the PTop has trivial solutions (Proposition 1): beyond the non-positive constraint, there is no penalty on negative instances. In particular, there is *no penalty for assigning all negative instances the same score*.

4.2 RectTop: the rectified PTop risk

Having traced the trivial Bayes-optimal solutions to the nature of the loss $\ell(-1, \cdot)$ on negatives (Equation 11), our strategy to rectify the risk is immediate: we will suitably modify $\ell(-1, \cdot)$ so that we discourage assigning the same score to all such instances. At the same time, we wish to encourage preferentially focussing on the "best" instances.

The former goal can be achieved by choosing $\ell(-1, v) = \phi(-v)$. This would make ℓ a standard margin loss, which has non-trivial Bayes-optimal scorers. To further achieve the latter goal, we additionally impose the non-positive score constraint of Equation 11, yielding:

$$\ell(+1, v) = \phi(v) \qquad \ell(-1, v) = \begin{cases} +\infty & \text{if } v > 0 \\ \phi(-v) & \text{if } v \leq 0. \end{cases} \tag{13}$$

Employing such a loss yields the *RectTop* (*rectified PTop*) ϕ-risk,

$$\begin{aligned} R_{\text{rtop}}(f; D, \phi) &\doteq R_{\text{bal}}(f; D, \ell) \\ &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} \phi(f(\mathsf{X})) + \mathop{\mathbb{E}}_{\mathsf{X} \sim Q} \begin{cases} +\infty & \text{if } f(\mathsf{X}) > 0 \\ \phi(-f(\mathsf{X})) & \text{else.} \end{cases} \end{aligned} \tag{14}$$

This can be seen as an amalgam of the balanced and PTop risks (Equations 3, 11). We illustrate the role of the additional $\phi(-v)$ loss for negatives in the case of 0-1 loss.

Example 3 For $\phi = \ell^{01}$, if our scorer has $\max_{x' \in \text{supp}(Q)} f(x') > 0$, the risk is trivially infinite. If instead $\max_{x' \in \text{supp}(Q)} f(x') \leq 0$,

$$R_{\text{rtop}}(f; D, \phi) = \mathbb{P}_{\mathsf{X} \sim P}\left(f(\mathsf{X}) < 0\right) + \frac{1}{2} \cdot \left(\mathbb{P}_{\mathsf{X} \sim P}\left(f(\mathsf{X}) = 0\right) + \mathbb{P}_{\mathsf{X} \sim Q}\left(f(\mathsf{X}) = 0\right)\right).$$

Observe now that making $\max_{x' \in \text{supp}(Q)} f(x') < 0$ would be sub-optimal for a non-separable distribution, since we would incur a penalty of $+1$ from the first term, as opposed to $+\frac{1}{2}$ from the last term. Assuming then that $\max_{x' \in \text{supp}(Q)} f(x') = 0$,

$$R_{\text{rtop}}(f; D, \phi) = R_{\text{ptop}}(f; D, \phi) + \frac{1}{2} \cdot \mathbb{P}_{\mathsf{X}' \sim Q}\left(f(\mathsf{X}') = 0\right). \tag{15}$$

Thus, compared to the PTop, *we penalise any negative scores equal to that of the highest ranked negative*. This important difference prevents trivial solutions, as we now see.

4.3 Bayes-optimal scorers for the RectTop

We now show that the Bayes-optimal RectTop scorers are non-trivial even for non-separable distributions. We begin with a counterpart to Proposition 1 for 0-1 loss.

Proposition 2 *Pick any distribution $D = (M, \eta)$, and let*

$$(\forall x \in \mathcal{X})\, f^*(x) \in \begin{cases} (0, \infty) & \textit{if } \eta(x) = 1 \\ \{0\} & \textit{if } \eta(x) \in (\pi, 1) \\ (-\infty, 0] & \textit{if } \eta(x) = \pi \\ (-\infty, 0) & \textit{if } \eta(x) \in [0, \pi). \end{cases}$$

Then, $f^* \in \mathop{\mathrm{argmin}}_{f \in \mathbb{R}^{\mathcal{X}}} R_{\text{rtop}}(f; D, \ell^{01})$.

Proposition 2 implies that under the RectTop risk, instances with η *greater than average* are assigned the score zero, with other instances assigned a lower score. This yields non-trivial solutions, while ensuring the highest negative is placed below the positives. For non-separable D, when η takes on two or more distinct values we *must* have at least one x with $\eta(x) \in [\pi, 1)$, and one with $\eta(x) \in (0, \pi)$: this is because $\pi = \mathbb{E}_{\mathsf{X} \sim M}\left[\eta(\mathsf{X})\right]$, and so $\eta(x)$ must be greater than its average at least once.

Next, we consider the case of surrogate ϕ. When this surrogate is strictly proper composite (capturing many commonly used losses such as logistic and exponential), as per §2, the optimal scorers have an intuitive form.

Proposition 3 *Pick any distribution $D = (M, \eta)$. Let ϕ be a convex, strictly proper composite loss with link Ψ. Let*

$$(\forall x \in \mathcal{X})\, f^*(x) = \begin{cases} +\infty & \textit{if } \eta(x) = 1 \\ \{0\} & \textit{if } \eta(x) \in [\pi, 1) \\ \bar{\Psi}_\pi(\eta(x)) & \textit{if } \eta(x) \in [0, \pi), \end{cases}$$

where $\Phi_\pi \doteq \Psi \circ g_\pi$ for $g_\pi(u) \doteq \frac{(1-\pi)\cdot u}{\pi + (1 - 2\cdot\pi)\cdot u}$. Then, $\underset{f}{\operatorname{argmin}}\, R_{\text{rtop}}(f; D, \phi) = \{f^*\}$.

When $\pi = 1/2$, $\Phi_\pi \equiv \Psi$, *viz.* the original link itself. For general π, the link appears more complicated, but in fact guarantees that we escape the issue observed in Proposition 1: an easy calculation reveals that $\Phi_\pi(\pi) = 0$, so that instances with η greater than average are assigned a score of 0, while the rest are assigned a strictly lower score. Thus, for non-separable D, optimising R_{rtop} for a strictly proper composite ϕ avoids the trivial solutions plaguing R_{ptop}.

Example 4 For $\phi(v) = e^{-v}$, one can verify that $\Phi_\pi(u) = \Psi(u) - \frac{1}{2}\log\frac{\pi}{1-\pi}$. Thus,

$$f^*(x) = \begin{cases} +\infty & \text{if } \eta(x) = 1 \\ \{0\} & \text{if } \eta(x) \in [\pi, 1) \\ \frac{1}{2}\log\frac{\eta(x)}{1-\eta(x)} - \frac{1}{2}\log\frac{\pi}{1-\pi} & \text{if } \eta(x) \in [0, \pi). \end{cases}$$

Instances with $\eta(x) \in [0, \pi)$ are thus scored strictly less than instances with $\eta(x) \in [\pi, 1)$, which are all clamped at 0. This is in contrast to the optimal solution for the PTop (Example 2). Note that a trivial modification to the loss ensures the opposite behaviour, i.e. the score for $\eta(x) \in [0, \pi)$ clamped to zero, and other instances accurately modelled; see Appendix C.

4.4 A differentiable approximation of the RectTop

Both the PTop and RectTop risks target good performance at the head of the ranked list. For the PTop, Rudin (2009) proposed a parametric family of approximations to the risk, known as the p-norm push. These provide a user-controlled parameter p, which as $p \to +\infty$ reduces to the PTop, and as $p \to 1$ reduces to the AUC. For the case of the exponential loss, Ertekin and Rudin (2011) further showed that the minimiser of the p-norm push is equivalent to that of a standard classification risk with the *p-classification loss*,

$$\ell(+1, v) = e^{-v} \qquad \ell(-1, v) = p^{-1} \cdot e^{vp}. \tag{16}$$

We now show construct a similar family of approximations for the RectTop. Suppose ϕ is some differentiable strictly proper composite margin loss. Following §4.2, define

$$\ell(+1, v) = \phi(v) \qquad \ell(-1, v; p) = \begin{cases} (\phi(-vp) - \phi(0))/p + \phi(0) & \text{if } v > 0 \\ \phi(-v) & \text{if } v \leq 0 \end{cases} \tag{17}$$

for some $p > 0$. Clearly, this loss approaches Equation 13 pointwise as $p \to \infty$. Further, the loss is differentiable as well as strictly proper composite. Thus, for finite p, we will approximate the Bayes-optimal scorers of Equation 21, and can minimise $R_{\mathrm{bal}}(f; \mathsf{S}, \ell)$ using gradient-based optimisation while remaining faithful to the original RectTop objective.

Example 5 For $\phi(v) = e^{-v}$, we have

$$\ell(+1, v) = e^{-v} \qquad \ell(-1, v; p) = \begin{cases} (1/p) \cdot (e^{vp} - 1) + 1 & \text{if } v > 0 \\ e^{v} & \text{if } v \leq 0. \end{cases}$$

In fact, the partial loss $(1/p) \cdot e^{vp}$ is exactly as per the p-classification loss (Equation 16), which as noted was shown to have equivalent minimiser to the p-norm push risk of Rudin (2009). Equation 17 is thus a translation of this family to the RectTop risk.

4.5 Generalisation bound

We conclude our analysis with a generalisation bound for the balanced risk $R(f; D, \ell) \doteq R_{\mathrm{bal}}(f - \max_{x'} f(x'); D, \ell)$; when the $\max(\cdot)$ term is 0, this is exactly the rectified PTop risk. The following builds on Agarwal (2011, Theorem 5.1), which was for the PTop.

Proposition 4 *Pick any distribution D, and $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. Let $\mathsf{S} \sim D^N$ with n_+ (n_-) positives (negatives). Then, for any $\epsilon, \gamma > 0$, uniformly over all scorers $f \in \mathcal{F}$, for $R(\cdot)$ as above,*

$$R(f; D, \ell^{01}) \leq R_\infty(f; \mathsf{S}, \ell^{\gamma}) + \epsilon$$

with probability at least $1 - \delta$ over the draw of S, where

$$\delta = \mathcal{N}(\mathcal{F}, \epsilon\gamma/8) \cdot \left((\rho(f, \gamma, \epsilon/6))^{n_-} + 2n_- \cdot e^{-\epsilon^2 n_+/18} + 2n_q \cdot e^{-\epsilon^2 n_-/18} \right)$$

for $\mathcal{N}(\mathcal{F}, \cdot)$ the ℓ_∞ covering number of $\mathcal{F}$, $n_q = |\mathrm{supp}(Q)|$, and

$$R_\infty(f; \mathsf{S}, \ell_\gamma) \doteq \max_{1 \leq j \leq n_-} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathsf{S}} \left[\ell_\gamma(\mathsf{Y}(f(\mathsf{X}) - f(x_j^-))) \right]$$

where $\ell_\gamma(v) = [\![v < \gamma]\!] + \frac{1}{2}[\![v = \gamma]\!]$.

Compared to Agarwal (2011, Theorem 5.1), we need some additional analysis to account for the tie-breaking term in RectTop compared to the classic PTop. As a result, our final bound includes an extra term which decays exponentially in the number of negative samples, but has a linear dependence on the support size of the negative class. Intuitively, this term arises from the additional penalty which ensures each negative instance has a score which is ≤ 0; in particular, it measures the estimation error that arises from using a sample of negative instances, versus the underlying population.

The bound implies that good sample *margin* performance guarantees good generalisation performance. However, as a caveat, recall that we assumed of $\mathcal{X}$; without this the bound may be trivial, as we could have $n_q = |\mathrm{supp}(Q)| = +\infty$. This limits the practical viability of the bound, since one typically expects $\mathcal{X}$ to be non-finite. We conjecture it is possible to extend such a bound to the case of non-finite $\mathcal{X}$, perhaps employing a suitable cover of $|\mathrm{supp}(Q)|$, at the expense of a penalty on the dependence on ϵ. Exploring such a generalisation would be of interest in future work, for which Proposition 4 may be a useful starting point. Alternate techniques for establish generalisation bounds, such as those based on similarity of train and test samples (Xu and Mannor, 2012), may also be fruitful.

