# Statistical Optimality in Multipartite Ranking and Ordinal Regression

Kazuki Uematsu, *Chemitox Inc., Japan*
Yoonkyung Lee, *The Ohio State University*

**Technical Report No. 873**

**August, 2013**

**Department of Statistics**
**The Ohio State University**
**1958 Neil Avenue**
**Columbus, OH 43210-1247**

# Statistical Optimality in Multipartite Ranking and Ordinal Regression

Kazuki Uematsu
Chemitox, Inc.
Yamanashi Testing Center
Hokuto Yamanashi 408-0103, Japan
uematsu.1@buckeyemail.osu.edu

Yoonkyung Lee[*]
Department of Statistics
The Ohio State University
Columbus, OH 43210
yklee@stat.osu.edu

### Abstract

Statistical optimality in multipartite ranking is investigated as an extension of bipartite ranking. We consider the optimality of ranking algorithms through minimization of the theoretical risk which combines pairwise ranking errors of ordinal categories with differential ranking costs. The extension shows that for a certain class of convex loss functions including exponential loss, the optimal ranking function can be represented as a ratio of weighted conditional probability of upper categories to lower categories, where the weights are given by the misranking costs. This result also bridges traditional ranking methods such as proportional odds model in statistics with various ranking algorithms in machine learning. Further, the analysis of multipartite ranking with different costs provides a new perspective on non-smooth ranking measures such as the discounted cumulative gain (DCG) and preference learning. We illustrate our findings with simulation study and real data analysis.

## 1 Introduction

The need for ranking given instances arises in a wide range of applications as in collaborative filtering, information retrieval, recommender systems, and computational biology. For example, given a query, we want to rank web pages according to their relevance to the query. As a supervised learning problem, ranking concerns how to learn a general rule to order instances from training data with attributes and associated labels that determine the desired preference between observed instances. Using the observed ordering of instances, we estimate a ranking (or scoring) function such that the resulting scores reflect the ordinal relation among the labels.

There are various forms of ranking problems in the machine learning literature, including bipartite ranking, multipartite ranking, preference learning, and multilabel ranking in the context of multilabel classification. Bipartite ranking is a special form of ranking, where instances have only binary labels (e.g. positive or negative). A desired ranking function, whose scores induce an ordering over the instance space, would assign higher scores to positive instances than negative instances. The standard loss used in bipartite ranking penalizes violation of the order of a pair of instances with known preference. Minimization of the proportion of misordered pairs is shown to be equivalent to maximization of the Area Under the ROC Curve (AUC) through the link between the AUC criterion and the Wilcoxon-Mann-Whitney statistic; see Hanley and McNeil (1982) and Cortes and Mohri (2004).

[*] Lee's research was supported in part by National Science Foundation grant DMS-12-09194.

There is a great parallel between binary classification and bipartite ranking. From this perspective, bipartite ranking has been studied quite extensively from computational to theoretical aspects. Several ranking algorithms are inspired by analogous classification methods. For example, RankBoost (Freund et al. 2003), RankNet (Burges et al. 2005), and AUC maximizing support vector machine (SVM) (Brefeld and Scheffer 2005, Rakotomamonjy 2004) are a ranking version of AdaBoost, logistic regression, and SVM. In general, the computational strategy of convex risk minimization with a surrogate loss for classification has been adopted for bipartite ranking.

On the theoretical front, the notion of the Bayes ranking function or the best ranking function with minimum ranking error has been established in parallel with the the Bayes classification rule in classification. Clémençon et al. (2008) and Uematsu and Lee (2011) showed that the theoretically optimal ordering over the instance space is determined by the likelihood ratio of the positive label to the negative label, and the best ranking functions under some convex loss criteria produce the same ordering. Uematsu and Lee (2011) further examined the ranking calibration condition for a family of convex surrogate loss criteria to ensure ranking consistency akin to the classification calibration condition in Bartlett et al. (2006) and identified the explicit form of optimal ranking functions under some loss criteria. With the theory of $U$-processes, Clémençon et al. (2008) studied the consistency of empirical risk minimizers in bipartite ranking. Agarwal et al. (2005) and Agarwal and Niyogi (2005) obtained generalization bounds for ranking, using the standard learning theory. Recently, Kotłowski et al. (2011) and Agarwal (2013) investigated bipartite ranking consistency through regret bounds when discriminant functions from binary classification are directly used as ranking functions.

The aim of this paper is to investigate the notion of optimal ranking functions under extension of the bipartite ranking error on the population level when there are more than two ordered labels, and to generalize the results in bipartite ranking to the multipartite case. This optimality is then used to define consistency in multipartite ranking.

Through the framework of pairwise ranking with differential costs, we show that for a certain class of convex loss functions, the optimal ranking function can be represented as a ratio of a weighted sum of conditional probabilities of upper categories to that of lower categories, where the weights are given by the misranking costs. Based on the results, we investigate the link between statistical methods including subset ranking using regression in Cossock and Zhang (2008), ordinal regression in Chu and Keerthi (2007) and Shashua and Levin (2003), proportional odds model in McCullagh (1980) and other pairwise ranking methods as in Agarwal and Niyogi (2009), and show that the optimal function in each method can be represented as such a ratio. Further we investigate the consistency of risk minimization with a convex surrogate loss in the multipartite case and propose several convex risk minimization techniques that can guarantee ranking consistency.

In addition, we consider non-smooth ranking measures that capture the practical need for accuracy of the instances near the top of the list in many ranking applications. There are several non-smooth ranking measures: for example, the average precision (AP), normalized discounted cumulative gain (NDCG) and others (Rudin 2009, Cossock and Zhang 2008, Clémençon and Vayatis 2007, Le and Smola 2007). Many papers have proposed optimization methods for such ranking measures; see Xu and Li (2007) and Yue et al. (2007), for instance, and Chen et al. (2009) for theoretical investigation. We provide a new perspective on non-smooth ranking measures in connection with pairwise ranking with differential costs.

Some earlier work on performance measures and optimal ranking functions in multipartite ranking includes Waegeman and Baets (2011), Waegeman et al. (2008), and Clémençon et al. (2011). These references are concerned about maximization of the Volume Under the ROC Surface (VUS) as the main evaluation metric, which differs from minimization of the expected pairwise ranking cost in this paper. The analysis of optimality in Clémençon et al. (2011) employs a stringent condition called the likelihood ratio monotonicity, which assumes that the order of instances given by likelihood ratio for any pair of lower and upper categories is the same. This assumption is so strong that it reduces multipartite ranking to a collection of bipartite ranking problems, which is in contrast with minimization of the average pairwise ranking cost and ordinal regression considered in this paper.

This paper is organized as follows. Section 2 reviews the problem setting and basic results about the optimal ranking functions and consistency in bipartite ranking. Section 3 introduces a pairwise loss function weighing misranking costs for multipartite ranking, and extends the theoretically optimal ranking function in bipartite ranking to the multipartite case as a minimizer of the ranking risk under the loss. Section 4 examines the relation between the optimal ranking function for the pairwise approach and the population version of the target ranking function for commonly used ordinal regression methods such as ordinal regression boosting, support vector ordinal regression, and proportional odds model. Alternatively, Section 5 considers convex risk minimization with surrogate loss functions, and discusses necessary adjustment to ensure ranking consistency. Further, connection between pairwise ranking risk minimization and optimization of non-smooth ranking measures is investigated. Section 6 presents numerical analysis using simulation data and MovieLens data for illustration, followed by conclusion in Section 7.

## 2 Review of Bipartite Ranking

Let $\mathcal{X}$ be the space of objects or instances that we want to rank and $\mathcal{Y}$ be the space of labels. Suppose that we have training data for ranking which consist of independent pairs of $(X, Y)$ from $\mathcal{X} \times \mathcal{Y}$. Each object has associated attributes $X$ for ranking and an ordinal response $Y$.

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. An object $x$ is considered to be preferred to $x'$ by $f$ if $f(x) > f(x')$. From the training data, we want to learn a ranking function $f$ such that objects with higher label are generally scored higher than those with lower label, that is, $f(x) > f(x')$ if $y > y'$ for a pair of $(x, y)$ and $(x', y')$.

For example, consider constructing a movie recommender system which produces an ordered list of movies to users based on movie attributes (e.g. genre and release year) and user characteristics (e.g. gender and age). In this case, a ranking function $f$ for the system assigns scores to movies for recommendation using the information about a movie and a user. Given training data of movie ratings in $\mathcal{Y}$ with attributes of movies and users in $\mathcal{X}$, we can learn a scoring function by considering two distinctive ratings $y$ and $y'$ with corresponding movie and user information $x$ and $x'$ in a pairwise manner. The scoring function is expected to preserve the observed orderings, satisfying the relation that $f(x) > f(x')$ if $y > y'$ for a pair of $(x, y)$ and $(x', y')$.

As a special case of ranking, bipartite ranking refers to the case when there are only two labels represented as $\mathcal{Y} = \{1, -1\}$. The labels are called positive and negative, respectively, analogous to binary classification. We review the concepts and results obtained for bipartite ranking first as the most basic form of ranking and discuss their extensions to multipartite ranking.

For each pair of a positive object $x$ and a negative object $x'$, the bipartite ranking loss of $f$ is defined as

$$L_0(f; x, x') = I(f(x) < f(x')) + \frac{1}{2} I(f(x) = f(x')), \tag{1}$$

where $I(\cdot)$ is the indicator function. Note that the loss is invariant under any order-preserving transformation of $f$. $L_0(f; x, x')$ can be expressed as $l_0(f(x) - f(x'))$, where $l_0(s) = I(s < 0) +$

$\frac{1}{2}I(s = 0)$. The theoretical ranking risk of $f$ is then defined as

$$\begin{aligned} R_0(f) &= E_{X,X'}[L_0(f; X, X')] \\ &= P(f(X) < f(X')|Y = 1, Y' = -1) + \frac{1}{2}P(f(X) = f(X')|Y = 1, Y' = -1). \end{aligned}$$

Ranking risk minimization is shown to be equivalent to the AUC maximization as the AUC is one minus the ranking risk.

Assuming that $X$ is a continuous random variable or vector, let $g_+$ be the pdf of $X$ for the positive category, and let $g_-$ be that for the negative category. We further assume that $0 < g_+(x) < \infty$ and $0 < g_-(x) < \infty$ for $x \in \mathcal{X}$ in this paper. Under this setting, the optimal ranking function for bipartite ranking is shown to be any order-preserving function of the likelihood ratio of $x$ under two categories $g_+(x)/g_-(x)$ as stated in the following theorem. See Clémençon et al. (2008), Uematsu and Lee (2011), Kotłowski et al. (2011) and Agarwal (2013).

**Theorem 1.** *Let $f_0^*(x) \equiv g_+(x)/g_-(x)$. For any ranking function $f$,*

$$R_0(f_0^*) \leq R_0(f).$$

Let $p_1(x) = P(Y = 1|X = x)$. Then, alternatively, the optimal ranking function can be represented by $p_1(x)/(1 - p_1(x))$, which is proportional to $g_+(x)/g_-(x)$. We will use a similar representation of the optimal ranking function for multipartite ranking in later sections.

It is worth noting that $p_1(x)/(1 - p_1(x))$ is essential not only in ranking but also regression and classification. It is known that $c^*(x) = \text{sign}[\log\{p_1(x)/(1 - p_1(x))\}]$ achieves the Bayes risk in classification (the minimum error rate). Hence, consistent classification methods that estimate the conditional probability $p_1(x)$ or equivalently logit can be used for ranking. In regression, the least squares estimator approximates $E[Y|x] = P(Y = 1|x) - P(Y = -1|x)$, which can be expressed as $1 - 2\{1 + p_1(x)/(1 - p_1(x))\}^{-1}$, a monotonic transformation of $p_1(x)/(1 - p_1(x))$. Thus the difference among regression, soft classification (probability model based classification) and ranking is not essential in bipartite ranking asymptotically, and it is justified to tackle ranking via soft classification or even regression on the population level.

## 2.1 Consistency in Bipartite Ranking

Given the training data with $n_+$ positive objects $\{x_i\}_{i=1}^{n_+}$ and $n_-$ negative ones $\{x_j'\}_{j=1}^{n_-}$, we can define the best ranking function on the sample level as the function $f$ minimizing the empirical ranking risk by considering all pairs of positive and negative instances in the training data:

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

However, $R_{n_+,n_-}$ is computationally hard to minimize over $f$ since $l_0$ is a non-convex function. To circumvent the computational issue, many researchers have proposed ranking algorithms that replace the bipartite ranking loss with a convex surrogate loss, for example, RankBoost (Freund et al. 2003), RankNet (Burges et al. 2005), and support vector ranking (Brefeld and Scheffer 2005).

Given a ranking function $f$ and a pair of a positive instance $x$ and a negative instance $x'$, consider a non-negative, non-increasing convex function $l : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ as a surrogate loss function, which defines $l(f(x) - f(x'))$ as a ranking loss. For example, the RankBoost algorithm takes the exponential loss, $l(s) = \exp(-s)$, and the support vector ranking takes the hinge loss, $l(s) = (1 - s)_+$ as a surrogate loss.

When a surrogate loss $l$ is used instead of $l_0$, there arises the issue of whether $f^*$ minimizing the surrogate risk $R_l(f) \equiv E[l(f(X) - f(X'))]$ among all measurable functions $f : \mathcal{X} \to \mathbb{R}$ may actually minimize the ranking error $R_0(f)$. A similar issue arises in binary classification, and conditions for surrogate loss functions to ensure classification consistency or *classification calibration* have been identified (see, for example, Bartlett et al. 2006). As a ranking analogue, Uematsu and Lee (2011) studied conditions of surrogate loss functions for ranking consistency or *ranking calibration* in bipartite ranking and further identified the forms of the optimal ranking functions under surrogate losses. See also Agarwal et al. (2005), Agarwal and Niyogi (2005), and Agarwal and Niyogi (2009) for AUC generalization bounds, and Kotłowski et al. (2011) and Agarwal (2013) for bipartite ranking consistency through regret bounds when discriminant functions from binary classification are directly used as ranking functions.

