On Theoretically Optimal Ranking Functions in Bipartite Ranking

Kazuki Uematsu
Department of Statistics
The Ohio State University
Columbus, OH 43210
uematsu.1@osu.edu
AND
Yoonkyung Lee *
Department of Statistics
The Ohio State University
Columbus, OH 43210
yklee@stat.osu.edu

Abstract

This paper investigates the theoretical relation between loss criteria and the optimal ranking functions driven by the criteria in bipartite ranking. In particular, the relation between AUC maximization and minimization of ranking risk under a convex loss is examined. We characterize general conditions for ranking-calibrated loss functions in a pairwise approach, and show that the best ranking functions under convex ranking-calibrated loss criteria produce the same ordering as the likelihood ratio of the positive category to the negative category over the instance space. The result illuminates the parallel between ranking and classification in general, and suggests the notion of consistency in ranking when convex ranking risk is minimized as in the RankBoost algorithm for instance. For a certain class of loss functions including the exponential loss and the binomial deviance, we specify the optimal ranking function explicitly in relation to the underlying probability distribution. In addition, we present an in-depth analysis of hinge loss optimization for ranking and point out that the RankSVM may produce potentially many ties or granularity in ranking scores due to the singularity of the hinge loss, which could result in ranking inconsistency. The theoretical findings are illustrated with numerical examples.

Keywords: Bipartite ranking, Convex loss, RankBoost, RankSVM, Ranking calibration, Ranking consistency

1 Introduction

How to order a set of given objects or instances reflecting their underlying utility, relevance or quality is a long standing problem. It has been a subject of interest in various fields, for example, the theory of choices or preferences in economics, and the theory of responses in psychometrics. Ranking as a statistical problem regards how to learn a real-valued scoring or ranking function from observed order relationships among the objects. Recently the ranking problem has gained great

*Lee's research was supported in part by National Science Foundation grant DMS-12-09194. The authors thank Shivani Agarwal for helpful discussions during the ASA Conference on Statistical Learning and Data Mining in Ann Arbor in 2012 and references.

interest in the machine learning community for information retrieval, web search, and collaborative filtering; see, for example, Goldberg et al. (1992), Shardanand and Maes (1995), Crammer and Singer (2001), Freund et al. (2003), Cao et al. (2006) and Zheng et al. (2008).

As a special form of ranking, bipartite ranking involves instances from two categories (say, positive or negative), and given observed instances from the two categories, the goal is to learn a ranking function which places positive instances ahead of negative instances; see Agarwal et al. (2005). For example, in document retrieval, documents are categorized as either relevant or irrelevant, and from the observed documents one wants to find a ranking function over the documents space which ranks relevant documents higher than irrelevant documents.

There exists notable similarity between bipartite ranking and binary classification. However, ranking aims at correct ordering of instances rather than correct prediction of the categories associated with them. This distinction is clear in the loss criterion used to measure the error of a ranking function as opposed to classification error; the former is the bipartite ranking loss indicating the misordering of a pair of instances with known preference or order relationship while the latter is the misclassification loss. As a result, while a given discriminant function for classification can be used as a ranking function, specification of a threshold or a cut-off value is not needed for ranking.

The performance of a ranking function is closely connected to the so-called Receiver Operating Characteristic (ROC) curve of the function, which has been used in radiology, psychological diagnostics, pattern recognition and medical decision making. Minimization of the expected bipartite ranking loss is shown to be equivalent to maximization of the Area Under the ROC Curve (AUC) by using the link between the AUC criterion and the Wilcoxon-Mann-Whitney statistic as in Hanley and McNeil (1982). Cortes and Mohri (2004) further investigate the relation between the AUC and the classification accuracy, contrasting the two problems.

On the other hand, the similarity has prompted a host of applications of classification techniques such as boosting and support vector machines to ranking problems. For example, RankBoost proposed by Freund et al. (2003) is an adaptation of AdaBoost to combine preferences. Application of the large margin principle in classification to ranking has led to the procedures that aim to maximize the AUC by minimizing the ranking risk under a convex surrogate function of the bipartite ranking loss for computational efficiency. See Joachims (2002), Herbrich et al. (2000), Brefeld and Scheffer (2005) and Rakotomamonjy (2004) for optimization of AUC by support vector learning.

Further, to capture the practical need for accuracy of the instances near the top of the list in many ranking applications, various modifications of the standard formulation for bipartite ranking and alternative ranking criteria have been proposed (Rudin 2009, Cossock and Zhang 2008, Clémençon and Vayatis 2007, Le and Smola 2007).

Theoretical developments in the current literature regarding bipartite ranking in part center around the convergence of the empirical ranking risk to the minimal risk achievable within the class of ranking functions as an application of the standard learning theory for generalization bounds. See Agarwal et al. (2005), Agarwal and Niyogi (2005) and Clémençon et al. (2008).

With particular focus on the Bayes ranking consistency, we investigate the theoretical relation between a loss criterion used for pairwise ranking and the optimal ranking function driven by the criterion in this paper. Motivated by considerable developments of ranking algorithms and procedures in connection with classification, we examine how the optimal ranking functions defined by a family of convex surrogate loss criteria are related to the underlying probability distribution for data, and identify the explicit form of optimal ranking functions for some loss criteria. In doing so, we draw a parallel between binary classification and bipartite ranking and establish the minimal notion of consistency in ranking, namely *ranking calibration*, analogous to the notion of classification calibration or Fisher consistency, which has been studied quite extensively in the classification literature (see, for example, Bartlett et al. 2006, Zhang 2004). Recently, Kotlowski et al. (2011)

and Agarwal (2012) also study ranking consistency in connection with classification by examining the risk bounds for bipartite ranking when discriminant functions from binary classification are directly used as ranking functions. Duchi et al. (2010) examine ranking consistency of general loss functions defined over preference graphs in label ranking and show negative results about convex loss functions, which are largely due to the generality of label ranking in contrast to bipartite ranking.

In this paper we employ a pairwise ranking approach, which is standard in learning to rank. We show that the theoretically optimal ordering over the instance space is determined by the likelihood ratio of the positive category to the negative category, and the best ranking functions under some convex loss criteria produce the same ordering. In particular, the RankBoost algorithm with the exponential loss is shown to target a half of the log likelihood ratio on the population level. This result reveals the theoretical relationship between the ranking function from RankBoost and the discriminant function from AdaBoost. Rudin and Schapire (2009) arrive at a qualitatively similar conclusion through an algebraic proof of the finite sample equivalence of AdaBoost and RankBoost algorithms. The binomial deviance loss used in RankNet (Burges et al. 2005) also preserves the optimal ordering for ranking consistency. Further, the result suggests that discriminant functions for classification that are order-preserving transformations of the likelihood ratio (e.g. logit function) can be used as a consistent ranking function in general. We establish general conditions for ranking-calibrated losses, and show that they are stricter than those conditions for classification calibration as optimal ranking requires more information about the underlying conditional probability (a transformation of the likelihood ratio) than classification. Some classification-calibrated loss functions are also ranking-calibrated (e.g. exponential loss and binomial deviance loss), but the hinge loss for support vector ranking is a notable exception among the commonly used loss functions. We prove in this paper that the support vector ranking with the hinge loss may produce potentially many ties or granularity in ranking scores due to the singularity of the loss, and this could result in ranking inconsistency.

As a related work, Clémençon et al. (2008) partially investigate the theoretical aspects in ranking by defining a statistical framework that transforms the bipartite ranking problem into a pairwise binary classification problem. By minimizing empirical convex risk functionals in the framework, the authors study *ranking rules* which specify preference between two instances instead of real-valued *ranking or scoring functions* in our approach, and draw on the connection to convex risk minimization in binary classification. However, a ranking rule for a pair of instances or the associated function that induces the rule, in general, does not define a ranking function consistently in the form of pairwise difference as assumed in bipartite ranking. This fact yields significantly different results for the two formulations. Equivalence between the two formulations depends largely on the loss, and we specify the condition for loss in convex risk minimization under which the equivalence holds (see Theorem 7). Here, the functional equivalence means that the convex risk minimizer that induces the optimal ranking rule in pairwise binary classification is identical to the pairwise difference of the optimal bipartite ranking function. The condition implies that not all ranking-calibrated loss functions result in such functional equivalence. This result renders the relevance of the pairwise binary classification approach rather limited in practice because standard ranking algorithms such as RankBoost, RankNet, and RankSVM are designed to produce a scoring function instead of a ranking rule in the framework of bipartite ranking. Moreover, the pairwise binary classification approach does not make the important distinction between classification calibration and ranking calibration for loss functions. Theorem 3 and Section 3.3 highlight the difference in the two approaches. Some comments are made further on the link between our results and pertinent discussion in the paper later.

The rest of this paper is organized as follows. Section 2 introduces the problem setting and

specifies the best ranking function that maximizes the AUC. The properties of the theoretically optimal ranking functions under convex loss criteria are discussed together with conditions for ranking calibration, and the form of the optimal ranking function for a certain class of loss criteria is specified in Section 3. The general relation between the optimal bipartite ranking function and the optimal ranking rule in pairwise binary classification approach is investigated in Section 4, followed by numerical illustration of the results in Section 5. Conclusion and further discussions are in Section 6.

2 Bipartite ranking

2.1 Basic setting

Let $\mathcal{X}$ denote the space of objects or instances we want to rank. Suppose each object $X \in \mathcal{X}$ comes from two categories, either positive or negative, $\mathcal{Y} = \{1, -1\}$, and let Y indicate the category associated with X. Training data for ranking consist of independent pairs of (X, Y) from $\mathcal{X} \times \mathcal{Y}$. Suppose the data set has n_+ positive objects $\{x_i\}_{i=1}^{n_+}$ and n_- negative ones $\{x'_j\}_{j=1}^{n_-}$. From the data, we want to learn a ranking function such that positive objects are generally ranked higher than negative objects. A ranking function is a real-valued function defined on $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$, whose values determine the ordering of instances. For x and $x' \in \mathcal{X}$, x is preferred to x' by f if $f(x) > f(x')$.

For each pair of a positive object x and a negative object x', the bipartite ranking loss of f is defined as $l_0(f; x, x') = I(f(x) < f(x')) + \frac{1}{2} I(f(x) = f(x'))$, where $I(\cdot)$ is the indicator function. Note that the loss is invariant under any order-preserving transformation of f. The best ranking function can then be defined as the function f minimizing the empirical ranking risk over the training data

$$R_{n_+,n_-}(f) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} l_0(f; x_i, x'_j)$$

by considering all pairs of positive and negative instances from the data.

The ROC curve of a ranking function f shows the trade-off between true positive rates (TPR) and false positive rates (FPR) over a range of threshold values, where $\mathrm{TPR}(r) = |\{x_i | f(x_i) > r\}|/n_+$, and $\mathrm{FPR}(r) = |\{x'_j | f(x'_j) > r\}|/n_-$ for threshold value r. The AUC is shown to be equivalent to a U-statistic, the Wilcoxon-Mann-Whitney statistic (Hanley and McNeil 1982). Thus, the empirical ranking risk $R_{n_+,n_-}(f)$ is given by one minus the AUC of f, and minimization of the bipartite ranking risk is equivalent to maximization of the AUC.

2.2 Optimal ranking function

Theoretically the AUC of a ranking function is the probability that the function ranks a positive instance higher than a negative instance when they are drawn at random. Casting the AUC maximization problem in the context of statistical inference, consider hypothesis testing of $H_0 : Y = -1$ versus $H_a : Y = 1$ based on a 'test statistic' $f(x)$. For critical value r, the test rejects H_0 if $f(x) > r$, and retains H_0 otherwise. Then the size of the test is $P(f(X) > r | Y = -1)$, which is, in fact, the theoretical $\mathrm{FPR}(r)$, and the power of the test is $P(f(X) > r | Y = 1)$, which is the theoretical $\mathrm{TPR}(r)$ of f. Hence, the relationship between the false positive rate and the true positive rate of a ranking function f is the same as that between the size and the power of a test based on f in statistical hypothesis testing. This dual interpretation of ranking also appears in Clémençon et al. (2008) and Clémençon and Vayatis (2009b). By the Neyman-Pearson lemma, the

most powerful test at any fixed size is based on the likelihood ratio of x under the two hypotheses. This implies that the best ranking function which maximizes the theoretical AUC is a function of the likelihood ratio.

Let g_+ be the pdf or pmf of X for positive category, and let g_- be the pdf or pmf of X for negative category. For simplicity, we further assume that $0 < g_+(x) < \infty$ and $0 < g_-(x) < \infty$ for $x \in \mathcal{X}$ in this paper. The following theorem states that the optimal ranking function for bipartite ranking is any order-preserving function of the likelihood ratio $f_0^*(x) \equiv g_+(x)/g_-(x)$. For notational convenience, let $R_0(f) \equiv E(l_0(f; X, X'))$ denote the ranking error rate of f under the bipartite ranking loss, where X and X' are, respectively, a positive instance and a negative instance randomly drawn from the distributions with g_+ and g_-. The proof of the theorem is omitted here, but interested readers can refer to Uematsu and Lee (2011).

Theorem 1. *For any ranking function f, $R_0(f_0^*) \leq R_0(f)$.*

To see the connection of ranking with classification, let $\pi = P(Y = 1)$ and verify that the posterior probability

$$P(Y = 1|X = x) \quad = \quad \frac{\pi g_+(x)}{\pi g_+(x) + (1-\pi) g_-(x)} = \frac{f_0^*(x)}{f_0^*(x) + (1-\pi)/\pi}$$

is a monotonic transformation of $f_0^*(x)$. Indeed, through a different formulation of ranking, Clémençon et al. (2008) and Clémençon and Vayatis (2009b) showed the equivalent result that a class of optimal ranking functions should be strictly increasing transformations of the posterior probability. The difference in the formulation is described in Section 4. This fact implies that those discriminant functions from classification methods estimating the posterior probability consistently, for example, logistic regression, may well be used as a ranking function for minimal ranking error in practice.

3 Ranking with convex loss

Since minimization of ranking error involves non-convex optimization with discontinuous l_0 loss function, direct maximization of the AUC is not computationally advisable just as direct minimization of classification error under the misclassification loss is not. For computational advantages of convex optimization, many researchers have applied successful classification algorithms such as boosting and support vector machines to ranking for the AUC maximization by replacing the bipartite ranking loss with a convex surrogate loss.

In this section, we identify the form of the minimizers of convex ranking risks and examine the properties of the optimal ranking functions. Consider non-negative, non-increasing convex loss functions $l : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ which define $l(f(x) - f(x'))$ as a ranking loss, given a ranking function f and a pair of a positive instance x and a negative instance x'. For example, the RankBoost algorithm in Freund et al. (2003) takes the exponential loss, $l(s) = \exp(-s)$, for learning the best ranking function, and the support vector ranking in Joachims (2002), Herbrich et al. (2000) and Brefeld and Scheffer (2005) takes the hinge loss, $l(s) = (1 - s)_+$.

To understand the relation between a convex loss function l used to define a ranking loss and the minimizer of the ranking risk on the population level, let $R_l(f) \equiv E[l(f(X) - f(X'))]$ be the convex ranking risk and let f^* be the optimal ranking function minimizing $R_l(f)$ among all measurable functions $f : \mathcal{X} \to \mathbb{R}$. When f^* preserves the ordering of the likelihood ratio f_0^*, we call the loss l *ranking-calibrated*, which is analogous to the notion of classification-calibration of l for the Bayes error consistency in classification.

3.1 Special case

The following theorem states some special conditions on the loss function under which the theoretically best ranking function can be specified explicitly.

Theorem 2. *Suppose that l is convex, differentiable, $l'(s) < 0$ for all $s \in \mathbb{R}$, and $l'(-s)/l'(s) = \exp(s/\alpha)$ for some positive constant α. Then the optimal ranking function minimizing $E[l(f(X) - f(X'))]$ is of the form*

$$f^*(x) = \alpha \log(g_+(x)/g_-(x)) + \beta,$$

where β is an arbitrary constant.