We also remark here that the bound does *not* address the issue of the comparative difficulty of optimising the RectTop over the differentiable approximation in §4.4. This issue is however not fully resolved even for the original PTop risk. The original analysis of Rudin (2009) provided a generalisation bound on the p-norm push risk with sample complexity exponential in p. This indicates that increasing p makes generalisation harder, which is plausible: by increasingly focussing on the head of the ranked list, one works with fewer effective samples and makes the algorithm more sensitive to its inputs. Interestingly, the bound in Rudin (2009) is vacuous for the limiting case $p = +\infty$ (i.e., the PTop risk), and so does not provide a guarantee for PTop minimisation being able to generalise. Subsequently, Agarwal (2011) provided a non-trivial bound for this risk, which implies there is some looseness in the bound of Rudin (2009). It is still plausible that optimising the p-norm push risk for finite p is easier; however, this issue was not considered in Agarwal (2011).

As a final remark, we note that Li et al (2014b) provided a different generalisation bound for PTop minimisation, which looked at the probability of a positive instance being ranked below *most* negative instances. It was shown that provided the empirical surrogate PTop risk is sufficiently small, this probability can be non-trivially bounded. Deriving an analogous bound for our risk would be of interest, though again would not directly address the relative difficulty of RectTop versus approximate RectTop minimisation.

4.6 Discussion of results

Our results show that our rectification of the PTop avoids trivial asymptotic solutions. While this rectification has a simple final form, the connection between the PTop and balanced risks (Lemma 1) is the crucial insight from which our modifications derived naturally. A few comments on the RectTop risk are prudent.

First, the focus of the RectTop is not to *measure* top ranking performance, but rather to *attain* good top ranking performance. When the concern is to measure performance, there may be little difference in using the PTop and RectTop. This is because for the 0-1 loss, Equation 15 implies that the *empirical* RectTop risk may be identical (upto translation) to the empirical PTop risk, since the maximum negative score may be uniquely attained. However, when the concern is to attain good performance, there is a non-trivial difference in the two risks. This is because for a general surrogate ϕ, the two risks will be fundamentally different, even on a finite sample. Further, our Proposition 3 implies that for proper composite ϕ, the RectTop will result in a non-trivial asymptotic solution, unlike the PTop.

Second, the RectTop risk has an implicit score anchor (i.e. 0) that distinguishes positives and negatives, and thus forgoes translation invariance. Strictly, then, it is not a ranking risk; however, its theoretical minimisers are nonetheless sensible in the top ranking regime.

Third, the correction applied by RectTop can be viewed as a form of tie-breaking: it ensures that the negative instances are not all trivially assigned the same score. As noted in Equation 7, the original PTop already includes one form of tie-breaking: this however ensures that the *positive* instances are not all trivially assigned the same score. Without the latter correction, it would be optimal to trivially assign all instances the same score, regardless of whether $1 \in \mathrm{Im}(\eta)$ or not. The correction rules out such solutions when $1 \in \mathrm{Im}(\eta)$; however, we need our additional correction to rule out trivial solutions when $1 \notin \mathrm{Im}(\eta)$.

Fourth, one would hope that low RectTop ϕ-risk also ensures good performance with respect to other top ranking measures, such as discounted cumulative gain and average precision. The form of optimal scorers in Proposition 3 suggest this is plausible. While we have not confirmed this intuition theoretically, we do confirm this empirically in §6.3.

Fifth, the surrogate RectTop ϕ-risk is only a means to an end: in practice, one's goal is to ensure that the RectTop 0-1 risk is small, so that most positives are ranked above all negatives. For computational reasons, it is preferable to work with a (convex) surrogate risk. Our analysis justifies such a risk in an asymptotic regime, where one has sufficiently many samples and a rich function class. However, when one or both of these assumptions fail, surrogate minimisation may fail to produce desirable results; for example, a counterexample in Rudin and Wang (2018) demonstrates that exponential loss minimisation may not maximise the AUC. The brittleness of convex surrogates is not unique to ranking, and plagues their use in standard binary classification as well (Long and Servedio, 2010; Ben-David et al, 2012). As with classification problems, one might ameliorate the problem by using a non-convex surrogate. Exploring such losses for the RectTop would be of interest, and would be accommodated by our Proposition 3 which simply requires ϕ to be strictly proper composite.[3]

Sixth, our Bayes-optimal analysis does not provide a means of discriminating amongst different surrogates. Such analysis simply verifies that the asymptotic target of surrogate minimisation is sensible, and leaves untold how different surrogate minimisers behave under finite samples and restricted function classes. In practice, one often finds that different surrogates can yield quite different performance. This raises the non-trivial issue of choosing *which* surrogate ϕ should be employed, which is again largely unresolved in binary classification (Reid and Williamson, 2010, Appendix A). Nonetheless, one does have some (competing) guidance on this issue by means of minimax analysis (Ben-David et al, 2012), appeals to noise robustness (Ghosh et al, 2015), and empirical surrogate tuning (Nock and Nielsen, 2009). We believe that similar ideas might be useful in the top ranking setting.

Seventh, the RectTop is only one possible rectification of the PTop. As noted, our modification allows one to use the connection between the PTop and standard balanced classification risks. Exploring other rectifications is of interest for future work.

5 Optimising the RectTop risk

Recall that one appeal of using the PTop ϕ-risk for top ranking problems is that it admits a simple convex optimisation. We now show the same is true for the RectTop ϕ-risk. Given $\mathsf{S} = \{x_i^+\}_{i=1}^{n_+} \cup \{x_j^-\}_{j=1}^{n_-} \sim D^N$, consider optimising $R_{\mathrm{rtop}}(f; \mathsf{S}, \phi)$ with a linear $f_{w,b} \colon x \mapsto \langle w, x \rangle + b$. By Equations 13, 14, this is (c.f. Equation 12)

$$\min_{w,b} \frac{1}{n_+} \sum_{i=1}^{n_+} \phi(\langle w, x_i^+ \rangle + b) + \frac{1}{n_-} \sum_{j=1}^{n_-} \phi(-(\langle w, x_j^- \rangle + b)) \colon (\forall j) \langle w, x_j^- \rangle + b \leq 0. \tag{18}$$

Note that the constraint is trivially feasible since for any w, we may pick $b = -\max_j \langle w, x_j^- \rangle$. While convex, Equation 18 does not permit standard gradient-based optimisation due to the constraint. Fortunately, like the PTop, the dual objective is amenable to efficient optimisation.

[3] Strictly proper composite losses are allowed to be non-convex; see, e.g., Buja et al (2005); Reid and Williamson (2010).

Proposition 5 *For any convex differentiable ϕ with conjugate ϕ^* and sample* $\mathsf{S} = \{(x_i^+, +1)\}_{i=1}^{n_+} \cup \{(x_j^-, -1)\}_{j=1}^{n_-}$, *the dual objective of Equation 18 with regulariser $(\lambda/2)\|w\|_2^2$ is*

$$\min_{(\alpha,\beta,\gamma)\in\Theta} \frac{1}{2\lambda} \left\|\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)\right\|^2 + \frac{1}{n_+}\sum_{i=1}^{n_+} \phi^*\left(-n_+ \cdot \alpha_i\right) + \frac{1}{n_-}\sum_{j=1}^{n_-} \phi^*\left(-n_- \cdot \beta_j\right) \tag{19}$$

for positive and negative feature matrices $\mathbf{X}^+, \mathbf{X}^-$, and

$$\Theta = \left\{(\alpha,\beta,\gamma)\colon \sum_{i=1}^{n_+}\alpha_i = \sum_{j=1}^{n_-}(\beta_j + \gamma_m), \alpha \in -\mathrm{dom}(\phi^*), \beta \in -\mathrm{dom}(\phi^*), \gamma \geq 0\right\}.$$

Constraining $\beta \equiv 0$, would yield exactly the R_{ptop} dual of Li et al (2014b, Theorem 1). The similarity between the two problems means that for smooth ϕ^*, we can adopt the same optimisation scheme as in Li et al (2014b), using Nesterov's method (Nesterov, 2004) with a minor augmentation to additionally optimise for the variables β present in our objective; see Appendix B for details.

Three comments are prudent. First, as per the objective of Li et al (2014b), optimisation of Equation 19 requires complexity *linear in the number of training samples*. Second, one can equally work with kernelised scorers, as Equation 19 only involves inner products between instances. Third, for the differentiable risk approximations in §4.4, one can employ standard unconstrained gradient-based optimisation; trivially, this also has linear time complexity.

6 Experimental illustration of results

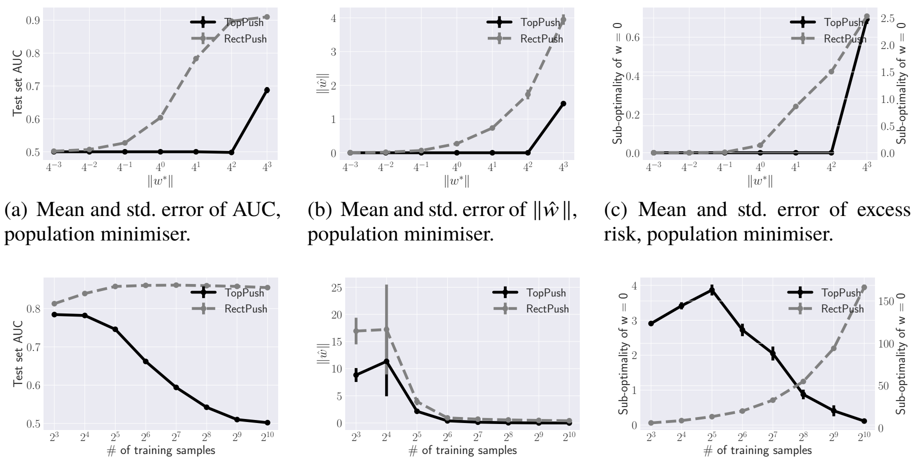
We now validate our theoretical analyses empirically: we show in §6.1, 6.2 that the PTop minimiser will be trivial for certain non-separable distributions (per **C1**), while the RectTop minimiser avoids such solutions (per **C2**). We further show in §6.3 that on real-world datasets, the RectTop minimiser is competitive or superior to its PTop counterpart. In sum, RectTop yields *comparable or superior results to PTop in the average case, while ensuring there are no trivial solutions in the worst case*.

6.1 Illustration of trivial population PTop minimisers

We first validate that the PTop ϕ-risk minimiser will be trivial for certain non-separable distributions. We fix a finite $\mathcal{X}$ comprising N_{atom} points in $\mathbb{R}^2$, and use a distribution D over $\mathcal{X} \times \{\pm 1\}$ with a uniform marginal M and class-probability $\eta(x) = 1/(1+e^{-\langle w^*, x\rangle})$ for some $w^* \in \mathbb{R}^2$. As $\mathcal{X}$ is finite, we can explicitly compute the PTop risk as $R_{\mathrm{ptop}}(w; D, \phi) \propto \sum_{x\in\mathcal{X}} \eta(x) \cdot \phi\left(\langle w, x\rangle - \max_{x'\in\mathcal{X}}\langle w, x'\rangle\right)$. Since $\mathrm{Im}(\eta) \subseteq (0,1)$, by Corollary 1 we expect the PTop risk minimiser to be the all-zeros vector.