The following theorem from Uematsu and Lee (2011) states conditions for ranking calibration, and shows that for a *ranking-calibrated* loss $l$, the optimal ranking function $f^*$ under $l$ preserves the order of the likelihood ratio $f_0^*$.

**Theorem 2.** *Suppose that $l$ is convex, non-increasing, differentiable, and $l'(0) < 0$. Let $f^* \equiv \arg\min_f R_l(f)$. For almost every $(x, z) \in \mathcal{X} \times \mathcal{X}$, $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$ implies $f^*(x) > f^*(z)$.*

The theorem provides justification for RankBoost and RankNet since $l(s) = \exp(-s)$ and $l(s) = \log(1+\exp(-s))$ satisfy the conditions above and thus the optimal ranking function $f^*$ derived under $l$ preserves the order of the likelihood ratio without ties.

On the other hand, $l(s) = (1 - s)_+$, the hinge loss in support vector ranking has a singularity point at $s = 1$ and does not satisfy the conditions. Uematsu and Lee (2011) have shown that the optimal ranking function under the hinge loss can produce ties between objects with distinct values of the likelihood ratio, which leads to ranking inconsistency. Further, the theoretical analysis of ranking under the hinge loss in the paper suggests that the support vector ranking could produce granularity in ranking scores.

## 3 Extension to Multipartite Ranking

For extension of the results in bipartite ranking, we begin with loss functions for multipartite ranking. Several evaluation metrics for ranking error have been proposed; see, for example, Waegeman et al. (2008), and Hand and Till (2001). We briefly review some of the evaluation criteria designed for multipartite ranking.

Consider a $K$-partite ranking problem where the ordinal responses $y_i$ in the training data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, are in $\mathcal{Y} = \{1, \ldots, K\}$. Let $n_j$ be the number of observations in the category $j \in \{1, \ldots, K\}$. Waegeman et al. (2008) considered the following evaluation metrics for a given ranking function $f$:

$$\hat{U}_{\text{pairs}}(f) = \frac{1}{\sum_{l'<l} n_{l'} n_l} \sum_{y_i < y_j} I(f(x_i) < f(x_j)). \tag{2}$$

$$\hat{U}_{\text{ovo}}(f) = \frac{2}{K(K-1)} \sum_{l'<l} \hat{A}_{l'l}, \quad \text{where} \quad \hat{A}_{l'l} = \frac{1}{n_{l'} n_l} \sum_{y_i = l'} \sum_{y_j = l} I(f(x_i) < f(x_j)). \tag{3}$$

$$\hat{U}_{\text{cons}}(f) = \frac{1}{K-1} \sum_{l=1}^{K-1} \hat{B}_l, \quad \text{where} \quad \hat{B}_l = \frac{1}{\sum_{i=1}^{l} n_i \sum_{j=l+1}^{K} n_j} \sum_{y_i \le l} \sum_{y_j > l} I(f(x_i) < f(x_j)). \tag{4}$$

Given a pair of $(x, y)$ and $(x', y')$ with $y > y'$, let $L_t(f; x, x') = I(f(x) \le f(x'))$, where a tie is considered as an error, for simplicity. For the evaluation metrics, define $\hat{R}_{\text{pairs}}(f) = 1 - \hat{U}_{\text{pairs}}(f)$, $\hat{R}_{\text{ovo}}(f) = 1 - \hat{U}_{\text{ovo}}(f)$, and $\hat{R}_{\text{cons}}(f) = 1 - \hat{U}_{\text{cons}}(f)$, respectively.

Then we see that $\hat{R}_{\text{pairs}}(f)$ is the empirical ranking risk under the simple loss function $L_t$ evaluated over all the pairs of $(y_i, y_j)$ with $y_i > y_j$. On the other hand, $\hat{R}_{\text{ovo}}$ (Hand and Till 2001) can be regarded as the average of the (1−AUC) values by considering all pairs of the categories in the "one versus one" (abbreviated as "ovo") fashion, where $\hat{A}_{l'l}$ equals the AUC for the ranking function $f$ when $l'$ and $l$ are considered. Just as the AUC is related to Mann-Whitney statistic for a two-sample location problem, $\hat{U}_{\text{ovo}}$ is related to Jonckheere-Terpstra statistic for a $K$-sample location problem. The rationale behind $\hat{R}_{\text{cons}}$ is to transform multipartite ranking into $(K-1)$ bipartite ranking problems, where two categories are formed by taking the first "consecutive" $l$ categories, $\{1, \dots, l\}$ as negative, and the rest $\{l+1, \dots, K\}$ as positive for $l = 1, \dots, K-1$. Then $\hat{R}_{\text{cons}}$ is the average of the (1−AUC) values for the $(K-1)$ bipartite ranking problems, where $\hat{B}_l$ is the AUC of $f$ for the $l$th problem.

These metrics are easy to interpret, and their differences mainly lie in the way the weights are assigned to the ranking risk of $f$ for each pair of categories, as a function of sample sizes $\{n_i\}_{i=1}^K$. On the other hand, these evaluation metrics do not reflect potentially different consequence of misranking categories. For example, misranking of instances in categories 1 and $K$ is more serious than misranking of instances in categories 1 and 2 or categories $K-1$ and $K$. To take such differential costs into account, we consider a general representation of a loss function for ranking analogous to multicategory classification with differential costs.

For $K \ge 2$ and a pair of $(x, y)$ and $(x', y')$ with $y > y'$, we define a loss $L_0$ for ranking function $f$ as

$$L_0(f; (x, y), (x', y')) = c_{y'y}\left[I(f(x) < f(x')) + \frac{1}{2}I(f(x) = f(x'))\right], \tag{5}$$

where $c_{y'y}$ is a non-negative cost of incorrectly ranking $y'$ above $y$. $L_0$ is an extension of the bipartite ranking loss in (1). It encompasses existing loss functions for multipartite ranking.

Let $\mathbf{c} = \{c_{ji} | j < i \text{ for } i, j = 1, \dots, K\}$ represent a set of misranking costs. Then under the loss $L_0$, the ranking risk of $f$ is given as

$$\begin{aligned}
R_0(f; \mathbf{c}) &= E[L_0(f; (X, Y), (X', Y'))|Y > Y'] \\
&= \sum_{1 \le j < i \le K} c_{ji} P(f(X) < f(X')|Y = i, Y' = j, Y > Y')P(Y = i, Y' = j|Y > Y') \\
&\quad + \frac{1}{2} \sum_{1 \le j < i \le K} c_{ji} P(f(X) = f(X')|Y = i, Y' = j, Y > Y')P(Y = i, Y' = j|Y > Y') \\
&= \sum_{1 \le j < i \le K} c_{ji} P(f(X) < f(X')|Y = i, Y' = j)P(Y = i, Y' = j)/P(Y > Y') \\
&\quad + \frac{1}{2} \sum_{1 \le j < i \le K} c_{ji} P(f(X) = f(X')|Y = i, Y' = j)P(Y = i, Y' = j)/P(Y > Y').
\end{aligned}$$

If $c_{ji} = I(j < i)$, then $R_0(f; \mathbf{c})$ is the probability of misranking a pair of objects by $f$. Minimizing $R_0(f; \mathbf{c})$ is equivalent to maximizing the expected pairwise AUC.

When ties occur with probability zero, the evaluation criteria, $\hat{R}_{\text{pairs}}(f), \hat{R}_{\text{ovo}}(f)$ and $\hat{R}_{\text{cons}}(f)$ can be viewed as estimates of $R_0$ under $L_0$ with a specific cost structure $\mathbf{c}$. Define $\hat{R}_0(f; \mathbf{c})$ as the empirical ranking risk over the training data by assigning the equal weight of $\dfrac{1}{\sum_{l' < l} n_{l'} n_l}$ to each pair of $(x_j, y_j)$ and $(x_i, y_i)$ with $y_j < y_i$. When ties occur with probability zero,

$$\hat{R}_0(f;\mathbf{c}) = \sum_{1\le l'<l\le K} c_{l'l} \sum_{y_i=l'}\sum_{y_j=l} I(f(x_i) > f(x_j)) / \sum_{l'<l} n_{l'}n_l.$$

When $c_{ji} = I(j < i)$, $\hat{R}_0(f;\mathbf{c}) = \hat{R}_{\text{pairs}}(f)$. Similarly, if $c_{ji} = \frac{2}{K(K-1)}\big(\frac{\sum_{l'<l} n_{l'}n_l}{n_i n_j}\big)$, $\hat{R}_0(f;\mathbf{c}) = \hat{R}_{\text{ovo}}(f)$, and if $c_{ji} = \frac{1}{K-1}\big(\sum_{l'<l} n_{l'}n_l\big)\sum_{l=j}^{i-1}(\sum_{k=1}^{l} n_k \sum_{k'=l+1}^{K} n_{k'})^{-1}$, $\hat{R}_0(f;\mathbf{c}) = \hat{R}_{\text{cons}}(f)$.

Agarwal and Niyogi (2009) have considered the general notion of “ranking preference” to represent the misranking cost for a pair of $(x, y)$ and $(x', y')$, and used $|y - y'|$ as the preference. This corresponds to the cost of $c_{y'y} = y - y'$ for $y > y'$.

Now we identify the theoretical ranking function that minimizes the risk defined above under some mild conditions on misranking costs. The result can be viewed as a generalized version of Theorem 1 to multipartite ranking. Proof of the theorem is given in Appendix. It uses a similar argument as in bipartite ranking based on structured decomposition of the difference of the risks for an arbitrary ranking function and the optimal function.

**Theorem 3.**

*(i) When $K = 3$, let $f_0^*(x) := \dfrac{c_{12}P(Y=2|x) + c_{13}P(Y=3|x)}{c_{13}P(Y=1|x) + c_{23}P(Y=2|x)}$. Then for any ranking function $f$,*

$$R_0(f_0^*;\mathbf{c}) \le R_0(f;\mathbf{c}).$$

*(ii) When $K > 3$, let*

$$f_0^*(x) := \frac{\sum_{i=2}^{K} c_{1i}P(Y=i|x)}{\sum_{j=1}^{K-1} c_{jK}P(Y=j|x)}.$$

*If $c_{1K}c_{ji} = c_{1i}c_{jK} - c_{1j}c_{iK}$ for all $1 \le j < i \le K$, then for any ranking function $f$,*

$$R_0(f_0^*;\mathbf{c}) \le R_0(f;\mathbf{c}).$$

Theorem 3 implies that for pairwise multipartite ranking, optimal ranking depends on the ratio of conditional probabilities weighted by misranking costs. Under the natural assumption of $c_{11} = c_{KK} = 0$, the optimal ranking function can be also expressed as

$$f_0^*(x) = \frac{E(c_{1Y}|X=x)}{E(c_{YK}|X=x)},$$

which is the ratio of the expected misranking costs when an instance $x$ from category $Y$ is compared with the lowest category and the highest category. When higher categories are more likely for $x$, $f_0^*$ tends to produce a high ranking score as $c_{1y}$ increases in $y$ while $c_{yK}$ decreases in $y$. By contrast, when lower categories are more likely for $x$, the ratio of the expected costs tends to be small.

The condition on the ranking costs in Theorem 3 (ii) is not too restrictive, and it is satisfied by a variety of reasonable cost schemes. For example, $c_{ji} = (i - j)I(i > j)$ considered by Agarwal and Niyogi (2009) meets the condition.

For illumination of the condition, consider a "normalized" cost $c_{ij}/c_{1K}$, where we regard $c_{1K}$ (usually expected to be the maximum cost) as a normalization term. Then, the condition above is equivalent to the requirement that for $1 \le j < i \le K-1$,

$$\frac{c_{ji}}{c_{1K}} = \left(\frac{c_{jK}}{c_{1K}}\right)\left(\frac{c_{iK}}{c_{1K}}\right)\left(\frac{c_{1i}}{c_{iK}} - \frac{c_{1j}}{c_{jK}}\right).$$

Note the exclusion of $i = K$ in the restatement of the condition as the restriction in the theorem becomes void when $i = K$. Here, $\frac{c_{jK}}{c_{1K}}$ and $\frac{c_{iK}}{c_{1K}}$ can be considered as relative weights attached to labels $j$ and $i$, and $\left(\frac{c_{1i}}{c_{iK}} - \frac{c_{1j}}{c_{jK}}\right)$ can be viewed as relative spacing between the two labels in terms of cost. Observe that the relative spacing induced in the cost space is additive in the sense that for $1 \le k < j < i \le K-1$,

$$\left(\frac{c_{1i}}{c_{iK}} - \frac{c_{1j}}{c_{jK}}\right) + \left(\frac{c_{1j}}{c_{jK}} - \frac{c_{1k}}{c_{kK}}\right) = \frac{c_{1i}}{c_{iK}} - \frac{c_{1k}}{c_{kK}}.$$

If we take $(c_{1i}/c_{iK})$ for $i = 1, \ldots, K-1$ as a scale on the label space, then the relative spacing in the cost can be determined by the corresponding difference on the scale.

Letting $w_i$ and $s_i$ ($i = 1, \ldots, K$) denote weights and a scale on the label space, respectively, we can treat the cost scheme in Theorem 3 generally given in the form of $c_{ji} = w_i w_j (s_i - s_j) I(i > j)$. Application of Theorem 3 to the cost scheme yields the following corollary.

**Corollary 1.** *Suppose that* $c_{ji} = w_i w_j (s_i - s_j) I(i > j)$ *for some increasing* $\{s_j\}_{j=1}^K$ *and non-negative* $\{w_j\}_{j=1}^K$. *Let*

$$f_0^*(x) = \frac{\sum_{i=1}^K s_i w_i P(Y = i|x)}{\sum_{j=1}^K w_j P(Y = j|x)}.$$

*Then for any ranking function* $f$,

$$R_0(f_0^*; \mathbf{c}) \le R_0(f; \mathbf{c}).$$

Corollary 1 implies that optimal ranking under the cost scheme with equal weights preserves the ordering of the expected scale $s_Y$ given $x$. For example, when $s_j = j$ (a linear scale) and $w_j = 1$, $c_{ji} = (i - j) I(i > j)$ and hence $f_0^*(x) = E[Y|X = x]$, which is known as the "expected relevance" given $x$ (Li et al. 2007). Equivalently, the regression approach to ranking in Cossock and Zhang (2008) by minimization of squared loss $\sum_{i=1}^n (f(x_i) - y_i)^2$ leads to estimation of the expected relevance.