Proof. Given (X, X'), let (V, V') be an unordered pair of (X, X') taking (X, X') or (X', X) with equal probability. Then the conditional probability of (X, X') given (V, V') is

$$\begin{cases} \dfrac{g_+(v)g_-(v')}{g_+(v)g_-(v') + g_+(v')g_-(v)} & \text{if } (x, x') = (v, v'), \\ \dfrac{g_+(v')g_-(v)}{g_+(v)g_-(v') + g_+(v')g_-(v)} & \text{if } (x, x') = (v', v), \end{cases}$$

and 0 otherwise. With (V, V'), $R_l(f) = E[l(f(X) - f(X'))]$ can be expressed as

$$\begin{aligned} & E_{V,V'}(E_{X,X'}[l(f(X) - f(X'))|V, V']) \\ = \; & E_{V,V'}\left[\frac{l(f(V) - f(V'))g_+(V)g_-(V') + l(f(V') - f(V))g_+(V')g_-(V)}{g_+(V)g_-(V') + g_+(V')g_-(V)}\right]. \end{aligned}$$

To find the minimizer of $R_l(f)$, it is sufficient to find f such that $E_{X,X'}[l(f(X) - f(X'))|V, V']$ is minimized for each $(V, V') = (v, v')$. For fixed (v, v'), let $s = f(v) - f(v')$. Noting that $g_+(v)g_-(v') + g_+(v')g_-(v)$ is fixed, consider minimizing $l(s)g_+(v)g_-(v') + l(-s)g_+(v')g_-(v)$ with respect to s. Under the assumption that l is differentiable, the above expression is minimized when $l'(s)g_+(v)g_-(v') - l'(-s)g_+(v')g_-(v) = 0$, or equivalently

$$\frac{l'(-s)}{l'(s)} = \frac{g_+(v)}{g_-(v)}\frac{g_-(v')}{g_+(v')}. \tag{1}$$

Define $G(s) \equiv l'(-s)/l'(s)$. Then clearly $G(0) = 1$ and $G(s)$ is increasing. Let v_0 be the point at which $g_+(v_0) = g_-(v_0)$. Using the relation $G(s) = \frac{g_+(v)}{g_-(v)}\frac{g_-(v')}{g_+(v')}$, taking $v' = v_0$, and solving for s, we get

$$f^*(v) - f^*(v_0) = G^{-1}\left(\frac{g_+(v)}{g_-(v)}\frac{g_-(v_0)}{g_+(v_0)}\right) = G^{-1}\left(\frac{g_+(v)}{g_-(v)}\right).$$

From the assumption of $G(s) = \exp(s/\alpha)$, we have $f^*(v) = f^*(v_0) + \alpha \log(g_+(v)/g_-(v))$, which completes the proof. □

Remark 1. *For the RankBoost algorithm, $l(s) = \exp(-s)$, and $l'(-s)/l'(s) = \exp(2s)$. Hence, Theorem 2 applies, and it gives $f^*(x) = \frac{1}{2}\log(g_+(x)/g_-(x))$ as the optimal ranking function up to an additive constant. Similarly, when $l(s) = \log(1 + \exp(-s))$, the negative log likelihood in logistic regression as in RankNet (Burges et al. 2005), $l'(-s)/l'(s) = \exp(s)$, and the optimal ranking function under the loss is given by $f^*(x) = \log(g_+(x)/g_-(x))$ up to a constant. Thus, the loss functions for RankBoost and RankNet are ranking-calibrated.*

3.2 General case

To deal with a general loss l beyond those covered by Theorem 2, we consider convex loss criteria. The next theorem specifies general conditions for ranking calibration, and states the general relation between the best ranking function f^* under convex ranking loss criteria and the likelihood ratio (g_+/g_-) in terms of the relative ordering of a pair of instances when $\mathcal{X}$ is continuous. The proof of the theorem is given in Appendix. Similar arguments can be made to establish the corresponding results for discrete $\mathcal{X}$.

Theorem 3. *Suppose that l is convex, non-increasing, differentiable, and $l'(0) < 0$. Let $f^* \equiv \arg\min_f R_l(f)$.*

(i) For almost every $(x, z) \in \mathcal{X} \times \mathcal{X}$, $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$ implies $f^*(x) > f^*(z)$.

(ii) If l' is one-to-one, then for almost every $(x, z) \in \mathcal{X} \times \mathcal{X}$, $\frac{g_+(x)}{g_-(x)} = \frac{g_+(z)}{g_-(z)}$ implies $f^*(x) = f^*(z)$.

Remark 2. *As an interesting example, consider $l(s) = (1 - s)_+$, the hinge loss in support vector ranking. It is differentiable at 0 with $l'(0) = -1$, but it has a singularity point at $s = 1$. Thus, Theorem 3 does not apply. In comparison, $l(s) = [(1-s)_+]^2$, the squared hinge loss, is differentiable everywhere, and Theorem 3 (i) implies that the optimal ranking function f^* under l preserves the order of the likelihood ratio without ties.*

3.3 Support vector ranking

To cover the hinge loss for support vector ranking, we resort to results in convex analysis. A subderivative of a convex function l at point s_0 is a real number c such that $l(s) - l(s_0) \geq c(s - s_0)$ for all $s \in \mathbb{R}$. The set of all subderivatives is called the subdifferential of l at s_0 and denoted by $\partial l(s_0)$. It can be shown that the subdifferential is a nonempty closed interval. For example, the subdifferential of hinge loss $l(s) = (1-s)_+$ at $s = 1$ is $[-1, 0]$. For a convex function l, its derivative may not be defined at some points, but a subderivative is always defined though it may not be unique. At differentiable points of the function, the subderivative is uniquely determined and the same as the derivative. The function is globally minimized at s_0 if and only if zero is contained in the subdifferential at s_0. For more details of convex analysis, see Rockafellar (1997).

First we illustrate potential ties in the optimal ranking function under the hinge loss with a toy example. We can derive explicitly the conditions under which ties can occur in this simple example.

3.3.1 Toy example

Suppose $\mathcal{X} = \{x_1^*, x_2^*, x_3^*\}$, and without loss generality, assume that $\frac{g_+(x_1^*)}{g_-(x_1^*)} < \frac{g_+(x_2^*)}{g_-(x_2^*)} < \frac{g_+(x_3^*)}{g_-(x_3^*)}$ for the pmfs of X and X', g_+ and g_-. Let f^* be a minimizer of the ranking risk under the hinge loss $l(s) = (1 - s)_+$. Define $s_1 = f(x_2^*) - f(x_1^*)$ and $s_2 = f(x_3^*) - f(x_2^*)$ for a ranking function f. Then we can express the risk $R_l(f)$ in terms of s_1 and s_2 as follows:

$$
\begin{aligned}
& R_l(s_1, s_2) \\
= \; & l(s_1)g_+(x_2^*)g_-(x_1^*) + l(-s_1)g_+(x_1^*)g_-(x_2^*) + l(s_2)g_+(x_3^*)g_-(x_2^*) + l(-s_2)g_+(x_2^*)g_-(x_3^*) \\
& + l(s_1 + s_2)g_+(x_3^*)g_-(x_1^*) + l(-s_1 - s_2)g_+(x_1^*)g_-(x_3^*) + \sum_{i=1}^{3} g_+(x_i^*)g_-(x_i^*).
\end{aligned}
$$

It is easy to check that truncation of s_1 and s_2 to 1 always reduces the risk if they are greater than 1. So, it is sufficient to consider $s_1 \leq 1$ and $s_2 \leq 1$ for the minimizer f^*. Letting $s_i^* =$

$f^*(x^*_{i+1}) - f^*(x^*_i)$ for $i = 1, 2$, we can show that the minimizer f^* has non-negative increments, that is, $s^*_i \geq 0$. Otherwise, there are three cases: i) $s^*_i < 0$ $(i = 1, 2)$, ii) $s^*_1 < 0$ and $s^*_2 \geq 0$, and iii) $s^*_1 \geq 0$ and $s^*_2 < 0$. From the assumption of ordering of x^*_i, for all $i = 1, 2$ and $k = 1, \ldots, 3 - i$

$$g_+(x^*_{i+k})g_-(x^*_i) > g_+(x^*_i)g_-(x^*_{i+k}). \tag{2}$$

Using the inequalities and the fact that for positive constants p and q with $p > q$, $p \cdot l(s) + q \cdot l(-s)$ is strictly decreasing in $(-\infty, 1]$ with minimum at $s = 1$, we can verify that i) $(s_1, s_2) = (0, 0)$, ii) $(s_1, s_2) = (0, 1)$, and iii) $(s_1, s_2) = (1, 0)$ yield strictly smaller risk values than (s^*_1, s^*_2), respectively. Thus, we can restrict the region for the increments of f^* to $0 \leq s_1 \leq 1$ and $0 \leq s_2 \leq 1$.

Taking the risk as a function of s_1 and a function of s_2, we have its subderivatives as

$$\begin{aligned}
\partial_{s_1} R_l(s_1, s_2) &= l'(s_1)g_+(x^*_2)g_-(x^*_1) - l'(-s_1)g_+(x^*_1)g_-(x^*_2) + l'(s_1 + s_2)g_+(x^*_3)g_-(x^*_1) \\
&\quad -l'(-s_1 - s_2)g_+(x^*_1)g_-(x^*_3) \\
\partial_{s_2} R_l(s_1, s_2) &= l'(s_2)g_+(x^*_3)g_-(x^*_2) - l'(-s_2)g_+(x^*_2)g_-(x^*_3) + l'(s_1 + s_2)g_+(x^*_3)g_-(x^*_1) \\
&\quad -l'(-s_1 - s_2)g_+(x^*_1)g_-(x^*_3).
\end{aligned} \tag{3}$$

Since s_1 and s_2 for f^* are non-negative, $l'(-s_1) = -1$ and $l'(-s_2) = -1$. Furthermore, we can prove $s_1 + s_2 \geq 1$. Otherwise, $s_1 < 1$ and $s_2 < 1$, which simplifies the subderivatives above to

$$-g_+(x^*_2)g_-(x^*_1) + g_+(x^*_1)g_-(x^*_2) - g_+(x^*_3)g_-(x^*_1) + g_+(x^*_1)g_-(x^*_3), \text{ and}$$
$$-g_+(x^*_3)g_-(x^*_2) + g_+(x^*_2)g_-(x^*_3) - g_+(x^*_3)g_-(x^*_1) + g_+(x^*_1)g_-(x^*_3).$$

However, (2) implies that the subderivatives then can not be zero. This contradicts that f^* is a risk minimizer. In summary, we have the constraints on s_1 and s_2 for f^* that $0 \leq s_1 \leq 1$, $0 \leq s_2 \leq 1$, and $s_1 + s_2 \geq 1$.

With the constraints, we show that the optimal increments s_1 and s_2 of a minimizer f^* can be identified explicitly in some cases. Let $a = g_+(x^*_1)g_-(x^*_2) - g_+(x^*_2)g_-(x^*_1) + g_+(x^*_1)g_-(x^*_3)$ and $b = g_+(x^*_2)g_-(x^*_3) - g_+(x^*_3)g_-(x^*_2) + g_+(x^*_1)g_-(x^*_3)$. Then the subderivatives in (3) are re-expressed as

$$g_-(x^*_1)[(l'(s_1) + 1)g_+(x^*_2) + l'(s_1 + s_2)g_+(x^*_3)] + a, \text{ and}$$
$$g_+(x^*_3)[(l'(s_2) + 1)g_-(x^*_2) + l'(s_1 + s_2)g_-(x^*_1)] + b.$$

If $a < 0$, then the term in the bracket has to be positive in order to have zero subderivative. This, in turn, implies that $l'(s_1) + 1 > 0$ since $l'(s_1 + s_2) \leq 0$. Hence, $s_1 \geq 1$ and together with the condition $s_1 \leq 1$, we have $s_1 = 1$. When $s_1 = 1$, the ranking risk $R_l(f)$ is simplified to a function of s_2 only up to a constant (ignoring $\sum_{i=1}^3 g_+(x^*_i)g_-(x^*_i)$):

$2g_+(x^*_1)g_-(x^*_2) + (1 - s_2)g_+(x^*_3)g_-(x^*_2) + (1 + s_2)g_+(x^*_2)g_-(x^*_3) + (2 + s_2)g_+(x^*_1)g_-(x^*_3)$
$= bs_2 + 2g_+(x^*_1)g_-(x^*_2) + g_+(x^*_2)g_-(x^*_3) + g_+(x^*_3)g_-(x^*_2) + 2g_+(x^*_1)g_-(x^*_3)$. Therefore, if $b < 0$, the risk is minimized at $s_2 = 1$, and if $b > 0$, it is minimized at $s_2 = 0$. When $b = 0$, s_2 can be any value between 0 and 1.

If $a = 0$, then either $[l'(s_1) + 1 = 0$ and $l'(s_1 + s_2) = 0]$ or $[l'(s_1) + 1 > 0$ and $l'(s_1 + s_2) < 0]$. The former leads to $s_1 \leq 1$ and $s_1 + s_2 \geq 1$ while the latter leads to $s_1 = 1$ and $s_2 = 0$. Similarly, if $a > 0$, then $l'(s_1 + s_2) < 0$, implying $s_1 + s_2 \leq 1$. However, from the constraint $s_1 + s_2 \geq 1$, we conclude that $s_1 + s_2 = 1$. When $s_1 + s_2 = 1$, again the ranking risk can be simplified to a function of one variable (say, s_1) only up to a constant:

$$\begin{aligned}
&(1 - s_1)g_+(x^*_2)g_-(x^*_1) + (1 + s_1)g_+(x^*_1)g_-(x^*_2) + 2g_+(x^*_1)g_-(x^*_3) \\
&+s_1 g_+(x^*_3)g_-(x^*_2) + (2 - s_1)g_+(x^*_2)g_-(x^*_3) \\
= \;& (a - b)s_1 + g_+(x^*_1)g_-(x^*_2) + g_+(x^*_2)g_-(x^*_1) + 2g_+(x^*_1)g_-(x^*_3) + 2g_+(x^*_2)g_-(x^*_3).
\end{aligned}$$

It is minimized at $s_1 = 0$ if $a > b$ or at $s_1 = 1$ if $a < b$. If $a = b$, then the minimizer s_1 is indefinite, and it can be any value between 0 and 1.

Combining with the same consideration of the subderivative with respect to s_2 in terms of b, we arrive at the following summary of the values of $s_1^* = f^*(x_2^*) - f^*(x_1^*)$ and $s_2^* = f^*(x_3^*) - f^*(x_2^*)$ for a risk minimizer f^*. Note that for some values of a and b, s_1^* and s_2^* are not uniquely determined and neither is f^*.

a	$-$	$-$	$-$	0	0	0	$+$	$+$		$+$	
b	$-$	0	$+$	$-$	0	$+$	$-$	0		$+$	
									$b < a$	$b = a$	$b > a$
s_1^*	1	1	1	[0,1]	[0,1]	1	0	0	0	[0,1]	1
s_2^*	1	[0,1]	0	1	[0,1]	0	1	1	1	[0,1]	0
$s_1^* + s_2^*$	2	[1,2]	1	[1,2]	[1,2]	1	1	1	1	1	1

The table above shows that the optimal increments for the support vector ranking function could be zero with the only exception of $a < 0$ and $b < 0$ case. Another notable fact is that the maximum increment is 1, which clearly stems from the singularity point of the hinge loss. Having at least one of s_1^* and s_2^* equal to zero means that the theoretically optimal ranking function f^* produces ties for the pair of x_1^* and x_2^* or x_2^* and x_3^*. Such ties make the ranking error rate of f^* strictly greater than the minimal ranking error rate, and thus f^* is not consistent with f_0^*. This toy example demonstrates that in general, ranking by risk minimization under the hinge loss could lead to inconsistency due to ties when the sample space is discrete.

To understand the ideal case when the optimal increments are both 1 and hence ties in ranking by f^* do not occur, we examine the conditions $a < 0$ and $b < 0$. Expressing $a < 0$ equivalently as $g_+(x_2^*)g_-(x_1^*) > g_+(x_1^*)(g_-(x_2^*) + g_-(x_3^*))$ and $b < 0$ as $g_+(x_3^*)g_-(x_2^*) > g_-(x_3^*)(g_+(x_1^*) + g_+(x_2^*))$, we can describe the conditions alternatively as

$$\frac{g_-(x_1^*)}{g_+(x_1^*)} > \frac{g_-(x_2^*)}{g_+(x_2^*)} + \frac{g_-(x_3^*)}{g_+(x_2^*)} \qquad \text{and} \qquad \frac{g_+(x_3^*)}{g_-(x_3^*)} > \frac{g_+(x_2^*)}{g_-(x_2^*)} + \frac{g_+(x_1^*)}{g_-(x_2^*)}.$$

From the inequalities, $a < 0$ can be interpreted as a condition for the gap between the likelihood ratios of g_-/g_+ at x_1^* and x_2^* being sufficiently large (more precisely, larger than $g_-(x_3^*)/g_+(x_2^*)$) and likewise, $b < 0$ means that the gap between the likelihood ratios of g_+/g_- at x_2^* and x_3^* is large enough. In other words, when the elements in the sample space are sufficiently apart in terms of their likelihood ratio, we expect theoretically optimal rankings by the function f^* to be without any ties. Otherwise, f^* could yield ties.

In addition, we see from the table that $a < b$ is a sufficient condition for $s_1^* = 1$ and similarly $b < a$ is a sufficient condition for $s_2^* = 1$. Note that $a < b$ is equivalent to $g_-(x_2^*) < g_+(x_2^*)$. From the result displayed in the table, by choosing either 0 or 1 arbitrarily for s_i^* in the indefinite cases, we can partition the space of all probability distributions g_+ and g_- on $\mathcal{X}$ into three cases: $(s_1^*, s_2^*) = (1,1)$, $(1,0)$ or $(0,1)$. Letting

$$\Delta_{12} \equiv \frac{g_-(x_1^*)}{g_+(x_1^*)} - \left(\frac{g_-(x_2^*)}{g_+(x_2^*)} + \frac{g_-(x_3^*)}{g_+(x_2^*)}\right) \quad \text{and} \quad \Delta_{23} \equiv \frac{g_+(x_3^*)}{g_-(x_3^*)} - \left(\frac{g_+(x_2^*)}{g_-(x_2^*)} + \frac{g_+(x_1^*)}{g_-(x_2^*)}\right),$$

we can verify the following relations:

(i) if $\Delta_{12} > 0$ and $\Delta_{23} > 0$, $(s_1^*, s_2^*) = (1,1)$.

(ii) if $\Delta_{23} < 0$ and $g_-(x_2^*) < g_+(x_2^*)$, $(s_1^*, s_2^*) = (1, 0)$.