We fix the optimal $w^* = W/\sqrt{2} \cdot (1,1)$, where $W \in \{4^{-3}, \ldots, 4^3\}$. We fix $N_{\mathrm{atom}} = 4$, and draw the elements of $\mathcal{X}$ uniformly from $[-1,1]^2$. For each choice of $(\mathcal{X}, w^*)$, we minimise the *population* risks $R_{\mathrm{ptop}}(w; D, \phi)$ and $R_{\mathrm{rtop}}(w; D, \phi)$ for ϕ the square-hinge loss[4] using MATLAB's `fmincon` function. We compute the two minimisers' norm and AUC for 100 random draws of $\mathcal{X}$. (The use of AUC is sufficient to illustrate that the PTop solution is tantamount to random guessing.)

[4] Strictly, this makes a linear model misspecified, since the true η involves a sigmoid link. We nonetheless find that the TopPush solution is trivial, as argued in §3.4.



(a) Mean and std. error of AUC, population minimiser.

(b) Mean and std. error of $\|\hat{w}\|$, population minimiser.

(c) Mean and std. error of excess risk, population minimiser.

(d) Mean and std. error of AUC, empirical minimiser.

(e) Mean and std. error of $\|\hat{w}\|$, empirical minimiser.

(f) Mean and std. error of excess risk, empirical minimiser.

Fig. 1 Comparison of R_{ptop} (TopPush) versus R_{rtop} (RectPush) minimisers on population (§6.1) and finite sample (§6.2), synthetic data. As predicted by the theory, the TopPush converges to a trivial solution, unlike the RectPush.

Figures 1(a), 1(b) show that for smaller W the R_{rtop} separator is reasonable, while the R_{ptop} minimiser is the trivial solution (evidenced by having zero norm). For larger W, the performance of the two is comparable, with the RectTop solution being slightly better. Note that for larger W, η will be close, but not exactly equal to $\{0, 1\}$; conversely, for smaller W, η will be close to $1/2$. Finally, Figure 1(c) studies the excess risk (or suboptimality gap) of the trivial zero solution, i.e. computes the excess risk of this solution over that of the risk minimiser. As expected, the zero solution is in fact optimal for the TopPush for all but the largest choice of W, while for the RectPush, this solution quickly has non-trivial excess risk.

6.2 Illustration of trivial sample PTop minimisers

We next confirm that for the above example, the PTop minimiser on a finite sample also performs poorly. For the same discrete D as above, and $W = 1$, we draw a sample $\mathsf{S} \sim D^N$, and compute the empirical minimisers for the PTop and the RectTop. We report the mean AUC on D of these solutions over 100 random draws of $\mathcal{X}$ as N is varied.

Figure 1(d) confirms that while for small N the TopPush solution performs reasonably – a consequence of the empirical distribution often being separable – as N increases there is a steady decrease in performance. When N is suitably large, the TopPush solution approaches the trivial population minimiser from the previous section, as evidenced by the AUC converging to 0.5. By contrast, as N increases the RectPush solution improves its performance, as is desirable for any learning algorithm. Figure 1(e) further confirms that the trivial solution is increasingly close to optimal for the empirical TopPush risk, but increasing sub-optimal for the empirical RectPush risk.

6.3 Illustration of real-world RectTop performance

We conclude with results on real-world datasets used in Li et al (2014b), plus some additional ones from the UCI repository. We summarise the statistics of the datasets in Table 3. For the high dimensional `real-sim` and `news20-forsale` datasets, we performed an SVD projection to 100 dimensions and used this as input to all methods.

We compare logistic regression (**Logistic**), the PTop risk (**TopPush**) Li et al (2014b), our rectified PTop risk of §4.2 (**RectPush**), and its differentiable approximation of §4.4 (**RectPush**$_{\text{App}}$). In Li et al (2014b), TopPush was shown superior to a number of other top-ranking approaches such as those of Boyd et al (2012); Agarwal (2011); Narasimhan and Agarwal (2013), as well as standard ranking approaches such as SVMRank (Joachims, 2005). For all methods, we used a regularised linear scorer f; for all methods other than logistic, we set ϕ to be the square-hinge loss.

Each dataset was randomly split 10 times in the ratio 2:1, with all instances normalised so that $||x||_2 \leq 1$. We measure average test performance across these splits using the average reciprocal rank (ARR), discounted cumulative gain (DCG)[5], average precision (AP), positives at the top[6] (PTop), and precision at 10 (Prec@10). For each split, 5-fold CV was used to tune the regularisation strength $\lambda \in \{2^{-20}, 2^{-19}, \ldots, 2^{15}\}$ based on AP.

Dataset	n	d	**Dataset**	n	d
`german`	1000	24	`covtype-binary`	38501	54
`abalone`	4177	9	`w8a`	64700	300
`spambase`	4601	57	`real-sim`	72309	20958
`magic`	19020	10	`ijcnn1`	141691	22
`news20-forsale`	19928	62061	`nsl-kdd`	148517	119
`skin`	245057	3	`kddcup98`	191779	15
`activity`	14704	561	`kddcup04`	50000	70

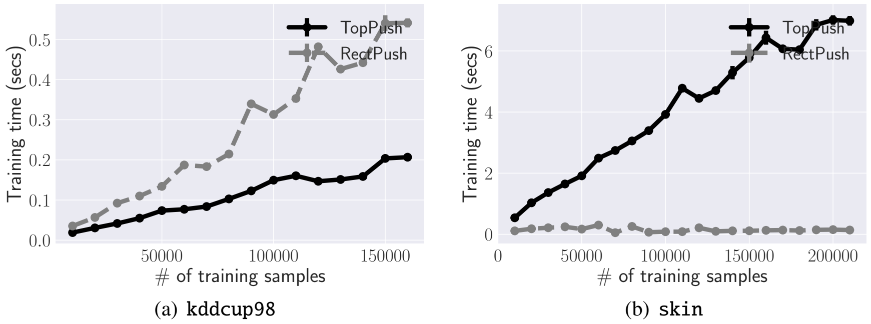
Table 3 Statistics of # of samples (n) and dimensions (d) for datasets used in experiments.

6.3.1 Performance comparison

Table 4 shows that both RectPush and its differentiable approximation RectPush$_{\text{App}}$ offer consistent (if sometimes modest) improvements over TopPush and logistic regression. We reiterate that the value of the RectPush is that it guards against trivial solutions, as shown in §6.1, 6.2. Put another way, RectPush yields comparable or superior results to TopPush in the average case, while ensuring that there are no trivial solutions in the worst case. From a practical perspective, one could make a case for using logistic regression for top ranking problems, given its ubiquity and its respectable showing in the results. Translating the theoretical gains from RectTop into more visible practical gains would nonetheless be of interest for future work.

[5] We scaled this by the number of positives to produce scores in [0, 1].

[6] While our analysis suggests this measure favours trivial solutions for non-separable distributions, we present the scores here as they have been reported in previous work. Recall also that from §4.6, the empirical RectTop for 0-1 loss is often identical to the empirical PTop for 0-1 loss.



(a) `kddcup98`

(b) `skin`

Fig. 2 Training times (in seconds) of RectPush versus TopPush as training set size increases.

A Friedman test confirms that on all but Prec@10 and ARR, the difference in ranks is statistically significant at the 5% level. A post-hoc Holm test confirms the following significant differences for the best performing methods on individual measures:

- **AUC**: RectPush$_{\text{App}}$ is significantly better than TopPush and RectPush
- **DCG**: RectPush$_{\text{App}}$ is significantly better than Logistic
- **AP**: RectPush$_{\text{App}}$ is significantly better than Logistic

Of note is that a Nemenyi test fails to reveal the differences for TopPush over other methods, when they exist, to be significant.

As a final remark, we note that the RectTop generally produces better scores with respect to the PTop. While we have argued for the latter being a misleading measure of performance, we re-iterate that per Section 4.6, the issue of ties is less prominent for the empirical PTop with 0-1 loss; thus, it is unsurprising that as a performance measure, this does not overly penalise the method which is best performing on the other metrics.

6.3.2 Timing comparison

Table 5 presents a comparison of the training times for the TopPush, RectPush, and RectPush$_{\text{App}}$ on all datasets. (Times are as computed on 2.2GHz Intel Core i7.) We find that the TopPush is generally slightly faster, but RectPush training is never prohibitive. While the $\mathcal{O}(\cdot)$ complexity of RectPush is identical to the TopPush, the training times for the two are unsurprisingly different, which is owing to the objectives employing different optimal regularisation, and thus requiring different numbers of iterations to converge.

To further study the scalability of RectPush, we look at how the training time grows as the number of training instances increases. For the larger `skin` and `kddcup98` datasets, we subsample the number of training instances at various levels, and compare the training times of the RectPush and TopPush methods. Figure 2 shows that for both methods, the scalability is empirically linear in the number of instances. The slopes of the methods however vary in a problem-specific manner.

7 Conclusion and future work

We showed that the fraction of positives of the top (PTop), a popular measure for bipartite top ranking, may be trivially optimised by predicting the same score for all instances. This

