On the Consistency of Ranking Algorithms

John Duchi
Lester Mackey
Michael Jordan

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-56
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-56.html

May 9, 2010

Acknowledgement

John Duchi and Lester Mackey were supported by the Department of Defense (DoD) through the National Defense Science and Engineering (NDSEG) graduate fellowship program. We also thank the anonymous reviewers for their helpful feedback.

On the Consistency of Ranking Algorithms

John C. Duchi JDUCHI@CS.BERKELEY.EDU
Lester W. Mackey LMACKEY@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA

Abstract

We present a theoretical analysis of supervised ranking, providing necessary and sufficient conditions for the asymptotic consistency of algorithms based on minimizing a surrogate loss function. We show that many commonly used surrogate losses are inconsistent; surprisingly, we show inconsistency even in low-noise settings. We present a new value-regularized linear loss, establish its consistency under reasonable assumptions on noise, and show that it outperforms conventional ranking losses in a collaborative filtering experiment.

The goal in ranking is to order a set of inputs in accordance with the preferences of an individual or a population. In this paper we consider a general formulation of the supervised ranking problem in which each training example consists of a *query* q, a set of inputs $\boldsymbol{x}$, sometimes called *results*, and a weighted graph G representing *preferences* over the results. The learning task is to discover a function that provides a query-specific ordering of the inputs that best respects the observed preferences. This query-indexed setting is natural for tasks like web search in which a different ranking is needed for each query. Following existing literature, we assume the existence of a *scoring function* $\boldsymbol{f}(\boldsymbol{x}, q)$ that gives a score to each result in $\boldsymbol{x}$; the scores are sorted to produce a ranking (Herbrich et al., 2000; Freund et al., 2003). We assume simply that the observed preference graph G is a directed acyclic graph (DAG). Finally, we cast our work in a decision-theoretic framework in which ranking procedures are evaluated via a loss function $L(\boldsymbol{f}(\boldsymbol{x}, q), G)$.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

It is important to distinguish between the loss function used for evaluating learning procedures from the loss-like functions used to define specific methods (generally via an optimization algorithm). In prior work the former (evaluatory) loss has often been taken to be a pairwise 0-1 loss that sums the number of misordered pairs of results. Recent work has considered losses that penalize errors on more highly ranked instances more strongly. Järvelin & Kekäläinen (2002) suggest using discounted cumulative gain, which assumes that each result x_i is given a score y_i and that the loss is a weighted sum of the y_i of the predicted order. Rudin (2009) uses a p-norm to emphasize the highest ranked instances. Here we employ a general graph-based loss $L(\boldsymbol{f}(\boldsymbol{x}, q), G)$ which is equal to zero if $\boldsymbol{f}(q, \boldsymbol{x})$ obeys the order specified by G—that is, $f_i(\boldsymbol{x}, q) > f_j(\boldsymbol{x}, q)$ for each edge $(i \to j) \in G$, where $f_i(\boldsymbol{x}, q)$ is the score assigned to the ith object in $\boldsymbol{x}$—and is positive otherwise. We make the assumption that L is *edgewise*, meaning that L depends only on the relative order of $f_i(\boldsymbol{x}, q)$ rather than on its values. Such losses are natural in settings with feedback in the form of ordered preferences, for example when learning from click data.

Although we might wish to base a learning algorithm on the direct minimization of the loss L, this is generally infeasible due to the non-convexity and discontinuity of L. In practice one instead employs a *surrogate loss* that lends itself to more efficient minimization. This issue is of course familiar from the classification literature, where a deep theoretical understanding of the statistical and computational consequences of the choices of various surrogate losses has emerged (Zhang, 2004; Bartlett et al., 2006). There is a relative paucity of such understanding for ranking. In the current paper we aim to fill this gap, taking a step toward bringing the ranking literature into line with that for classification. We provide a general theoretical analysis of the consistency of ranking algorithms that are based on a surrogate loss function.