(iii) if $\Delta_{12} < 0$ and $g_-(x_2^*) > g_+(x_2^*)$, $(s_1^*, s_2^*) = (0, 1)$.

Figure 1 illustrates such partitions of g_- distributions for various g_+ distributions that are given. Obeying the inherent restriction that $\frac{g_+(x_1^*)}{g_-(x_1^*)} < \frac{g_+(x_2^*)}{g_-(x_2^*)} < \frac{g_+(x_3^*)}{g_-(x_3^*)}$ and $\sum_{i=1}^3 g_+(x_i^*) = \sum_{i=1}^3 g_-(x_i^*) = 1$, it shows a partition of the feasible region of $(g_-(x_1^*), g_-(x_2^*))$ in each panel, based on (s_1^*, s_2^*) given $(g_+(x_1^*), g_+(x_2^*), g_+(x_3^*))$.

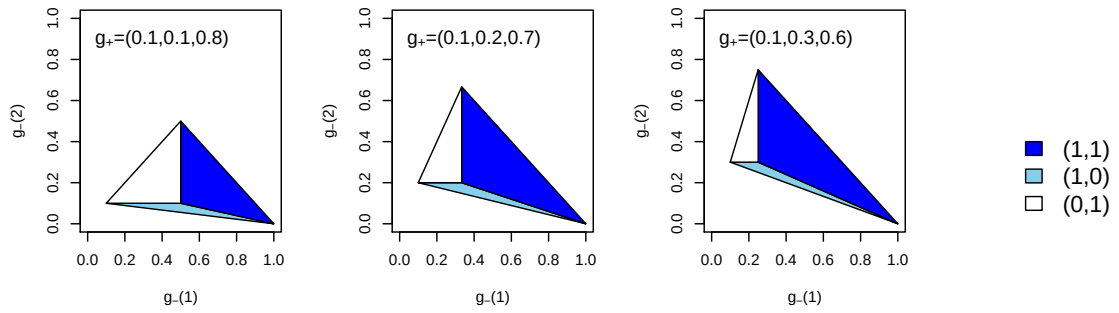

Figure 1: Partitions of feasible $(g_-(x_1^*), g_-(x_2^*))$ based on (s_1^*, s_2^*) for various g_+ distributions $(g_+(x_1^*), g_+(x_2^*), g_+(x_3^*))$. $g_-(x_i^*)$ is abbreviated as $g_-(i)$ for $i = 1, 2$.

3.3.2 General properties

By using the constructive way of identifying ranking functions with increments in the toy example, we derive general properties of optimal ranking functions under the hinge loss.

As illustrated in the toy example, the ranking risk is minimized by a unique set of optimal increments except for some degenerate cases of negligible measure. Without loss of generality in practical sense, we derive the following result under the assumption that the risk minimizer f^* is unique (up to an additive constant).

Theorem 4. *For discrete space $\mathcal{X} = \{x_i^*\}_{i=1}^N$ (with $N = \infty$ allowed), let $f^* = \arg\min_f E(1 - (f(X) - f(X'))_+$. Suppose that f^* is unique up to an additive constant. Then for every $(x, z) \in \mathcal{X} \times \mathcal{X}$, $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$ implies $f^*(x) \geq f^*(z)$.*

Proof. Without loss of generality, assume $\frac{g_+(x_1^*)}{g_-(x_1^*)} < \cdots < \frac{g_+(x_N^*)}{g_-(x_N^*)}$. To prove the theorem, we have only to show that the optimal increments $s_i^* \equiv f^*(x_{i+1}^*) - f^*(x_i^*)$ for $i = 1, \ldots, N-1$ are non-negative.

Consider a sequence of "smoothed" version of hinge loss $l(s) = (1 - s)_+$ defined as

$$l_n(s) = \begin{cases} 1 - s & \text{if} \quad s \leq 1 - 1/n \\ (n(s-1) - 1)^2/4n & \text{if} \quad 1 - 1/n < s \leq 1 + 1/n \\ 0 & \text{if} \quad s > 1 + 1/n. \end{cases} \tag{4}$$

Each l_n is differentiable, $0 \leq l_n(s) - l(s) \leq 1/4n$ for every $s \in \mathbb{R}$, and hence $l_n(s) \to l(s)$ as $n \to \infty$. Define $R(f) \equiv E[l(f(X) - f(X'))]$, $R_n(f) \equiv E[l_n(f(X) - f(X'))]$, $f_n^* = \text{argmin}_f R_n(f)$ and $f^* = \text{argmin}_f R(f)$.

Letting $s_i = f(x^*_{i+1}) - f(x^*_i)$ for given ranking function f, we can represent f as $\{s_i\}_{i=1}^{N-1}$. Likewise, we represent f^*_n and f^* as $\{s^n_i = f^*_n(x^*_{i+1}) - f^*_n(x^*_i)\}_{i=1}^{N-1}$ and $\{s^*_i\}_{i=1}^{N-1}$, respectively. Taking $R(f)$ as a function of $\{s_i\}_{i=1}^{N-1}$, we have $R(\{s^n_i\}_{i=1}^{N-1}) \to R(\{s^*_i\}_{i=1}^{N-1})$ as $n \to \infty$. From the assumption that $\{s^*_i\}_{i=1}^{N-1}$ is unique, we can show that $s^n_i \to s^*_i$ for $i = 1, \ldots, N-1$.

The following proof is based on similar arguments in Pollard (1991) and Hjort and Pollard (1993). Given $\delta > 0$, define $h(\delta) \equiv \inf_{\|s-s^*\|=\delta} R(s) - R(s^*)$, where $s = \{s_i\}_{i=1}^{N-1}$ and $s^* = \{s^*_i\}_{i=1}^{N-1}$ for short. The uniqueness of s^* implies that $h(\delta) > 0$ for every $\delta > 0$. Since $R_n(s) \to R(s)$ uniformly, there is N such that for all $n \geq N$, $\sup_{\|s-s^*\|\leq\delta} |R_n(s) - R(s)| < h(\delta)/2$.

Let s be any point outside the ball around s^* with radius δ, say $s = s^* + l \cdot u$ for a unit vector u with $l > \delta$. Convexity of R_n implies $(\delta/l)R_n(s) + (1-\delta/l)R_n(s^*) \geq R_n(s^* + \delta u)$. Then, from the inequality, we have

$$\begin{aligned} & (\delta/l)(R_n(s) - R_n(s^*)) \geq R_n(s^* + \delta u) - R_n(s^*) \\ = \;& [R(s^* + \delta u) - R(s^*)] + [R_n(s^* + \delta u) - R(s^* + \delta u)] - [R_n(s^*) - R(s^*)] \\ \geq \;& h(\delta) - 2 \sup_{\|s-s^*\|\leq\delta} |R_n(s) - R(s)|. \end{aligned}$$

Thus, for all $n \geq N$, if $\|s - s^*\| > \delta$, then $R_n(s) - R_n(s^*) > 0$, and hence the minimizer of R_n, s^*_n should be inside the ball, $\|s^*_n - s^*\| \leq \delta$. This means that $s^*_n \to s^*$ as $n \to \infty$.

Since l_n satisfies the conditions in Theorem 3, the discrete version of the theorem implies that $s^n_i > 0$ for each n and i. Then, as a limit of the sequence $\{s^n_i\}_{n=1}^{\infty}$, $s^*_i \geq 0$. □

The following theorem shows more specific results of optimal ranking under the hinge loss. They reveal the undesirable property of potential ties in ranking when the hinge loss is used, extending the phenomenon observed in the toy example to general case. Detailed proof is given in Appendix.

Theorem 5. *Let $f^* = \arg\min_f E(1 - (f(X) - f(X'))_+$. Suppose that f^* is unique up to an additive constant.*

(i) For discrete $\mathcal{X} = \{x^*_i\}_{i=1}^{N}$ (with $N = \infty$ allowed), if elements are ordered by the likelihood ratio g_+/g_- such that $\frac{g_+(x^*_1)}{g_-(x^*_1)} < \cdots < \frac{g_+(x^*_N)}{g_-(x^*_N)}$, then the increments of f^* can not be any value other than 0 or 1, that is, $\{f^*(x^*_{i+1}) - f^*(x^*_i)\}_{i=1}^{N-1} = \{0, 1\}$. Thus, a version of f^* is integer-valued.

(ii) For continuous $\mathcal{X}$, there exists an integer-valued ranking function whose risk is arbitrarily close to the minimum risk.

3.3.3 Integer-valued ranking functions

As an implication of Theorem 5, it is sufficient to consider only integer-valued functions in order to find a risk minimizer f^* under the hinge loss.

Let K be the number of distinct values that f takes (possibly ∞) and re-define $A_i(f)$ as $\{x|\ f(x) = i\}$ slightly different from that in the proof of Theorem 5 (ii). For $\mathcal{A}(f) = \{A_i\}_{i=1}^{K}$, using the same definition of $\hat{g}_+$ and $\hat{g}_-$ as in the proof, we have $\frac{\hat{g}_+(A_1)}{\hat{g}_-(A_1)} < \cdots < \frac{\hat{g}_+(A_K)}{\hat{g}_-(A_K)}$. Emphasizing the connection between the partition $\mathcal{A}(f)$ and f, let $f_{\mathcal{A}}(x) = \sum_{i=1}^{K} i \cdot I(x \in A_i)$ represent f. Then the

ranking risk of $f_{\mathcal{A}}$ is explicitly given by

$$\begin{aligned} & E[(1-(f_{\mathcal{A}}(X)-f_{\mathcal{A}}(X')))_+] \\ = & \sum_{i=1}^{K}\sum_{j=i}^{K}(1-i+j)\hat{g}_+(A_i)\hat{g}_-(A_j) = \sum_{i=1}^{K}\hat{g}_+(A_i)\sum_{l=i}^{K}\hat{g}_-(\cup_{j=l}^{K}A_j) \\ = & \sum_{l=1}^{K}\sum_{i=1}^{l}\hat{g}_+(A_i)\hat{g}_-(\cup_{j=l}^{K}A_j) = \sum_{l=1}^{K}\hat{g}_+(\cup_{i=1}^{l}A_i)\hat{g}_-(\cup_{j=l}^{K}A_j). \end{aligned} \tag{5}$$

To examine the effect of the number of distinct values K or the number of steps $(K-1)$ on the minimal ranking risk, define $\mathcal{F}_K$ as the set of all integer-valued functions with $(K-1)$ steps only. Let $R_K = \inf_{f\in\mathcal{F}_K} R_l(f)$ be the minimal risk achieved by ranking functions within $\mathcal{F}_K$. The following results show that if the likelihood ratio g_+/g_- is unbounded, the ranking risk R_K is non-increasing in K and strictly decreasing as long as g_- has a positive probability for diminishing tails of the likelihood ratio where it diverges to ∞. See Section 5 for a concrete example illustrating Theorem 6 and Corollary 1.

Theorem 6. *If* $\inf_{x\in\mathcal{X}}\left(g_-(x)/g_+(x)\right)=0$, *then* $R_K \geq R_{K+1}$ *for* $K=1,2,\ldots$.

Proof. Suppose that $f_{\mathcal{A}}(x)=\sum_{i=1}^{K} i\cdot I(x\in A_i)$ is a risk minimizer in $\mathcal{F}_K$. Given $f_{\mathcal{A}}$, we construct a simple function with K values by splitting A_K into two sets B_K and B_{K+1} and setting B_i equal to A_i for $i=1,\ldots,K-1$. For the new partition $\mathcal{B}=\{B_i\}_{i=1}^{K+1}$, define a simple function

$$f_{\mathcal{B}}(x)=\sum_{i=1}^{K+1} i\cdot I(x\in B_i).$$

Using the relation that $\hat{g}_-(A_K)=\hat{g}_-(B_K)+\hat{g}_-(B_{K+1})$ and $\sum_{i=1}^{K}\hat{g}_+(A_i)=\sum_{i=1}^{K+1}\hat{g}_+(B_i)=1$, and the identity $R_l(f_{\mathcal{B}})=\sum_{l=1}^{K+1}\hat{g}_+(\cup_{i=1}^{l}B_i)\hat{g}_-(\cup_{j=l}^{K+1}B_j)$ from (5), we can verify that

$$\begin{aligned} R_l(f_{\mathcal{B}}) &= \sum_{l=1}^{K+1}\left(\sum_{i=1}^{l}\hat{g}_+(B_i)\right)\left(\sum_{j=l}^{K+1}\hat{g}_-(B_j)\right) \\ &= \sum_{l=1}^{K-1}\left(\sum_{i=1}^{l}\hat{g}_+(A_i)\right)\left(\sum_{j=l}^{K}\hat{g}_-(A_j)\right)+(1-\hat{g}_+(B_{K+1}))\hat{g}_-(A_K)+\hat{g}_-(B_{K+1}) \\ &= \sum_{l=1}^{K}\left(\sum_{i=1}^{l}\hat{g}_+(A_i)\right)\left(\sum_{j=l}^{K}\hat{g}_-(A_j)\right)-\hat{g}_+(B_{K+1})\hat{g}_-(A_K)+\hat{g}_-(B_{K+1}). \end{aligned}$$

Therefore $R_l(f_{\mathcal{B}})-R_l(f_{\mathcal{A}})=-\hat{g}_+(B_{K+1})\hat{g}_-(A_K)+\hat{g}_-(B_{K+1})$.

Given $\hat{g}_-(A_K)$, define $B_{K+1}=\{x|\ g_-(x)/g_+(x)<\hat{g}_-(A_K)\}$. If $\hat{g}_-(A_K)>0$, B_{K+1} is not empty since $\inf_{x\in\mathcal{X}} g_-(x)/g_+(x)=0$. Furthermore, we can show that $\hat{g}_-(B_{K+1})<\hat{g}_-(A_K)\hat{g}_+(B_{K+1})$, which implies that $R_l(f_{\mathcal{B}})<R_l(f_{\mathcal{A}})$. If $\hat{g}_-(A_K)=0$, then $\hat{g}_-(B_{K+1})=0$, and consequently $R_l(f_{\mathcal{B}})=R_l(f_{\mathcal{A}})$. Hence the risk of $f_{\mathcal{B}}$ is at most that of $f_{\mathcal{A}}$, and this proves the desired result. □

Corollary 1. *Let* $C_\epsilon=\{x|\ g_-(x)/g_+(x)<\epsilon\}$. *If* $\hat{g}_-(C_\epsilon)>0$ *for all* $\epsilon>0$, *then* $R_K>R_{K+1}$ *for each* K *under the assumption of Theorem 6. Hence the optimal* K *is infinity.*

Proof. Given a risk minimizer $f_{\mathcal{A}}$ with $(K-1)$ steps, there exists ϵ_K such that $A_K = \{x|\, g_-(x)/g_+(x) < \epsilon_K\} = C_{\epsilon_K}$. By the assumption, $\hat{g}_-(A_K) = \hat{g}_-(C_{\epsilon_K}) > 0$. As shown in Theorem 6, when $\hat{g}_-(A_K) > 0$, there exists a simple function $f_{\mathcal{B}}$ with K values such that $R_l(f_{\mathcal{B}}) < R_l(f_{\mathcal{A}})$. Hence $R_K > R_{K+1}$ for each K. □

Remark 3. *By reversing the role of g_+ and g_- and redefining A_i, we can establish similar results as Theorem 6 and Corollary 1 when* $\inf_{x\in\mathcal{X}} \left(g_+(x)/g_-(x)\right) = 0$.

4 Comments on related results

As a related work, Clémençon et al. (2008) provide a rigorous statistical framework for studying the ranking problem and also discuss convex risk minimization methods for ranking. We explain the connection between the two approaches and point out the differences.

Their formulation considers a ranking rule $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$ directly, instead of a ranking (or scoring) function $f : \mathcal{X} \to \mathbb{R}$ as in our formulation. If $r(x, x') = 1$, then x is ranked higher than x'. A ranking rule r represents a partial order or preference between two instances while a ranking function f represents a total order over the instance space. A real-valued function $h(x, x')$ on $\mathcal{X} \times \mathcal{X}$ can induce a ranking rule via $r(x, x') \equiv \mathrm{sgn}(h(x, x'))$.

Covering more general ranking problems with a numerical response Y, for each independent and identically distributed pair of (X, Y) and (X', Y') from a distribution on $\mathcal{X} \times \mathbb{R}$, they define a variable $Z = (Y - Y')/2$, and consider X being better than X' if $Z > 0$. Then by directly relating (X, X') with $\mathrm{sgn}(Z)$, they transform the ranking problem to a pairwise binary classification problem and examine the implications of the formulation to ranking.

Note that in the transformed classification framework, the bipartite ranking loss corresponds to the 0-1 loss. As a result, when $Y = 1$ or -1, the best ranking rule $r^*(x, x')$ is given by the Bayes decision rule for the classification problem:

$$\phi^*(x, x') = \mathrm{sgn}\left(\log \frac{P(Z = 1|X = x, X' = x')}{P(Z = -1|X = x, X' = x')}\right).$$

Although it was not explicitly stated in the paper, we can infer from $\phi^*(x, x')$ that the best ranking rule in bipartite ranking can be expressed as

$$r^*(x, x') = \mathrm{sgn}\left(g_+(x)/g_-(x) - g_+(x')/g_-(x')\right)$$

since

$$\begin{aligned}\frac{P(Z = 1|X, X')}{P(Z = -1|X, X')} &= \frac{P(X, X'|Z = 1)}{P(X, X'|Z = -1)} \times \frac{P(Z = 1)}{P(Z = -1)} \\ &= \frac{g_+(X)g_-(X')}{g_-(X)g_+(X')} \times \frac{P(Z = 1)}{P(Z = -1)} = \frac{g_+(X)g_-(X')}{g_-(X)g_+(X')}.\end{aligned}$$

Hence, with this different formulation, we can arrive at the same conclusion of Theorem 1 that the theoretically optimal rankings over the instance space $\mathcal{X}$ are given by the likelihood ratio (g_+/g_-).