In addition, Corollary 1 indicates that the weights $w_j$ have the effect of adjusting the conditional probability of each category on optimal ranking. When there is discrepancy in the proportions of categories in training data and the target population, differential weights can be used for desired adjustment.

## 4 Relation to Ordinal Regression

The mathematical formulation of multipartite ranking through pairs of low and high labels in the previous section is conceptually simple and straightforward. However, it is general consensus that pairwise ranking requires more computational resources than itemwise ranking as the number of pairs of instances to consider increases in the order of $n^2$, where $n$ is the sample size. Instead, ordinal regression is commonly used in practice to analyze data with multiple ordinal categories as a way of itemwise ranking. Many ranking algorithms are formulated as ordinal regression, for instance, ordinal regression boosting (ORBoost) and support vector ordinal regression (SVOR).

In this section we consider the relationship between the optimal ranking functions in ordinal regression and the minimizers of pairwise ranking risk. We show that the solutions to some ordinal regression methods can be viewed as a special case of the optimal function in multipartite ranking.

Ordinal responses are often modeled as discretized outcomes of a continuous latent variable. In this case, modeling amounts to finding a real-valued function $f(x)$ associated with the tail probability of the latent variable given $x$ and thresholds $\{\theta_i\}_{i=1}^{K-1}$. Given $f$ and the thresholds, the estimated response of an instance $x$ is taken as $i$ if $\theta_{i-1} < f(x) \leq \theta_i$, where $\theta_0 = -\infty$ and $\theta_K = \infty$. Based on these facts, a typical form of loss in ordinal regression for $f$ with thresholds $\{\theta_i\}_{i=0}^{K}$ is given by

$$l(f, \{\theta_i\}_{i=0}^{K}\,; x, y) = l(f(x) - \theta_{y-1}) + l(\theta_y - f(x)),$$

and it is often expected that $\theta_1 \leq \ldots \leq \theta_{K-1}$.

For given loss function $l$, the risk of $f$ is represented as

$$E_{X,Y}[l(\theta_Y - f(X)) + l(f(X) - \theta_{Y-1})] = \sum_{j=1}^{k} E_X[p_j(X)(l(\theta_j - f(X)) + l(f(X) - \theta_{j-1}))],$$

where $p_j(x) = P(Y = j|X = x)$.

To find $f^*$ that minimizes the risk, consider the conditional risk given $X = x$. For fixed $x$,

$$\sum_{j=1}^{K} p_j(x)\{l(\theta_j - f(x)) + l(f(x) - \theta_{j-1})\} = p_1 l(f(x) + \infty) + p_K l(\infty - f(x)) + \sum_{j=1}^{K-1} p_j l(\theta_j - f(x)) + p_{j+1} l(f(x) - \theta_j).$$

If $|f(x)| < \infty$, the first and the second terms are $\lim_{x\to\infty} l(x)$, which does not depend on $f(x)$. Then it is necessary that $f^*(x) \equiv t^*$, the value of the optimal ranking function for fixed $\theta_1, \ldots, \theta_{K-1}$, satisfies

$$\frac{d}{dt}\left(\sum_{j=1}^{K-1} p_j l(\theta_j - t) + p_{j+1} l(t - \theta_j)\right)\Bigg|_{t=t^*} = \sum_{j=1}^{K-1} p_{j+1} l'(t^* - \theta_j) - \sum_{j=1}^{K-1} p_j l'(\theta_j - t^*) = 0. \qquad (6)$$

## 4.1 Ordinal Regression Boosting

Ordinal regression boosting in Lin and Li (2006) takes exponential loss, $l(s) = \exp(-s)$. The necessary condition in (6) with the loss leads to

$$f^*(x) = \frac{1}{2}\log\frac{\sum_{i=2}^{K} P(Y = i|x)\exp(\theta_{i-1}^*)}{\sum_{j=1}^{K-1} P(Y = j|x)\exp(-\theta_j^*)}. \qquad (7)$$

To see the relationship between ORBoost and multipartite ranking in Corollary 1, given $\theta_j^*$, let $s_j = [1 + \exp(-\theta_j^* - \theta_{j-1}^*)]^{-1}$ and $w_j = \exp(\theta_{j-1}^*)(1 + \exp(-\theta_j^* - \theta_{j-1}^*))$ with $\theta_0^* = -\infty$ and $\theta_K^* = \infty$. Then $c_{ji} = \exp(\theta_{i-1}^* - \theta_j^*) - \exp(\theta_{j-1}^* - \theta_i^*)$ for $i > j$. The ordering restriction on the thresholds ($\theta_i^* \geq \theta_j^*$) yields non-negative ranking costs ($c_{ji} \geq 0$).

From the following relation

$$\frac{\sum_{i=1}^{K} s_i w_i P(Y = i|x)}{\sum_{j=1}^{K} w_j P(Y = j|x)} = \left(\frac{\sum_{j=1}^{K-1} \exp(-\theta_j^*) P(Y = j|x)}{\sum_{i=2}^{K} \exp(\theta_{i-1}^*) P(Y = i|x)} + 1\right)^{-1},$$

we see that the optimal ranking function $f^*$ in (7) is a monotonic transformation of the function $f_0^*$ in Corollary 1, and hence the optimal ranking by ORBoost provides a solution to pairwise ranking risk.

It is hard to specify thresholds analytically. They should be specified numerically in practice. At least it is guaranteed that there exist $\{\theta_i^*\}_{i=1}^{K-1}$ such that the ranking function in (7) is well-defined.

### 4.2 Proportional Odds Model

One of classical ordinal regression methods in statistics is the proportional odds model. Take the "nonparametric" version of cumulative logits model proposed by McCullagh (1980), which is defined as

$$\log \frac{P(Y > j|x)}{P(Y \le j|x)} = f(x) - \theta_j,$$

where $-\infty = \theta_0 < \theta_1 \le \ldots \le \theta_{K-1} < \theta_K = \infty$. The log likelihood of $\{y_i\}_{i=1}^n$ given $\{x_i\}_{i=1}^n$ under the model is

$$\begin{aligned}
&\sum_{i=1}^n \log\left(\frac{1}{1+\exp(\theta_{y_i-1} - f(x_i))} - \frac{1}{1+\exp(\theta_{y_i} - f(x_i))}\right)\\
&= \sum_{i=1}^n \log(1-\exp(\theta_{y_i-1} - \theta_{y_i})) - \left[\log\left(1+\exp(-f(x_i)+\theta_{y_i-1})\right) + \log\left(1+\exp(-\theta_{y_i}+f(x_i))\right)\right].
\end{aligned}$$

Hence, given $\{\theta_i\}_{i=0}^K$, maximizing the log likelihood amounts to ordinal regression with $l(s) = \log(1+e^{-s})$; see Rennie (2006) for further discussions. By considering the population version of the likelihood above, we arrive at the following equation for optimal ranking $f^*$ :

$$\frac{d}{dt}\left(\sum_{j=1}^{K-1} p_{j+1}(x)\log\left(1+\exp(\theta_j - t)\right) + p_j(x)\log\left(1+\exp(t-\theta_j)\right)\right)\Bigg|_{t=t^*} = 0,$$

where $f^*(x) \equiv t^*$. This is equivalent to the condition:

$$\sum_{j=1}^{K-1} \frac{p_j(x) + p_{j+1}(x)}{1+\exp(\theta_j - t^*)} = 1 - p_1(x).$$

The equation above is valid for general $K$, but an explicit solution may not be obtainable. When $K = 3$, $1 - p_1(x) = p_2(x) + p_3(x)$. Thus, the equation above is equivalent to

$$(1-p_3(x))(1+\exp(f^*(x)-\theta_2)) = (1-p_1(x))(1+\exp(\theta_1 - f^*(x))),$$

and since $e^{f^*(x)} > 0$, it has an explicit solution given by

$$e^{f^*(x)} = \frac{q(x) - 1 + \sqrt{(q(x)-1)^2 + 4e^{\theta_1-\theta_2}q(x)}}{2e^{-\theta_2}}, \tag{8}$$

where

$$q(x) = \frac{p_2(x) + p_3(x)}{p_1(x) + p_2(x)}.$$

This implies that $f^*$ for the proportional odds model preserves the ordering of $q$. Note that $q$ is the ranking function minimizing pairwise ranking risk when $c_{12} = c_{13} = c_{23} = 1$. Hence the proportional odds model corresponds to pairwise ranking risk minimization with a special cost scheme. The implicit ranking cost scheme associated with the proportional odds model, however, may not be deemed desirable.

### 4.3 Support Vector Ordinal Regression

Support vector ordinal regression (SVOR) is an ordinal regression method using the "large margin principle", which stems from Support Vector Machine for classification; see Herbrich et al. (2000), Shashua and Levin (2003) and Chu and Keerthi (2007). It takes the hinge loss $l(s) = (1-s)_+$. We look at two versions of SVOR.

The first version in Shashua and Levin (2003) aims to minimize the empirical ranking risk under

$$l(f, \{\theta_j\}_{j=0}^K; x, y) = (1 - (f(x) - \theta_{y-1}))_+ + (1 - (\theta_y - f(x)))_+.$$

Note that this loss function does not always guarantee the monotonicity of $\{\theta_j\}$. Chu and Keerthi (2007) showed a numerical experiment where the thresholds are not properly ordered.

To discuss some special characteristics of the theoretically optimal ranking function under hinge loss, we focus on $K = 3$. The conditional risk given an instance $x$ is

$$p_1(x)l(\theta_1 - f(x)) + p_2(x)[l(f(x) - \theta_1) + l(\theta_2 - f(x))] + p_3(x)l(f(x) - \theta_2).$$

Piecewise linearity of the risk makes identification of $f^*$ in this case somewhat tedious and enumerative.

With abbreviation of $p_i(x)$ as $p_i$, the conditional risk is represented as follows:

Case 1: $\theta_1 \le \theta_2$ and $\theta_1 + 1 \ge \theta_2 - 1$,

$$\begin{cases} p_2(1+\theta_1) + p_3(1+\theta_2) - (p_2+p_3)f(x) & \text{if} \quad f(x) \le \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(1+\theta_1) + p_3(1+\theta_2) + (p_1-p_2-p_3)f(x) & \text{if} \quad \theta_1 - 1 \le f(x) \le \theta_2 - 1. \\ p_1(1-\theta_1) + p_2(2+\theta_1-\theta_2) + p_3(1+\theta_2) + (p_1-p_3)f(x) & \text{if} \quad \theta_2 - 1 \le f(x) \le \theta_1 + 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + p_3(1+\theta_2) + (p_1+p_2-p_3)f(x) & \text{if} \quad \theta_1 + 1 \le f(x) \le \theta_2 + 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + (p_1+p_2)f(x) & \text{if} \quad \theta_2 + 1 \le f(x). \end{cases}$$

Case 2: $\theta_1 \le \theta_2$ and $\theta_1 + 1 \le \theta_2 - 1$,

$$\begin{cases} p_2(1+\theta_1) + p_3(1+\theta_2) - (p_2+p_3)f(x) & \text{if} \quad f(x) \le \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(1+\theta_1) + p_3(1+\theta_2) + (p_1-p_2-p_3)f(x) & \text{if} \quad \theta_1 - 1 \le f(x) \le \theta_1 + 1. \\ p_1(1-\theta_1) + p_3(1+\theta_2) + (p_1-p_3)f(x) & \text{if} \quad \theta_1 + 1 \le f(x) \le \theta_2 - 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + p_3(1+\theta_2) + (p_1+p_2-p_3)f(x) & \text{if} \quad \theta_2 - 1 \le f(x) \le \theta_2 + 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + (p_1+p_2)f(x) & \text{if} \quad \theta_2 + 1 \le f(x). \end{cases}$$

Hence the minimum is achieved when $f(x)$ is either $\theta_1 - 1$, $\theta_2 - 1$, $\theta_1 + 1$ or $\theta_2 + 1$ according to the sign of $p_1 - p_2 - p_3$, $p_1 - p_3$ and $p_1 + p_2 - p_3$ in both Cases 1 and 2.

Case 3: $\theta_2 \le \theta_1$ and $\theta_2 + 1 \ge \theta_1 - 1$,

$$\begin{cases} p_2(1+\theta_1) + p_3(1+\theta_2) - (p_2+p_3)f(x) & \text{if} \quad f(x) \le \theta_2 - 1. \\ p_2(2+\theta_1-\theta_2) + p_3(1+\theta_2) - p_3 f(x) & \text{if} \quad \theta_2 - 1 \le f(x) \le \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(2+\theta_1-\theta_2) + p_3(1+\theta_2) + (p_1-p_3)f(x) & \text{if} \quad \theta_1 - 1 \le f(x) \le \theta_2 + 1. \\ p_1(1-\theta_1) + p_2(2+\theta_1-\theta_2) + p_1 f(x) & \text{if} \quad \theta_2 + 1 \le f(x) \le \theta_1 + 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + (p_1+p_2)f(x) & \text{if} \quad \theta_1 + 1 \le f(x). \end{cases}$$

In this case the minimum is attained when $f(x)$ is either $\theta_1 - 1$ or $\theta_2 + 1$ depending on the sign of $p_1 - p_3$.