The paper is organized as follows. In Section 1, we define the consistency problem formally and present a theorem that provides conditions under which consistency is achieved for ranking algorithms. In Section 2 we show that finding consistent surrogate losses is difficult in general, and we establish results showing that many commonly used ranking loss functions are inconsistent, even in low-noise settings. We complement this in Section 3 by presenting losses that are consistent in these low-noise settings. We finish with experiments and conclusions in Sections 4 and 5.

1. Consistency for Surrogate Losses

Our task is to minimize the risk of the scoring function $\boldsymbol{f}$. The risk is the expected loss of $\boldsymbol{f}$ across all queries q, result sets $\boldsymbol{x}$, and preference DAGs G:

$$R(\boldsymbol{f}) = \mathbb{E}_{\boldsymbol{X},Q,G} L(\boldsymbol{f}(\boldsymbol{X}, Q), G). \tag{1}$$

Given a query q and result set $\boldsymbol{x}$, we define $\mathcal{G}$ to be the set of possible preference DAGs and $\boldsymbol{p}$ to be (a version of) the vector of conditional probabilities of each DAG. That is, $\boldsymbol{p} = [p_G]_{G\in\mathcal{G}} = [\mathrm{P}(G \mid \boldsymbol{x}, q)]_{G\in\mathcal{G}}$. In what follows, we suppress dependence of $\boldsymbol{p}$, $\mathcal{G}$, and G on the query q and results $\boldsymbol{x}$, as they should be clear from context. We assume that the cardinality of any result set $\boldsymbol{x}$ is bounded above by $M < \infty$. We further define the conditional risk of $\boldsymbol{f}$ given $\boldsymbol{x}$ and q to be

$$\begin{aligned}\ell(\boldsymbol{p}, \boldsymbol{f}(\boldsymbol{x}, q)) &= \sum_{G\in\mathcal{G}} p_G L(\boldsymbol{f}(\boldsymbol{x}, q), G) \\ &= \sum_{G\in\mathcal{G}} \mathrm{P}(G \mid \boldsymbol{x}, q) L(\boldsymbol{f}(\boldsymbol{x}, q), G). \end{aligned} \tag{2}$$

With this definition, we see the risk of $\boldsymbol{f}$ is equal to

$$\mathbb{E}_{\boldsymbol{X},Q}\left[\sum_{G\in\mathcal{G}} \mathrm{P}(G \mid \boldsymbol{X}, Q) L(\boldsymbol{f}(\boldsymbol{X}, Q), G)\right] = \mathbb{E}_{\boldsymbol{X},Q}\ell(\boldsymbol{p}, \boldsymbol{f}).$$

We overload notation so that $\boldsymbol{\alpha}$ takes the value of $\boldsymbol{f}(\boldsymbol{x}, q)$ in $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$. The minimal risk, or Bayes' risk, is the minimal risk over all measurable functions,

$$R^* = \inf_{\boldsymbol{f}} R(\boldsymbol{f}) = \mathbb{E}_{\boldsymbol{X},Q} \inf_{\boldsymbol{\alpha}} \ell(\boldsymbol{p}, \boldsymbol{\alpha}).$$

It is infeasible to directly minimize the true risk in Eq. (1), as it is non-convex and discontinuous. As is done in classification (Zhang, 2004; Bartlett et al., 2006), we thus consider a bounded-below surrogate φ to minimize in place of L. For each G, we write $\varphi(\cdot, G) : \mathbb{R}^{|G|} \to \mathbb{R}$. The φ-risk of the function $\boldsymbol{f}$ is

$$\begin{aligned} R_\varphi(\boldsymbol{f}) &= \mathbb{E}_{\boldsymbol{X},Q,G}\left[\varphi(\boldsymbol{f}(\boldsymbol{X}, Q), G)\right] \\ &= \mathbb{E}_{\boldsymbol{X},Q}\left[\sum_{G\in\mathcal{G}} \mathrm{P}(G \mid \boldsymbol{X}, Q)\varphi(\boldsymbol{f}(\boldsymbol{X}, Q), G)\right], \end{aligned}$$

while the optimal φ-risk is $R^*_\varphi = \inf_{\boldsymbol{f}} R_\varphi(\boldsymbol{f})$.