As an extension, for the ranking rules minimizing convex risk functionals, Clémençon et al. (2008) invoke the results of Bartlett et al. (2006) on the consistency of classification with convex loss functions. Again, directly aiming at the optimal ranking rule rather than the scoring function, they discuss theoretical implications of minimization of the risk $E[l(\mathrm{sgn}(Z)h(X, X'))]$ for a convex loss l.

Considering only a positive instance X and a negative instance X', we can describe the difference between our approach and theirs being whether one finds the optimal ranking rule induced by a real-valued function h, $\mathrm{argmin}_h E[l(h(X, X'))]$ or the optimal ranking function f, $\mathrm{argmin}_f E[l(f(X) - f(X'))]$. In practice, ranking algorithms such as the RankBoost algorithm produce a ranking function, not a ranking rule, which makes our approach more natural and pertinent. More importantly, a ranking rule does not define a ranking function consistently in general, and Clémençon et al. (2008) have overlooked the fact when applying the classification results to ranking.

On the other hand, in some special cases, if there exists f such that $h^*(x, x') = f(x) - f(x')$, h^* can be used to specify the optimal f. Theorem 2 regards those special cases. For example, in the case of the exponential loss, its population minimizer in the classification problem is known as $(1/2)$ times the logit. Therefore, the best ranking rule $r^*(x, x')$ is induced by

$$h^*(x, x') = \frac{1}{2} \log \frac{P(Z = 1|X = x, X' = x')}{P(Z = -1|X = x, X' = x')} = \left(\frac{1}{2} \log \frac{g_+(x)}{g_-(x)} + \beta \right) - \left(\frac{1}{2} \log \frac{g_+(x')}{g_-(x')} + \beta \right),$$

where β is an arbitrary constant. h^* then identifies $f^* = (1/2) \log(g_+/g_-)$ as the optimal ranking function, the same conclusion as Theorem 2.

The following theorem states that the conditions for ranking loss in Theorem 2 are indeed necessary for the existence of a ranking (or scoring) function consistent with $h^*(x, x')$, and therefore, the equivalence between the two formulations.

Theorem 7. *Suppose that l is convex, differentiable, $l'(s) < 0$ for all $s \in \mathbb{R}$, and $l'(-s)/l'(s)$ is strictly increasing in s. Let h^* be the optimal function on $\mathcal{X} \times \mathcal{X}$ minimizing $E[l(h(X, X'))]$. Then h^* is of the form, $h^*(x, x') = f(x) - f(x')$ for some function f on $\mathcal{X}$ if and only if $l'(-s)/l'(s) = \exp(s/\alpha)$ for some positive constant α.*

Proof. For x and x', let $s \equiv h^*(x, x')$. From (1) in the proof of Theorem 2, we have

$$\frac{l'(-s)}{l'(s)} = \frac{g_+(x)}{g_-(x)} \frac{g_-(x')}{g_+(x')}.$$

Note that this intermediate result is obtained using the convexity and differentiability of l only. Let $G(s) \equiv l'(-s)/l'(s)$. Solving the above equation for $s = h^*(x, x')$, we get

$$h^*(x, x') = G^{-1}\left(\frac{g_+(x)}{g_-(x)} \frac{g_-(x')}{g_+(x')} \right).$$

Suppose that $h^*(x, x') = f(x) - f(x')$ for some function f on $\mathcal{X}$. Then

$$\begin{aligned} G^{-1}\left(\frac{g_+(x)}{g_-(x)} \frac{g_-(x')}{g_+(x')} \right) &= f(x) - f(x') = [f(x) - f(x_0)] + [f(x_0) - f(x')] \\ &= G^{-1}\left(\frac{g_+(x)}{g_-(x)} \frac{g_-(x_0)}{g_+(x_0)} \right) + G^{-1}\left(\frac{g_+(x_0)}{g_-(x_0)} \frac{g_-(x')}{g_+(x')} \right), \end{aligned}$$

which implies that $G^{-1}(a \cdot b) = G^{-1}(a) + G^{-1}(b)$ for any $a, b > 0$, in general. This condition for G^{-1} further implies that $G^{-1}(t) = \alpha \log(t)$ for some α. Since G is increasing, α has to be positive, and thus $G(s) = \exp(s/\alpha)$, which completes the proof of the necessity. Theorem 2 proves the sufficiency. □

Remark 4. *When the conditions in Theorem 7 hold for loss l, the two formulations for ranking are equivalent in the sense that $h^*(x, x') = f^*(x) - f^*(x')$ by Theorem 2. As mentioned before, the exponential loss for RankBoost and deviance loss for RankNet satisfy the conditions, leading to the functional correspondence between h^* and f^*. Note that the regret bound for bipartite ranking through $h(\cdot, \cdot)$ in Clémençon et al. (2008) (see p.864) is based on the result in Bartlett et al. (2006) for binary classification, and its translation to the regret bound for ranking through $f(\cdot)$ is valid only when $h^*(x, x') = f^*(x) - f^*(x')$.*

Remark 5. *Kotlowski et al. (2011) and Agarwal (2012) also assume the relation between a pairwise ranking rule $sgn(h(x, x'))$ and a scoring function $f(x)$ from binary classification as $sgn(f(x)-f(x'))$ in their regret bounds analysis.*

Remark 6. *The assumption in Theorem 7 about $l'(-s)/l'(s)$ is related to the necessary and sufficient condition for proper composite losses in Reid and Williamson (2010) for binary classification.*

In contrast to those loss functions that yield the functional correspondence between h^* and f^*, other loss functions call for careful distinction between the two approaches. For example, the squared hinge loss l is ranking-calibrated (also classification-calibrated), but $l'(-s)/l'(s) \neq \exp(s/\alpha)$ for any positive constant α. Hence $h^*(x, x')$ can not be expressed as $f(x)-f(x')$ for any f including f^*, although $\text{sgn}(h^*(x, x')) = \text{sgn}(f^*(x) - f^*(x'))$.

Another case in point which illustrates the difference between the two approaches clearly is the hinge loss, $l(s) = (1 - s)_+$. Application of the well-known result in the classification literature (for example, Bartlett et al. 2006) about the population minimizer of the hinge loss gives

$$\begin{aligned} h^*(x, x') &= \text{sgn}\left(P(Z = 1|X = x, X' = x') - \frac{1}{2}\right) \\ &= \text{sgn}\left(\log \frac{P(Z = 1|X = x, X' = x')}{P(Z = -1|X = x, X' = x')}\right) = \text{sgn}\left(\log \frac{g_+(x)g_-(x')}{g_-(x)g_+(x')}\right). \end{aligned}$$

It is easy to argue that there exists no ranking function f such that $h^*(x, x') = f(x) - f(x')$. If there existed f such that

$$\text{sgn}\left(\log \frac{g_+(x)g_-(x')}{g_-(x)g_+(x')}\right) = f(x) - f(x'),$$

then for x_0 with $g_+(x_0) = g_-(x_0)$, the ranking function would be given by

$$f(x) = f(x_0) + \text{sgn}\left(\log \frac{g_+(x)}{g_-(x)}\right).$$

However, the functional form leads to the equation

$$\text{sgn}\left(\log \frac{g_+(x)}{g_-(x)} - \log \frac{g_+(x')}{g_-(x')}\right) = \text{sgn}\left(\log \frac{g_+(x)}{g_-(x)}\right) - \text{sgn}\left(\log \frac{g_+(x')}{g_-(x')}\right),$$

which is not generally true.

In contrast, Theorem 4 implies at least that the optimal ranking function under the hinge loss preserves the order of the likelihood ratio, but not strictly with some possible ties. Although the explicit form of f^* may not be specified, Theorem 5 further describes that f^* could exhibit the characteristic of a step function. As the toy example illustrates, such ties could lead to ranking inconsistency.

5 Numerical illustration

5.1 Simulation study

To illustrate the theoretical results pertaining to the large sample characteristics of estimated ranking functions under different loss criteria, we carried out a numerical experiment under a simple setting.

With binary Y (1 or -1), the distribution of X for the positive category is set to $N(1,1)$, and that for the negative category is set to $N(-1,1)$. From $g_+(x) = \frac{1}{\sqrt{2\pi}}\exp(-(x-1)^2/2)$ and $g_-(x) = \frac{1}{\sqrt{2\pi}}\exp(-(x+1)^2/2)$ in this case, we have $\log(g_+(x)/g_-(x)) = 2x$. Thus the theoretically best ranking function with minimum ranking error should be an order-preserving transformation of x. A training data set of (X,Y) pairs was generated from the distributions with 2000 instances in each category ($n_+ = n_- = 2000$).

First, the RankBoost algorithm in Freund et al. (2003) was applied to the training sample by considering 2000×2000 positive and negative pairs. It aims to attain the minimum ranking error by minimizing the empirical risk under the exponential loss. In the boosting algorithm for ranking, a weak learner is a ranking function whose performance in terms of the AUC is slightly better than random assignment. In our experiment, a stump $f_\theta(x) \equiv I(x > \theta)$ or $I(x \le \theta)$ with a threshold $\theta \in \mathbb{R}$ was used as a weak learner, and the threshold θ was taken from the observed values, $\{x_i\}_{i=1}^{n_+} \cup \{x'_j\}_{j=1}^{n_-}$. At each iteration, a weak ranking function was chosen and added to the current ranking function with weight determined to minimize the ranking risk over the positive and negative pairs from the training data. We iterated the boosting process for 400 times to combine weak rankings and obtained the final ranking function $\hat{f}$. It is depicted in the left panel of Figure 2. The $\hat{f}$ in the figure is centered to zero in the y axis. The dotted line is the theoretically optimal ranking function, $f^*(x) = (1/2)\log(g_+(x)/g_-(x)) = x$, under the exponential loss as indicated by Theorem 2. The centered ranking function from boosting appears to approximate f^* closely, especially over $[-2,2]$, where the density is relatively high as marked by the rug plot of a subset of the observed values sampled at the rate of 1/20. The flat part of the function on either end is an artifact due to the form of the weak learners used in boosting. Increasing the number of iterations further did not change the visual appearance of the ranking function. In fact, after fewer than 20 iterations, the AUC values of boosted rankings over the training data became stable as shown in the right panel, and the changes afterwards were only incremental.

Second, the AUC maximizing support vector machine (SVM) in Brefeld and Scheffer (2005) was applied to the training data. In general, the AUCSVM (also known as RankSVM) finds a ranking function $f \in \mathcal{H}_K$ minimizing

$$C\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}(1-(f(x_i)-f(x'_j)))_+ + \|f\|^2_{\mathcal{H}_K},$$

where C is a tuning parameter and $\mathcal{H}_K$ is a reproducing kernel Hilbert space with a kernel K. The solution $\hat{f}(x)$ takes the form of $\sum_{i,j} c_{ij}(K(x_i,x) - K(x'_j,x))$. As the data involve four million pairs, which would make exact computation almost prohibitive, a clustering approach was proposed in the paper for approximate computation. Since X is univariate in this example, we could streamline the clustering step by taking sample quantiles, instead of relying on general k-means clustering as suggested in the paper. We first selected a certain number of quantiles of pairwise differences $(x_i - x'_j)$ for $i = 1,\ldots,n_+$ and $j = 1,\ldots,n_-$, and used only the corresponding pairs for an approximate solution. To allow a rich space with sufficiently local basis functions for approximation of the optimal ranking functions, the Gaussian kernel $K(x,x') = \exp(-(x-x')^2/2\sigma^2)$ with parameter

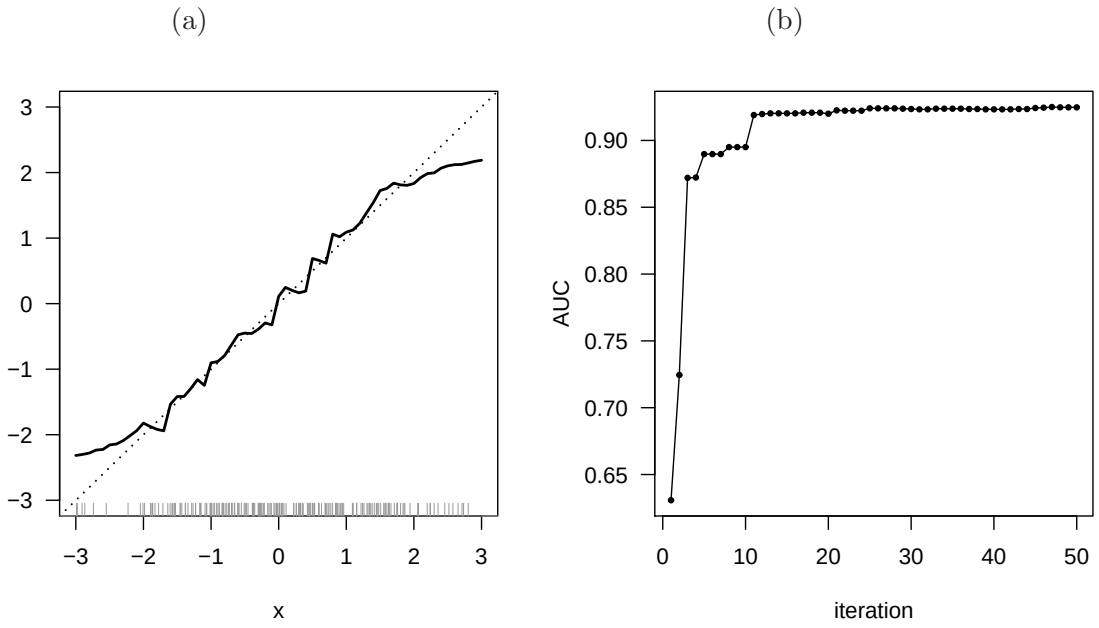

Figure 2: (a) Ranking function given by the RankBoost algorithm. The solid line is the estimated ranking function $\hat{f}$ centered to 0, and the dotted line is the theoretically optimal ranking function $f^*(x) = x$. (b) AUC values of boosted rankings over the training data as the iteration proceeds.

$\sigma^2 = 0.15$ was used. To illuminate the implications of Theorem 5, we also considered a range of other sample sizes $n = n_+ = n_-$ and tuning parameter C.

Figure 3 shows approximate ranking functions $\hat{f}$ (solid lines) obtained by the AUC maximizing SVM for some combinations of n and C. For approximation, we selected 400 pairs based on quantiles of the pairwise differences when $n_+ = n_- = 30$ and 1500 pairs when $n_+ = n_- = 2000$ or 3000. As expected from Theorem 5, the estimated ranking functions appear to approximate step functions increasing in x roughly over the region of high density $[-2, 2]$ as indicated by the visible bumps. The reverse trend on either end is again an artifact due to the form of the Gaussian kernel used as a basis function and the fact that there are relatively few observations near the end. On the whole, the ranking functions attempt to provide the same ordering as the likelihood ratio as indicated by Theorem 4, however, with potentially many ties. In general, ties are considered undesirable in ranking. The dotted lines in Figure 3 are the optimal step functions that are theoretically identified when the numbers of steps are 1, 2, 4, and 5, respectively. Explanation of how to characterize the optimal step functions in general is given in the next subsection. We empirically chose the step function that matches each of the estimated ranking functions most closely in terms of the number of steps. For better alignment, we shifted the step function in each panel vertically so that the values of the pair of functions at $x = 0$ are identical.