Dataset	Loss	AUC	ARR	DCG	AP	PTop	Prec@10
german	Logistic	0.8051 ± 0.0080 (1)	0.0369 ± 0.0016 (4)	0.1846 ± 0.0015 (4)	0.6087 ± 0.0192 (2)	0.0224 ± 0.0083 (4)	0.7000 ± 0.0333 (3)
	TopPush	0.8007 ± 0.0088 (3)	0.0378 ± 0.0014 (3)	0.1850 ± 0.0013 (3)	0.6056 ± 0.0198 (4)	0.0261 ± 0.0089 (3)	0.7200 ± 0.0291 (2)
	RectPush	0.7953 ± 0.0083 (4)	0.0392 ± 0.0015 (2)	0.1860 ± 0.0013 (2)	0.6057 ± 0.0164 (3)	0.0367 ± 0.0101 (1)	0.7300 ± 0.0367 (1)
	RectPush$_{\text{App}}$	0.8048 ± 0.0080 (2)	0.0393 ± 0.0012 (1)	0.1864 ± 0.0013 (1)	0.6136 ± 0.0191 (1)	0.0334 ± 0.0107 (2)	0.6900 ± 0.0407 (4)
abalone	Logistic	0.8161 ± 0.0166 (1)	0.0106 ± 0.0010 (3)	0.1407 ± 0.0025 (3)	0.0317 ± 0.0022 (3)	0.0000 ± 0.0000 (1)	0.0000 ± 0.0000 (1)
	TopPush	0.7332 ± 0.0295 (4)	0.0119 ± 0.0008 (1)	0.1428 ± 0.0019 (2)	0.0355 ± 0.0029 (1)	0.0000 ± 0.0000 (1)	0.0000 ± 0.0000 (1)
	RectPush	0.7337 ± 0.0251 (3)	0.0091 ± 0.0011 (4)	0.1362 ± 0.0028 (4)	0.0289 ± 0.0031 (4)	0.0000 ± 0.0000 (1)	0.0000 ± 0.0000 (1)
	RectPush$_{\text{App}}$	0.8006 ± 0.0235 (2)	0.0112 ± 0.0008 (2)	0.1431 ± 0.0019 (1)	0.0348 ± 0.0024 (2)	0.0000 ± 0.0000 (1)	0.0000 ± 0.0000 (1)
spambase	Logistic	0.9658 ± 0.0011 (3)	0.0104 ± 0.0004 (4)	0.1336 ± 0.0004 (3)	0.9337 ± 0.0035 (3)	0.0202 ± 0.0084 (4)	0.9200 ± 0.0249 (3)
	TopPush	0.9476 ± 0.0017 (4)	0.0108 ± 0.0003 (3)	0.1336 ± 0.0003 (3)	0.9197 ± 0.0034 (4)	0.0566 ± 0.0163 (2)	0.9500 ± 0.0269 (1)
	RectPush	0.9671 ± 0.0012 (2)	0.0113 ± 0.0001 (1)	0.1345 ± 0.0002 (1)	0.9442 ± 0.0035 (1)	0.0677 ± 0.0249 (1)	0.9500 ± 0.0167 (1)
	RectPush$_{\text{App}}$	0.9676 ± 0.0011 (1)	0.0110 ± 0.0002 (2)	0.1343 ± 0.0003 (2)	0.9433 ± 0.0036 (2)	0.0443 ± 0.0137 (3)	0.9400 ± 0.0267 (2)
magic	Logistic	0.8418 ± 0.0011 (3)	0.0020 ± 0.0000 (2)	0.0961 ± 0.0000 (2)	0.8867 ± 0.0018 (4)	0.0018 ± 0.0005 (4)	0.9200 ± 0.0133 (3)
	TopPush	0.8357 ± 0.0011 (4)	0.0021 ± 0.0000 (1)	0.0963 ± 0.0000 (1)	0.9003 ± 0.0015 (1)	0.0085 ± 0.0040 (2)	0.9500 ± 0.0167 (2)
	RectPush	0.8421 ± 0.0014 (2)	0.0021 ± 0.0000 (1)	0.0963 ± 0.0000 (1)	0.8993 ± 0.0016 (2)	0.0090 ± 0.0042 (1)	0.9900 ± 0.0100 (1)
	RectPush$_{\text{App}}$	0.8427 ± 0.0011 (1)	0.0020 ± 0.0000 (2)	0.0961 ± 0.0000 (2)	0.8895 ± 0.0018 (3)	0.0023 ± 0.0008 (3)	0.9200 ± 0.0249 (3)
news20-forsale	Logistic	0.8016 ± 0.0033 (4)	0.0035 ± 0.0003 (4)	0.1068 ± 0.0004 (4)	0.1487 ± 0.0041 (4)	0.0003 ± 0.0003 (4)	0.1200 ± 0.0249 (4)
	TopPush	0.8333 ± 0.0057 (2)	0.0130 ± 0.0006 (1)	0.1256 ± 0.0013 (1)	0.3249 ± 0.0137 (1)	0.0148 ± 0.0028 (1)	0.7700 ± 0.0367 (2)
	RectPush	0.8275 ± 0.0041 (3)	0.0128 ± 0.0008 (2)	0.1240 ± 0.0013 (3)	0.2999 ± 0.0120 (3)	0.0147 ± 0.0032 (2)	0.7900 ± 0.0458 (1)
	RectPush$_{\text{App}}$	0.8589 ± 0.0055 (1)	0.0119 ± 0.0005 (3)	0.1249 ± 0.0011 (2)	0.3196 ± 0.0129 (2)	0.0079 ± 0.0020 (3)	0.6300 ± 0.0473 (3)
skin	Logistic	0.9475 ± 0.0003 (2)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9886 ± 0.0001 (2)	0.9146 ± 0.0003 (3)	1.0000 ± 0.0000 (1)
	TopPush	0.9470 ± 0.0003 (3)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9886 ± 0.0001 (2)	0.9171 ± 0.0003 (1)	1.0000 ± 0.0000 (1)
	RectPush	0.9466 ± 0.0003 (4)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9885 ± 0.0001 (3)	0.9165 ± 0.0003 (2)	1.0000 ± 0.0000 (1)
	RectPush$_{\text{App}}$	0.9479 ± 0.0003 (1)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9887 ± 0.0001 (1)	0.9111 ± 0.0004 (4)	1.0000 ± 0.0000 (1)
activity	Logistic	0.8978 ± 0.0012 (3)	0.0041 ± 0.0000 (2)	0.1085 ± 0.0000 (2)	0.8975 ± 0.0017 (2)	0.0043 ± 0.0010 (4)	0.9200 ± 0.0291 (2)
	TopPush	0.8442 ± 0.0082 (4)	0.0038 ± 0.0002 (3)	0.1075 ± 0.0002 (3)	0.8504 ± 0.0057 (3)	0.0169 ± 0.0047 (1)	0.8700 ± 0.0803 (3)
	RectPush	0.8997 ± 0.0011 (2)	0.0042 ± 0.0000 (1)	0.1087 ± 0.0000 (1)	0.9020 ± 0.0013 (1)	0.0146 ± 0.0028 (3)	0.9800 ± 0.0133 (1)
	RectPush$_{\text{App}}$	0.9005 ± 0.0010 (1)	0.0042 ± 0.0000 (1)	0.1087 ± 0.0000 (1)	0.9020 ± 0.0013 (1)	0.0153 ± 0.0042 (2)	0.9800 ± 0.0133 (1)
covtype-binary	Logistic	0.9374 ± 0.0010 (2)	0.0054 ± 0.0003 (3)	0.1132 ± 0.0003 (4)	0.5492 ± 0.0051 (4)	0.0056 ± 0.0054 (4)	0.6900 ± 0.0482 (3)
	TopPush	0.9323 ± 0.0019 (3)	0.0074 ± 0.0000 (1)	0.1169 ± 0.0002 (3)	0.6446 ± 0.0054 (2)	0.0452 ± 0.0057 (3)	0.9900 ± 0.0100 (2)
	RectPush	0.9374 ± 0.0018 (2)	0.0074 ± 0.0001 (1)	0.1172 ± 0.0002 (1)	0.6514 ± 0.0044 (1)	0.0490 ± 0.0028 (2)	1.0000 ± 0.0000 (1)
	RectPush$_{\text{App}}$	0.9472 ± 0.0007 (1)	0.0073 ± 0.0001 (2)	0.1171 ± 0.0002 (2)	0.6422 ± 0.0046 (3)	0.0497 ± 0.0039 (1)	1.0000 ± 0.0000 (1)
ijcnn1	Logistic	0.9356 ± 0.0005 (1)	0.0014 ± 0.0000 (2)	0.0892 ± 0.0001 (2)	0.5651 ± 0.0025 (4)	0.0008 ± 0.0001 (3)	0.6600 ± 0.0427 (2)
	TopPush	0.9101 ± 0.0007 (4)	0.0015 ± 0.0000 (1)	0.0898 ± 0.0001 (1)	0.6113 ± 0.0016 (1)	0.0009 ± 0.0001 (2)	0.6500 ± 0.0428 (3)
	RectPush	0.9301 ± 0.0006 (2)	0.0015 ± 0.0000 (1)	0.0898 ± 0.0000 (1)	0.5935 ± 0.0022 (3)	0.0011 ± 0.0001 (1)	0.7000 ± 0.0365 (1)
	RectPush$_{\text{App}}$	0.9283 ± 0.0006 (3)	0.0015 ± 0.0000 (1)	0.0898 ± 0.0000 (1)	0.5953 ± 0.0021 (2)	0.0011 ± 0.0001 (1)	0.7000 ± 0.0365 (1)
w8a	Logistic	0.9676 ± 0.0009 (1)	0.0074 ± 0.0003 (4)	0.1232 ± 0.0003 (4)	0.6631 ± 0.0034 (4)	0.0002 ± 0.0002 (4)	0.6500 ± 0.0619 (4)
	TopPush	0.9219 ± 0.0061 (4)	0.0104 ± 0.0001 (3)	0.1252 ± 0.0005 (3)	0.6978 ± 0.0085 (3)	0.1131 ± 0.0267 (2)	0.9800 ± 0.0200 (3)
	RectPush	0.9639 ± 0.0015 (3)	0.0105 ± 0.0001 (2)	0.1279 ± 0.0002 (2)	0.7594 ± 0.0060 (2)	0.1072 ± 0.0357 (3)	0.9900 ± 0.0100 (2)
	RectPush$_{\text{App}}$	0.9655 ± 0.0011 (2)	0.0106 ± 0.0001 (1)	0.1285 ± 0.0003 (1)	0.7783 ± 0.0029 (1)	0.2174 ± 0.0293 (1)	1.0000 ± 0.0000 (1)
real-sim	Logistic	0.9852 ± 0.0001 (2)	0.0013 ± 0.0000 (1)	0.0896 ± 0.0000 (1)	0.9674 ± 0.0003 (3)	0.0927 ± 0.0064 (3)	1.0000 ± 0.0000 (1)
	TopPush	0.9804 ± 0.0002 (3)	0.0013 ± 0.0000 (1)	0.0894 ± 0.0000 (2)	0.9570 ± 0.0007 (4)	0.0403 ± 0.0087 (4)	1.0000 ± 0.0000 (1)
	RectPush	0.9857 ± 0.0001 (1)	0.0013 ± 0.0000 (1)	0.0896 ± 0.0000 (1)	0.9696 ± 0.0003 (2)	0.1097 ± 0.0157 (2)	1.0000 ± 0.0000 (1)
	RectPush$_{\text{App}}$	0.9857 ± 0.0001 (1)	0.0013 ± 0.0000 (1)	0.0896 ± 0.0000 (1)	0.9697 ± 0.0003 (1)	0.1107 ± 0.0139 (1)	1.0000 ± 0.0000 (1)
nsl-kdd	Logistic	0.9810 ± 0.0002 (3)	0.0004 ± 0.0000 (1)	0.0769 ± 0.0000 (2)	0.9803 ± 0.0003 (3)	0.3711 ± 0.0229 (1)	1.0000 ± 0.0000 (1)
	TopPush	0.9703 ± 0.0014 (4)	0.0004 ± 0.0000 (1)	0.0769 ± 0.0000 (2)	0.9750 ± 0.0013 (4)	0.2261 ± 0.0242 (2)	1.0000 ± 0.0000 (1)
	RectPush	0.9831 ± 0.0005 (2)	0.0004 ± 0.0000 (1)	0.0770 ± 0.0000 (1)	0.9875 ± 0.0003 (2)	0.0786 ± 0.0435 (4)	1.0000 ± 0.0000 (1)
	RectPush$_{\text{App}}$	0.9887 ± 0.0001 (1)	0.0004 ± 0.0000 (1)	0.0770 ± 0.0000 (1)	0.9892 ± 0.0002 (1)	0.1059 ± 0.0725 (3)	1.0000 ± 0.0000 (1)
kddcup98	Logistic	0.6083 ± 0.0021 (4)	0.0004 ± 0.0000 (1)	0.0731 ± 0.0001 (4)	0.0762 ± 0.0011 (4)	0.0000 ± 0.0000 (2)	0.1700 ± 0.0300 (2)
	TopPush	0.6129 ± 0.0016 (2)	0.0004 ± 0.0000 (1)	0.0734 ± 0.0001 (3)	0.0796 ± 0.0010 (3)	0.0000 ± 0.0000 (2)	0.0800 ± 0.0291 (4)
	RectPush	0.6124 ± 0.0016 (3)	0.0004 ± 0.0001 (1)	0.0735 ± 0.0001 (2)	0.0797 ± 0.0010 (2)	0.0001 ± 0.0001 (1)	0.2000 ± 0.0516 (1)
	RectPush$_{\text{App}}$	0.6142 ± 0.0015 (1)	0.0004 ± 0.0000 (1)	0.0736 ± 0.0001 (1)	0.0804 ± 0.0010 (1)	0.0000 ± 0.0000 (2)	0.1600 ± 0.0371 (3)
kdd04	Logistic	0.7938 ± 0.0008 (2)	0.0011 ± 0.0000 (1)	0.0862 ± 0.0000 (2)	0.7936 ± 0.0008 (3)	0.0051 ± 0.0017 (4)	0.9300 ± 0.0260 (3)
	TopPush	0.7494 ± 0.0030 (4)	0.0011 ± 0.0000 (1)	0.0859 ± 0.0000 (3)	0.7697 ± 0.0024 (4)	0.0183 ± 0.0029 (1)	0.9900 ± 0.0100 (1)
	RectPush	0.7906 ± 0.0014 (3)	0.0011 ± 0.0000 (1)	0.0863 ± 0.0000 (1)	0.7975 ± 0.0009 (1)	0.0111 ± 0.0042 (2)	0.9300 ± 0.0260 (3)
	RectPush$_{\text{App}}$	0.7939 ± 0.0008 (1)	0.0011 ± 0.0000 (1)	0.0863 ± 0.0000 (1)	0.7967 ± 0.0008 (2)	0.0065 ± 0.0025 (3)	0.9400 ± 0.0221 (2)
Average rank	Logistic	2.2857	2.3571	2.7143	3.2143	3.2143	2.3571
	TopPush	3.4286	1.5714	2.2143	2.6429	1.9286	1.9286
	RectPush	2.5714	1.4286	1.5714	2.1429	1.8571	1.2143
	RectPush$_{\text{App}}$	1.3571	1.4286	1.2857	1.6429	2.1429	1.7857