To develop a theory of consistency for ranking methods, we pursue a treatment that parallels that of Zhang (2004) for classification. Using the conditional risk in Eq. (2), we define a function to measure the discriminating ability of the surrogate φ. Let $\mathcal{G}(m)$ denote the set of possible DAGs G over m results, noting that $|\mathcal{G}(m)| \le 3^{\binom{m}{2}}$. Let $\Delta_{|\mathcal{G}(m)|} \subset \mathbb{R}^{|\mathcal{G}(m)|}$ denote the probability simplex. For $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathbb{R}^m$ we define

$$\begin{aligned} H_m(\varepsilon) = \inf_{\boldsymbol{p}\in\Delta, \boldsymbol{\alpha}} \Big\{ &\sum_{G\in\mathcal{G}(m)} p_G \varphi(\boldsymbol{\alpha}, G) - \inf_{\boldsymbol{\alpha}'} \sum_{G\in\mathcal{G}(m)} p_G \varphi(\boldsymbol{\alpha}', G) \\ &: \ell(\boldsymbol{p}, \boldsymbol{\alpha}) - \inf_{\boldsymbol{\alpha}'} \ell(\boldsymbol{p}, \boldsymbol{\alpha}') \ge \varepsilon \Big\}. \end{aligned} \tag{3}$$

H_m measures surrogate risk suboptimality as a function of true risk suboptimality. A reasonable surrogate loss should declare any setting of $\{\boldsymbol{p}, \boldsymbol{\alpha}\}$ suboptimal that the true loss declares suboptimal, which corresponds to $H_m(\varepsilon) > 0$ whenever $\varepsilon > 0$. We will see soon that this condition is the key to consistency.

Define $H(\varepsilon) = \min_{m\le M} H_m(\varepsilon)$. We immediately have $H \ge 0$, $H(0) = 0$, and $H(\varepsilon)$ is non-decreasing on $0 \le \varepsilon < \infty$, since individual $H_m(\varepsilon)$ are non-decreasing in ε. We have the following lemma (a simple consequence of Jensen's inequality), which we prove in Appendix A.

Lemma 1. *Let ζ be a convex function such that $\zeta(\varepsilon) \le H(\varepsilon)$. Then for all $\boldsymbol{f}$, $\zeta(R(\boldsymbol{f}) - R^*) \le R_\varphi(\boldsymbol{f}) - R^*_\varphi$.*

Corollary 26 from Zhang (2004) then shows as a consequence of Lemma 1 that if $H(\varepsilon) > 0$ for all $\varepsilon > 0$, there is a nonnegative concave function ξ, right continuous at 0 with $\xi(0) = 0$, such that

$$R(\boldsymbol{f}) - R^* \le \xi(R_\varphi(\boldsymbol{f}) - R^*_\varphi). \tag{4}$$

Clearly, if $\lim_n R_\varphi(\boldsymbol{f}_n) = R^*_\varphi$, we have consistency: $\lim_n R(\boldsymbol{f}_n) = R^*$. Though it is not our focus, it is possible to use Eq. (4) to get strong rates of convergence if ξ grows slowly. The remainder of this paper concentrates on finding conditions relating the surrogate loss φ to the risk ℓ to make $H(\varepsilon) > 0$ for $\varepsilon > 0$.

We achieve this goal by using conditions based on the edge structure of the observed DAGs. Given a probability vector $\boldsymbol{p} \in \mathbb{R}^{|\mathcal{G}|}$ over a set of DAGs $\mathcal{G}$, we recall Eq. (2) and define the set of optimal result scores $A(\boldsymbol{p})$ to be all $\boldsymbol{\alpha}$ attaining the infimum of $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$,

$$A(\boldsymbol{p}) = \{\boldsymbol{\alpha} : \ell(\boldsymbol{p}, \boldsymbol{\alpha}) = \inf_{\boldsymbol{\alpha}'} \ell(\boldsymbol{p}, \boldsymbol{\alpha}')\}. \tag{5}$$

The infimum is attained since ℓ is edgewise as described earlier, so $A(\boldsymbol{p})$ is not empty. The following definition captures the intuition that the surrogate loss