5.1.1 The optimal step functions for ranking under hinge loss

Although there is no explicit expression of the optimal ranking function under the hinge loss in this case, Theorem 5 (ii) suggests that there exists an integer-valued function whose risk is arbitrarily close to the minimum. In an attempt to find the best ranking function among integer-valued

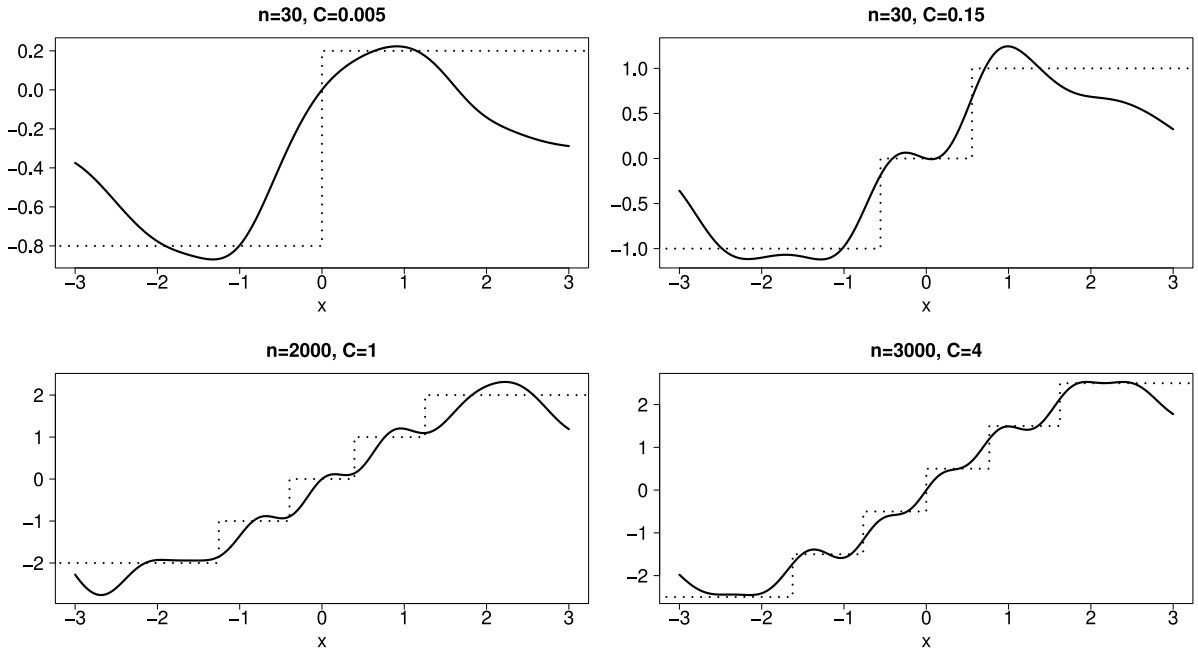

Figure 3: In each panel, the solid line is the estimated ranking function by the AUC maximizing SVM (RankSVM), and the dotted line is a step function obtained theoretically.

functions given the number of steps K, we consider a step function of the form

$$f_{\mathcal{A}}(x) = \sum_{i=1}^{K+1} i \cdot I(a_{i-1} < x \leq a_i)$$

with jump discontinuities $\{a_i\}_{i=1}^{K}$, $a_0 = -\infty$ and $a_{K+1} = \infty$. Note that $f_{\mathcal{A}}$ is non-decreasing in x as the likelihood ratio is.

Using (5) with $A_i = (a_{i-1}, a_i]$, we can explicitly calculate the risk of $f_{\mathcal{A}}$ as

$$E[(1 - (f_{\mathcal{A}}(X) - f_{\mathcal{A}}(X')))_+] = \sum_{i=1}^{K+1} G_+(a_i)(1 - G_-(a_{i-1})) = \sum_{i=1}^{K+1} \Phi(a_i - 1)(1 - \Phi(a_{i-1} + 1)),$$

where G_+ and G_- are the cdfs of X and X', and Φ is the cdf of the standard normal distribution. Given K, the necessary condition for risk minimization is then

$$\frac{\partial}{\partial a_i} \sum_{l=1}^{K+1} G_+(a_l)(1 - G_-(a_{l-1})) = g_+(a_i)(1 - G_-(a_{i-1})) - g_-(a_i)G_+(a_{i+1}) = 0 \tag{6}$$

for $i = 1, \ldots, K$. In the normal setting, (6) is simplified to

$$\exp(-a_i)\Phi(a_{i+1} - 1) = \exp(a_i)(1 - \Phi(a_{i-1} + 1)) \quad \text{for } i = 1, \ldots, K. \tag{7}$$

By solving for a_i analytically, we can identify the jump discontinuities of the step function with minimal risk given the number of steps. Table 1 displays the solutions to the equations for small K that are obtained numerically. For example, when $K = 1$, the optimal ranking function has a jump

Table 1: Jump discontinuities $\{a_i\}_{i=1}^K$ of the step functions with minimal risk in the normal setting given the number of steps

	Number of steps K										
	1	2	3	4	5	6	7	8	9	10	11
											7.0139
										6.1511	
									4.5196		4.5202
								3.4437		3.4437	
							2.6756		2.6759		2.6759
a_K						2.0887		2.0842		2.0842	
					1.6262		1.5943		1.5943		1.5943
a_{K-1}				1.2532		1.1613		1.1605		1.1605	
			0.9205		0.7662		0.7589		0.7589		0.7589
		0.5564		0.3947		0.3755		0.3755		0.3754	
$\vdots$	0		0		0		0		0		0
		-0.5564		-0.3947		-0.3755		-0.3755		-0.3754	
			-0.9205		-0.7662		-0.7589		-0.7589		-0.7589
a_2				-1.2532		-1.1613		-1.1605		-1.1605	
					-1.6262		-1.5943		-1.5943		-1.5943
a_1						-2.0887		-2.0842		-2.0842	
							-2.6756		-2.6759		-2.6759
								-3.4437		-3.4437	
									-4.5196		-4.5202
										-6.1511	
											-7.0139

at $a_1 = 0$, which coincides with the decision boundary of the Bayes rule if the problem was treated as binary classification. The sequences displayed in the table certainly reveal some symmetry in the solutions to (7).

To verify the observed symmetry analytically, first consider the case when K is odd, say, $2m+1$ for $m = 0, 1, \ldots$. Set $a_{m+1} = 0$. Then the equation (7) for $i = m+1$ becomes $\Phi(a_{m+2} - 1) = 1 - \Phi(a_m + 1)$, which implies $a_{m+2} = -a_m$. By using this fact and (7) for $i = m$ and $i = m+2$, we can derive

$$\Phi(a_{m+3} - 1) = \exp(-2a_m)\Phi(-1) = 1 - \Phi(a_{m-1} + 1),$$

which yields $a_{m+3} = -a_{m-1}$. In general, suppose that $a_{(m+1)+i} = -a_{(m+1)-i}$ for $i = 1, \ldots, k$. Then $\Phi(a_{(m+1)+(k+1)} - 1) = \exp(2a_{(m+1)+k})(1 - \Phi(a_{m+k} + 1))$
$= \exp(-2a_{(m+1)-k})\Phi(a_{(m+1)-(k-1)} - 1) = 1 - \Phi(a_{(m+1)-(k+1)} + 1)$. Thus, the symmetry $a_{(m+1)+i} = -a_{(m+1)-i}$ holds for $i = k+1$. By mathematical induction, we conclude that for $K = 2m+1$, a sequence $\{a_i\}_{i=1}^K$ symmetric around zero with $a_{m+1} = 0$ solves (7). As a special case, when $K = 3$, $-a_1 = a_3$ and $a_2 = 0$. Furthermore, from the optimality equation (7), we can determine $a_1 = \frac{1}{2}\log\Phi(-1) \approx -0.9205$.

Similarly when K is even, say, $2m$, we can show that a sequence $\{a_i\}_{i=1}^K$ symmetric around zero with $a_i = -a_{K+1-i}$ for $i = 1, \ldots, m$ solves (7). For example, when $K = 2$, $a_1 = -a_2$, and a_1 is the solution to the equation $\exp(2a_1) = \Phi(-a_1 - 1)$, which yields $a_1 \approx -0.5564$. In both cases, the symmetry halves the number of unknowns in the solution sequences.

Since $\log(g_+(x)/g_-(x)) = 2x$, by taking $C_\epsilon = \{x | x > -(\log\epsilon)/2\}$, we can check that $\hat{g}_-(C_\epsilon) = \Phi((\log\epsilon)/2 - 1) > 0$ for all $\epsilon > 0$. Hence, the optimal K on the population level is infinity by Corollary 1. To study the limiting sequence of jump discontinuities as $K \to \infty$, let us examine the relation between a_{i-1} and a_i when $\{a_i\}_{i=1}^K$ is generated sequentially by using (7). Given a_{i-1}, consider possible values of a_i. From $\Phi(a_{i+1} - 1) = \exp(2a_i)(1 - \Phi(a_{i-1} + 1))$, we have $\exp(2a_i)(1 - \Phi(a_{i-1} + 1)) \leq 1$ with the equality holding for $i = K$ only. Similarly, from $\exp(-2a_{i-1})\Phi(a_i - 1) = 1 - \Phi(a_{i-2} + 1)$, we have $\exp(-2a_{i-1})\Phi(a_i - 1) \leq 1$ with the equality holding for $i = 2$

only. In addition, by the monotonicity of the sequence, $\exp(-a_i)\Phi(a_i - 1) < \exp(-a_i)\Phi(a_{i+1} - 1) = \exp(a_i)(1 - \Phi(a_{i-1} + 1))$ and $\exp(a_{i-1})(1 - \Phi(a_{i-1} + 1)) < \exp(a_{i-1})(1 - \Phi(a_{i-2} + 1)) = \exp(-a_{i-1})\Phi(a_i - 1)$.

Figure 4 depicts the feasible region of (a_{i-1}, a_i) derived from the four inequalities. Examples of (a_{i-1}, a_i) pairs for some K are also drawn, which are the circles and the square falling into the shaded area. For instance, when $K = 2$, $a_2 = -a_1 \approx 0.5564$, and (a_1, a_2) lies on the line $y = -x$. When $K = 3$, $(a_1, a_2) = (a_1, 0)$ and $(a_2, a_3) = (0, -a_1)$ with $a_1 \approx -0.9205$, which are the points on the x-axis and the y-axis, respectively in the figure.

By the symmetry in $\{a_i\}_{i=1}^K$, we know that for even $K = 2m$, $a_m = -a_{m+1}$ and thus $(a_m, a_{m+1}) = (-a_{m+1}, a_{m+1})$ lies on the line $y = -x$, and for odd $K = 2m + 1$, $a_{m+1} = 0$ and $(a_{m+1}, a_{m+2}) = (0, a_{m+2})$ lies on the y-axis. Since the sequence of jump discontinuities is characterized by the recursive relation in (7), where a_{i-1} and a_i determine a_{i+1}, finding a_{m+1} for even K (or a_{m+2} for odd K) as $K \to \infty$ is sufficient to identify the limiting sequence. The discussion of the feasible region for (a_{i-1}, a_i) implies that $0 \le a_{m+1} \le 0.5564$ (the line segment of $y = -x$ in the region) for even K, and $0 \le a_{m+2} \le 0.9205$ (the segment of the y-axis in the region) for odd K. In each case, when the progression from a_{i-1} to a_i is made iteratively for the limiting sequence, each pair should remain in the feasible region. This internal dynamics allows us to devise a bisection algorithm to pin down the value of a_{m+1} (or a_{m+2}), the least positive element in the limiting sequence. Figure 5 illustrates the dynamic process of generating $\{a_i\}$ given several potential values of a_{m+1} for even K in the left panel and a_{m+2} for odd K in the right panel. The bisection algorithm results in $a_{m+1} \approx 0.37537$ for even K and $a_{m+2} \approx 0.75893$ for odd K. Observe that they are close to a_6 and a_7 in Table 1 when $K = 10$ and 11, respectively.

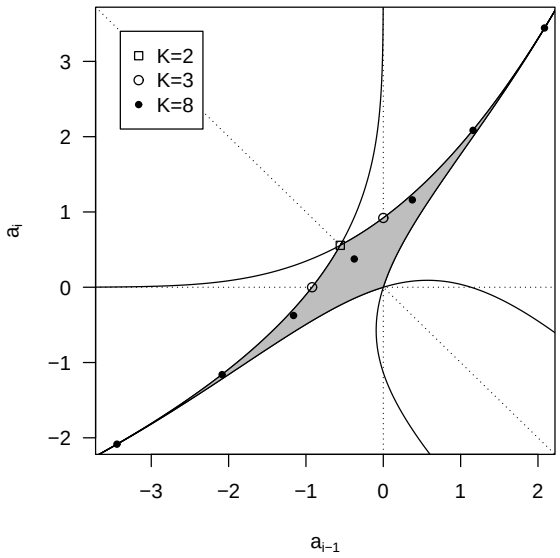

Figure 4: Feasible region for a pair of jump discontinuities (a_{i-1}, a_i) in the normal setting.

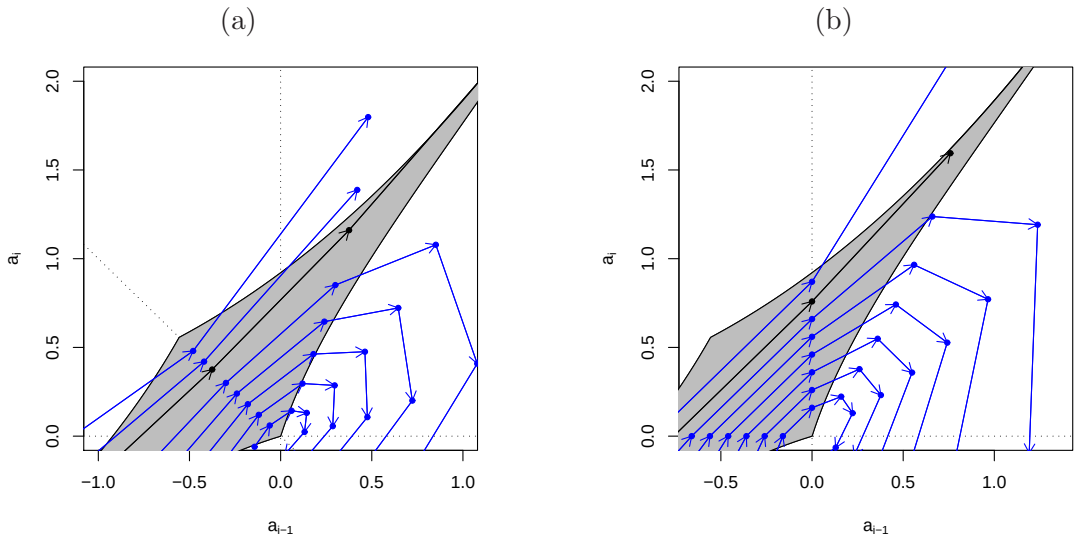

Figure 5: Progression of jump discontinuities $\{a_i\}$ for various values of (a) a_{m+1} (represented by the points on the line $y = -x$) for even $K = 2m$ and (b) a_{m+2} (represented by the points on the y-axis) for odd $K = 2m + 1$.

5.1.2 Ranking risk

For a step function of the form $f_{\mathcal{A}}(x) = \sum_{i=1}^{K+1} i \cdot I(a_{i-1} < x \le a_i)$, its theoretical ranking risk (or $1 - AUC$) is given by $\sum_{i=1}^{K+1} P(a_{i-1} < X \le a_i)\{P(X' > a_i) + \frac{1}{2}P(a_{i-1} < X' \le a_i)\}$. In the normal setting, $1 - AUC$ is then given as

$$\begin{aligned} & \sum_{i=1}^{K+1} \{\Phi(a_i - 1) - \Phi(a_{i-1} - 1)\}[1 - \Phi(a_i + 1) + \frac{1}{2}\{\Phi(a_i + 1) - \Phi(a_{i-1} + 1)\}] \\ = \quad & 1 - \frac{1}{2}\sum_{i=1}^{K+1} \{\Phi(a_i - 1) - \Phi(a_{i-1} - 1)\}\{\Phi(a_i + 1) + \Phi(a_{i-1} + 1)\}, \end{aligned}$$

and it can be calculated explicitly for a step function with specified jump discontinuities.

Corollary 1 says that the theoretical risk of the optimal ranking function with K steps decreases with K. Figure 6 illustrates the theoretical risk $(1 - AUC)$ for some K values and corresponding empirical risk of the RankSVM calculated over simulated data under the bipartite ranking loss (left panel) and the hinge loss (right panel). The values of the theoretical risk are indicated by the solid dots in the figure. They get smaller as K increases. For $K = 6$, the risk is 0.08908. For $K = 10$ and 11, it is approximately 0.08906, which can be taken as almost the limit of the minimal risk of the RankSVM from the consideration of the range of x and x'. Compared with the smallest risk (or the 'Bayes' ranking risk) of $P(X < X') = \Phi(-\sqrt{2}) \approx 0.07865$ in this setting, the RankSVM produces an extra error of 0.01041, and this clearly indicates the inconsistency of risk minimization under the hinge loss.

Along with the theoretical risk, the values of the empirical risk of RankSVM are plotted for various combinations of sample size and the tuning parameter C. The number next to each plotting symbol is the value of C used for the corresponding case. We note that determination of the number

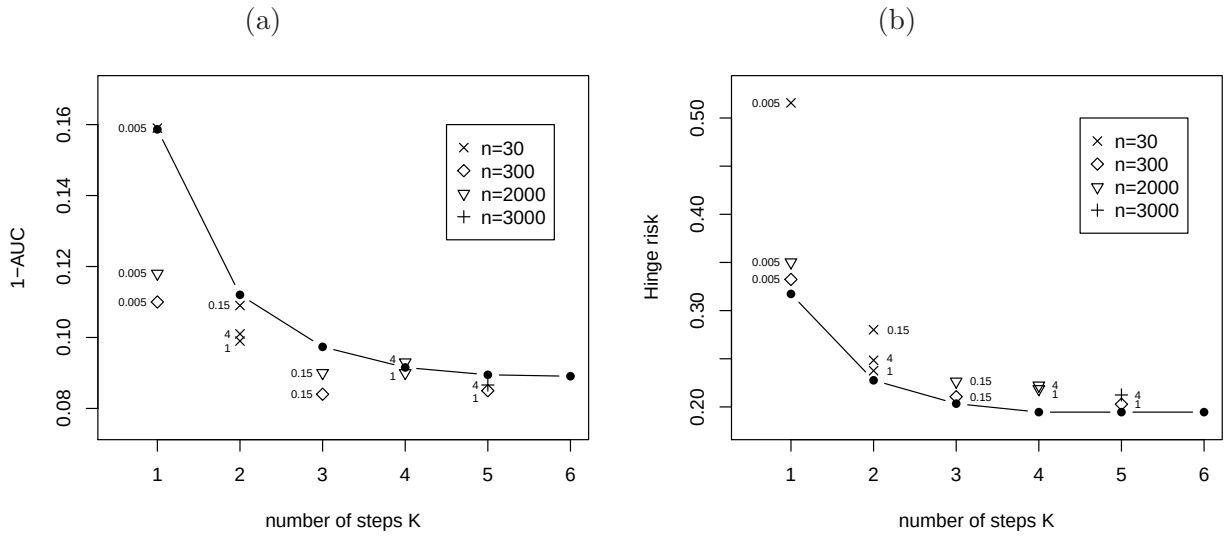

Figure 6: Theoretically minimal risk and empirical risk of RankSVM under (a) bipartite ranking loss and (b) hinge loss. The solid dots are for the theoretical risk while other symbols are for the empirical risk calculated under various combinations of sample size n and tuning parameter C (indicated by the number next to each symbol).

of steps K given an estimated ranking function was not always unambiguous, and subjective calls had to be made for some cases. Nonetheless, Figure 6 suggests that the empirical results are generally consistent with the theory. Sample size n and the tuning parameter C influence the data range and the range of ranking functions, which, in turn, affect the number of steps in finite sample results. In general, as n increases, $\hat{f}$ tends to have more steps, and for a fixed n, larger C (equivalently, smaller penalty parameter) produces more steps. Of course, σ^2 affects the visual characteristics of $\hat{f}$ as well. A systematic investigation would be necessary to understand the granularity of finite-sample solutions to the RankSVM further.