Table 4 Mean and standard error of performance measures over 10 trials on real-world datasets (selecting parameters with AP). The rank of each method on each dataset is shown in parentheses.

arises from the simple observation that under distributions with noise, there may not be any instances that are deterministically positive; consequently, each instance is required to score at least as much as every other instance, which results in a trivial solution. We proposed a simple rectification that dispels such trivial solutions, while being as simple to optimise. This arises from viewing the original PTop as a form of constrained loss minimisation; by

Dataset	**TopPush**	**RectPush**	**RectPush$_{\text{App}}$**
`german`	0.00 ± 0.00	0.02 ± 0.00	0.05 ± 0.00
`abalone`	0.01 ± 0.00	0.03 ± 0.00	0.07 ± 0.00
`spambase`	0.02 ± 0.00	0.94 ± 0.00	1.67 ± 0.01
`magic`	0.04 ± 0.00	0.53 ± 0.00	0.11 ± 0.00
`news20-forsale`	0.06 ± 0.00	12.02 ± 0.02	8.31 ± 0.01
`skin`	25.87 ± 0.29	0.16 ± 0.00	0.55 ± 0.01
`activity`	1.68 ± 1.45	11.36 ± 0.68	17.68 ± 1.21
`covtype-binary`	0.10 ± 0.00	11.68 ± 0.04	2.04 ± 0.01
`ijcnn1`	0.21 ± 0.00	5.15 ± 0.05	1.36 ± 0.01
`w8a`	61.02 ± 0.15	49.12 ± 0.33	47.03 ± 0.18
`real-sim`	0.14 ± 0.00	13.08 ± 0.18	3.13 ± 0.04
`nsl-kdd`	23.94 ± 0.29	32.84 ± 0.31	59.10 ± 0.46
`kddcup98`	1.66 ± 0.07	2.21 ± 0.13	3.30 ± 0.15
`kddcup04`	1.68 ± 1.45	11.36 ± 0.68	8.56 ± 1.24

Table 5 Training times (in seconds) for the TopPush, RectPush, and RectPush$_{\text{App}}$.

suitably modifying the underlying loss, we can penalise ties amongst instances, and thus avoid trivial solutions.

There are several possible directions for future work. Most immediately, it is of interest to study finite sample behaviour of the PTop and RectTop minimisers more carefully. In particular, removing the need for finiteness of the instance-space would make the generalisation bounds more practically relevant. It would also be of interest to establish formally the finite-sample probabilities of obtaining a trivial solution from PTop minimisation; this would theoretically ground our empirical findings in Figure 1(d). Further, establishing regret bounds relating RectTop performance to performance with respect to other top ranking measures would justify its use when optimising such measures. This would be of interest owing to such measures typically being harder to optimise than the PTop and RectTop. Finally, we hope our result motivates the study of other top ranking risks that avoid trivial solutions.

References

Agarwal S (2011) 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)

Agarwal S, Niyogi P (2005) Stability and generalization of bipartite ranking algorithms. In: Conference on Learning Theory (COLT)

Agarwal S, Graepel T, Herbrich R, Har-Peled S, Roth D (2005) Generalization bounds for the area under the ROC curve. Journal of Machine Learning Research 6

Bartlett PL, Jordan MI, Mcauliffe JD (2006) Convexity, classification, and risk bounds. Journal of the American Statistical Association 101(473)

Ben-David S, Loker D, Srebro N, Sridharan K (2012) Minimizing the misclassification error rate using a surrogate convex loss. In: Langford J, Pineau J (eds) Proceedings of the 29th International Conference on Machine Learning (ICML-12), Omnipress, New York, NY, USA, ICML '12, pp 1863–1870

Boyd SP, Cortes C, Mohri M, Radovanovic A (2012) Accuracy at the top. In: Advances In Neural Information Processing Systems (NIPS)

Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. In: International Conference on Pattern Recognition (ICPR)

Buja A, Stuetzle W, Shen Y (2005) Loss functions for binary class probability estimation and classification: Structure and applications

Chan PK, Stolfo SJ (1998) Learning with non-uniform class and cost distributions: Effects and a multi-classifier approach. In: KDD 1998 Workshop on Distributed Data Mining

Clémençon S, Vayatis N (2007) Ranking the best instances. Journal of Machine Learning Research 8

Clémençon S, Lugosi G, Vayatis N (2008) Ranking and empirical minimization of U-statistics. The Annals of Statistics 36(2)

Ertekin c, Rudin C (2011) On equivalence relationships between classification and ranking algorithms. Journal of Machine Learning Research 12

Freund Y, Iyer R, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research 4

Gao W, Zhou Z (2015) On the consistency of AUC pairwise optimization. In: International Joint Conference on Artificial Intelligence (IJCAI)

Ghosh A, Manwani N, Sastry PS (2015) Making risk minimization tolerant to label noise. Neurocomputing 160:93–107

Järvelin K, Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: ACM Conference on Research and Development in Information Retrieval, SIGIR ’00

Joachims T (2005) A support vector method for multivariate performance measures. In: Proceedings of the 22Nd International Conference on Machine Learning, ACM, New York, NY, USA, ICML ’05, pp 377–384

Kar P, Narasimhan H, Jain P (2015) Surrogate functions for maximizing precision at the top. In: International Conference on Machine Learning (ICML)

Li N, Jin R, Zhou Z (2014a) Top rank optimization in linear time. CoRR abs/1410.1462

Li N, Jin R, Zhou ZH (2014b) Top rank optimization in linear time. In: Advances in Neural Information Processing Systems

Liu LP, Dietterich TG, Li N, Zhou ZH (2015) Transductive optimization of top k precision. In: International Joint Conference on Artificial Intelligence (IJCAI)

Long PM, Servedio RA (2010) Random classification noise defeats all convex potential boosters. Machine Learning 78(3):287–304

Menon AK, Ong CS (2016) Linking losses for density ratio and class-probability estimation. In: International Conference on Machine Learning (ICML)

Narasimhan H, Agarwal S (2013) SVMpAUC: a new support vector method for optimizing partial AUC based on a tight convex upper bound. In: ACM International Conference on Knowledge discovery and data mining (KDD)

Nesterov Y (2004) Introductory lectures on convex optimization: A basic course. Kluwer Academic Publishers

Nock R, Nielsen F (2009) On the efficient minimization of classification calibrated surrogates. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in Neural Information Processing Systems 21, Curran Associates, Inc., pp 1201–1208

Rakotomamonjy A (2012) Sparse support vector infinite push. In: International Conference on Machine Learning (ICML), Omnipress, USA, ICML’12, pp 339–346

Reid MD, Williamson RC (2010) Composite binary losses. Journal of Machine Learning Research 11

Rudin C (2009) The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. Journal of Machine Learning Research 10

Rudin C, Wang Y (2018) Direct learning to rank and rerank. In: Storkey A, Perez-Cruz F (eds) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, Playa Blanca, Lanzarote, Canary Islands, Proceedings of Machine Learning Research, vol 84, pp 775–783

Scott C (2012) Calibrated asymmetric surrogate losses. Electronic Journal of Statistics 6

Uematsu K, Lee Y (2012) On theoretically optimal ranking functions in bipartite ranking. Unpublished manuscript

Xu H, Mannor S (2012) Robustness and generalization. Machine Learning 86(3):391–423

Yue Y, Finley T, Radlinski F, Joachims T (2007) A support vector method for optimizing average precision. In: ACM Conference on Research and Development in Information Retrieval (SIGIR)

Yun H, Raman P, Vishwanathan SVN (2014) Ranking via robust binary classification. In: Advances in Neural Information Processing Systems (NIPS)

Zălinescu C (2002) Convex Analysis in General Vector Spaces. World Scientific

A Proofs of results in the main body

Proof (Proof of Proposition 1)
Let

$$M(f; D) \doteq \max_{x' \in \mathrm{supp}(Q)} f(x').$$

The desired objective is

$$\begin{aligned}
&\min_f \max_{x' \in \mathrm{supp}(Q)} \mathbb{E}_{\mathsf{X} \sim P} \left[\phi(f(\mathsf{X}) - f(x')) \right] \\
&= \min_f \mathbb{E}_{\mathsf{X} \sim P} \left[\phi(f(\mathsf{X}) - M(f; D)) \right] \text{ since } \phi \text{ is non-increasing} \\
&= \min_f \sum_{P(x) > 0, Q(x) = 0} P(x) \cdot \phi(f(x) - M(f; D)) + \sum_{P(x) > 0, Q(x) > 0} P(x) \cdot \phi(f(x) - M(f; D)).
\end{aligned}$$

Clearly, the set of optimal scorers is closed under translation. Thus, without loss of generality, we can assume that $M(f; D) = 0$, giving the objective

$$\min_f \sum_{P(x) > 0, Q(x) = 0} P(x) \cdot \phi(f(x)) + \sum_{P(x) > 0, Q(x) > 0} P(x) \cdot \phi(f(x))$$

subject to the constraint that $f(x) \leq 0$ for $Q(x) > 0$ (so that indeed $M(f) = 0$ is an upper bound on all negative scores), and $(\exists x)\, Q(x) > 0 \wedge f(x) = 0$ (so that indeed $M(f) = 0$ is attained).

Suppose that $\eta(x) = 1$, i.e. $P(x) > 0, Q(x) = 0$. Then, the only appearance of $f(x)$ is in the corresponding ϕ term in the first summation above. To minimise this term, we need to choose $f(x) \in \operatorname{argmin}_v \phi(v)$.

Suppose that $\eta(x) \in (0, 1)$, i.e. $P(x) > 0, Q(x) > 0$. Then, $f(x)$ appears in the ϕ term in the second summation above, and must satisfy $f(x) \leq 0$ by the constraint. Choosing $f(x) < 0$ would be suboptimal, because for the appearance in the second summation, we would be considering ϕ of a negative quantity, and thus be attaining a higher loss, by assumption that ϕ is non-increasing and $\phi(0) < \phi(0^-)$. Thus, we need to choose $f(x) = 0$ i.e. $f(x)$ is identical for *all* such x.

Suppose that $\eta(x) = 0$, i.e. $P(x) = 0, Q(x) > 0$. Then, the only appearance of $f(x)$ is in the constraint $f(x) \leq 0$. We can choose any such $f(x)$; the precise choice does not matter, as all such choices result in the same $M(f; D)$, and thus do not affect the final objective.

To summarise, we have

$$\eta(x) = 1 \implies f(x) \in \operatorname{argmin}_v \phi(v)$$

$$\begin{aligned}\eta(x) \in (0, 1) &\implies f(x) = 0\\ \eta(x) = 0 &\implies f(x) \leq 0\end{aligned}$$

and by translation invariance of the objective, the result follows.

Proof (Proof of Corollary 1) We have $\mathrm{Im}(\eta) \subseteq (0, 1)$ since $\mathrm{Im}(u) \subseteq (0, 1)$, and by Cauchy-Schwartz and the assumption on $\|w^*\|_2$ and[7] $\|x\|_2$ we have $|\langle w^*, x\rangle| < +\infty$ (so that we do not consider e.g. the possibility that $u(+\infty) = 1$). Thus, by Proposition 1, the set of Bayes-optimal scorers for $R_{\mathrm{ptop}}(f; D, \phi)$ is the set of constant functions. This has non-empty intersection with the set of linear scorers, since we can just use f_w for $w = \mathbf{0}$. Thus, $w = \mathbf{0}$ this must be a risk minimiser when using linear scorers.