Case 4: $\theta_2 \le \theta_1$ and $\theta_2 + 1 \le \theta_1 - 1$,

$$\begin{cases} p_2(1+\theta_1) + p_3(1+\theta_2) - (p_2+p_3)f(x) & \text{if} \quad f(x) \le \theta_2 - 1. \\ p_2(2+\theta_1-\theta_2) + p_3(1+\theta_2) - p_3 f(x) & \text{if} \quad \theta_2 - 1 \le f(x) \le \theta_2 + 1. \\ p_2(2+\theta_1-\theta_2) & \text{if} \quad \theta_2 + 1 \le f(x) \le \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(2+\theta_1-\theta_2) + p_1 f(x) & \text{if} \quad \theta_1 - 1 \le f(x) \le \theta_1 + 1. \\ p_1(1-\theta_1) + p_2(1-\theta_2) + (p_1+p_2)f(x) & \text{if} \quad \theta_1 + 1 \le f(x). \end{cases}$$

Hence the minimum risk is achieved as long as $f(x)$ is between $\theta_2 + 1$ and $\theta_1 - 1$.

Define $r(x) \equiv (1 - 2p_1(x))/p_2(x)$, thresholding of which is very conducive to expressing the signs of the coefficients of linear terms in the conditional risk. Since $p_1 + p_2 + p_3 = 1$, the following equivalence holds:

$$\begin{aligned} p_1 - p_2 - p_3 = 2p_1 - 1 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless 0, \\ p_1 - p_3 = 2p_1 + p_2 - 1 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless 1, \\ p_1 + p_2 - p_3 = 2p_1 + 2p_2 - 1 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless 2. \end{aligned}$$

Then for instance, when $\theta_1 - \theta_2 \le 0$ (Cases 1 and 2), there is the following correspondence between $f^*$ and $r$:

| $r(x)$ | $(-\infty, 0)$ | $(0, 1)$ | $(1, 2)$ | $(2, \infty)$ |
|---|---|---|---|---|
| $f^*(x)$ | $\theta_1 - 1$ | $\min(\theta_1 + 1, \theta_2 - 1)$ | $\max(\theta_1 + 1, \theta_2 - 1)$ | $\theta_2 + 1$ |

Similarly, when $0 \le \theta_1 - \theta_2 \le 2$ (Case 3), we observe the following relation:

| $r(x)$ | $(-\infty, 1)$ | $(1, \infty)$ |
|---|---|---|
| $f^*(x)$ | $\theta_1 - 1$ | $\theta_2 + 1$ |

Thus the optimal ranking function $f^*$ is a step function of $r$, which takes a very different form than the minimizers of multipartite ranking in our framework as in Theorem 3.

On the other hand, we can see that when $\theta_1 \le \theta_2$, $r$ corresponds to the median of $Y$ given $x$ from the relation:

| $r(x)$ | $(-\infty, 0)$ | $(0, 2)$ | $(2, \infty)$ |
|---|---|---|---|
| Probability | $p_1(x) \ge \frac{1}{2}$ | $p_1(x) < \frac{1}{2} \le p_1(x) + p_2(x)$ | $p_1(x) + p_2(x) < \frac{1}{2}$ |
| Median$(Y \mid X = x)$ | 1 | 2 | 3 |

The second version of SVOR proposed by Chu and Keerthi (2007) is called implicit constraints method. The new loss function as a modified version of Shashua and Levin (2003)'s is defined as

$$l(f, \{\theta_j\}_{j=0}^{K}; x, y) = \sum_{j=1}^{y-1} (1 - (f(x) - \theta_j))_+ + \sum_{j=y}^{K-1} (1 - (\theta_j - f(x)))_+.$$

The loss function does not have the "typical" form of ordinal regression, but it is shown that the monotonicity of $\{\theta_j\}$ is always satisfied. For fixed instance $x$, the conditional risk given $x$ is

$$p_1[l(\theta_1 - f(x)) + l(\theta_2 - f(x))] + p_2[l(f(x) - \theta_1) + l(\theta_2 - f(x))] + p_3[l(f(x) - \theta_1) + l(f(x) - \theta_2)].$$

Again, due to piecewise linearity of the risk, case-by-case consideration is necessary to find $f^*(x)$. The conditional risk is represented as follows:

Case 1: $\theta_1 + 1 \geq \theta_2 - 1$

$$\begin{cases} p_2(1+\theta_1) + p_3(2+\theta_1+\theta_2) - (p_2+2p_3)f(x) & \text{if} \quad f(x) \leq \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(1+\theta_1) & \\ \qquad +p_3(2+\theta_1+\theta_2) + (p_1-p_2-2p_3)f(x) & \text{if} \quad \theta_1 - 1 \leq f(x) \leq \theta_2 - 1. \\ p_1(2-\theta_1-\theta_2) + p_2(2+\theta_1-\theta_2) & \\ \qquad +p_3(2+\theta_1+\theta_2) + (2p_1-2p_3)f(x) & \text{if} \quad \theta_2 - 1 \leq f(x) \leq \theta_1 + 1. \\ p_1(2-\theta_1-\theta_2) + p_2(1-\theta_2) & \\ \qquad +p_3(1+\theta_2) + (2p_1+p_2-p_3)f(x) & \text{if} \quad \theta_1 + 1 \leq f(x) \leq \theta_2 + 1. \\ p_1(2-\theta_1-\theta_2) + p_2(1-\theta_2) + (2p_1+p_2)f(x) & \text{if} \quad \theta_2 + 1 \leq f(x). \end{cases}$$

Hence the minimum is achieved when $f(x)$ is either $\theta_1 - 1$, $\theta_2 - 1$, $\theta_1 + 1$ or $\theta_2 + 1$ depending on the sign of $p_1 - p_2 - 2p_3$, $2p_1 - 2p_3$ and $2p_1 + p_2 - p_3$.

Case 2: $\theta_1 + 1 \leq \theta_2 - 1$

$$\begin{cases} p_2(1+\theta_1) + p_3(2+\theta_1+\theta_2) - (p_2+2p_3)f(x) & \text{if} \quad f(x) \leq \theta_1 - 1. \\ p_1(1-\theta_1) + p_2(1+\theta_1) & \\ \quad +p_3(2+\theta_1+\theta_2) + (p_1-p_2-2p_3)f(x) & \text{if} \quad \theta_1 - 1 \leq f(x) \leq \theta_1 + 1. \\ p_1(1-\theta_1) + p_3(1+\theta_2) + (p_1-p_3)f(x) & \text{if} \quad \theta_1 + 1 \leq f(x) \leq \theta_2 - 1. \\ p_1(2-\theta_1-\theta_2) + p_2(1-\theta_2) & \\ \quad +p_3(1+\theta_2) + (2p_1+p_2-p_3)f(x) & \text{if} \quad \theta_2 - 1 \leq f(x) \leq \theta_2 + 1. \\ p_1(2-\theta_1-\theta_2) + p_2(1-\theta_2) + (2p_1+p_2)f(x) & \text{if} \quad \theta_2 + 1 \leq f(x). \end{cases}$$

In this case the minimum risk is achieved when $f(x)$ is either $\theta_1 - 1$, $\theta_1 + 1$, $\theta_2 - 1$ or $\theta_2 + 1$ according to the sign of $p_1 - p_2 - 2p_3$, $p_1 - p_3$ and $2p_1 + p_2 - p_3$.

Consider $r(x) = \{1 - p_1(x)\}/\{1 - p_3(x)\}$ for succinct expressions of the signs of linear coefficients in the conditional risk of the second version:

$$\begin{aligned} p_1 - p_2 - 2p_3 = 2p_1 - p_3 - 1 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless \frac{1}{2}, \\ p_1 - p_3 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless 1, \\ 2p_1 + p_2 - p_3 = 1 + p_1 - 2p_3 \lesseqgtr 0 &\Leftrightarrow r(x) \gtreqless 2. \end{aligned}$$

Using the equivalence, we can express the relation between the optimal ranking function $f^*$ and $r$ as follows:

| $r(x)$ | $(0, \frac{1}{2})$ | $(\frac{1}{2}, 1)$ | $(1, 2)$ | $(2, \infty)$ |
|---|---|---|---|---|
| $f^*(x)$ | $\theta_1 - 1$ | $\min(\theta_1 + 1, \theta_2 - 1)$ | $\max(\theta_1 + 1, \theta_2 - 1)$ | $\theta_2 + 1$ |

Hence the optimal ranking function $f^*$ is a step function that is monotonically increasing in

$$\frac{1 - p_1}{1 - p_3} = \frac{p_2 + p_3}{p_1 + p_2}.$$

Asymptotically, implicit constraints method and proportional odds model are based on the same function $(p_2+p_3)/(p_1+p_2)$ for their ranking. However, there is a critical difference between them. The ranking function from implicit constraints method is a step function that produces many ties among instances. This could increase the risk of misranking and cause poor performance when used in applications. On the other hand, the ranking function from proportional odds model is a strictly increasing function of $(p_2+p_3)/(p_1+p_2)$.

### 4.4 Relation to Multicategory Classification

Finally we briefly discuss the case where $l(x) = I(x < 0)$. The corresponding conditional risk is

$$\sum_{j=1}^{K-1} p_j I(\theta_j < f(x)) + \sum_{j=2}^{K} p_j I(f(x) < \theta_{j-1}).$$

If $\theta_{i-1} < f(x) < \theta_i$, the conditional risk above is $\sum_{j=1}^{i-1} p_j + \sum_{j=i+1}^{K} p_j = 1 - p_i$. Hence $f^*(x)$ can be any value in $(\theta_{j^*-1}, \theta_{j^*})$ for $j^* = \text{argmax}_j\ p_j(x)$, and so the case is essentially the same as multiclass classification. Dembczyński et al. (2008) pointed out the same result.

When $K = 3$, there is no $\{c_{ji}\}$ such that $f_0^*(x) = \dfrac{c_{12}p_2(x) + c_{13}p_3(x)}{c_{13}p_1(x) + c_{23}p_2(x)}$ is a monotonic transformation of $\text{argmax}_j\ p_j(x)$. Thus, there is essential difference between the use of an indicator function and a convex loss function in ordinal regression unlike bipartite ranking and binary classification.

## 5 Convex Risk Minimization and Other Extensions

### 5.1 Convex Risk Minimization for Multipartite Ranking

As in bipartite ranking, one may consider convex risk minimization by replacing the misranking loss with a convex loss. This approach derives methods different from ordinal regression for multipartite ranking. As expected from multicategory classification (Lee et al. 2004, Zhang 2004, Tewari and Bartlett 2007), proper extension of convex risk minimization from bipartite to multipartite ranking may not be straightforward, and some care has to be taken to ensure ranking consistency of such extensions.

To examine the issue for the simple extension of replacing the indicator in the misranking loss with a convex loss $l$, define a convex loss function for ranking function $f$ as follows:

$$L(f; (x, y), (x', y')) = c_{y'y} l(f(x) - f(x')).$$

The corresponding risk is

$$\begin{aligned} R_l(f) &= E[L(f; (X, Y), (X', Y'))|Y > Y'] \\ &= \sum_{1 \le j < i \le K} c_{ji} E[l(f(X) - f(X'))|Y = i, Y' = j, Y > Y'] P(Y = i, Y' = j|Y > Y') \\ &= \sum_{1 \le j < i \le K} c_{ji} E[l(f(X) - f(X'))|Y = i, Y' = j] P(Y = i, Y' = j)/P(Y > Y'). \end{aligned}$$

It turns out that minimizing the convex risk above alone does not lead to the best ranking function $f_0^*$ with the minimum pairwise ranking risk characterized in Theorem 3. Duchi et al. (2010) also proved inconsistency of convex risk minimization in more general setting of label ranking with graph-based losses defined over preference graphs; see Lemma 10 in the paper.

The next theorem shows that some modification of the risk is needed to obtain a desirable ranking function that is consistent with $f_0^*$ in multipartite ranking. The optimal ranking function $f^*$ of the modified risk is easily derived as a simple application of the results in bipartite ranking.

**Theorem 4.** *Suppose that $l$ satisfies the condition for ranking calibration in Theorem 2.*

*(i) When $K = 3$, $f^*$ minimizing*

$$R_l(f) + \frac{c_{12}c_{23}}{c_{13}} E[l(f(X) - f(X'))|Y = 2, Y' = 2] \frac{P(Y = 2, Y' = 2)}{P(Y > Y')}$$

*is a monotonic transformation of*

$$f_0^*(x) = \frac{c_{12}P(Y = 2|x) + c_{13}P(Y = 3|x)}{c_{13}P(Y = 1|x) + c_{23}P(Y = 2|x)}.$$

*(ii) When $K > 3$ and $c_{1K}c_{ji} = c_{1i}c_{jK} - c_{1j}c_{iK}$ for all $1 \leq j < i \leq K$, $f^*$ minimizing*

$$\begin{aligned} R_l(f) &+ \sum_{1\leq j<i\leq K} \frac{c_{1j}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y = i, Y' = j] \frac{P(Y = i, Y' = j)}{P(Y > Y')} \\ &+ \sum_{i=2}^{K-1} \frac{c_{1i}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y = i, Y' = i] \frac{P(Y = i, Y' = i)}{P(Y > Y')} \\ &+ \sum_{1\leq j<i\leq K} \frac{c_{1j}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y = j, Y' = i] \frac{P(Y = j, Y' = i)}{P(Y > Y')} \end{aligned}$$

*is a monotonic transformation of*

$$f_0^*(x) = \frac{\sum_{i=2}^{K} c_{1i}P(Y = i|x)}{\sum_{j=1}^{K-1} c_{jK}P(Y = j|x)}.$$

The key point of Theorem 4 is that an extra term (e.g. $\frac{c_{12}c_{23}}{c_{13}} E[l(f(X) - f(X'))|Y = 2, Y' = 2]P(Y = 2, Y' = 2)/P(Y > Y')$ when $K = 3$) is necessary to guarantee consistency under ranking-calibrated loss $l$ in the pairwise framework. This is a major difference from the bipartite case. In practice, the extra term has to be estimated from the data, but it is straightforward to include the extra term in the empirical version of a given convex risk.