5.1.3 Multivariate extension

Analytical comparisons between different ranking procedures in the univariate normal setting can be generalized to the multivariate case. Suppose that a p-dimensional attribute vector X for positive category follows $N_p(\mu_+, \Sigma)$ and X' for negative category follows $N_p(\mu_-, \Sigma)$ as in the classical linear discriminant analysis setting with a common covariance matrix Σ. The log likelihood ratio, $\log(g_+(\mathbf{x})/g_-(\mathbf{x}))$ is given by $(\mu_+ - \mu_-)'\Sigma^{-1}\mathbf{x}$ up to an additive constant. Hence, the optimal ranking function is an order-preserving transformation of $r(\mathbf{x}) \equiv (\mu_+ - \mu_-)'\Sigma^{-1}\mathbf{x}$, and the Bayes ranking risk is $P(r(\mathbf{X}) < r(\mathbf{X}')) = \Phi\left(-\sqrt{(\Delta_\mu'\Sigma^{-1}\Delta_\mu)/2}\right)$, where $\Delta_\mu \equiv \mu_+ - \mu_-$.

Once the linear transform of the attributes is taken via $r(\mathbf{x})$, this multivariate ranking problem becomes essentially the same as the univariate normal problem where X and X' follow $N(\Delta_\mu'\Sigma^{-1}\mu_+, \Delta_\mu'\Sigma^{-1}\Delta_\mu)$ and $N(\Delta_\mu'\Sigma^{-1}\mu_-, \Delta_\mu'\Sigma^{-1}\Delta_\mu)$, respectively. For example, the univariate setting in the foregoing section is equivalent to the p-variate setting where $X \sim N_p(\mathbf{1}/\sqrt{p}, I)$ and $X' \sim N_p(-\mathbf{1}/\sqrt{p}, I)$.

Extending the discussions in the univariate case, we can identify $r(\mathbf{x})$ itself as $f^*(\mathbf{x})$ under the binomial deviance loss and $r(\mathbf{x})/2$ as that under the exponential loss. Similarly, we infer that $f^*(\mathbf{x})$ under the hinge loss can be an integer-valued function which is non-decreasing in $r(\mathbf{x})$.

For illustration, we simulated $n_+ = 1000$ and $n_- = 1000$ data pairs from the multivariate setting with $p = 10$, $\mu_+ = \mathbf{1}/\sqrt{p}$, $\mu_- = -\mathbf{1}/\sqrt{p}$, and $\Sigma = I$. In this case, $r(\mathbf{x}) = (2/\sqrt{10})\sum_{j=1}^{10} x_j$. We applied RankBoost, LogitBoost (a boosting algorithm for logistic regression; see Friedman et al. 2000), and RankSVM to the data. As in the univariate case, we used stumps as weak learners. For both boosting algorithms, stumps were generated by randomly choosing a variable from the 10 variables and its threshold $\theta \in \mathbb{R}$. The number of iterations was determined by minimizing the ranking error over test data of the same size as the training data.

To examine the estimated ranking functions, we display the main effects of the 10 variables derived from the RankBoost and the LogitBoost outputs in Figure 7. The black lines are for RankBoost and the red lines are for LogitBoost. For direct comparison with LogitBoost, the estimated main effects of RankBoost are doubled in the figure. The dotted lines indicate the theoretical components in the optimal ranking function. Figure 7 shows that the corresponding main effects of the two ranking functions, once scaled, coincide mostly over the high density regions described by the rug plots of the marginal distributions.

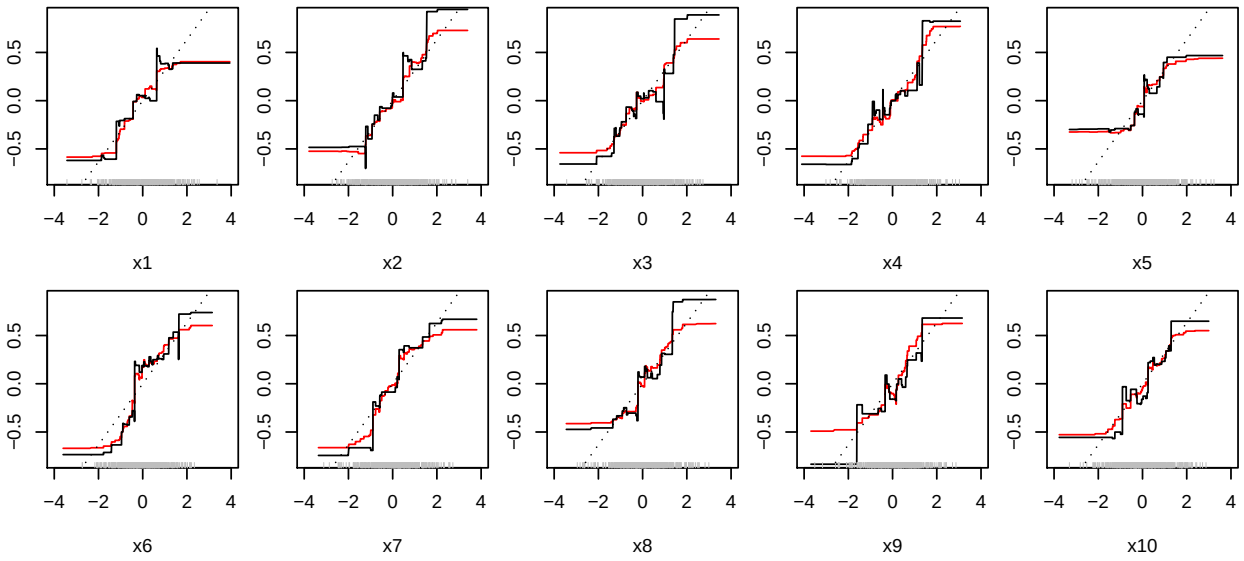

Figure 7: Main effects of the 10 variables (X_1-X_{10}) derived from RankBoost and LogitBoost outputs with stumps. Two times the main effects of RankBoost are in black, and the main effects of LogitBoost are in red.

5.2 Application to Movie-Lens data

To examine the implications of the theoretical findings to real applications, we took one of the Movie-Lens data sets (GroupLens-Research 2006). The data set consists of 100,000 ratings for 1682 movies by 943 users. The ratings are on a scale of 1 to 5. In addition to the ratings, the data set includes information about the movies such as release dates and genres and some demographic information about the users, which can be used as predictors. Among the features of a movie, we used its release date (year, month, and day) and genres (a movie can be in several genres at once) as explanatory variables for ranking. We also included age, gender, and occupation as user-specific factors. For simplicity, in our analysis, ratings are taken as the sampling unit instead of movies or users as in collaborative filtering (Bell et al. 2010, Koren et al. 2009).

We first excluded 10 ratings with incomplete data. A quick examination of the scatterplot of the proportion of movies rated 1 or 2 versus the number of movies rated for each user revealed that six users (id: 13, 181, 405, 445, 655, 774) are outliers in one of the metrics; either extremely critical or having rated unusually many movies. Exercising caution in modeling typical patterns in ratings, we further excluded their ratings from our analysis, which led to a total of 97,139 ratings. To turn this rating prediction into bipartite ranking, we dichotomized the ratings into 'high' (4 or 5) and 'low' (3 or below), which yields 54,806 of high ratings. We standardized the predictors before conducting numerical experiments.

We compared three methods: RankBoost, RankSVM, and LogitBoost. As in the simulation studies, stumps were generated by selecting a variable and a threshold from the observed values at random, and used as weak learners for boosting. To handle large data and at the same time to put the three procedures on an equal footing except for the loss function employed, we devised a boosting (as forward stagewise additive modeling) algorithm for the hinge loss, dubbed 'HingeBoost' in this paper. The loss criteria determine the weights attached to the weak learners in boosting, and thus they drive the main difference between RankBoost and HingeBoost in the numerical comparisons. Since the hinge loss is not strictly convex, depending on weak learners used, the optimal weight may not be determined uniquely. For HingeBoost, weight optimization was done by grid search, and when multiple minima existed, the smallest value was chosen. In addition, we examined the effect of sample size and input space on ranking functions and their accuracy by varying the combination of variables used in ranking and the number of training pairs ($n_+ \times n_-$) from small (200×200) to large (1000×1000).

In each experiment, we first set aside test data (of 10^6 pairs) chosen at random from the MovieLens data for evaluation. The same test data set was used across different training sample sizes and variable combinations. High-low pairs in training data were formed by selecting equal number of cases from each category at random from the remaining data, and additional 10^6 pairs were randomly chosen from the rest for determining the number of iterations in boosting by taking the given loss criterion as the corresponding validation criterion. Ranking accuracy was then evaluated over the test pairs. This process was repeated 50 times. In each replicate, test data as well as validation data were fixed and only training sample sizes and variable combinations were changed. Across replicates, test data and validation data were varied.

Table 2 provides a summary of the results with the mean AUC value and the standard error in parentheses for each setting. As the training set size increases, the ranking accuracy generally increases for all three methods. Among the main variables, the movie release year turns out to be a stronger predictor than genres, the user's occupation and age. In terms of the ranking accuracy, LogitBoost and RankBoost performed better than HingeBoost in general, and the differences become more pronounced as the number of training pairs increases. LogitBoost produced the highest mean AUC value for each setting.

Figure 8 depicts the main effect of the movie release year in the ranking functions when fitted to a training set of million pairs by the three methods with all the variables. Here, the main effect means the additive component of the ranking scores attributed to the corresponding variable, and it is taken to be centered to zero. Each panel contains sample curves of the estimated main effect from ten replicates, which are distinguished by different colors, and the solid black line indicates their mean curve. Overall, old films in the data set tend to be rated high. Probably, they are those films that survive for several decades for good reasons. The estimated effect of the movie release year peaks around 1940-50, which includes such film classics as *Citizen Kane*, *Vertigo*, *Casablanca*, *Rear Window*, and *The Seventh Seal*, just to name a few. The main effects from RankBoost and LogitBoost in the figure are very similar except for the scale factor of 2 as suggested by the theory. In contrast, HingeBoost provides a crude approximation to the ranking scores from the other two

Table 2: Mean AUC values over test set of 10^6 pairs from Movie-Lens data and their standard errors in parentheses when weak learners are stumps. The highest AUC value in each row is boldfaced.

Variable	Training pairs ($n_+ \times n_-$)	LogitBoost	RankBoost	HingeBoost
Age only	200×200	**0.5146** (0.0018)	0.5128 (0.0020)	0.5123 (0.0014)
	500×500	**0.5186** (0.0019)	0.5155 (0.0019)	0.5106 (0.0018)
	1000×1000	**0.5233** (0.0017)	0.5204 (0.0019)	0.5115 (0.0017)
Year only	200×200	**0.5788** (0.0021)	0.5698 (0.0027)	0.5678 (0.0023)
	500×500	**0.5846** (0.0020)	0.5786 (0.0024)	0.5673 (0.0020)
	1000×1000	**0.5881** (0.0015)	0.5840 (0.0019)	0.5665 (0.0017)
Genre and Occupation only	200×200	**0.5332** (0.0038)	0.5158 (0.0034)	0.5220 (0.0034)
	500×500	**0.5647** (0.0022)	0.5482 (0.0039)	0.5474 (0.0025)
	1000×1000	**0.5776** (0.0024)	0.5695 (0.0036)	0.5531 (0.0028)
All	200×200	**0.5659** (0.0033)	0.5324 (0.0046)	0.5444 (0.0040)
	500×500	**0.6018** (0.0020)	0.5895 (0.0040)	0.5772 (0.0029)
	1000×1000	**0.6194** (0.0015)	0.6160 (0.0017)	0.5909 (0.0023)

procedures, removing some fine details captured by the two. The granularity in the main effect of HingeBoost (clearly visible in the individual main effect curves) is due to the singularity of the hinge loss and its particular preference toward integer-valued scores. When it is coupled with such discrete weak learners as the stumps, the extent of granularity becomes stronger.

Similarly, Figure 9 displays the main effect of the user's age from the ranking functions fitted to the same replicates of training sets. Comparison among the three methods is qualitatively the same as in Figure 8. The replicates of the main effect of age from HingeBoost strongly exhibit features of step functions. Small effect sizes exacerbate the extent of such features.

For comparison of the overall ranking scores from the three procedures, scatter plots of the scores from RankBoost and HingeBoost versus those from LogitBoost are shown in Figure 10 for a replicate. Two times the score of RankBoost is very close to the score of LogitBoost in the figure, empirically confirming the theoretical findings in this paper. HingeBoost with the stumps as base learners produces integer-valued ranking scores.

To examine the effect of weak learners on ranking functions, we take the Gaussian kernel functions centered at data points as alternative smooth weak learners. Depending on the variables, they were taken either as a univariate or as a multivariate function. When taken as univariate for each of the standardized predictors, σ^2 of the Gaussian kernel was set to 1. In an attempt to handle the dummy variables for genre and occupation as a group and model potential interactions among the categories, we alternatively took multivariate Gaussian kernels as weak learners. With 19 movie genres and 21 occupations, the σ^2 in this case was set to 20 for proper normalization. At each iteration of boosting, a data point was chosen at random to specify the Gaussian kernel as a weak learner.

Table 3 shows the mean AUC values of the three methods with Gaussian kernels in parallel with Table 2. Comparisons between the two tables indicate that using the Gaussian kernels as weak learners improved the ranking accuracy of the three methods on the whole. In particular, the accuracy of the ranking functions with year only improved with the Gaussian kernels across all the training sample sizes and the methods. On the other hand, the smooth weak learners

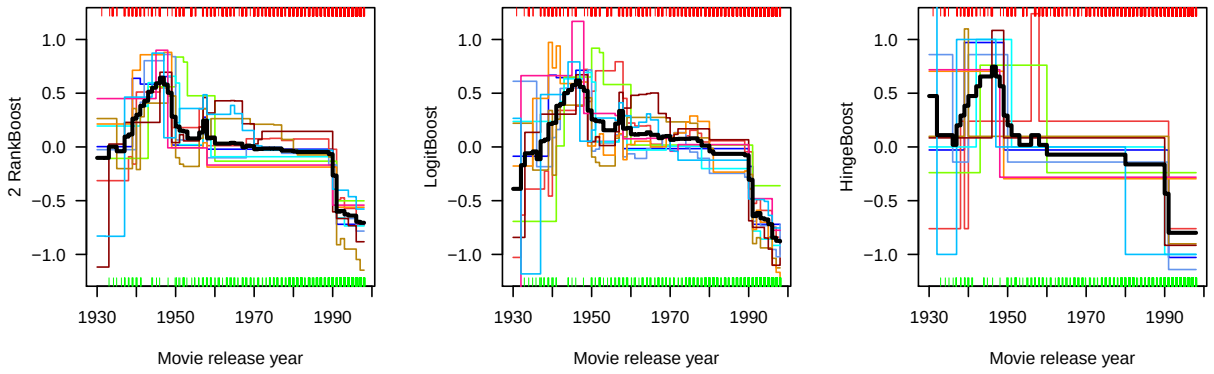

Figure 8: The main effect of the movie release year in the ranking functions fitted to a million training pairs by RankBoost, LogitBoost, and HingeBoost with stumps as base learners. Sample curves from ten replicates are drawn with different colors in each panel, and the solid black line indicates the mean curve. The rug plots in red and green are for a random sample from the movie data with labels 1 and -1, respectively.

improved small to modest-sample performance of those with age only and with all the variables except for HingeBoost. The adverse effect of the granularity of HingeBoost with stumps was mitigated by the smooth weak learners, which, in turn, improved the overall accuracy of HingeBoost significantly. Moreover, using the multivariate Gaussian kernels for the categorical variables (genre and occupation) further improved the accuracy as seen in the comparison of the rows for all vs all* in Table 3 and the corresponding rows for genre and occupation only in Tables 2 and 3.