φ should maintain ordering information. For this definition and the remainder of the paper, we use the following shorthand for the conditional φ-risk:

$$W(\boldsymbol{p}, \boldsymbol{\alpha}) \triangleq \sum_{G \in \mathcal{G}} p_G \varphi(\boldsymbol{\alpha}, G). \tag{6}$$

Definition 2. *Let φ be a bounded-below surrogate loss with $\varphi(\cdot, G)$ continuous for all G. φ is* edge-consistent *with respect to the loss L if for all $\boldsymbol{p}$,*

$$W^*(\boldsymbol{p}) \triangleq \inf_{\boldsymbol{\alpha}} W(\boldsymbol{p}, \boldsymbol{\alpha}) < \inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\}.$$

Definition 2 captures an essential property for the surrogate loss φ: if $\boldsymbol{\alpha}$ induces an edge $(i \to j)$ via $\alpha_i > \alpha_j$ so that the conditional risk $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$ is not minimal, then the conditional surrogate risk $W(\boldsymbol{p}, \boldsymbol{\alpha})$ is not minimal.

We now provide three lemmas and a theorem that show that if the surrogate loss φ satisfies edge-consistency, then its minimizer asymptotically minimizes the Bayes risk. As the lemmas are direct analogs of results in Tewari & Bartlett (2007) and Zhang (2004), we put their proofs in Appendix A.

Lemma 3. *$W^*(\boldsymbol{p})$ is continuous on Δ.*

Lemma 4. *Let φ be edge-consistent. Then $W(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \to W^*(\boldsymbol{p})$ implies that $\ell(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \to \inf_{\boldsymbol{\alpha}} \ell(\boldsymbol{p}, \boldsymbol{\alpha})$ and $\boldsymbol{\alpha}^{(n)} \in A(\boldsymbol{p})$* eventually.

Lemma 5. *Let φ be edge-consistent. For every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\boldsymbol{p} \in \Delta$, $\ell(\boldsymbol{p}, \boldsymbol{\alpha}) - \inf_{\boldsymbol{\alpha}'} \ell(\boldsymbol{p}, \boldsymbol{\alpha}') \geq \varepsilon$ implies $W(\boldsymbol{p}, \boldsymbol{\alpha}) - W^*(\boldsymbol{p}) \geq \delta$.*

Theorem 6. *Let φ be a continuous, bounded-below loss function and assume that the size of the result sets is upper bounded by a constant M. Then φ is edge-consistent if and only the following holds: Whenever $\boldsymbol{f}_n$ is a sequence of scoring functions such that*

$$R_\varphi(\boldsymbol{f}_n) \xrightarrow{p} R^*_\varphi, \quad \text{then} \quad R(\boldsymbol{f}_n) \xrightarrow{p} R^*.$$

Proof We begin by proving that if φ is edge-consistent, the implication holds. By Lemma 5 and the definition of H_m in Eq. (3), we have that if $\varepsilon > 0$, then there is some $\delta > 0$ such that $H_m(\varepsilon) \geq \delta > 0$. Thus $H(\varepsilon) = \min_{m \leq M} H_m(\varepsilon) > 0$, and Eq. (4) then immediately implies that $R(\boldsymbol{f}_n) \xrightarrow{p} R^*$.

Now suppose that φ is not edge-consistent, that is, there is some $\boldsymbol{p}$ so that $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{\alpha}}\{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\}$. Let $\boldsymbol{\alpha}^{(n)} \notin A(\boldsymbol{p})$ be a sequence such that $W(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \to W^*(\boldsymbol{p})$. If we simply define the risk to be the expected loss on one particular example $\boldsymbol{x}$ and set $\boldsymbol{f}_n(\boldsymbol{x}) = \boldsymbol{\alpha}^{(n)}$, then $R_\varphi(\boldsymbol{f}_n) = W(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)})$. Further, by assumption there is some $\varepsilon > 0$ such that $\ell(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \geq \inf_{\boldsymbol{\alpha}} \ell(\boldsymbol{p}, \boldsymbol{\alpha}) + \varepsilon$ for all n. Thus $R(\boldsymbol{f}_n) = \ell(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \not\to R^* = \inf_{\boldsymbol{\alpha}} \ell(\boldsymbol{p}, \boldsymbol{\alpha})$. □