## 5.2 Extension to Non-Smooth Ranking Measures

As explained before, the ranking accuracy defined through the pairwise loss function we have discussed is an extension of AUC. Alternatively, non-smooth ranking measures such as average precision (AP) (Yue et al. 2007), and normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen 2000) are frequently used in multipartite ranking. They are for measuring the accuracy of top-ranked instances and essentially different from AUC. In this section, we try to find the connection between the pairwise ranking loss and non-smooth ranking measures. Especially we show that the optimization of non-smooth ranking measures can be cast in the framework of minimizing pairwise ranking risk.

First we investigate the general structure of non-smooth ranking measures. Given a ranking function $f$, we have a permutation based on the rankings of $x_i$ by $f$, $\pi_f : \{1, \ldots, n\} \mapsto \{1, \ldots, n\}$

Table 1: Ranking measures for the bipartite case

| | $G(y)$ | $D(i)$ |
|---|---|---|
| AUC | $I(y=2)$ | $n-i$ |
| $p-$Norm Push | $I(y=2)$ | $(n-i)^p$ |
| Precision at $m$ (P@$m$) | $I(y=2)$ | $I(i \leq m)/m$ |
| Mean Reciprocal Rank (MRR) | $I(y=2)$ | $1/i$ |

such that $f(x_{\pi_f(1)}) \geq f(x_{\pi_f(2)}) \geq \cdots \geq f(x_{\pi_f(n)})$. Given $\{x_i\}_{i=1}^n$, we can specify $\pi_f(i) \in \{1, \ldots, n\}$ for each $i$ and regard $\pi_f(i)$ as fixed. Non-Smooth ranking measures are generally of the following form:

$$\hat{R}_n^{G,D}(f) \equiv \sum_{i=1}^n G(y_{\pi_f(i)})D(i), \tag{9}$$

where $G : \mathcal{Y} \mapsto \mathbb{R}^+$ is a gain function and $D : \{1, \ldots, n\} \mapsto \mathbb{R}^+$ is a discount function decreasing in $i$. For example, Table 1 shows commonly used ranking measures for the bipartite case with corresponding gain and discount functions for each measure.

Given a gain function $G$ and a discount function $D$, consider finding $f$ that maximizes the expected ranking measure by examining the conditional measure

$$E(\hat{R}_n^{G,D}(f)|\{X_i\}_{i=1}^n) = E\left[\sum_{i=1}^n G(Y_{\pi_f(i)})D(i) \middle| \{X_i\}_{i=1}^n\right] = \sum_{i=1}^n C_f(i)D(i),$$

where $C_f(i) = E[G(Y_{\pi_f(i)})|\{X_i\}_{i=1}^n] = \sum_{j=1}^K G(j)P(Y_{\pi_f(i)} = j|\{X_i\}_{i=1}^n)$. Note that $C_f : \{1, \ldots, n\} \to \mathbb{R}^+$ and it depends on $f$. Since $D$ is non-negative and decreasing, to maximize the conditional measure as a weighted sum of $D(i)$'s, $f$ must be a function such that $C_f(1) \geq C_f(2) \geq \cdots \geq C_f(n)$.

Let $C(i) \equiv E[G(Y_i)|\{X_i\}_{i=1}^n]$, which is identical to $E[G(Y_i)|X_i]$ due to the independence assumption. As explained above, for each $i$, $\pi_f(i)$ is regarded as fixed given $\{X_i\}_{i=1}^n$. This means that $C_f(i) = E[G(Y_{\pi_f(i)})|\{X_i\}_{i=1}^n] \in \{C(1), \ldots, C(n)\}$. Hence ordering of $\{C_f(i)\}_{i=1}^n$ is equivalent to ordering of $\{C(i)\}_{i=1}^n$. Consequently, the optimal ranking function should be based on

$$f^*(x) = \sum_{j=1}^K G(j)P(Y = j|X = x).$$

Note that the discount component $D(i)$ does not affect $f^*$ at all as long as it is decreasing and non-negative. Similar results have been discussed in the literature, for instance, Theorem 1 in Cossock and Zhang (2008) for the special case of $G(y) = y$, Proposition 7 in Clémençon and Vayatis (2008) for the bipartite case only, and Corollary 1 in Duchi et al. (2012) for NDCG.

For example, the discounted cumulative gain (DCG) can be interpreted as a ranking measure with $G(y) = 2^y - 1$ and $D(i) = 1/\log(1+i)$. Application of the result above indicates that the optimal ranking function for DCG preserves the ordering of $\sum_{j=1}^K (2^j - 1)P(Y = j|X = x)$, which is the same as $f_0^*$ in Corollary 1 with $s_i = 2^i - 1$ and $w_j = 1$. Hence, the optimal ranking with respect to DCG is equivalent to pairwise ranking with $c_{ji} = (2^i - 2^j)I(i > j)$. Li et al. (2007) also considered maximization of DCG by relating the DCG measure to multiple classification error, and derived an empirical version of $f^*$ as a ranking function.

Other than the ranking measures in Table 1 for the bipartite case, the average precision (AP) is commonly used, which is given by

$$\mathrm{AP}(f) = \sum_{i=1}^{n} I(y_{\pi_f(i)} = 2)\frac{1}{i}\sum_{j=1}^{i} I(y_{\pi_f(j)} = 2) = \sum_{i=1}^{n} \frac{1}{i} I(y_{\pi_f(i)} = 2)\left(1 + \sum_{j=1}^{i-1} I(y_{\pi_f(j)} = 2)\right).$$

Notice that AP does not depend only on $y_{\pi_f(i)}$ but also on $y_{\pi_f(1)}, \ldots, y_{\pi_f(i-1)}$ for the $i$th rated item, so the representation in (9) is not available. However, it can be shown that the optimal ranking for AP also depends on $P(Y = 2|X = x)$ as with other measures in Table 1.

To this end, examine the conditional expectation of AP:

$$\begin{aligned} E[\mathrm{AP}(f)|\{X_i\}_{i=1}^n] &= \sum_{i=1}^{n} \frac{1}{i} P(Y_{\pi_f(i)} = 2|\{X_i\}_{i=1}^n)\left(1 + \sum_{j=1}^{i-1} P(Y_{\pi_f(j)} = 2|\{X_i\}_{i=1}^n)\right) \\ &= \sum_{i=1}^{n} \frac{p_f(i)}{i}\left(1 + \sum_{j=1}^{i-1} p_f(j)\right), \end{aligned}$$

where $p_f(i) \equiv P(Y_{\pi_f(i)} = 2|\{X_i\}_{i=1}^n)$. Note that $p_f(i) \in \{P(Y_i = 2|\{X_i\}_{i=1}^n)\}_{i=1}^n = \{P(Y_i = 2|X_i)\}_{i=1}^n$. For $k = 2, \ldots, n$, fix $p_f(k+1), \ldots, p_f(n)$ and consider the conditional expectation of AP as a function of $p_f(1), \ldots, p_f(k)$, simply denoted by $\mathrm{AP}_k(p_1, \ldots, p_k)$ with $(p_1, \ldots, p_k) \equiv (p_f(1), \ldots, p_f(k))$. To find the optimal ordering of $p_i$ for maximal $\mathrm{AP}_k$, take the following $k$ permutations of $(p_1, \ldots, p_k)$: for $l = 1, \ldots, k$, let $\mathbf{p}_l = (p_1, \ldots, p_{l-1}, p_k, p_l, \ldots, p_{k-1})$, where $\mathbf{p}_1 = (p_k, p_1, \ldots, p_{k-1})$ and $\mathbf{p}_k = (p_1, \ldots, p_{k-1}, p_k)$. Then for $l = 1, \ldots, k-1$,

$$\mathrm{AP}_k(\mathbf{p}_{l+1}) - \mathrm{AP}_k(\mathbf{p}_l) = (p_l - p_k)\left(1 + \sum_{i=1}^{l-1} p_i\right)\frac{1}{l(l+1)}.$$

Hence the necessary condition that $\mathrm{AP}_k(\mathbf{p}_k) \geq \mathrm{AP}_k(\mathbf{p}_{k-1}) \geq \cdots \geq \mathrm{AP}_k(\mathbf{p}_1)$ is $p_l \geq p_k$ for any $l = 1, \ldots, k-1$ and $k = 2, \ldots, n$. This implies that a ranking function $f$ with $p_f(1) \geq p_f(2) \geq \cdots \geq p_f(n)$ maximizes $E[\mathrm{AP}(f)|\{X_i\}_{i=1}^n]$, and thus the optimal ranking should be based on $P(Y = 2|X = x)$, which is the same conclusion as other ranking measures in bipartite ranking.

# 6 Numerical Illustration

## 6.1 Simulation Study

First we consider tripartite ranking to illustrate the results on ordinal regression. A balanced sample with three categories (500 observations in each category) was generated from a mixture of normal distributions: $X|Y = 1 \sim N(-2, 1)$, $X|Y = 2 \sim N(0, 1)$ and $X|Y = 3 \sim N(2, 1)$. For estimation of a ranking function, we applied four methods to the sample: pairwise ranking risk minimization, proportional odds model, ORBoost and SVOR with implicit constraints in Chu and Keerthi (2007).

For pairwise ranking risk minimization, exponential loss, $l(s) = \exp(-s)$, was employed as a convex surrogate loss with the cost scheme of $c_{12} = c_{23} = c_{13} = 1$. See Theorem 4 for theoretical justification of this choice for ranking consistency. In the training process of each of the first three methods, we adopted a gradient descent algorithm (boosting) with a weak learner. We set weak learners to be Gaussian kernel, $f_\theta(x) = \exp(-(x-\theta)^2/2\sigma^2)$, where the parameter $\theta$ was taken from the observed values $x_i$ and $\sigma^2 = 0.5$. At each iteration, a weak ranking function was chosen and added to the current ranking function with weight determined to minimize the respective empirical risk (pairwise ranking risk or risk under ordinal regression loss). We iterated the boosting process for 300 times to combine weak rankings and obtained the final ranking function. For validation of the number of iterations and testing, we generated another sample of the same size as the training data.

In ordinal regression, thresholds have to be estimated. When $K = 3$, there are two thresholds $\theta_1$ and $\theta_2$ ($\theta_1 \leq \theta_2$), but without loss of generality we can fix one of them. We fixed $\theta_2$ to be 0 and estimated $\theta_1 (\leq 0)$ only in our experiment. The estimated threshold in this study was $\hat{\theta}_1 = -2.06$ in ORBoost, and $\hat{\theta}_1 = -4.37$ in proportional odds model. For SVOR, Gaussian kernel with $\sigma^2 = 0.5$ was used and the tuning parameter $C$ was set to 10. The estimated thresholds in this case were $\hat{\theta}_1 = -1.2$ and $\hat{\theta}_2 = 1.1$, which fall into Case 2 in Section 4.3.

Theoretically, the optimal ranking function via pairwise ranking risk minimization and that derived from proportional odds model are monotonically related to

$$f_0^*(x) = \frac{P(Y = 2|X = x) + P(Y = 3|X = x)}{P(Y = 1|X = x) + P(Y = 2|X = x)} = \frac{e^{2x} + e^2}{e^{-2x} + e^2}.$$

So is the population minimizer of SVOR with implicit constraints in Chu and Keerthi (2007). In particular, from the result in Uematsu and Lee (2011), $f^*$ for pairwise ranking risk minimization under exponential loss is given by $\frac{1}{2} \log f_0^*$ up to a constant. On the other hand, $f^*$ for ORBoost is approximately given as

$$\frac{1}{2} \log \frac{e^{2x} + e^{2+\hat{\theta}_1}}{e^{-2x-\hat{\theta}_1} + e^2}$$

and that for proportional odds model is approximately

$$\log \frac{f_0^*(x) - 1 + \sqrt{(f_0^*(x) - 1)^2 + 4e^{\hat{\theta}_1} f_0^*(x)}}{2}.$$

Figure 1 shows the estimated ranking functions (solid) and their theoretical counterparts (dotted). They are centered to zero except for SVOR with implicit constraints. The estimated ranking function for implicit constraints method has steps at about $-1$, 0 and 1, which correspond to the theoretical values of $-1.02$, 0 and 1.02 that yield $f_0^*(x) = \frac{1}{2}, 1$ and 2, respectively. All the ranking functions are increasing in $x$ although there is some difference in the functional form. The dotted lines lie close to the solid lines over $[-3, 3]$, where the data density is relatively high. This implies that the theoretical results describe actual ranking scores very well. As discussed, SVOR with implicit constraints minimizes hinge risk and the optimal ranking function is a step function of $f_0^*$ theoretically. Clearly, the estimated ranking function by the method exhibits features of a step function as suggested by the theory. This feature could increase pairwise ranking risk compared with the other ranking functions. Similar discussions can be found in Uematsu and Lee (2011) for the bipartite case.

Comparison of the ranking function for ORBoost with those for pairwise ranking or proportional odds model in Figure 1 indicates that the difference in cost schemes ($c_{13} = e^{-\hat{\theta}_1} = e^{2.06} = 7.846$ versus $c_{13} = 1$) has little effect on the actual ordering of instances in this case.