Proof (Proof of Lemma 1) We can rewrite the minimisation of the PTop risk as follows:

$$\begin{aligned}\min_{f\in\mathcal{F}} R_{\mathrm{ptop}}(f; D, \phi) &= \min_{f\in\mathcal{F}} \max_{x'\in\mathrm{supp}(Q)} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(f(\mathsf{X}) - f(x')\right)\right]\\ &= \min_{f\in\mathcal{F}} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(f(\mathsf{X}) - \max_{x'\in\mathrm{supp}(Q)} f(x')\right)\right]\\ &= \min_{g\in\mathcal{F}} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(g(\mathsf{X})\right)\right] : \max_{x'\in\mathrm{supp}(Q)} g(x') = 0,\end{aligned}$$

where the last line uses the translation invariance of $\mathcal{F}$ to define a new scorer $g\colon x \mapsto f(x) - \max_{x'\in\mathrm{supp}(Q)} f(x')$.

To argue that the equality constraint $\max_{x'\in\mathrm{supp}(Q)} g(x') = 0$ can be replaced by an inequality, suppose

$$g^* \in \operatorname*{argmin}_{g} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(g(\mathsf{X})\right)\right] : \max_{x'\in\mathrm{supp}(Q)} g(x') \leq 0,$$

and $M(g; D) < 0$. We show that g^* can be equivalently converted to an optimal solution with $M(g; D) = 0$: replace g^* by $\tilde{g}$, where all instances in $\mathrm{supp}(Q)$ have score increased by $-M(g; D)$. The resulting $\tilde{g}$ has $M(\tilde{g}; D) = 0$; further, it cannot result in an increase to the objective, since ϕ is non-increasing. Hence, there is always an optimal solution to the inequality constrained objective that satisfies $M(g; D) = 0$.

Note now that we can rewrite the constraint

$$\begin{aligned}\min_{f} R_{\mathrm{ptop}}(f; D, \phi) &= \min_{g} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(g(\mathsf{X})\right)\right] : \max_{x'\in\mathrm{supp}(Q)} g(x') = 0\\ &= \min_{g} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(g(\mathsf{X})\right)\right] : \max_{x'\in\mathrm{supp}(Q)} g(x') \leq 0\\ &= \min_{g} \mathbb{E}_{\mathsf{X}\sim P}\left[\phi\left(g(\mathsf{X})\right)\right] + \mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell(-1, g(\mathsf{X}'))\right]\\ &= \min_{g} R_{\mathrm{bal}}(g; D, \ell).\end{aligned}$$

Proof (Proof of Proposition 2) The risk is

$$\begin{aligned}R_{\mathrm{rtop}}(f; D, \phi) &= R_{\mathrm{bal}}(f; D, \ell)\\ &= \mathbb{E}_{\mathsf{X}\sim P}\left[\ell(+1, f(\mathsf{X}))\right] + \mathbb{E}_{\mathsf{X}'\sim Q}\left[\ell(-1, f(\mathsf{X}'))\right]\\ &= \mathbb{E}_{\mathsf{X}\sim M}\left[\pi^{-1}\cdot\eta(\mathsf{X})\cdot\ell(+1, f(\mathsf{X})) + (1-\pi)^{-1}\cdot(1-\eta(\mathsf{X}))\cdot\ell(-1, f(\mathsf{X}))\right].\end{aligned}$$

The Bayes-optimal scorers may be computed pointwise. Pick any $x \in \mathcal{X}$, and consider the minimiser of the inner expression:

$$f^*(x) \in \operatorname*{argmin}_{v\in\mathbb{R}} \pi^{-1}\cdot\eta(x)\cdot\phi(v) + (1-\pi)^{-1}\cdot(1-\eta(x))\cdot\begin{cases}+\infty & \text{if } v > 0\\ \phi(-v) & \text{if } v \leq 0.\end{cases} \tag{20}$$

When $\eta(x) = 1$, the minimiser is evidently just any element of $\operatorname{argmin}_{\,v\in\mathbb{R}}\phi(v)$, which is $(0, \infty)$. When $\eta(x) < 1$, by the implicit constraint on the negative scores owing to the partial loss ℓ_{-1},

$$f^*(x) \in \operatorname*{argmin}_{v\in(-\infty,0]} \pi^{-1}\cdot\eta(x)\cdot\phi(v) + (1-\pi)^{-1}\cdot(1-\eta(x))\phi(-v)$$

[7] We assumed $\mathcal{X} \subseteq \mathbb{R}^d$, rather than $\mathbb{R}^d_*$.

$$
\begin{aligned}
&= \operatorname*{argmin}_{v \in (-\infty, 0]} \pi^{-1} \cdot \eta(x) \cdot \phi(v) + (1-\pi)^{-1} \cdot (1 - \eta(x))(1 - \phi(v)) \\
&= \operatorname*{argmin}_{v \in (-\infty, 0]} (\pi^{-1} \cdot \eta(x) + (1-\pi)^{-1} \cdot (1 - \eta(x))) \cdot \phi(v) \\
&= \operatorname*{argmin}_{v \in (-\infty, 0]} (\eta(x) - \pi) \cdot \phi(v).
\end{aligned}
$$

Thus, when $\eta(x) \in (\pi, 1)$, the optimal solution is just $\{0\}$; when $\eta(x) = \pi$, the optimal solution is any element of $(-\infty, 0]$; and when $\eta(x) \in [0, \pi)$, the optimal solution is any element of $(-\infty, 0)$.

Proof (Proof of Proposition 3) We proceed as per the proof of Proposition 20, and begin the form of the minimiser in Equation 20. When $\eta(x) = 1$, the minimiser is evidently just $f^*(x) = \operatorname{argmin}_{v \in \mathbb{R}} \phi(v)$, which is[8] $f^*(x) = \Psi(1)$. When $\eta(x) < 1$, by the implicit constraint on the negative scores owing to the partial loss ℓ_{-1},

$$
f^*(x) \in \operatorname*{argmin}_{v \in (-\infty, 0]} \pi^{-1} \cdot \eta(x) \cdot \phi(v) + (1-\pi)^{-1} \cdot (1 - \eta(x)) \phi(-v)
$$

Now, the unconstrained minimiser is $f^*_{\mathrm{unc}}(x) = \bar{\Psi}_\pi(\eta(x))$, since ϕ is strictly proper composite with link Ψ, and a simple calculation reveals that the balanced version of such a loss is also strictly proper composite with link $\bar{\Psi}_\pi$ (for a proof, see e.g. Menon and Ong (2016, Lemma 5)). If $f^*_{\mathrm{unc}}(x) \leq 0$, clearly this remains the minimiser. If $f^*_{\mathrm{unc}}(x) > 0$, then the minimiser must be 0, because of convexity of $L(\eta(x), v)$ in v (being the convex combination of two convex functions). Thus, the minimiser is $\min(0, \bar{\Psi}_\pi \circ \eta(x))$, and

$$
\operatorname*{argmin}_f R_{\mathrm{rtop}}(f; D, \phi) = \begin{cases} \Psi(1) & \text{if } \eta(x) = 1 \\ \bar{\Psi}_\pi(\eta(x)) \wedge 0 & \text{else.} \end{cases} \tag{21}
$$

We can be more precise about the case $\eta(x) < 1$: any strictly proper composite margin loss has $\Psi(1/2) = 0$ by Reid and Williamson (2010, Section 4.3). By definition of $\bar{\Psi}_\pi$ and invertibility of Ψ, we have

$$
\begin{aligned}
\bar{\Psi}_\pi(u) \geq 0 &\iff (\Psi \circ g_\pi)(u) \geq 0 \\
&\iff g_\pi(u) \geq \Psi^{-1}(0) \\
&\iff g_\pi(u) \geq 1/2 \\
&\iff 2 \cdot (1 - \pi) \cdot u \geq \pi + (1 - 2 \cdot \pi) \cdot u \\
&\iff u \geq \pi.
\end{aligned}
$$

Thus, $\bar{\Psi}_\pi \circ \eta$ is non-negative whenever $\eta \in [\pi, 1]$. From Equation 21, this means that when $\eta(x) \in [\pi, 1)$, the Bayes-optimal scorer is 0. The argument for this case being attained by some instance for non-separable D where η takes on more than two distinct values is as per Example 4.

Proof (Proof of Proposition 4) We follow closely the proof of Agarwal (2011, Theorem 5.1). First,

$$
\begin{aligned}
R(f; D, \phi) &= \mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - \max_{x \in \mathrm{supp}(Q)} f(x)\right)\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(\max_{x \in \mathrm{supp}(Q)} f(x) - f(\mathsf{X})\right)\right] \\
&= \max_{x \in \mathrm{supp}(Q)} \mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - f(x)\right)\right] + \min_{x \in \mathrm{supp}(Q)} \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(f(x) - f(\mathsf{X})\right)\right] \\
&\leq \max_{x \in \mathrm{supp}(Q)} \left(\mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - f(x)\right)\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(f(x) - f(\mathsf{X})\right)\right]\right).
\end{aligned}
$$

The inequality above is since, for any functions f, g, if $x^* \in \operatorname{argmax}_x f(x)$

$$
\begin{aligned}
\max_x f(x) + \min_x g(x) &= f(x^*) + g(x^*) + \min_x g(x) - g(x^*) \\
&\leq f(x^*) + g(x^*) \\
&\leq \max_x \left(f(x) + g(x)\right).
\end{aligned}
$$

Now define

$$
\begin{aligned}
L(f; P, Q, x, \phi) &\doteq \mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - f(x)\right)\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(f(x) - f(\mathsf{X})\right)\right] \\
R_\infty(f; P, Q, \phi) &\doteq \max_{x \in \mathrm{supp}(Q)} L(f; P, Q, x),
\end{aligned} \tag{22}
$$

[8] In fact, it is easy to check that $\Psi(1) = \bar{\Psi}_\pi(1)$.

so that $R(f; D, \phi) \leq R_\infty(f; P, Q, \phi)$.

We are interested in bounding $R(f; D, \ell^{01})$. Following Agarwal (2011, Theorem 5.1), we consider the margin loss at $\gamma > 0$,

$$\ell^\gamma(v) \doteq [\![v < \gamma]\!] + \frac{1}{2} \cdot [\![v = \gamma]\!].$$

Clearly, $R_\infty(f; P, Q, \ell^{01}) \leq R_\infty(f; P, Q, \ell^\gamma)$.