2. The Difficulty of Consistency

In this section, we explore the difficulty of finding edge-consistent ranking losses in practice. We first show that unless $P = NP$ many useful losses cannot be edge-consistent in general. We then show that even in low-noise settings, common losses used for ranking are not edge-consistent. We focus our attention on pairwise losses, which impose a separate penalty for each edge that is ordered incorrectly; this generalizes the disagreement error described by Dekel et al. (2004). We assume we have a set of non-negative penalties a_{ij}^G indexed by edge $(i \to j)$ and graph G so that

$$L(\boldsymbol{\alpha}, G) = \sum_{i<j} a_{ij}^G 1_{(\alpha_i \leq \alpha_j)} + \sum_{i>j} a_{ij}^G 1_{(\alpha_i < \alpha_j)}. \tag{7}$$

We distinguish the cases $i < j$ and $i > j$ to avoid minor technical issues created by doubly penalizing $1_{(\alpha_i = \alpha_j)}$. If we define $a_{ij} \triangleq \sum_{G \in \mathcal{G}} a_{ij}^G p_G$, then

$$\ell(\boldsymbol{p}, \boldsymbol{\alpha}) = \sum_{i<j} a_{ij} 1_{(\alpha_i \leq \alpha_j)} + \sum_{i>j} a_{ij} 1_{(\alpha_i < \alpha_j)}. \tag{8}$$

2.1. General inconsistency results

Finding an efficiently minimizable surrogate loss that is also consistent for Eq. (8) for all $\boldsymbol{p}$ is unlikely, as indicated by the next lemma. The result is a consequence of the fact that the feedback arc-set problem is NP-complete (Karp, 1972); we defer its proof to Appendix A.

Lemma 7. *Define $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$ as in Eq. (8). Finding an $\boldsymbol{\alpha}$ minimizing ℓ is NP-hard.*

Since many convex functions are minimizable in polynomial time or can be straightforwardly transformed into a formulation that is minimizable in polylogarithmic time (Ben-Tal & Nemirovski, 2001), most convex surrogates are inconsistent unless $P = NP$.

2.2. Low-noise inconsistency

In this section we show that, surprisingly, many common convex surrogates are inconsistent even in low-noise settings. Inspecting Eq. (7), a natural choice for a surrogate loss is one of the form (Herbrich et al., 2000; Freund et al., 2003; Dekel et al., 2004)

$$\varphi(\boldsymbol{\alpha}, G) = \sum_{(i \to j) \in G} h(a_{ij}^G) \phi(\alpha_i - \alpha_j) \tag{9}$$

where $\phi \geq 0$ is a non-increasing function, and h is a function of the penalties a_{ij}^G. In this case, the conditional surrogate risk is $W(\boldsymbol{p}, \boldsymbol{\alpha}) = \sum_{i \neq j} h_{ij} \phi(\alpha_i - \alpha_j)$, where we define $h_{ij} \triangleq \sum_{G \in \mathcal{G}} h(a_{ij}^G) p_G$.

If φ from Eq. (9) is edge-consistent, then ϕ must be differentiable at 0 with $\phi'(0) < 0$. This is a consequence of Bartlett et al.'s (2006) analysis of binary classification and the correspondence between binary classification and pairwise ranking; for the binary case, consistency requires $\phi'(0) < 0$. Similarly, we must have $h \geq 0$ on $\mathbb{R}_+$ and strictly increasing. For the remainder of this section, we make the unrestrictive assumption that ϕ decreases more slowly in the positive direction than it increases in the negative. Formally, we use the recession function (Rockafellar, 1970, Thm. 8.5) of ϕ,

$$\phi'_\infty(d) \triangleq \sup_{t>0} \frac{\phi(td) - \phi(0)}{t} = \lim_{t\to\infty} \frac{\phi(td) - \phi(0)}{t}.$$

The assumption, satisfied for bounded below ϕ, is

Assumption A. *$\phi'_\infty(1) \geq 0$ or $\phi'_\infty(-1) = \infty$.*

We now define precisely what we mean by a low-noise setting. For any $(\mathcal{G}, \boldsymbol{p})$, let $\widetilde{G}$ be the *difference graph*, that is, the graph with edge weights $\max\{a_{ij} - a_{ji}, 0\}$ on edges $(i \to j)$, where $a_{ij} = \sum_{G\in\mathcal{G}} a_{ij}^G p_G$, and if $a_{ij} \leq a_{ji}$ then the edge $(i \to j) \notin \widetilde{G}$ (see Fig. 1). We define the following *low-noise condition* based on self-reinforcement of edges in the difference graph.