To further examine the effect of ranking costs, let $c_{12} = c_{23} = 1$ and $c_{13} \geq 1$ in $K = 3$ case, and compare the contours of the underlying ranking function

$$\frac{P(Y = 2|x) + c_{13} P(Y = 3|x)}{c_{13} P(Y = 1|x) + P(Y = 2|x)} = \frac{p_2(x) + c_{13}(1 - p_1(x) - p_2(x))}{c_{13} \cdot p_1(x) + p_2(x)}$$

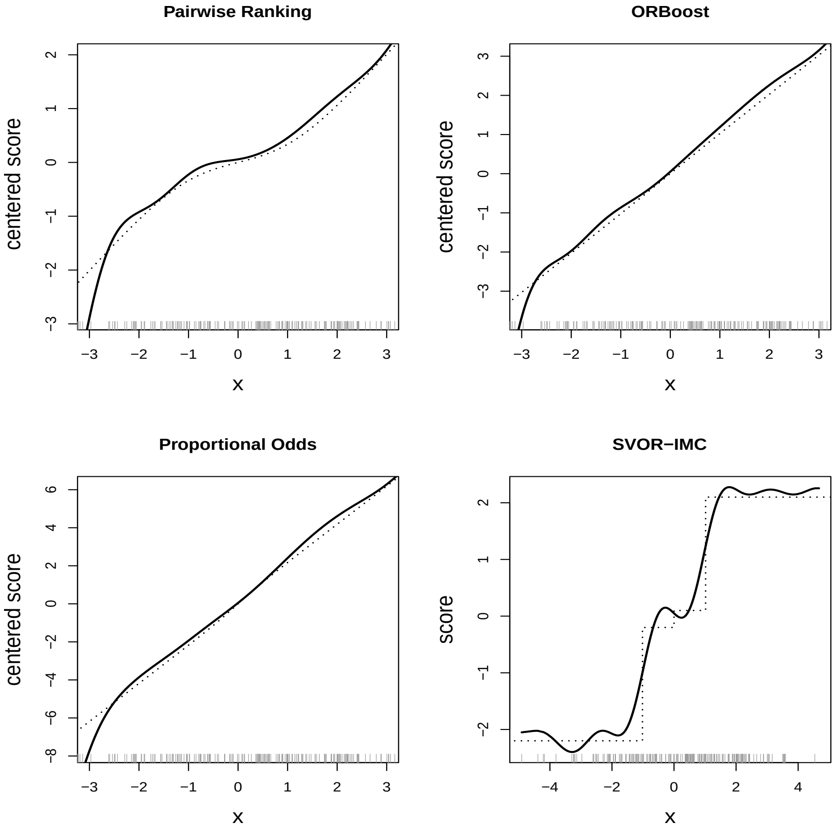

Figure 1: Theoretical ranking function (dotted line) and estimated ranking function (solid line) for pairwise ranking risk minimization with exponential loss, ORBoost, proportional odds model and SVOR with implicit constraints.

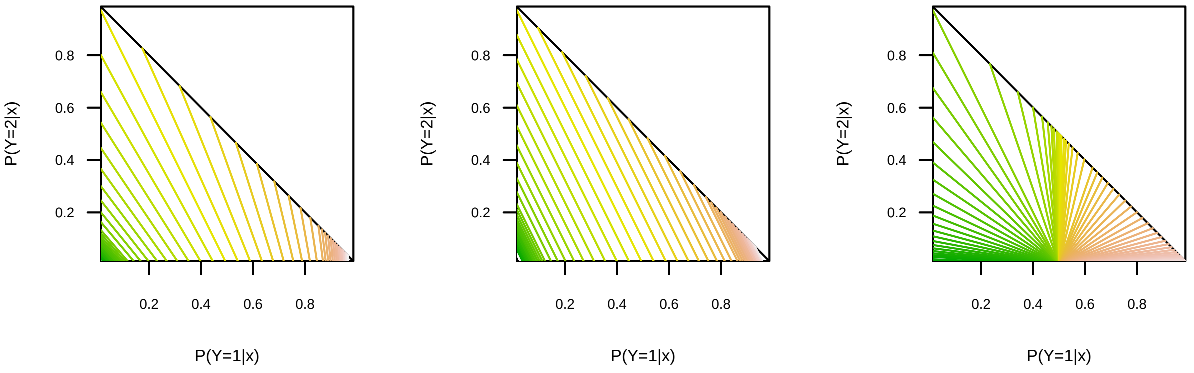

Figure 2: Contours of the underlying ranking functions, $\frac{p_2(x)+p_3(x)}{p_1(x)+p_2(x)}$ for proportional odds model and SVOR with implicit constraints (left), $\frac{p_2(x)+2p_3(x)}{2p_1(x)+p_2(x)}$ for expected relevance (middle) and $\frac{1-2p_1(x)}{p_2(x)}$ for Shashua and Levin's SVOR (right). The change in color from green to brown indicates the change in score from high to low.

over the space of $(p_1(x), p_2(x))$ for the simulation setting as $c_{13}$ varies. Figure 2 shows such contours when $c_{13} = 1$ in the left panel and when $c_{13} = 2$ in the middle panel where the equivalence sets form parallel lines. Depending on $c_{13}$ (either 1 or 2), two objects can be ranked differently. However, the sets in the left and the middle are fairly similar to each other, implying that it is less likely to observe substantial differences in ranking. By contrast, the right panel shows the contours of $\frac{1-2p_1(x)}{p_2(x)}$, the underlying function for support vector ordinal regression in Shashua and Levin (2003). Figure 2 suggests that the rankings derived from Shashua and Levin (2003)'s method may be quite different from the other two methods.

## 6.2 Application to Movie-Lens Data

We applied ordinal regression and ranking methods to Movie-Lens data (GroupLens-Research 2006) collected for building movie recommender systems. The main focus of our analysis is in the comparison of pairwise ranking and ordinal regression methods and investigation of the effect of the cost schemes implicitly assumed by those methods in real applications.

The data set consists of 100,000 ratings for 1,682 movies by 943 users. The ratings are on a scale of 1 to 5. In addition to the ratings, the data include content information about the movies such as release date and genres and demographic information about the users such as age, gender and occupation. We followed the preprocessing steps taken in Uematsu and Lee (2011) for exclusion of ratings with incomplete data and ratings from six unusual users. The remaining ratings are 97,139 in total, and there are 5,039, 10,846, 26,448, 33,842 and 20,964 cases for 1 to 5 ratings in the respective order. We standardized the predictors for ranking.

### Three Categories

There are five categories in the original data, but we transformed them into three categories first ("Low", "Middle" and "High") to compare numerical results with more specific analytical results in the tripartite case. “Low” corresponds to ratings of 1, 2 and 3, “Middle” to 4 and “High” to 5. We consider the following six ranking methods in this analysis:

- PAIRRANK1: pairwise ranking risk minimization with $(c_{12}, c_{13}, c_{23}) = (1, 1, 1)$
- PAIRRANK2: pairwise ranking risk minimization with $(c_{12}, c_{13}, c_{23}) = (1, 2, 1)$
- REG: regression (squared error minimization)
- PROPODDS: (nonparametric) proportional odds model
- ORBOOST: ordinal regression boosting
- SVOR-IMC: support vector ordinal regression with implicit constraints in Chu and Keerthi (2007)

Except for support vector ordinal regression, we implemented the first five ranking methods using a gradient descent algorithm with *univariate* Gaussian kernels defined for each predictor as weak learners as described in the simulation study. Boosting involved 500 iterations initially and the optimal number of iterations was determined later by validation data. For support vector ordinal regression, we used the algorithm in Chu and Keerthi (2007) with a *multivariate* Gaussian kernel. We set the bandwidth of the multivariate Gaussian kernel to $\sigma^2 = 44 \times 0.5 = 22$ with 44 standardized variables included in regression and set the tuning parameter $C$ to 500 for effective hinge risk minimization.

The theoretical results in the foregoing sections suggest particular functional relations among ranking functions from the methods in consideration. For example, the proof of Corollary 1 implies that a ranking score of instance $x$ from PAIRRANK2 is represented as

$$\frac{1}{2} \log \frac{\sum_{j=1}^{3} jP(Y = j|x) - 1}{3 - \sum_{j=1}^{3} jP(Y = j|x)} = \frac{1}{2} \log \frac{s_r(x) - 1}{3 - s_r(x)}$$

up to a constant, where $s_r(x) = E(Y|x)$ is the theoretical ranking score from REG. This identity shows the expected relation between scores from the two methods. Given thresholds $\theta_1$ and $\theta_2$, we can also specify the relation between ORBoost and pairwise ranking. As in the simulation, $\theta_2$ was fixed at zero, and only $\theta_1$ was estimated for ordinal regression in our experiment. This yields the cost scheme of $c_{12} = c_{23} = 1$ and $c_{13} = e^{-\theta_1}$ for ORBoost. Hence, ranking scores from ORBoost can be replicated with pairwise ranking theoretically by setting $c_{12} = c_{23} = 1$ and $c_{13} = e^{-\theta_1}$. Similarly, scores from proportional odds model can be replicated with PAIRRANK1 from the equation (8) and the fact that a ranking function of PAIRRANK1 converges to $\frac{1}{2} \log \frac{p_2(x)+p_3(x)}{p_1(x)+p_2(x)}$ up to a constant.

We empirically verify the relations using ranking scores from those methods. Figure 3 shows scatter plots of the ranking scores derived from the methods when they were applied to the MovieLens data. The scores are from a training sample of size 3,000 chosen at random while the proportions of the three categories in the sample are approximately kept at those in the full data. In each panel, ranking scores from a given method are plotted against those from pairwise ranking with the value of $c_{13}$ specified for their correspondence, and the solid line indicates the theoretical relation. Note that the solid lines are determined up to an additive constant in the horizontal direction and scores from pairwise ranking were shifted by matching the average ranking scores for a pair of methods in each panel. The estimated threshold $\theta_1$ for ORBoost and proportional odds model was $-0.183$ and $-1.615$, respectively, which amounts to $e^{-\theta_1} = 1.201$ as the corresponding ranking cost $c_{13}$ for ORBoost. Generally, the observed relations among the ranking scores show good agreement with their theoretical counterparts for ORBoost, regression, and proportional odds model in relation to pairwise ranking.

For support vector ordinal regression, the estimates of $\theta_1$ and $\theta_2$ are $-1.39$ and 0.61 and satisfy the relation $\theta_1 + 1 = \theta_2 - 1$, which implies three clusters of scores around $\theta_1 - 1$, $\theta_1 + 1 (= \theta_2 - 1)$ and $\theta_2 + 1$ or $(-2.39, -0.39, 1.61)$ theoretically. However, due to the smoothness of Gaussian kernel used as basis functions in SVOR and a limited range of pairwise ranking scores, the stepwise feature expected from the theory is hardly discernible in the scores from SVOR. Nevertheless, the plot for SVOR exhibits a noticeably different pattern from other methods as indicated by the theoretical analysis. In general, regularization and the type of a basis function (its smoothness or monotonicity) will affect the empirical relation between SVOR and pairwise ranking scores.

Using the same data, we further examine the effect of differential cost $c_{13}$ in tripartite ranking on pairwise ranking scores when $c_{12} = c_{23} = 1$. Figure 4 displays scatter plots of pairwise ranking scores when the cost $c_{13}$ varies from 1 to 10. Clearly, they exhibit a strongly positive linear relation and increasing $c_{13}$ has the effect of widening the range of scores with little change in the rankings.

### Five Categories

We consider analysis with the original five ratings. Since theoretical results available for the five category case are less specific than the three category case, we focus on comparisons of the methods in terms of ranking error.

The theoretical results in the previous sections indicate that the class-conditional probabilities $P(Y = k|x)$ are the key factors for optimal ranking, which depend on the distribution of instances within each label and the proportion of each label, $\pi_k = P(Y = k)$. To examine the impact of the proportions, we experimented with two samples of size 2,500 with different proportions. A balanced sample contains 500 observations chosen at random for each rating, and an unbalanced sample of the same size contains 130, 280, 680, 870, and 540 cases for 1 to 5 ratings, approximately reflecting the proportions (5.18%, 11.17%, 27.23%, 34.84%, 21.58%) in the original data. Random sampling was done 50 times, yielding two sets of 50 replicates. The same sampling process was used to generate two sets (balanced and unbalanced) of validation data of the same size as training and test data of size 5,000. Three ranking methods (pairwise ranking, proportional odds model, and ORBoost) were then applied to each of the training samples. The setting for training was the same as in the tripartite case. Univariate Gaussian kernels were used as weak learners and the number of iterations was 300 initially and determined by validation data later. The final ranking functions were evaluated over test data. The exponential loss $l(s) = e^{-s}$ was employed as a convex surrogate loss for pairwise ranking, and a linear cost scheme of $c_{ji} = i - j$ for $i > j$ was used as given below:

| $j\backslash i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 1 | 2 | 3 | 4 |
| 2 | | | 1 | 2 | 3 |
| 3 | | | | 1 | 2 |
| 4 | | | | | 1 |
| 5 | | | | | |

Cost schemes affect the expression of the optimal ranking function in multipartite ranking theoretically, and could change actual rankings of instances practically. For comparison, the misranking cost implied by ORBoost can be calculated using its relation to thresholds: $c_{ji} = \exp(\theta_{i-1} - \theta_j) - \exp(\theta_{j-1} - \theta_i)$ for $i > j$. In our experiment, $\theta_4$ was fixed at 0 without losing generality and $\theta_1$, $\theta_2$, and $\theta_3$ ($\theta_1 \leq \theta_2 \leq \theta_3$) were estimated. On average, the estimates for $\theta_1, \theta_2$ and $\theta_3$ were $-0.0932$ (0.0015), $-0.0656$ (0.0014), $-0.0386$ (0.0009) in the balanced case, and $-0.7742$

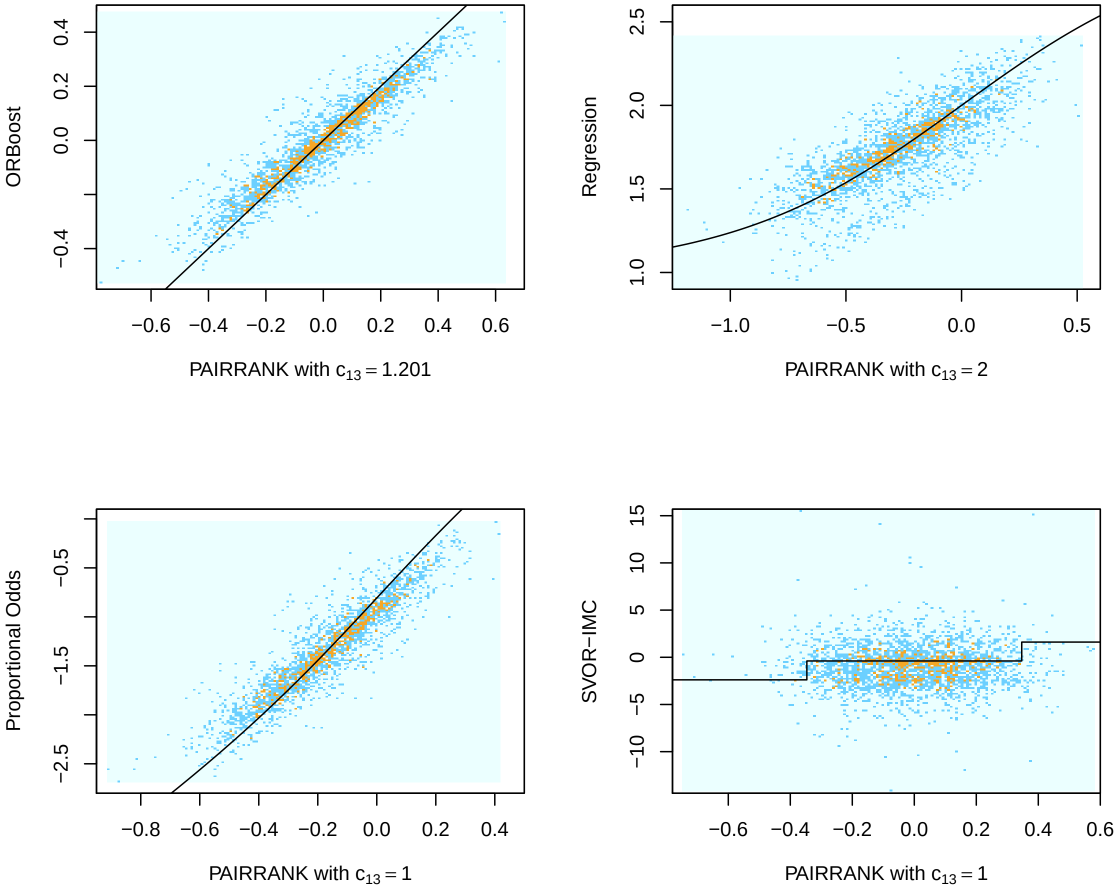

Figure 3: Scatter plots of ranking scores from ORBoost, regression, proportional odds model, and SVOR against pairwise ranking scores with matching cost $c_{13}$ for MovieLens data with three categories. The solid lines indicate theoretical relation between ranking scores.