We now focus on generalisation bounds for $R_\infty(f; P, Q, \ell^\gamma)$, which imply generalisation bounds for $R(f; D, \ell^{01})$. (For clarity, in the sequel we will drop the dependence on ℓ^γ.) Specifically, suppose we have a sample $\mathsf{S} \sim D^{n_+ + n_-}$, which we will think of as a positive sample $\mathsf{S}^+ \sim P^{n_+}$ and negative sample $\mathsf{S}^- \sim Q^{n_-}$. We are interested in the uniform deviation bound

$$\sup_{f \in \mathcal{F}} R_\infty(f; P, Q) - R_\infty(f; \mathsf{S}^+, \mathsf{S}^-),$$

where, following the notation of Equation 22, we have the natural empirical risks

$$\begin{aligned} L(f; \mathsf{S}^+, \mathsf{S}^-, x, \phi) &\doteq \frac{1}{n_+} \sum_{i=1}^{n_+} \phi\left(f(x_i^+) - f(x)\right) + \frac{1}{n_-} \sum_{j=1}^{n_-} \phi\left(f(x) - f(x_j^-)\right) \\ R_\infty(f; \mathsf{S}^+, \mathsf{S}^-, \phi) &\doteq \max_{1 \leq j \leq n_-} L(f; \mathsf{S}^+, \mathsf{S}^-, x_j). \end{aligned}$$

To establish our bound, we will also need the following hybrid risk,

$$L(f; P, \mathsf{S}^-, x, \phi) \doteq \mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - f(x)\right)\right] + \frac{1}{n_-} \sum_{j=1}^{n_-} \phi\left(f(x) - f(x_j^-)\right).$$

As in Agarwal (2011, Theorem 5.1), we decompose the quantity of interest, but with one additional term to account for our risk having an additional expectation:

$$\begin{aligned} R_\infty(f; P, Q) - R_\infty(f; \mathsf{S}^+, \mathsf{S}^-) = R_\infty(f; P, Q) &- \max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-) + \\ &\max_{1 \leq j \leq n_-} L(f; P, \mathsf{S}^-, x_j^-) - R_\infty(f; \mathsf{S}^+, \mathsf{S}^-) + \\ &\max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-) - \max_{1 \leq j \leq n_-} L(f; P, \mathsf{S}^-, x_j^-). \end{aligned}$$

We bound the deviations for each of these terms for a fixed $f \in \mathcal{F}$.

First term. We are interested in

$$R_\infty(f; P, Q) - \max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-) = \max_{x \in \mathrm{supp}(Q)} L(f; P, Q, x) - \max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-).$$

Here, the difference arises simply from taking an empirical versus distributional maximum. Following the strategy of Agarwal (2011, Theorem 5.1), we have

$$\begin{aligned} P_1 &\doteq \mathbb{P}_{\mathsf{S}^-}\left(R_\infty(f; P, Q) - \max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-) > \epsilon/6\right) \\ &= \mathbb{P}_{\mathsf{S}^-}\left(\max_{1 \leq j \leq n_-} L(f; P, Q, x_j^-) < R_\infty(f; P, Q) - \epsilon/6\right) \\ &= \mathbb{P}_{\mathsf{S}^-}\left((\forall 1 \leq j \leq n_-) L(f; P, Q, x_j^-) < R_\infty(f; P, Q) - \epsilon/6\right) \\ &= \prod_{j=1}^{n_-} \mathbb{P}_{x_j^-}\left(L(f; P, Q, x_j^-) < R_\infty(f; P, Q) - \epsilon/6\right) \text{ since } x_j^- \text{ are drawn iid} \\ &= (\rho(f, \gamma, \epsilon/6))^n Neg, \end{aligned}$$

where

$$\rho(f, \gamma, \epsilon/6) \doteq \mathbb{P}_{x \sim Q}\left(L(f; P, Q, x) < R_\infty(f; P, Q) - \epsilon/6\right).$$

Second term. We are interested in

$$\max_{1 \leq j \leq n_-} L(f; P, \mathsf{S}^-, x_j^-) - R_\infty(f; \mathsf{S}^+, \mathsf{S}^-) = \max_{1 \leq j \leq n_-} L(f; P, \mathsf{S}^-, x_j^-) - \max_{1 \leq j \leq n_-} L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-).$$

Here, the difference arises simply from replacing the positive distribution with its empirical version. Following the strategy of Agarwal (2011, Theorem 5.1),

$$\begin{aligned}
P_2 &\doteq \mathbb{P}_{\mathsf{S}}\left(\max_{1 \le j \le n_-} L(f; P, \mathsf{S}^-, x_j^-) - \max_{1 \le j \le n_-} L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-) > \epsilon/6\right) \\
&\le \mathbb{P}_{\mathsf{S}}\left(\max_{1 \le j \le n_-} \left|L(f; P, \mathsf{S}^-, x_j^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-)\right| > \epsilon/6\right) \text{ since } \max f(x) - \max g(x) \le \max(f(x) - g(x)) \\
&\le \mathbb{P}_{\mathsf{S}}\left((\exists 1 \le j \le n_-)\left|L(f; P, \mathsf{S}^-, x_j^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-)\right| > \epsilon/6\right) \\
&\le \sum_{j=1}^{n_-} \mathbb{P}_{\mathsf{S}}\left(\left|L(f; P, \mathsf{S}^-, x_j^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-)\right| > \epsilon/6\right) \text{ by the union bound.}
\end{aligned}$$

Observe now that

$$L(f; P, \mathsf{S}^-, x_j^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-) = \mathbb{E}_{\mathsf{X} \sim P}\left[\phi\left(f(\mathsf{X}) - f(x_j^-)\right)\right] - \frac{1}{n_+}\sum_{i=1}^{n_+} \phi\left(f(x_i^+) - f(x_j^-)\right),$$

so that the term is independent of the negative sample S^-. Further, the difference is between a random variable and its expectation. We can thus write

$$\begin{aligned}
P_2 &= \sum_{j=1}^{n_-} \mathbb{P}_{\mathsf{S}^+, x_j^-}\left(\left|L(f; P, \mathsf{S}^-, x_j^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x_j^-)\right| > \epsilon/6\right) \\
&\le n_- \cdot \sup_{x^- \in \mathrm{supp}(Q)} \mathbb{P}_{\mathsf{S}^+}\left(\left|L(f; P, \mathsf{S}^-, x^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x^-)\right| > \epsilon/6\right) \\
&\le 2n_- \cdot e^{-\epsilon^2 n_+/18} \text{ by Hoeffding's inequality.}
\end{aligned}$$

Third term. We are interested in

$$\max_{1 \le j \le n_-} L(f; P, Q, x_j^-) - \max_{1 \le j \le n_-} L(f; P, \mathsf{S}^-, x_j^-).$$

Here, the difference arises simply from replacing the negative distribution with its empirical version. We follow an identical strategy to bounding the second term, but require an additional step of replacing the maximum over the sample with the maximum over the negative support:[9]

$$\begin{aligned}
P_2 &\doteq \mathbb{P}_{\mathsf{S}}\left(\max_{1 \le j \le n_-} L(f; P, Q, x_j^-) - \max_{1 \le j \le n_-} L(f; P, \mathsf{S}^-, x_j^-) > \epsilon/6\right) \\
&\le \mathbb{P}_{\mathsf{S}}\left(\max_{1 \le j \le n_-} \left|L(f; P, Q, x_j^-) - L(f; P, \mathsf{S}^-, x_j^-)\right| > \epsilon/6\right) \\
&\le \mathbb{P}_{\mathsf{S}}\left(\max_{x^- \in \mathrm{supp}(Q)} |L(f; P, Q, x^-) - L(f; P, \mathsf{S}^-, x^-)| > \epsilon/6\right) \\
&\le \mathbb{P}_{\mathsf{S}}\left((\exists x^- \in \mathrm{supp}(Q))\left|L(f; P, \mathsf{S}^-, x^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x^-)\right| > \epsilon/6\right) \\
&\le \sum_{x^- \in \mathrm{supp}(Q)} \mathbb{P}_{\mathsf{S}}\left(\left|L(f; P, \mathsf{S}^-, x^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x^-)\right| > \epsilon/6\right).
\end{aligned}$$

Observe now that

$$L(f; P, Q, x^-) - L(f; P, \mathsf{S}^-, x^-) = \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(f(x^-) - f(\mathsf{X})\right)\right] - \frac{1}{n_-}\sum_{j=1}^{n_-} \phi\left(f(x^-) - f(x_j^-)\right),$$

so that the term is independent of the positive distribution. Further, for fixed x^-, the difference is between a random variable and its expectation. We can thus write

$$P_3 = \sum_{x^- \in \mathrm{supp}(Q)} \mathbb{P}_{\mathsf{S}}\left(\left|L(f; P, \mathsf{S}^-, x^-) - L(f; \mathsf{S}^+, \mathsf{S}^-, x^-)\right| > \epsilon/6\right)$$

[9] Without this step, we would no longer be comparing a random variable with its expectation, as each negative instance features both in the maximum and empirical risk.

$$\leq 2n_q \cdot e^{-\epsilon^2 n_-/18} \text{ by Hoeffding's inequality.}$$

We remark that the dependence on n_q could be improved to one on a suitable covering number for $\operatorname{supp}(Q)$, since the max term for a suitable cover will be close to that studied above. In such a bound we would have to consider $\epsilon - \epsilon_0$, where ϵ_0 is the discrepancy between the two max terms, which will depend on the granularity of the cover.

Final bound. As per , Theorem 5.1 a covering number argument to $\mathcal{F}$ applied to the above yields that for any $\epsilon, \gamma > 0$,

$$R_{\mathrm{rtop}}(f; D) \leq R_\infty(f; \mathsf{S}, \ell^\gamma) + \epsilon$$

with probability at least $1 - \delta$, where

$$\delta = \mathcal{N}(\mathcal{F}, \epsilon\gamma/8) \cdot \left((\rho(f, \gamma, \epsilon/6))^{n_-} + 2n_- \cdot e^{-\epsilon^2 n_+/18} + 2n_q \cdot e^{-\epsilon^2 n_-/18} \right).$$

Proof (Proof of Proposition 5) Let $F(w, b; \mathsf{S}, \phi, \lambda)$ denote the regularised primal objective. For simplicity, let $\phi_{-1}(v) = \phi(-v)$. Starting from Equation 18, by rewriting ϕ in terms of its convex conjugate[10] ϕ^* and appealing to strong duality, we have

$$\begin{aligned}
\min_{w,b} F(w, b; \mathsf{S}, \phi, \lambda) &= \min_{w,b} \quad \frac{\lambda}{2}||w||^2 + \frac{1}{n_+}\sum_{i=1}^{n_+} \phi\left(w^T x_i^+ + b\right) + \frac{1}{n_-}\sum_{j=1}^{n_-} \phi_{-1}\left(w^T x_j^- + b\right) : w^T x_j^- + b \leq 0. \\
&= \min_{w,b} \max_{\bar{\alpha},\bar{\beta}\in\mathbb{R},\bar{\gamma}\geq\mathbf{0}} \frac{\lambda}{2}||w||^2 + \frac{1}{n_+}\sum_{i=1}^{n_+} \left(\bar{\alpha}_i \cdot (w^T x_i^+ + b) - \phi^*(\bar{\alpha}_i)\right) + \\
&\qquad \frac{1}{n_-}\sum_{n=1}^{n_-} \left(\bar{\beta}_j \cdot (w^T x_j^- + b) - \phi_{-1}^*(\bar{\beta}_j)\right) + \frac{1}{n_-}\sum_{j=1}^{n_-} \bar{\gamma}_j \cdot (w^T x_j^- + b) \\
&= \max_{\bar{\alpha},\bar{\beta}\in\mathbb{R},\bar{\gamma}\geq\mathbf{0}} \min_{w,b} \frac{\lambda}{2}||w||^2 + \frac{1}{n_+}\sum_{i=1}^{n_+} \left(\bar{\alpha}_i \cdot (w^T x_i^+ + b) - \phi^*(\bar{\alpha}_i)\right) + \\
&\qquad \frac{1}{n_-}\sum_{n=1}^{n_-} \left(\bar{\beta}_j \cdot (w^T x_j^- + b) - \phi_{-1}^*(\bar{\beta}_j)\right) + \frac{1}{n_-}\sum_{j=1}^{n_-} \bar{\gamma}_j \cdot (w^T x_j^- + b) \\
&= \max_{\bar{\alpha},\bar{\beta}\in\mathbb{R},\bar{\gamma}\geq\mathbf{0}} \min_{w} \frac{\lambda}{2}||w||^2 + w^T \left(\sum_{i=1}^{n_+}(1/n_+)\bar{\alpha}_i x_i^+ + \sum_{j=1}^{n_-}(1/n_-)(\bar{\beta}_j + \bar{\gamma}_j)x_j^-\right) + \\
&\qquad \min_b \left(\sum_{i=1}^{n_+}(1/n_+)\bar{\alpha}_i + (1/n_-)\sum_{j=1}^{n_-}(\bar{\beta}_j + \bar{\gamma}_j)\right) \cdot b - \frac{1}{n_+}\sum_{i=1}^{n_+} \phi^*(\bar{\alpha}_i) - \frac{1}{n_-}\sum_{j=1}^{n_-} \phi_{-1}^*(\bar{\beta}_j).
\end{aligned}$$