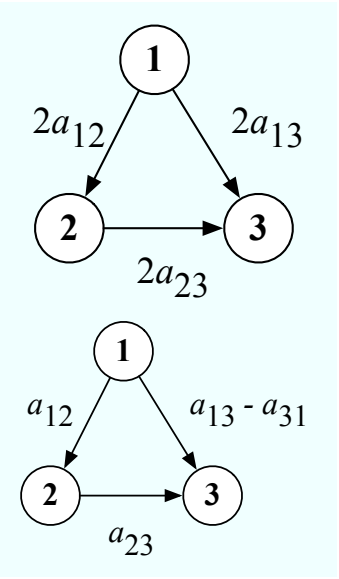
Definition 8. *We say $(\mathcal{G}, \boldsymbol{p})$ is* low-noise *when the corresponding difference graph $\widetilde{G}$ satisfies the following reverse triangle inequality: whenever there is an edge $(i \to j)$ and an edge $(j \to k)$ in $\widetilde{G}$, then the weight $a_{ik} - a_{ki}$ on $(i \to k)$ is greater than or equal to the path weight $a_{ij} - a_{ji} + a_{jk} - a_{kj}$ on $(i \to j \to k)$.*

It is not difficult to see that if $(\mathcal{G}, \boldsymbol{p})$ satisfies Def. 8, its difference graph $\widetilde{G}$ is a DAG. Indeed, the definition ensures that all global preference information in $\widetilde{G}$ (the sum of weights along any path) conforms with and reinforces local preference information (the weight on a single edge). Reasonable ranking methods should be consistent in this setting, but this is not trivial.

In the lemmas to follow, we consider simple 3-node DAGs that admit unique minimizers for their conditional risks. In particular, we consider DAGs on nodes 1, 2, and 3 that induce only the four penalty values a_{12}, a_{13}, a_{23}, and a_{31} (see Fig. 1). In this case, if $a_{13} > a_{31}$, any $\boldsymbol{\alpha}$ minimizing $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$ clearly will have $\alpha_1 > \alpha_2 > \alpha_3$. We now show under some very general conditions that if φ is edge-consistent, ϕ is non-convex.

Let $\phi'(x)$ denote an element of the subgradient set $\partial\phi(x)$. The subgradient conditions for optimality of

$$\begin{aligned} W(\boldsymbol{p}, \boldsymbol{\alpha}) = \; & h_{12}\phi(\alpha_1 - \alpha_2) + h_{13}\phi(\alpha_1 - \alpha_3) \\ & + h_{23}\phi(\alpha_2 - \alpha_3) + h_{31}\phi(\alpha_3 - \alpha_1) \end{aligned} \tag{10}$$



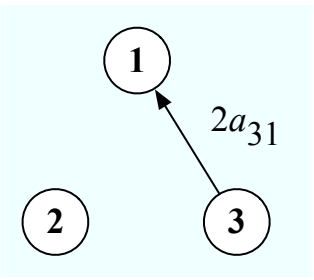


Figure 1. The two DAGs above occur with probability $\frac{1}{2}$, giving the difference graph $\widetilde{G}$ on the left, assuming that $a_{13} > a_{31}$.

are that

$$\begin{aligned} 0 &= h_{12}\phi'(\alpha_1 - \alpha_2) + h_{13}\phi'(\alpha_1 - \alpha_3) - h_{31}\phi'(\alpha_3 - \alpha_1) \\ 0 &= -h_{12}\phi'(\alpha_1 - \alpha_2) + h_{23}\phi'(\alpha_2 - \alpha_3). \end{aligned} \tag{11}$$

We begin by showing that under Assumption A on ϕ, there is a finite minimizer of $W(\boldsymbol{p}, \boldsymbol{\alpha})$. The lemma is technical and its proof is in Appendix A.