Inspection of the estimated main effect of movie release year with the Gaussian kernels as base learners reveals that the smoothness of the base learners provides more stable fits across replicates, yet it leads to some loss of interesting details evident in Figure 8 with stumps. In addition, Figure 11 shows scatter plots of the ranking scores from the three procedures with the Gaussian kernels (multivariate for genre and occupation) for the same replicate used in Figure 10. A notable change from the plots with stumps is that the ranking scores from HingeBoost are now continuous and nearly in line with those from LogitBoost. More details of the analysis can be found in Uematsu and Lee (2011).

6 Conclusion and discussion

We have investigated the properties of the best ranking functions under convex loss criteria on the population level in bipartite ranking problems, and have specified general conditions for ranking calibrated loss functions. Our results show that the best ranking functions under convex ranking-calibrated loss criteria produce the same ordering as the likelihood ratio. The best ranking function specified for a certain class of loss functions including the exponential loss provides justification for boosting method in maximizing the AUC.

For the AUC maximizing SVM (or the RankSVM), the result points to the undesirable property of potential ties in ranking, which could lead to inconsistency. Numerical results confirm these theoretical findings. In particular, it was observed that the ranking scores from the RankSVM exhibit granularity. Our result offers much improved understanding of the RankSVM, and at the same time, provides due caution that contrary to the current practice and widespread belief

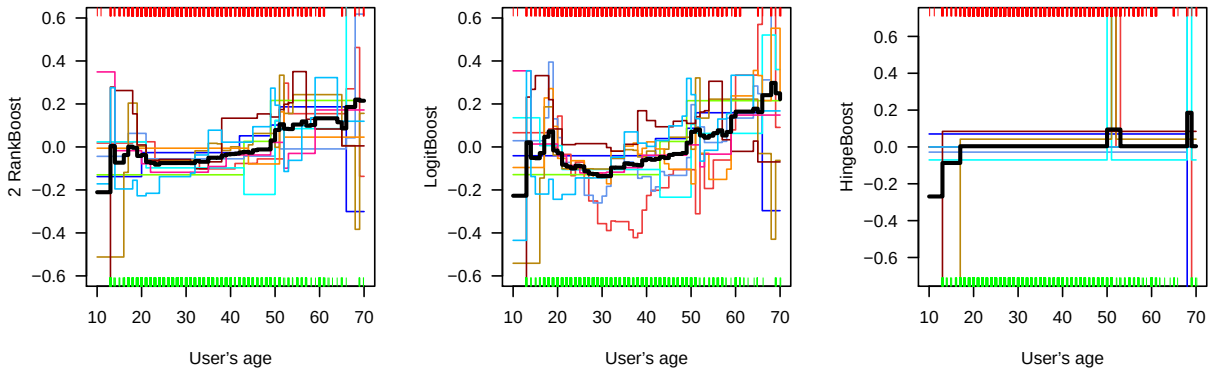

Figure 9: The main effect of the user's age in the fitted ranking functions by RankBoost, LogitBoost, and HingeBoost.

regarding the utility of the hinge loss in machine learning, ranking with the hinge loss is not consistent.

As for practical implications of the theoretical findings about the RankSVM, we need to carefully examine the effect of some of the factors involved in the ranking algorithm on ranking scores. As observed in the numerical examples, weak learners in boosting and kernel parameters such as bandwidth for the Gaussian kernel or degree for polynomial kernels are expected to be critical in determining the extent of granularity in rankings. A systematic study will be necessary to understand the operational relation between the factors in the algorithm and notable features of the resulting ranking function. In practice, such knowledge of the relation can be utilized to minimize potential ties in ranking.

Study of the theoretical relation between a loss criterion and the optimal ranking function is important not only for understanding of consistency, but also for appropriate modification of ranking procedures to achieve different goals in ranking other than minimization of the overall ranking error. For example, many ranking applications in web search and recommender systems focus on those instances ranked near the top only. Clémençon and Vayatis (2009a) investigate the relation between AUC maximization and optimization of linear rank statistics including mean reciprocal rank (MRR). It is worth extending the current results to study the impact of certain modifications proposed for specific aims in ranking, and further develop a principled framework for proper modification.

As another direction for extension, our on-going research shows that the results in bipartite ranking can be generalized to multipartite ranking, providing a new perspective on ordinal regression methods in machine learning and the proportional odds model in statistics.

Appendix

Proof of Theorem 3

Proof. Recall that the risk of a ranking function f under the loss l is defined as

$$R_l(f) = \int_{\mathcal{X}} \int_{\mathcal{X}} l(f(x) - f(x'))g_+(x)g_-(x')dxdx'.$$

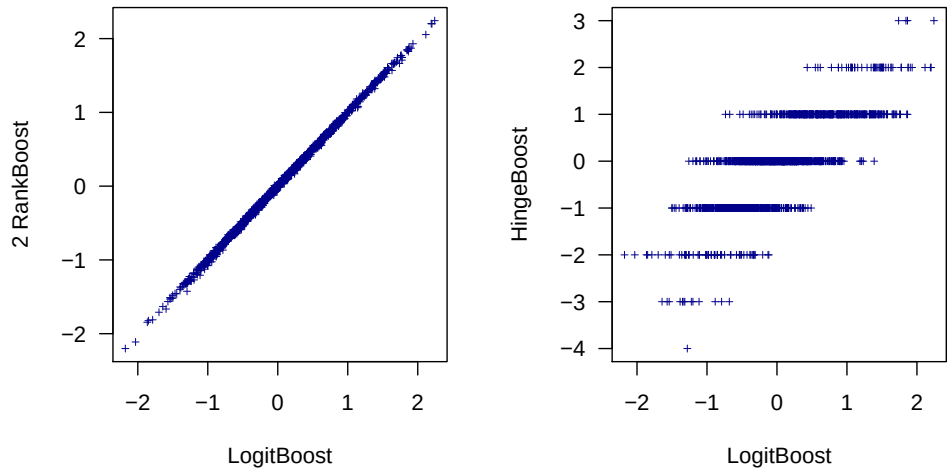

Figure 10: Scatter plots of the scores from the ranking functions with all the variables fitted to a training set of a million pairs by RankBoost and HingeBoost plotted against the ranking scores from LogitBoost when stumps are used as weak learners.

For a ranking function $f \neq f^*$, consider $h = f - f^*$. For a real number δ and a ranking function $(f^* + \delta h)$, define $s(\delta) \equiv R_l(f^* + \delta h)$. Since l is convex, $s(\cdot)$ is a convex function of δ. As f^* minimizes the risk, we have

$$s'(0) = \int_{\mathcal{X}} \int_{\mathcal{X}} (h(x) - h(x'))l'(f^*(x) - f^*(x'))g_+(x)g_-(x')dxdx' = 0.$$

Since f is arbitrary, the equation above holds for any h. This means that

$$\int_{\mathcal{X}} \int_{\mathcal{X}} h(x)(l'(f^*(x) - f^*(x'))g_+(x)g_-(x') - l'(f^*(x') - f^*(x))g_-(x)g_+(x'))dxdx' = 0,$$

and for almost every $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} (l'(f^*(x) - f^*(x'))g_+(x)g_-(x') - l'(f^*(x') - f^*(x))g_-(x)g_+(x'))dx' = 0. \tag{8}$$

First, for z satisfying (8), we verify that $\int_{\mathcal{X}} l'(f^*(z) - f^*(x'))g_-(x')dx' < 0$. Since $l'(s) \leq 0$ for all s, the above integral is either strictly negative or zero. However, having zero for the integral leads to contradiction. The optimality condition for z implies

$$g_+(z) \int_{\mathcal{X}} l'(f^*(z) - f^*(x'))g_-(x')dx' = g_-(z) \int l'(f^*(x') - f^*(z))g_+(x')dx'.$$

If $\int_{\mathcal{X}} l'(f^*(z) - f^*(x'))g_-(x')dx' = 0$, then the right hand side of the above equation can be shown to be nonzero while the left hand side is zero. Let $t_0 \equiv \inf\{t|\ l'(t) = 0\}$. First note that $t_0 > 0$ from the assumption that $l'(0) < 0$. Since l' is non-positive and non-decreasing, $l'(t) < 0$ if $t < t_0$, and $l'(t) = 0$ if $t > t_0$. If $\int_{\mathcal{X}} l'(f^*(z) - f^*(x'))g_-(x')dx' = 0$, then $f^*(z) - f^*(x') \geq t_0$ for almost all x'. This implies $f^*(x') - f^*(z) \leq -t_0$, and hence $l'(f^*(x') - f^*(z)) < 0$ for almost all x' given z, which then makes the integral on the right hand side strictly negative.

Table 3: Mean AUC values over test set of 10^6 pairs from Movie-Lens data and their standard errors in parentheses when weak learners are Gaussian kernels. The highest AUC value in each row is boldfaced.

Variable	Training pairs $(n_+ \times n_-)$	LogitBoost	RankBoost	HingeBoost
Age only	200×200	0.5153 (0.0023)	0.5153 (0.0026)	**0.5169** (0.0023)
	500×500	0.5135 (0.0025)	0.5118 (0.0028)	**0.5144** (0.0024)
	1000×1000	**0.5174** (0.0018)	0.5147 (0.0019)	0.5155 (0.0021)
Year only	200×200	0.5909 (0.0015)	0.5862 (0.0027)	**0.5910** (0.0016)
	500×500	0.5920 (0.0015)	**0.5925** (0.0015)	0.5923 (0.0015)
	1000×1000	**0.5931** (0.0015)	0.5928 (0.0014)	0.5925 (0.0015)
Genre and Occupation only*	200×200	**0.5377** (0.0036)	0.5353 (0.0030)	0.5223 (0.0031)
	500×500	**0.5669** (0.0032)	0.5516 (0.0025)	0.5385 (0.0038)
	1000×1000	**0.5915** (0.0025)	0.5671 (0.0024)	0.5671 (0.0037)
All	200×200	0.5809 (0.0034)	**0.5900** (0.0022)	0.5649 (0.0050)
	500×500	0.6026 (0.0020)	**0.6063** (0.0020)	0.6014 (0.0024)
	1000×1000	0.6189 (0.0015)	0.6140 (0.0015)	**0.6198** (0.0016)
All*	200×200	0.5862 (0.0027)	**0.5885** (0.0023)	0.5757 (0.0034)
	500×500	**0.6112** (0.0022)	0.6083 (0.0021)	0.6088 (0.0023)
	1000×1000	**0.6246** (0.0020)	0.6183 (0.0018)	0.6235 (0.0021)

Note: * indicates that multivariate Gaussian kernels are used for genre and occupation.

Now consider a pair of x and z satisfying (8).
(i) Suppose that $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$, yet $f^*(x) \le f^*(z)$. Since l' is non-decreasing,

$$\begin{aligned}
&\int_{\mathcal{X}} (l'(f^*(x) - f^*(x'))g_+(x)g_-(x') - l'(f^*(x') - f^*(x))g_-(x)g_+(x'))dx' \\
&\le \int_{\mathcal{X}} (l'(f^*(z) - f^*(x'))g_+(x)g_-(x') - l'(f^*(x') - f^*(z))g_-(x)g_+(x'))dx' \\
&< \frac{g_-(x)}{g_-(z)} \left(\int_{\mathcal{X}} (l'(f^*(z) - f^*(x'))g_+(z)g_-(x') - l'(f^*(x') - f^*(z))g_-(z)g_+(x'))dx' \right). \qquad (9)
\end{aligned}$$

The last strict inequality comes from the fact that $\int_{\mathcal{X}} l'(f^*(z) - f^*(x'))g_-(x')dx' < 0$. Since the lower and upper bounds are both 0 by the optimality condition (8), the above inequality leads to contradiction. Hence $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$ implies $f^*(x) > f^*(z)$.
(ii) If $\frac{g_+(x)}{g_-(x)} = \frac{g_+(z)}{g_-(z)}$ and $f^*(z) > f^*(x)$, similar derivation as in (9) shows that

$$\begin{aligned}
&\int_{\mathcal{X}} (l'(f^*(x) - f^*(x'))\frac{g_+(x)}{g_-(x)}g_-(x') - l'(f^*(x') - f^*(x))g_+(x'))dx' \\
&< \int_{\mathcal{X}} (l'(f^*(z) - f^*(x'))\frac{g_+(z)}{g_-(z)}g_-(x') - l'(f^*(x') - f^*(z))g_+(x'))dx' \\
&= \frac{1}{g_-(z)} \int_{\mathcal{X}} (l'(f^*(z) - f^*(x'))g_+(z)g_-(x') - l'(f^*(x') - f^*(z))g_-(z)g_+(x'))dx'.
\end{aligned}$$

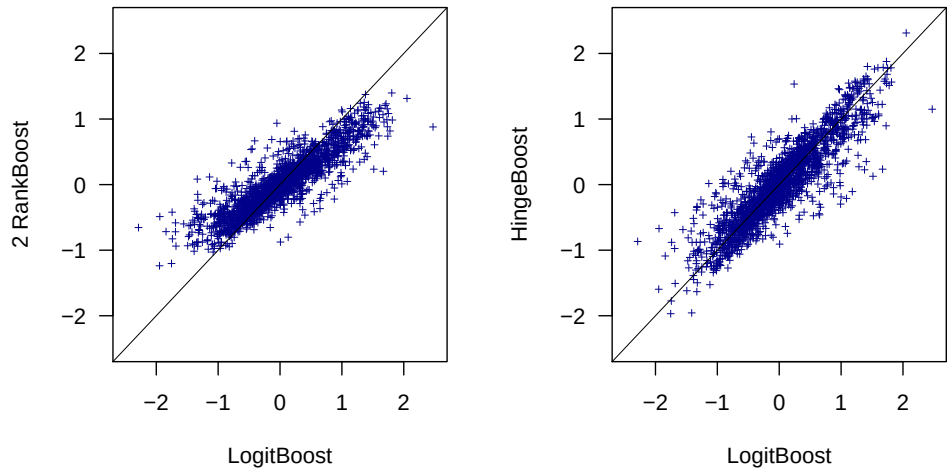

Figure 11: Scatter plots of the scores from the ranking functions with all the variables fitted to a million training pairs by RankBoost and HingeBoost plotted against the ranking scores from LogitBoost when the Gaussian kernels are used as weak learners.

The strict inequality holds because l' is one-to-one. Again both bounds are 0, leading to contradiction. Similarly, assuming $f^*(z) < f^*(x)$ yields the same contradiction. Consequently, $f^*(z) = f^*(x)$. □

Proof of Theorem 5

Proof. For part (i), we show that a set of the increments of the optimal ranking function can be replaced with an alternative set of either 0 or 1 only without increasing the risk, which contradicts the uniqueness of the optimal function. The alternative set of increments is constructed by using the fact that the subdifferential of hinge loss changes only at 1, and solving 0-1 valued equations for the partial sums of new increments. For part (ii), we approximate the optimal ranking function by a sequence of simple functions that depend on the likelihood ratio, akin to such approximation in real analysis, and apply the result in (i) to the simple functions.

(i) For $l(s) = (1-s)_+$, the risk $R_l(f)$ is given by $\sum_{i=1}^{N} g_+(x_i^*)g_-(x_i^*)$

$$+\sum_{i=2}^{N}\sum_{k=1}^{i-1}\left[(1-(f(x_i^*)-f(x_k^*)))_+ g_+(x_i^*)g_-(x_k^*) + (1-(f(x_k^*)-f(x_i^*)))_+ g_+(x_k^*)g_-(x_i^*)\right].$$

Letting $s_i = f(x_{i+1}^*) - f(x_i^*)$, we can express $R_l(f)$ as a function of s_i. $R_l(f)$ is

$$\sum_{i=2}^{N}\sum_{k=1}^{i-1}\left[\left(1-\sum_{l=k}^{i-1} s_l\right)_+ g_+(x_i^*)g_-(x_k^*) + \left(1+\sum_{l=k}^{i-1} s_l\right)_+ g_+(x_k^*)g_-(x_i^*)\right]$$

up to a constant. Then its minimizer f^* can be identified by $s_i^* = f^*(x_{i+1}^*) - f^*(x_i^*)$.

From Theorem 4, we know that $s_i^* \geq 0$. On the other hand, if $s_i^* > 1$, truncation of s_i^* to 1 does not change the first term in the risk while the second term gets smaller. Hence, $0 \leq s_i^* \leq 1$ for all $i = 1, \ldots, N-1$.

In the following we show that if there were s_i^* with $0 < s_i^* < 1$, then they could be replaced with either 0 or 1 without changing the risk. This contradicts the assumption that the minimizer is unique. Therefore, the increments of f^* have to be either 0 or 1.

For given sequence of $\{s_i^*\}_{i=1}^{N-1}$, consider blocks of s_i^* divided by $s_i^* = 1$. Let $I = \{i_1, \ldots, i_m\} = \{i \mid s_i^* = 1\}$, where m is the cardinality of I, and $i_1 < \cdots < i_m$. Defining $i_0 = 0$ and $i_{m+1} = N$, we can form $(m+1)$ blocks of indices, $B_j = \{i \mid i_j < i < i_{j+1}\}$ for $j = 0, \ldots, m$. Depending on i_j, some blocks may be empty.