This is an appropriate juncture to pause. Observe that the optimal w is

$$w^* = -\frac{1}{\lambda}\left(\sum_{i=1}^{n_+}(1/n_+)\bar{\alpha}_i x_i^+ + \sum_{j=1}^{n_-}(1/n_-)(\bar{\beta}_j + \bar{\gamma}_j)x_j^-\right),$$

and that for the objective to be bounded away from $-\infty$, we must have $\frac{1}{n_+}\sum_{i=1}^{n_+} \bar{\alpha}_i + \frac{1}{n_-}\sum_{j=1}^{n_-}(\bar{\beta}_j + \bar{\gamma}_j) = 0$. Let us rescale the variables so that $\alpha_i = -(1/n_+) \cdot \bar{\alpha}_i, \beta_j = (1/n_-) \cdot \bar{\beta}_j$, and $\gamma_j = (1/n_-) \cdot \bar{\gamma}_j$. Then,

$$w^* = -\frac{1}{\lambda}\left(\sum_{i=1}^{n_+} \alpha_i x_i^+ - \sum_{j=1}^{n_-}(\beta_j + \gamma_j)x_j^-\right),$$

and the constraint is $\sum_{i=1}^{n_+} \alpha_i = \sum_{j=1}^{n_-}(\beta_j + \gamma_j)$. Thus,

$$\min_{w,b} F(w, b; \mathsf{S}, \lambda) = \max_{\alpha,\beta,\gamma} -\frac{1}{2\lambda}\left\|\sum_{i=1}^{n_+} \alpha_i x_i^+ - \sum_{j=1}^{n_-}(\beta_j + \gamma_j)x_j^-\right\|^2 - \frac{1}{n_+}\sum_{i=1}^{n_+} \phi^*(-n_+ \cdot \alpha_i) - \frac{1}{n_-}\sum_{j=1}^{n_-} \phi_{-1}^*(n_- \cdot \beta_j) :$$

$$\sum_{i=1}^{n_+} \alpha_i = \sum_{j=1}^{n_-}(\beta_j + \gamma_j), \alpha \in -\operatorname{dom}(\phi^*), \beta \in \operatorname{dom}(\phi_{-1}^*), \gamma \geq \mathbf{0}.$$

[10] This is allowed by the Fenchel-Moreau theorem Zălinescu (2002, Theorem 2.3.4) since ϕ is assumed differentiable, and hence lower semi-continuous, and is implicitly assumed to be a proper convex function.

Recalling that $\phi^*_{-1}(y) = (\phi(-v)^*)(y) = \phi^*(-y)$, we get

$$\min_{w,b} F(w,b;\mathsf{S},\lambda) = \max_{\alpha,\beta,\gamma} -\frac{1}{2\lambda}\left\|\sum_{i=1}^{n_+}\alpha_i x_i^+ - \sum_{j=1}^{n_-}(\beta_j+\gamma_j)x_j^-\right\|^2 - \frac{1}{n_+}\sum_{i=1}^{n_+}\phi^*\left(-n_+\cdot\alpha_i\right) - \frac{1}{n_-}\sum_{j=1}^{n_-}\phi^*\left(-n_-\cdot\beta_j\right) :$$
$$\sum_{i=1}^{n_+}\alpha_i = \sum_{j=1}^{n_-}(\beta_j+\gamma_j), \alpha \in -\mathrm{dom}(\phi^*), \beta \in -\mathrm{dom}(\phi^*), \gamma \geq \mathbf{0}.$$

Suppose for example that $\phi(v) = \max(0, 1-v)^2$. Then an elementary calculation reveals that $\phi^*(y) = y + y^2/4$ when $y \leq 0$, and $+\infty$ otherwise, so that $\phi^*(-y) = -y + y^2/4$ when $y \geq 0$, and $+\infty$ otherwise. Thus, in this case we would have the dual problem and objective

$$\min_{w,b} F(w,b;\mathsf{S},\lambda) = \max_{\alpha,\beta,\gamma} -\frac{1}{2\lambda}\left\|\sum_{i=1}^{n_+}\alpha_i x_i^+ - \sum_{j=1}^{n_-}(\beta_j+\gamma_j)x_j^-\right\|^2 + \mathbf{1}^T\alpha - \frac{n_+}{4}\cdot\alpha^T\alpha + \mathbf{1}^T\beta - \frac{n_-}{4}\cdot\beta^T\beta :$$
$$\sum_{i=1}^{n_+}\alpha_i = \sum_{j=1}^{n_-}(\beta_j+\gamma_j)$$
$$\alpha,\beta,\gamma \geq \mathbf{0}.$$

B Details of dual optimisation

We can optimise the dual objective of Equation 19 by leveraging Li et al (2014a, Algorithm 1). Specifically, we have the problem

$$\min_{(\alpha,\beta,\gamma)\in\Theta} g(\alpha,\beta,\gamma)$$

where

$$g(\alpha,\beta,\gamma) = \frac{1}{2\lambda}\left\|\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)\right\|^2 + \frac{1}{n_+}\sum_{n=1}^{n_+}\phi^*\left(-n_+\cdot\alpha_n\right) + \frac{1}{n_-}\sum_{m=1}^{n_-}\phi^*\left(-n_-\cdot\beta_m\right).$$

and

$$\Theta = \{(\alpha,\beta,\gamma) \colon \mathbf{1}^T\alpha = \mathbf{1}^T(\beta+\gamma)\}.$$

We mimic the key steps of Li et al (2014a, Algorithm 1), following the notations of that paper:

- *Auxiliary solutions*. We keep iterates of auxiliary solutions $\{s_k^\alpha, s_k^\beta, s_k^\gamma\}$, updated via

$$s_k^\alpha = (1+\omega_k)\cdot\alpha_k - \omega_k\cdot\alpha_{k-1}$$
$$s_k^\beta = (1+\omega_k)\cdot\beta_k - \omega_k\cdot\beta_{k-1}$$
$$s_k^\gamma = (1+\omega_k)\cdot\gamma_k - \omega_k\cdot\gamma_{k-1}$$

  where ω_k is as per Li et al (2014a, Algorithm 1, Line 4).
- *Gradient computation*. We compute the gradients of the objective with respect to each variable:

$$\begin{aligned}
\nabla_\alpha g(\alpha,\beta,\gamma) &= \frac{1}{\lambda}\cdot(\mathbf{X}^+)^T(\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)) - \sum_{n=1}^{n_+}(\phi^*)'(-n_+\cdot\alpha_n) \\
\nabla_\beta g(\alpha,\beta,\gamma) &= -\frac{1}{\lambda}\cdot(\mathbf{X}^-)^T(\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)) - \sum_{m=1}^{n_-}(\phi^*)'(-n_-\cdot\beta_n) \qquad (23)\\
\nabla_\gamma g(\alpha,\beta,\gamma) &= -\frac{1}{\lambda}\cdot(\mathbf{X}^+)^T(\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)).
\end{aligned}$$

When $\phi(v) = \max(0, 1-v)^2$ for example, we have

$$\nabla_\alpha g(\alpha,\beta,\gamma) = \frac{1}{\lambda}\cdot(\mathbf{X}^+)^T(\mathbf{X}^+\alpha - \mathbf{X}^-(\beta+\gamma)) - \mathbf{1} + \frac{n_+}{2}\cdot\alpha$$

$$\nabla_\beta g(\alpha, \beta, \gamma) = -\frac{1}{\lambda} \cdot (\mathbf{X}^-)^T (\mathbf{X}^+ \alpha - \mathbf{X}^-(\beta + \gamma)) - \mathbf{1} + \frac{n_-}{2} \cdot \beta$$
$$\nabla_\gamma g(\alpha, \beta, \gamma) = -\frac{1}{\lambda} \cdot (\mathbf{X}^+)^T (\mathbf{X}^+ \alpha - \mathbf{X}^-(\beta + \gamma)).$$

- *Projection.* We compute a projection of a tentative solution $(\alpha_0, \beta_0, \gamma_0)$ onto the feasible set Θ by solving

$$\min_{\alpha,\beta,\gamma} \frac{1}{2} \|\alpha - \alpha_0\|_2^2 + \frac{1}{2} \|\beta - \beta_0\|_2^2 + \frac{1}{2} \|\gamma - \gamma_0\|_2^2 :$$
$$\mathbf{1}^T \alpha = \mathbf{1}^T (\beta + \gamma), \alpha, \beta, \gamma \geq \mathbf{0}$$

  which is equivalently

$$\min_{\alpha,\delta} \frac{1}{2} \|\alpha - \alpha_0\|_2^2 + \frac{1}{2} \|\delta - \delta_0\|_2^2 + :$$
$$\mathbf{1}^T \alpha = \mathbf{1}^T \delta, \alpha, \delta \geq \mathbf{0}$$

  where $\delta = [\beta; \gamma] \in \mathbb{R}^{2 \cdot n_-}$ and $\delta_0 = [\beta_0; \gamma_0] \in \mathbb{R}^{2 \cdot n_-}$. We note that the above can thus be solved by directly invoking the projection subroutine of Li et al (2014a, Section 3.2.2), where we simply read off β^* and γ^* from the augmented solution δ^*.
- *Stopping condition.* We compute as part of the stopping condition the squared norms $\|\nabla_\alpha g(\alpha, \beta, \gamma)\|_2^2$, $\|\nabla_\beta g(\alpha, \beta, \gamma)\|_2^2$ and $\|\nabla_\gamma g(\alpha, \beta, \gamma)\|_2^2$ using Equation 23.

C The reverse RectTop risk

Proposition 3 shows that minimisation of the rectified PTop accurately orders the negatives, while collapsing all positives into a single score. However, Clémençon and Vayatis (2007, Section 3.1) suggested the gold-standard scorer for top ranking problems should have the *opposite* behaviour.

In fact, this can be achieved with a simple variant of Equation 13: suppose

$$\ell(+1, v) = \begin{cases} \phi(v) & \text{if } v \geq 0 \\ +\infty & \text{if } v < 0 \end{cases} \qquad \ell(-1, v) = \phi(-v).$$

Now define the *reverse RectTop ϕ-risk* as

$$R_{\text{rbot}}(f; D, \phi) \doteq R_{\text{bal}}(f; D, \ell) = R_{\text{rtop}}(-f; D_{\text{rev}}, \ell), \tag{24}$$

where $D_{\text{rev}} = (Q, P, \pi)$ reverses the class-conditionals in D. This rectifies the following *negatives at the bottom* risk, per Rudin (2009, Section 7):

$$R_{\text{nbot}}(f; D, \phi) = \max_{x \in \text{supp}(P)} \mathbb{E}_{\mathsf{X}' \sim Q} \left[\phi(-(f(x) - f(\mathsf{X}'))) \right],$$

which pushes negatives to score lower than *every positive*.

Example 6 Combining Equation 24 with Proposition 3, the Bayes-optimal scorer for $\phi(v) = e^{-v}$ is

$$f^*(x) = \begin{cases} +\infty & \text{if } \eta(x) = 1 \\ \frac{1}{2} \log \frac{\eta(x)}{1 - \eta(x)} - \frac{1}{2} \log \frac{\pi}{1 - \pi} & \text{if } \eta(x) \in [\pi, 1) \\ \{0\} & \text{if } \eta(x) \in [0, \pi) \end{cases}$$

so that, as intended, only instances with η greater than average are accurately modelled.