Lemma 9. *There is a constant $C < \infty$ and a vector $\boldsymbol{\alpha}^*$ minimizing $W(\boldsymbol{p}, \boldsymbol{\alpha})$ with $\|\boldsymbol{\alpha}^*\|_\infty \leq C$.*

We use the following lemma to prove our main theorem about inconsistency of pairwise convex losses.

Lemma 10 (Inconsistency of convex losses)**.** *Suppose that $a_{13} > a_{31} > 0$, $a_{12} > 0$, $a_{23} > 0$. Let $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$ be*

$$a_{12}1_{(\alpha_1\leq\alpha_2)} + a_{13}1_{(\alpha_1\leq\alpha_3)} + a_{23}1_{(\alpha_2\leq\alpha_3)} + a_{31}1_{(\alpha_3<\alpha_1)}$$

and $W(\boldsymbol{p}, \boldsymbol{\alpha})$ be defined as in Eq. (10). For convex ϕ with $\phi'(0) < 0$, $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\}$ whenever either of the following conditions is satisfied:

$$\textsc{Cond 1: } h_{23} < \frac{h_{31}h_{12}}{h_{13} + h_{12}} \quad \textsc{Cond 2: } h_{12} < \frac{h_{31}h_{23}}{h_{13} + h_{23}}.$$

Proof Lemma 9 shows that the optimal $W^*(\boldsymbol{p})$ is attained by some finite $\boldsymbol{\alpha}$. Thus, we fix an $\boldsymbol{\alpha}^*$ satisfying Eq. (11), and let $\delta_{ij} = \alpha_i^* - \alpha_j^*$ and $g_{ij} = \phi'(\delta_{ij})$ for $i \neq j$. We make strong use of the monotonicity of subgradients, that is, $\delta_{ij} > \delta_{kl}$ implies $g_{ij} \geq g_{kl}$ (e.g. Rockafellar, 1970, Theorem 24.1). By Eq. (11),

$$g_{13} - g_{12} = \frac{h_{31}}{h_{13}} g_{31} - \left(1 + \frac{h_{12}}{h_{13}}\right) g_{12} \tag{12a}$$

$$g_{13} - g_{23} = \frac{h_{31}}{h_{13}} g_{31} - \left(1 + \frac{h_{23}}{h_{13}}\right) g_{23}. \tag{12b}$$

Suppose for the sake of contradiction that $\boldsymbol{\alpha}^* \in A(\boldsymbol{p})$. As $\delta_{13} = \delta_{12} + \delta_{23}$, we have that $\delta_{13} > \delta_{12}$ and $\delta_{13} > \delta_{23}$. The convexity of ϕ implies that if $\delta_{13} > \delta_{12}$, then $g_{13} \geq g_{12}$. If $g_{12} \geq 0$, we thus have that $g_{13} \geq 0$ and by Eq. (11), $g_{31} \geq 0$. This is a contradiction since $\delta_{31} < 0$ gives $g_{31} \leq \phi'(0) < 0$. Hence, $g_{12} < 0$. By identical reasoning, we also have that $g_{23} < 0$.

Now, $\delta_{23} > 0 > \delta_{31}$ implies that $g_{23} \geq g_{31}$, which combined with Eq. (12a) and the fact that $g_{23} = (h_{12}/h_{23})g_{12}$ (by Eq. (11)) gives

$$\begin{aligned} g_{13} - g_{12} &\leq \frac{h_{31}}{h_{13}} g_{23} - \left(1 + \frac{h_{12}}{h_{13}}\right) g_{12} \\ &= \left(\frac{h_{31}h_{12}}{h_{23}} - h_{13} - h_{12}\right) \frac{g_{12}}{h_{13}}. \end{aligned}$$

Since $g_{12}/h_{13} < 0$, we have that $g_{13} - g_{12} < 0$ whenever $h_{31}h_{12}/h_{23} > h_{13} + h_{12}$. But when $\delta_{13} > \delta_{12}$, we must have $g_{13} \geq g_{12}$, which yields a contradiction under Condition 1.