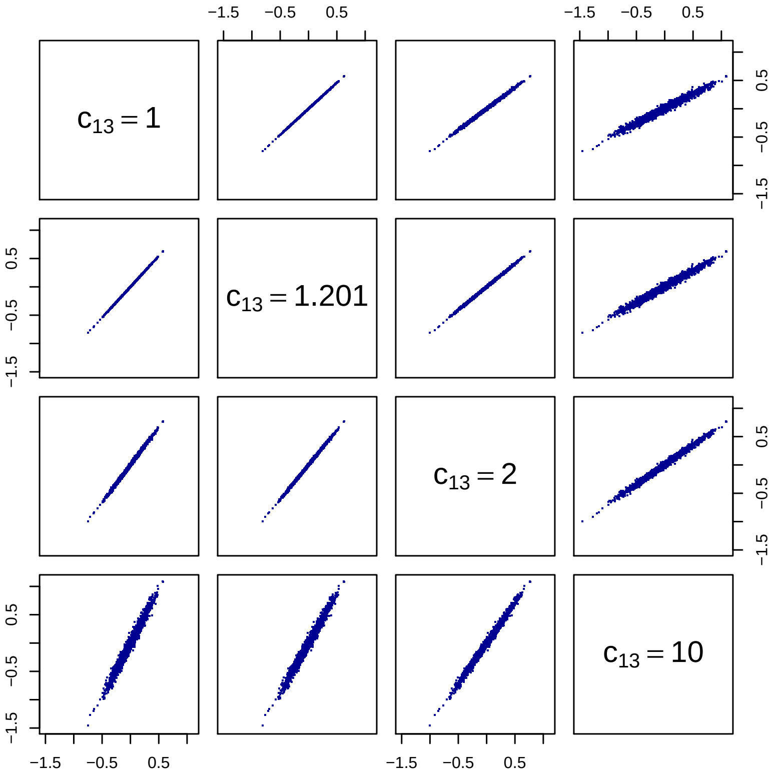

Figure 4: Scatter plots of pairwise ranking scores (centered to zero) with different ranking cost $c_{13}$ for MovieLens data when $c_{12} = c_{23} = 1$.

(0.0013), $-0.7742$ (0.0013), $-0.4183$ (0.0008) in the unbalanced case. The values in parentheses are their standard errors. It was observed that $\hat{\theta}_1 = \hat{\theta}_2$ for every replicate in the unbalanced case. Table 2 shows the implicit cost schemes corresponding to the average estimated thresholds in both cases. Compared to the linear scheme, ORBoost induces an exponential scale in general, which could penalize misranking of instances in nonadjacent labels increasingly more than adjacent ones, depending on the scale of thresholds. However, in this analysis, the estimated thresholds for ORBoost have a rather narrow range, which actually results in the opposite effect.

Table 2: Implicit cost schemes for ORBoost in the balanced (left) and unbalanced (right) cases

| $j\backslash i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 1 | 1.028 | 1.056 | 1.098 |
| 2 | | | 0.053 | 0.116 | 1.068 |
| 3 | | | | 0.063 | 1.039 |
| 4 | | | | | 1 |
| 5 | | | | | |

| $j\backslash i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 1 | 1 | 1.427 | 2.169 |
| 2 | | | 0.299 | 0.966 | 2.169 |
| 3 | | | | 0.539 | 1.519 |
| 4 | | | | | 1 |
| 5 | | | | | |

Table 3 gives the average ranking error rates evaluated over test sets along with their standard errors. Difference in the error rates of the three methods is not substantial. Comparison of pairwise ranking and ORBoost suggests that the difference in the cost scheme has little effect on the error rate. The difference between the balanced and unbalanced cases is relatively large for pairs of low ratings (e.g. 1 vs 2). The balanced case gave lower error rates than the unbalanced case except for pairs of high ratings. Since lower ratings have smaller proportions, the observed merit of balancing labels could be attributed to the increase in sample size for low ratings. In general, the pairwise approach requires relatively more data for stable results than the itemwise approach in ordinal regression. The numerical results show that for the ratings with smaller sample sizes, pairwise ranking is less effective than proportional odds model, but it gains relative efficiency as the sample size increases.

The alternative expression of $f_0^*$ in Theorem 3 in terms of the proportions $\pi_j$ and the distribution function of $X$ in each label gives a dual interpretation that the proportions of a training sample modify a given cost scheme. Large proportions of labels $i$ and $j$ have the effect of increasing the nominal cost of $c_{1i}$ and $c_{jK}$ as multiplicative factors. For example, in pairwise ranking, $c_{12} = c_{45} = 1$ and the proportions for ratings 2 and 4 in the unbalanced case are approximately 11.17% and 34.84%. Since $c_{12}\pi_2 = 0.1117 < 0.2$ while $c_{45}\pi_4 = 0.3484 > 0.2$, more ranking errors are expected for (1, 2) pair in the unbalanced case than the balanced case while the opposite is expected for (4, 5) pair. The numerical result in Table 3 for (1, 2) and (4, 5) pairs might be explained by this effect of imbalance in proportions apart from the sample size effect.

# 7 Conclusion

We have considered multipartite ranking as an extension of bipartite ranking by employing a ranking loss which combines pairwise ranking errors of ordinal categories with differential ranking costs. The extension shows that the optimal ranking function can be represented as a ratio of a weighted sum of conditional probability functions of upper categories to that of lower categories, and the optimal ranking derived from ORBoost, proportional odds model and a version of support vector ordinal regression is related to a ratio of the form. This result enhances our understanding of

Table 3: Mean ranking error rates for a pair of ratings and their standard errors in parentheses from 50 replicates of training samples of size 2,500. The smallest error rate for each pair is highlighted in bold.

| | Balanced | | | Unbalanced | | |
|---|---|---|---|---|---|---|
| Pair | PAIR | PROP | OR | PAIR | PROP | OR |
| | RANK | ODDS | BOOST | RANK | ODDS | BOOST |
| 1 vs 2 | 0.4485 | **0.4413** | 0.4431 | 0.4545 | **0.4529** | 0.4533 |
| | (0.0020) | (0.0019) | (0.0020) | (0.0025) | (0.0023) | (0.0023) |
| 1 vs 3 | 0.3875 | **0.3833** | 0.3853 | **0.4006** | 0.4014 | 0.4012 |
| | (0.0019) | (0.0018) | (0.0019) | (0.0030) | (0.0026) | (0.0028) |
| 1 vs 4 | 0.3193 | **0.3164** | 0.3195 | **0.3341** | 0.3346 | 0.3343 |
| | (0.0016) | (0.0016) | (0.0016) | (0.0029) | (0.0027) | (0.0027) |
| 1 vs 5 | **0.2607** | 0.2620 | 0.2637 | **0.2740** | 0.2768 | 0.2746 |
| | (0.0018) | (0.0015) | (0.0015) | (0.0028) | (0.0027) | (0.0027) |
| 2 vs 3 | **0.4368** | 0.4395 | 0.4399 | **0.4443** | 0.4472 | 0.4464 |
| | (0.0017) | (0.0018) | (0.0017) | (0.0023) | (0.0023) | (0.0024) |
| 2 vs 4 | **0.3652** | 0.3687 | 0.3707 | **0.3749** | 0.3780 | 0.3768 |
| | (0.0018) | (0.0018) | (0.0018) | (0.0024) | (0.0023) | (0.0024) |
| 2 vs 5 | **0.3023** | 0.3094 | 0.3103 | **0.3107** | 0.3167 | 0.3135 |
| | (0.0019) | (0.0019) | (0.0018) | (0.0021) | (0.0022) | (0.0022) |
| 3 vs 4 | **0.4266** | 0.4280 | 0.4295 | **0.4295** | 0.4300 | 0.4297 |
| | (0.0017) | (0.0018) | (0.0018) | (0.0016) | (0.0016) | (0.0015) |
| 3 vs 5 | **0.3601** | 0.3655 | 0.3658 | **0.3622** | 0.3665 | 0.3640 |
| | (0.0019) | (0.0019) | (0.0019) | (0.0016) | (0.0016) | (0.0015) |
| 4 vs 5 | **0.4313** | 0.4351 | 0.4340 | **0.4307** | 0.4345 | 0.4323 |
| | (0.0018) | (0.0018) | (0.0019) | (0.0015) | (0.0016) | (0.0016) |

commonly used ranking algorithms by pinning down the limiting target function for each ranking algorithm and further allows us to see their relations. In addition, it sheds light on multipartite ranking consistency and proper ranking calibration with convex losses for minimization of pairwise ranking risk.

The numerical results presented in this paper indicate that the effect of differential costs on ranking is practically little while sample size and proportions could have more substantial effect on the finite-sample performance of ranking methods. In particular, ordinal regression appears relatively more effective than the pairwise ranking approach in small samples. The situation seems reverse as the sample size gets large at the expense of increased computational load for the pairwise approach. Other factors also influence the properties of ranking functions estimated from a finite sample and in turn affect their performance. Basis functions, regularization, and selection of tuning parameters all affect the quality of estimated ranking functions different from the limiting target functions that we have theoretically characterized. For instance, a regularized ranking function from support vector ordinal regression with a linear kernel would considerably deviate from the undesirable stepwise function identified in Section 4.3.

The first-order analysis of the theoretical ranking functions of various algorithms for multipartite ranking in this paper shows that they (except for SVOR) have a monotone relation on the population level. However, the range of ranking scores differs, depending on the algorithm and the internal degree of difficulty of a given problem. To understand relative merits of ranking algorithms in terms of sampling variation and its effect on the ranking error rate, a second-order comparison will be called for. Further, extension of generalization bounds and regret bounds for bipartite ranking to the multipartite case will be another avenue for theoretical development.

# Appendix

## Proof of Theorem 3

For an arbitrary ranking function $f$ and the $f_0^*$ defined, consider the following partition of $\mathcal{X} \times \mathcal{X}$, which depends on $f$ and $f_0^*$:

| | $f_0^*(x) < f_0^*(x')$ | $f_0^*(x) = f_0^*(x')$ | $f_0^*(x) > f_0^*(x')$ |
|---|---|---|---|
| $f(x) < f(x')$ | $A_1$ | $A_2$ | $A_3$ |
| $f(x) = f(x')$ | $A_4$ | $A_5$ | $A_6$ |
| $f(x) > f(x')$ | $A_7$ | $A_8$ | $A_9$ |

That is, $\mathcal{X} \times \mathcal{X} = \cup_{j=1}^9 A_i$, where $A_1 = \{(x, x') | f(x) < f(x') \text{ and } f_0^*(x) < f_0^*(x')\}$, for instance.

(i) Let $g_i(x)$ be pdf or pmf of $X$ with label $i$. To express the expected misranking cost or the ranking risk, consider the following identity:

$$\begin{aligned} & \sum_{1 \le j < i \le 3} c_{ji} P(Y = i, Y' = j | Y > Y') g_i(x) g_j(x') \qquad (10) \\ = \; & -\frac{c_{12}c_{23}}{c_{13}} \frac{P(Y=2)P(Y'=2)}{P(Y > Y')} g_2(x) g_2(x') \\ & + \frac{1}{c_{13}P(Y > Y')} [c_{13} g_3(x) P(Y=3) + c_{12} g_2(x) P(Y=2)] \\ & \times [c_{13} g_1(x') P(Y'=1) + c_{23} g_2(x') P(Y'=2)] \\ \doteq \; & D(x, x') + C(x, x'). \end{aligned}$$

Then the ranking risks of $f$ and $f_0^*$ are given by

$$\begin{aligned} R_0(f) &= \iint_{A_1\cup A_2\cup A_3} [D(x,x') + C(x,x')]dxdx' + \frac{1}{2}\iint_{A_4\cup A_5\cup A_6} [D(x,x') + C(x,x')]dxdx'. \\ R_0(f_0^*) &= \iint_{A_1\cup A_4\cup A_7} [D(x,x') + C(x,x')]dxdx' + \frac{1}{2}\iint_{A_2\cup A_5\cup A_8} [D(x,x') + C(x,x')]dxdx'. \end{aligned}$$

The difference $R_0(f) - R_0(f_0^*)$ is then given by

$$\begin{aligned} &\iint_{A_3} [D(x,x') + C(x,x')]dxdx' - \iint_{A_7} [D(x,x') + C(x,x')]dxdx' \\ &\qquad + \frac{1}{2}\iint_{A_2\cup A_6} [D(x,x') + C(x,x')]dxdx' - \frac{1}{2}\iint_{A_4\cup A_8} [D(x,x') + C(x,x')]dxdx'. \end{aligned}$$

From $C(x,x')/C(x',x) = f_0^*(x)/f_0^*(x')$, $\iint_{A_3} C(x,x')dxdx' > \iint_{A_3} C(x',x)dxdx'$. By switching $x$ and $x'$, we can show that $\iint_{A_3} C(x',x)dxdx' = \iint_{A_7} C(x,x')dxdx'$. Hence, $\iint_{A_3} C(x,x')dxdx' > \iint_{A_7} C(x,x')dxdx'$. Similarly, we get $\iint_{A_6} C(x,x')dxdx' > \iint_{A_4} C(x,x')dxdx'$ and $\iint_{A_2} C(x,x')dxdx' = \iint_{A_8} C(x,x')dxdx'$.