To show that s_i^* in each block can be replaced with a sequence of 0 or 1 without changing the risk, define a new sequence $\{t_i\}_{i=1}^{N-1}$ as follows. If $s_i^* = 0$ or 1, set $t_i = s_i^*$. Otherwise, first find the block B_j that contains i. We can choose $t_i \in \{0, 1\}$ such that for every pair of (k, k') from B_j with $k \leq k'$, $\sum_{l=k}^{k'} t_l \leq 1$ (or ≥ 1) if $\sum_{l=k}^{k'} s_l^* \leq 1$ (or ≥ 1). We will show that such a choice of t_i is always feasible.

Let $C_j = \{k \in B_j \mid 0 < s_k^* < 1\} = \{\sigma_j(1), \ldots, \sigma_j(J_j)\}$, where J_j is the cardinality of the set. For each $\sigma_j(k) \in C_j$, consider $\tau_j(k) = \max_{k'}\{k' | \sum_{l=k}^{k'} s_{\sigma_j(l)}^* \leq 1,\ k' \geq k\}$.

Given the $\{\tau_j(k)\}_{k=1}^{J_j}$, we show that t_i for every $i \in C_j$ can be set to satisfy $\sum_{l=k}^{\tau_j(k)} t_{\sigma_j(l)} = 1$ for each $\sigma_j(k) \in C_j$ and $t_i \in \{0, 1\}$. Here we provide a proof by mathematical induction. If $J_j = 1$, then $t_i = t_{\sigma_j(1)} = 1$ follows immediately since $\tau_j(1) = 1$. Suppose that our assumption is true when $J_j = J$. For $J_j = J + 1$, consider the $(J+1)$ equations that t_i for $i \in C_j$ should satisfy. Depending on $\tau_j(k)$, they are of the following form:

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 1 & 1 & 0 & \ldots & 0 \\ & \vdots & & \ddots & & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} t_{\sigma_j(1)} \\ t_{\sigma_j(2)} \\ \vdots \\ t_{\sigma_j(J)} \\ t_{\sigma_j(J+1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix},$$

where the kth row of the matrix in the left hand side consists of ones only for the $t_{\sigma_j(l)}$, $l = k, \ldots, \tau_j(k)$, and zeros elsewhere. When we delete the first equation and remove $t_{\sigma_j(1)}$ from the rest, the equations above become

$$\begin{pmatrix} 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 1 & 1 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} t_{\sigma_j(2)} \\ \vdots \\ t_{\sigma_j(J)} \\ t_{\sigma_j(J+1)} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}.$$

By the assumption, we can set $t_{\sigma_j(2)}, \ldots, t_{\sigma_j(J+1)}$ such that they are either 0 or 1. When $\tau_j(1) = 1$, $t_{\sigma_j(1)} = 1$ trivially. When $\tau_j(1) \geq 2$, from $\sum_{l=2}^{\tau_j(2)} t_{\sigma_j(l)} = 1$, and $0 \leq \sum_{l=2}^{\tau_j(1)} t_{\sigma_j(l)} \leq \sum_{l=2}^{\tau_j(2)} t_{\sigma_j(l)} = 1$, $\sum_{l=2}^{\tau_j(1)} t_{\sigma_j(l)}$ as an integer should be either 0 or 1. From the equation that $\sum_{l=1}^{\tau_j(1)} t_{\sigma_j(l)} = 1$, it follows that $t_{\sigma_j(1)} = \sum_{l=1}^{\tau_j(1)} t_{\sigma_j(l)} - \sum_{l=2}^{\tau_j(1)} t_{\sigma_j(l)}$ is also either 0 or 1. Hence $t_{\sigma_j(1)}, \ldots, t_{\sigma_j(J+1)}$ are either 0 or 1. This completes the proof of the claim that $t_i \in \{0, 1\}$, $i = 1, \ldots, N-1$.

Now, based on $\{t_i\}_{i=1}^{N-1}$ in the proof, we verify that for every pair of (k, k') from B_j with $k \leq k'$, $\sum_{l=k}^{k'} t_l \leq 1$ (or ≥ 1) if $\sum_{l=k}^{k'} s_l^* \leq 1$ (or ≥ 1). If $k = k'$, then $\sum_{l=k}^{k'} s_l^* = s_k^* < 1$. As t_k for $s_k^* < 1$ is either 0 or 1, $t_k \leq 1$, and the statement clearly holds true. If $\sum_{l=k}^{k'} s_l^* = 0$ for

$k < k'$, then $s_l^* = 0$ for every $l = k, \ldots, k'$, and thus $t_l = 0$, retaining the inequality. Since $\sum_{l=k}^{k'} s_l^* = \sum_{l=k}^{k'} s_l^* I(s_l^* \neq 0)$, it is sufficient to consider (k, k') from C_j with $k < k'$. For such a pair (k, k') from C_j, there exist two indices i' and j' such that $k = \sigma_j(i')$ and $k' = \sigma_j(j')$, and $\sum_{l=k}^{k'} s_l^* = \sum_{l=i'}^{j'} s_{\sigma_j(l)}^*$ and $\sum_{l=k}^{k'} t_l = \sum_{l=i'}^{j'} t_{\sigma_j(l)}$. Since both $s_{\sigma_j(1)}^*, \ldots, s_{\sigma_j(J_j)}^*$ and $t_{\sigma_j(1)}, \ldots, t_{\sigma_j(J_j)}$ are nonnegative, for (i', j') with $1 \leq i' < j' \leq J_j$, $\sum_{l=i'}^{j'} s_{\sigma_j(l)}^* \leq 1$ (or ≥ 1) implies that $j' \leq \tau_j(i')$ (or $\geq \tau_j(i')$) and $\sum_{l=i'}^{j'} t_{\sigma_j(l)} \leq \sum_{l=i'}^{\tau_j(i')} t_{\sigma_j(l)} = 1$ (or ≥ 1) as desired.

Finally, we confirm that the new ranking function $\hat{f}$ defined by $t_i = \hat{f}(x_{i+1}^*) - \hat{f}(x_i^*)$ has the same risk as f^* with s_i^*. To the end, we demonstrate that $\hat{f}$ satisfies the optimality condition that the subdifferential of the risk of $\hat{f}$ includes zero. Letting l' denote a subderivative of l, we have the expression of the subderivative of $R_l(f)$ (as a function of s_i) taken with respect to s_j for $j = 1, \ldots, N-1$:

$$\sum_{i=1}^{j} \sum_{k=j-i+1}^{N-i} l'\Big(\sum_{l=i}^{k+i-1} s_l \Big) g_+(x_{i+k}^*) g_-(x_i^*) - l'\Big(- \sum_{l=i}^{k+i-1} s_l \Big) g_+(x_i^*) g_-(x_{i+k}^*).$$

If the index set $\{i, \ldots, k+i-1\}$ is contained in B_j for some j, then by the property of $\{t_i\}_{i=1}^{N-1}$, $\sum_{l=i}^{k+i-1} s_l^* \leq 1$ (or ≥ 1) implies $\sum_{l=i}^{k+i-1} t_l \leq 1$ (or ≥ 1), and in particular, the equality holds simultaneously for s_i^* and t_i. Hence $l'(\sum_{l=i}^{k+i-1} s_l^*) = l'(\sum_{l=i}^{k+i-1} t_l)$ in the case. If $\{i, \ldots, k+i-1\}$ includes any index from $I = \{i \mid s_i^* = 1\}$, then there are two possibilities: i) $\{s_j^*\}_{j=i}^{k+i-1}$ consists of either 0 or 1, and ii) for some j in $\{i, \ldots, k+i-1\}$, $0 < s_j^* < 1$. In the former, $\sum_{l=i}^{k+i-1} s_l^* = \sum_{l=i}^{k+i-1} t_l$, and thus $l'(\sum_{l=i}^{k+i-1} s_l^*) = l'(\sum_{l=i}^{k+i-1} t_l)$. In the latter, $\sum_{l=i}^{k+i-1} s_l^* > 1$ and $\sum_{l=i}^{k+i-1} t_l \geq 1$. Therefore $l'(\sum_{l=i}^{k+i-1} t_l)$ can be taken to be the same as $l'(\sum_{l=i}^{k+i-1} s_l^*) = 0$. As a result, if the subdifferential of $R_l(f)$ at f^* contains zero, then that of $R_l(f)$ at $\hat{f}$ contains zero as well. Due to the translation invariance of ranking functions, $\hat{f}$ can be taken to be integer-valued.

The assumption of strict ordering of x_i^* can be relaxed to allow some ties in the likelihood ratio. The same proof remains true if we consider equivalence classes defined by the likelihood ratio and relabel x_i^* as its equivalence class denoted by $[x_i^*]$. See Uematsu and Lee (2011) for detailed proof of this fact.

(ii) For any ranking function f defined on $\mathcal{X}$, consider a simple function f_n of the form:

$$f_n(x) = \sum_{i=1}^{2nM} \left(\frac{i}{n} - M \right) I(x \in A_i(f)) + M \cdot I(x \in A_{2nM+1}(f)) - M \cdot I(x \in A_0(f)),$$

where M is a positive constant, $A_i(f) = \{x \mid \frac{i-1}{n} - M < f(x) \leq \frac{i}{n} - M\}$, $A_0(f) = \{x \mid f(x) \leq -M\}$, and $A_{2nM+1}(f) = \{x \mid f(x) > M\}$. It is easy to see that

$$\begin{aligned}
& \big| E\{(1 - (f(X) - f(X')))_+\} - E\{(1 - (f_n(X) - f_n(X')))_+\} \big| \\
\leq\ & E\{I(|f(X)| \leq M, |f(X')| \leq M) \big| (f(X) - f(X')) - (f_n(X) - f_n(X')) \big| \} \\
& + 2E\{I(|f(X)| > M)(1 - (f(X) - f(X')))_+\} \\
& + 2E\{I(|f(X')| > M)(1 - (f(X) - f(X')))_+\} + \frac{2}{n} \\
\leq\ & \frac{4}{n} + 2E\{I(|f(X)| > M)(1 - (f(X) - f(X')))_+\} \\
& + 2E\{I(|f(X')| > M)(1 - (f(X) - f(X')))_+\}.
\end{aligned}$$

Without loss of generality, assume that $R_l(f) < \infty$. Then from the above inequality and the dominated convergence theorem, for any positive number ϵ, there exist n and M such that $\left|R_l(f) - R_l(f_n)\right| < \epsilon$.

Each f induces a partition of the sample space $\mathcal{X}$ through $\{A_i(f)\}_{i=0}^{2nM+1}$. Letting $\mathcal{A}(f)$ denote the partition, consider the discretized probability mass functions $\hat{g}_+$ and $\hat{g}_-$ on $\mathcal{A}(f)$ defined as $\hat{g}_\pm(A_i(f)) = \int_{A_i(f)} g_\pm(x)dx$.

Since the optimal ranking function f^* is a monotonic transformation of the likelihood ratio g_+/g_-, there exists a sequence $\{\alpha_i\}_{i=0}^{2nM+1}$ such that $A_i(f^*) = \left\{x \mid \alpha_{i-1} < \frac{g_+(x)}{g_-(x)} \leq \alpha_i\right\}$. $A_i(f^*)$ is in the order of the likelihood ratio $\hat{g}_+/\hat{g}_-$. For this, observe that

$$\begin{aligned}\hat{g}_+(A_{i+1})\hat{g}_-(A_i) &= \int_{A_{i+1}}\int_{A_i} g_+(x)g_-(x')dx'dx > \int_{A_{i+1}}\int_{A_i} g_-(x)g_+(x')dx'dx \\ &= \hat{g}_+(A_i)\hat{g}_-(A_{i+1})\end{aligned}$$

since $\alpha_{i-1} < \frac{g_+(x')}{g_-(x')} \leq \alpha_i < \frac{g_+(x)}{g_-(x)} \leq \alpha_{i+1}$, and thus $\frac{\hat{g}_+(A_{i+1})}{\hat{g}_-(A_{i+1})} > \frac{\hat{g}_+(A_i)}{\hat{g}_-(A_i)}$ for every i.

For the discretized version of the ranking problem with pmfs $\hat{g}_+$ and $\hat{g}_-$ on the countable $\mathcal{A}(f^*)$, by the result (i), there is an integer-valued function among the optimal ranking functions.

Given ϵ, let f_n^* be the simple function corresponding to f^* such that $\left|R_l(f^*) - R_l(f_n^*)\right| < \epsilon$, and let $\hat{f}_n$ be the integer-valued optimal ranking function on $\mathcal{A}(f^*)$. Then $R_l(\hat{f}_n) \leq R_l(f_n^*) < R_l(f^*) + \epsilon$. This completes the proof. □

References

Agarwal, S. (2012). Surrogate regret bounds for bipartite ranking via strongly proper losses. http://arxiv.org/abs/1207.0268.

Agarwal, S., Graepel, T., Herbrich, T., Har-Peled, T. and Roth, D. (2005). Generalization bounds for the area under the ROC curve, *Journal of Machine Learning Research* **6**: 393–425.

Agarwal, S. and Niyogi, P. (2005). Stability and Generalization of Bipartite Ranking Algorithms, *Lecture Notes in Computer Science*, Vol. 3559, Springer Berlin/Heidelberg, pp. 32–47.

Bartlett, P., Jordan, M. and McAuliffe, J. (2006). Convexity, classification, and risk bounds, *Journal of the American Statistical Association* **101**: 138–156.

Bell, R. M., Koren, Y. and Volinsky, C. (2010). All together now: A perspective on the Netflix prize, *Chance* **23**: 24–29.

Brefeld, U. and Scheffer, T. (2005). AUC maximizing support vector learning, *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. (2005). Learning to rank using gradient descent, *Proceedings of the 22nd international conference on Machine learning*, Vol. 119 of *ACM International Conference Proceeding Series*.

Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y. and Hon, H.-W. (2006). Adapting ranking SVM to document retrieval, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, pp. 186–193.

Clémençon, S. J. and Vayatis, N. (2009a). Empirical performance maximization for linear rank statistics, *in* D. Koller, D. Schuurmans, Y. Bengio and L. Bottou (eds), *Advances in Neural Information Processing Systems 21*, pp. 305–312.

Clémençon, S., Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics, *Annals of Statistics* **36**: 844–874.

Clémençon, S. and Vayatis, N. (2007). Ranking the best instances, *Journal of Machine Learning Research* **8**: 2671–2699.

Clémençon, S. and Vayatis, N. (2009b). Tree-based ranking methods, *IEEE Transactions on Information Theory* **55**: 4316–4336.

Cortes, C. and Mohri, M. (2004). AUC optimization vs. error rate minimization, *Advances in Neural Information Processing Systems*, Vol. 16, MIT Press, pp. 323–320.

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

Crammer, K. and Singer, Y. (2001). Pranking with ranking, *Advances in Neural Information Processing Systems 14*, MIT Press, pp. 641–647.

Duchi, J., Mackey, L. and Jordan, M. (2010). On the consistency of ranking algorithms, *Proceedings of the 27th International Conference on Machine Learning*.

Freund, Y., Iyer, R., Schapire, R. and Singer, Y. (2003). An efficient boosting algorithm for combining preferences, *Journal of Machine Learning Research* **4**: 933–969.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting, *The Annals of Statistics* **28**: 337–407.

Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *Communications of ACM* **35**: 61–70.

GroupLens-Research (2006). MovieLens Data Sets. http://grouplens.org/node/12.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology* **143**(1): 29–36.

Herbrich, R., Graepel, T. and Obermayer, K. (2000). Large margin rank boundaries for ordinal regression, *in* A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (eds), *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, pp. 115–132.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Statistical Research Report.

Joachims, T. (2002). Optimizing search engines using clickthrough data, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, ACM, New York, NY, pp. 133–142.

Koren, Y., Bell, R. and Volinsky, C. (2009). Matrix factorization techniques for recommender systems, *IEEE Computer* **42**(8): 30–37.

Kotlowski, W., Dembczynski, K. and Hüllermeier, E. (2011). Bipartite ranking through minimization of univariate loss, *in* L. Getoor and T. Scheffer (eds), *ICML*, pp. 1113–1120.

Le, Q. and Smola, A. (2007). Direct optimization of ranking measures. http://arxiv.org/abs/0704.3359.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory* **7**: 186–199.

Rakotomamonjy, A. (2004). Optimizing area under ROC curve with SVMs, *Proceedings of the First Workshop on ROC Analysis in Artificial Intelligence*.

Reid, M. and Williamson, R. (2010). Composite binary losses, *Journal of Machine Learning Research* **11**: 2387–2422.

Rockafellar, R. (1997). *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*, Princeton University Press.

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**: 2233–2271.

Rudin, C. and Schapire, R. (2009). Margin-based Ranking and an Equivalence between AdaBoost and RankBoost, *The Journal of Machine Learning Research* **10**: 2193–2232.

Shardanand, U. and Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth", *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, ACM Press/Addison-Wesley Publishing Co., New York, NY, pp. 210–217.

Uematsu, K. and Lee, Y. (2011). On theoretically optimal ranking functions in bipartite ranking, *Technical Report 863*, Department of Statistics, The Ohio State University. http://www.stat.osu.edu/∼yklee/mss/tr863.pdf.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization, *Annals of Statistics* **32**(1): 56–85.

Zheng, Z., Zha, H., Zhang, T., Chapelle, O., Chen, K. and Sun, G. (2008). A general boosting method and its application to learning ranking functions for web search, *in* J. Platt, D. Koller, Y. Singer and S. Roweis (eds), *Advances in Neural Information Processing Systems 20*, MIT Press, Cambridge, MA, pp. 1697–1704.