Since $D(x,x') = D(x',x)$, $\iint_{A_3} D(x,x')dxdx' = \iint_{A_3} D(x',x)dxdx' = \iint_{A_7} D(x,x')dxdx'$. Likewise $\iint_{A_6} D(x,x')dxdx' = \iint_{A_4} D(x,x')dxdx'$ and $\iint_{A_2} D(x,x')dxdx' = \iint_{A_8} D(x,x')dxdx'$.

Hence, $R_0(f) - R_0(f_0^*) > 0$ unless $f$ is an order-preserving transformation of $f_0^*$.

(ii) Consider the following generalization of the identity in (i).

$$\begin{aligned} &\sum_{1\le j<i\le K} c_{ji} P(Y=i, Y'=j|Y>Y') g_i(x) g_j(x') \qquad (11) \\ &= -\sum_{i=2}^{K-1} \frac{c_{1i}c_{iK}}{c_{1K}} \frac{P(Y=i)P(Y'=i)}{P(Y>Y')} g_i(x)g_i(x') \\ &+ \frac{1}{c_{1K}P(Y>Y')}\left(\sum_{i=2}^{K} c_{1i}g_i(x)P(Y=i)\right)\left(\sum_{j=1}^{K-1} c_{jK}g_j(x')P(Y'=j)\right) \\ &+ \sum_{1\le j<i\le K} \frac{P(Y'=j)P(Y=i)}{P(Y>Y')}\left(\frac{c_{ji}c_{1K} - c_{1i}c_{jK}}{c_{1K}} g_i(x)g_j(x') - \frac{c_{1j}c_{iK}}{c_{1K}} g_i(x')g_j(x)\right) \\ &\doteq D(x,x') + C(x,x') + \sum_{1\le j<i\le K} \frac{P(Y'=j)P(Y=i)}{P(Y>Y')} R_{ji}(x,x'). \end{aligned}$$

Since $D(x,x') = D(x',x)$, $\iint_{A_3} D(x,x')dxdx' = \iint_{A_7} D(x,x')dxdx'$, $\iint_{A_6} D(x,x')dxdx' = \iint_{A_4} D(x,x')dxdx'$ and $\iint_{A_2} D(x,x')dxdx' = \iint_{A_8} D(x,x')dxdx'$.

Since $C(x,x')/C(x',x) = f_0^*(x)/f_0^*(x')$ by a similar argument as in (i), $\iint_{A_3} C(x,x')dxdx' > \iint_{A_7} C(x,x')dxdx'$, $\iint_{A_6} C(x,x')dxdx' > \iint_{A_4} C(x,x')dxdx'$ and $\iint_{A_2} C(x,x')dxdx' = \iint_{A_8} C(x,x')dxdx'$.

For $R_{ji}(x,x')$, the condition on the misranking costs implies that

$$\iint_{A_3} R_{ji}(x,x')dxdx' - \iint_{A_7} R_{ji}(x,x')dxdx'$$
$$= \frac{c_{ji}c_{1K} - c_{1i}c_{jK} + c_{1j}c_{iK}}{c_{1K}} \left( \iint_{A_3} g_i(x)g_j(x')dxdx' - \iint_{A_7} g_i(x)g_j(x')dxdx' \right) = 0.$$

Likewise $\iint_{A_6} R_{ji}(x,x')dxdx' = \iint_{A_4} R_{ji}(x,x')dxdx'$, and $\iint_{A_2} R_{ji}(x,x')dxdx' = \iint_{A_8} R_{ji}(x,x')dxdx'$. Hence, $R_0(f) - R_0(f_0^*) > 0$ unless $f$ is an order-preserving transformation of $f_0^*$.

□

## Proof of Corollary 1

It is easy to check that $c_{y'y} = c_{1K}(s_y - s_{y'})t_y t_{y'} I(y > y')$ satisfies the condition in Theorem 3. Application of Theorem 3 leads to

$$f_0^*(x) = \frac{\sum_{i=2}^{K} c_{1i}P(Y=i|x)}{\sum_{j=1}^{K-1} c_{jK}P(Y=j|x)} = \frac{t_1}{t_K} \frac{\sum_{i=2}^{K}(s_i - s_1)w_iP(Y=i|x)}{\sum_{j=1}^{K-1}(s_K - s_j)w_jP(Y=j|x)}.$$

Further note that

$$\frac{\sum_{i=2}^{K}(s_i - s_1)w_iP(Y=i|x)}{\sum_{j=1}^{K-1}(s_K - s_j)w_jP(Y=j|x)} = \frac{\frac{\sum_{i=1}^{K} s_iw_iP(Y=i|x)}{\sum_{j=1}^{K} w_jP(Y=j|x)} - s_1}{s_K - \frac{\sum_{i=1}^{K} s_iw_iP(Y=i|x)}{\sum_{j=1}^{K} w_jP(Y=j|x)}}.$$

It is clearly a monotonic transformation of $f_0^*(x)$, and thus it also minimizes the risk.

□

## Proof of Theorem 4

The ranking risk of $f$ under loss $l$ is given by

$$R_l(f) = \int_{\mathcal{X}}\int_{\mathcal{X}} l(f(x) - f(x')) \sum_{1 \le j < i \le K} c_{ji}P(Y=i, Y'=j|Y > Y')g_i(x)g_j(x')dxdx'.$$

(i) When $K = 3$, from the equation (10), we have

$$R_l(f) + \frac{c_{12}c_{23}}{c_{13}} E[l(f(X) - f(X'))|Y=2, Y'=2] \frac{P(Y=2, Y'=2)}{P(Y > Y')}$$
$$= \frac{[c_{13}P(Y=3) + c_{12}P(Y=2)][c_{13}P(Y'=1) + c_{23}P(Y'=2)]}{c_{13}P(Y > Y')} \int_{\mathcal{X}}\int_{\mathcal{X}} l(f(x) - f(x'))h_+(x)h_-(x')dxdx',$$

where

$$h_+(x) = \frac{c_{12}g_2(x)P(Y=2) + c_{13}g_3(x)P(Y=3)}{c_{12}P(Y=2) + c_{13}P(Y=3)}$$

and

$$h_-(x') = \frac{c_{13}g_1(x')P(Y'=1) + c_{23}g_2(x')P(Y'=2)}{c_{13}P(Y'=1) + c_{23}P(Y'=2)}.$$

Since $h_{\pm} \geq 0$ and $\int_{\mathcal{X}} h_{\pm}(x)dx = 1$, we can interpret $\int_{\mathcal{X}}\int_{\mathcal{X}} l(f(x) - f(x'))h_+(x)h_-(x')dxdx'$ as $E[l(f(X_*) - f(X'_*))]$, where $h_{\pm}$ are the pdf or pmf of $X_*$ and $X'_*$, respectively. By applying Theorem 2, we can see that the optimal ranking function $f^*$ minimizing the risk above is a monotonic transformation of the modified likelihood ratio

$$\frac{h_+(x)}{h_-(x)} \propto \frac{c_{12}g_2(x)P(Y=2) + c_{13}g_3(x)P(Y=3)}{c_{13}g_1(x)P(Y'=1) + c_{23}g_2(x)P(Y'=2)} = \frac{c_{12}P(Y=2|x) + c_{13}P(Y=3|x)}{c_{13}P(Y'=1|x) + c_{23}P(Y'=2|x)}.$$

(ii) As in (i), we use the equation (11) to show that

$$\begin{aligned} R_l(f) &+ \sum_{1\leq j<i\leq K} \frac{c_{1j}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y=i, Y'=j]\frac{P(Y=i, Y'=j)}{P(Y>Y')} \\ &+ \sum_{i=2}^{K-1} \frac{c_{1i}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y=i, Y'=i]\frac{P(Y=i, Y'=i)}{P(Y>Y')} \\ &+ \sum_{1\leq j<i\leq K} \frac{c_{1j}c_{iK}}{c_{1K}} E[l(f(X) - f(X'))|Y=j, Y'=i]\frac{P(Y=j, Y'=i)}{P(Y>Y')} \\ = &\quad \frac{\left[\sum_{i=2}^{K} c_{1i}P(Y=i)\right]\left[\sum_{j=1}^{K-1} c_{jK}P(Y'=j)\right]}{c_{1K}P(Y>Y')} \int_{\mathcal{X}}\int_{\mathcal{X}} l(f(x) - f(x'))h_+(x)h_-(x')dxdx', \end{aligned}$$

where

$$h_+(x) = \frac{\sum_{i=2}^{K} c_{1i}g_i(x)P(Y=i)}{\sum_{i=2}^{K} c_{1i}P(Y=i)} \quad \text{and} \quad h_-(x') = \frac{\sum_{j=1}^{K-1} c_{jK}g_j(x')P(Y'=j)}{\sum_{j=1}^{K-1} c_{jK}P(Y'=j)}.$$

The same result as in (i) applies to $f^*$ for general $K$.

□

## References

Agarwal, S. (2013). Surrogate regret bounds for bipartite ranking via strongly proper losses, *in* S. Shalev-Shwartz and I. Steinwart (eds), *Journal of Machine Learning Research: Workshop and Conference Proceedings*, Vol. 30, pp. 1–16.

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.

Agarwal, S. and Niyogi, P. (2009). Generalization Bounds for Ranking Algorithms via Algorithmic Stability, *Journal of Machine Learning Research* **10**: 441–474.

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

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*.

Chen, W., Liu, T.-Y., Lan, Y., Ma, Z. and Li, H. (2009). Essential loss: Bridge the gap between ranking measures and loss functions in learning to rank, *Technical Report MSRTR-2009-141*, Microsoft Research.

Chu, W. and Keerthi, S. (2007). Support vector ordinal regression, *Neural computation* **19**(3): 792–815.

Clémençon, S. J., Robbiano, S. and Vayatis, N. (2011). Ranking Multi-Class Data: Optimality and Pairwise Aggregation. http://hal.archives-ouvertes.fr/hal-00630496/.

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. (2008). Empirical performance maximization for linear rank statistics, *Advances in Neural Information Processing Systems*, Vol. 21.

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.

Dembczyński, K., Kotłowski, W. and Słowiński, R. (2008). Ordinal classification with decision rules, *Lecture Notes in Computer Science*, Vol. 20.

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

Duchi, J., Mackey, L. and Jordan, M. (2012). The Asymptotics of Ranking Algorithms. http://arxiv.org/pdf/1204.1688v1.pdf.

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.

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

Hand, D. and Till, R. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, *Machine Learning* **45**(2): 171–186.

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.

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents, *Proceedings of the 23th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41–48.

Kotłowski, W., Dembczyński, 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.

Lee, Y., Lin, Y. and Wahba, G. (2004). Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data, *Journal of the American Statistical Association* **99**: 67–81.

Li, P., Wu, Q. and Burges, C. J. (2007). McRank: Learning to Rank Using Multiple Classification and Gradient Boosting, *in* J. Platt, D. Koller, Y. Singer and S. Roweis (eds), *Advances in Neural Information Processing Systems 20*, MIT Press, Cambridge, MA, pp. 897–904.

Lin, H. and Li, L. (2006). Large-Margin Thresholded Ensembles for Ordinal Regression: Theory and Practice, *Lecture Notes in Computer Science*, Vol. 4264, pp. 319–333.

McCullagh, P. (1980). Regression Models for Ordinal Data, *Journal of Royal Statistical Society (B)* **42**(2): 109–142.

Rakotomamonjy, A. (2004). Optimizing area under ROC curve with SVMs, *Proceedings of the First Workshop on ROC Analysis in Artificial Intelligence*, pp. 71–80.

Rennie, J. (2006). A comparison of McCullagh's proportional odds model to modern ordinal regression algorithms. http://people.csail.mit.edu/jrennie/writing/proportionalOdds.pdf.

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.

Shashua, A. and Levin, A. (2003). Ranking with large margin principle: Two approaches, *in* S. Becker, S. Thrun and K. Obermayer (eds), *Advances in Neural Information Processing Systems*, Vol. 15, MIT Press, pp. 961–968.

Tewari, A. and Bartlett, P. (2007). On the Consistency of Multiclass Classification Methods, *The Journal of Machine Learning Research* **8**: 1007–1025.

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.

Waegeman, W. and Baets, B. (2011). On the ERA ranking representability of pairwise bipartite ranking functions, *Artificial Intelligence* **175**: 1223–1250.

Waegeman, W., Baets, B. and Boullart, L. (2008). ROC analysis in ordinal regression learning, *Pattern Recognition Letters* **29**(1): 1–9.

Xu, J. and Li, H. (2007). AdaRank: a boosting algorithm for information retrieval, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.

Yue, Y., Finley, T., Radlinski, F. and Joachims, T. (2007). A support vector method for optimizing average precision, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods, *J. Mach. Learn. Res.* **5**: 1225–1251.