Similarly, $\delta_{12} > 0 > \delta_{31}$ implies that $g_{12} \geq g_{31}$, which with $g_{12} = (h_{23}/h_{12})g_{23}$ and Eq. (12b) gives

$$\begin{aligned} g_{13} - g_{23} &\leq \frac{h_{31}}{h_{13}} g_{12} - \left(1 + \frac{h_{23}}{h_{13}}\right) g_{23} \\ &= \left(\frac{h_{31}h_{23}}{h_{12}} - h_{13} - h_{23}\right) \frac{g_{23}}{h_{13}}. \end{aligned}$$

Since $g_{23}/h_{13} < 0$, we further have that $g_{13} - g_{23} < 0$ whenever $h_{31}h_{23}/h_{12} > h_{13} + h_{23}$. This contradicts $\delta_{13} > \delta_{23}$ under Condition 2. □

Lemma 10 allows us to construct scenarios under which arbitrary pairwise surrogate losses with convex ϕ are inconsistent. Assumption A only to specify an optimal $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_\infty < \infty$, and can be weakened to $W(\boldsymbol{p}, \boldsymbol{\alpha}) \to \infty$ as $(\alpha_i - \alpha_j) \to \infty$. The next theorem is our main negative result on the consistency of pairwise surrogate losses.

Theorem 11. *Let φ be a loss that can be written as*

$$\varphi(\boldsymbol{\alpha}, G) = \sum_{(i \to j) \in G} h(a_{ij}^G) \phi(\alpha_i - \alpha_j)$$

for h continuous and increasing with $h(0) = 0$. Even in the low-noise setting, for ϕ convex and satisfying Assumption A, φ is not edge-consistent.

Proof Assume for the sake of contradiction that φ is edge-consistent. Recall that for ϕ convex, $\phi'(0) < 0$, and we can construct graphs G_1 and G_2 so that the resulting expected loss satisfies Condition 1 of Lemma 10. Let $\mathcal{G} = \{G_1, G_2\}$ where $G_1 = (\{1,2,3\}, \{(1 \to 2), (1 \to 3)\})$ and $G_2 = (\{1,2,3\}, \{(2 \to 3), (3 \to 1)\})$. Fix any weights $a_{12}^{G_1}, a_{13}^{G_1}, a_{31}^{G_2}$ with $a_{13}^{G_1} > a_{12}^{G_1} > 0$ and $a_{13}^{G_1} > a_{31}^{G_2} > 0$, and let $\boldsymbol{p} = (.5, .5)$. As h is continuous with $h(0) = 0$, there exists some $\varepsilon > 0$ such that $h(\varepsilon) < 2h_{31}h_{12}/(h_{13} + h_{12})$, where $h_{ij} = \sum_{G \in \mathcal{G}} h(a_{ij}^G) p_G$. Take $a_{23}^{G_2} = \min\{\varepsilon, (a_{13}^{G_1} - a_{12}^{G_1})/2\}$. Then we have $h_{23} = h(a_{23}^{G_2})/2 \leq h(\varepsilon)/2 < h_{31}h_{12}/(h_{13} + h_{12})$. Hence Condition 1 of Lemma 10 is satisfied, so φ is not edge-consistent. Moreover, $a_{23}^{G_2} \leq (a_{13}^{G_1} - a_{12}^{G_1})/2 < a_{13}^{G_1} - a_{12}^{G_1}$ implies that $\widetilde{G}$ is a DAG satisfying the low-noise condition. □

2.3. Margin-based inconsistency

Given the difficulties encountered in the previous section, it is reasonable to consider a reformulation of our surrogate loss. A natural alternative is a margin-based loss, which encodes a desire to separate ranking scores by a large margins dependent on the preferences in a graph. Similar losses have been proposed, e.g., by Shashua & Levin (2002). In particular, we now consider losses of the form

$$\varphi(\boldsymbol{\alpha}, G) = \sum_{(i \to j) \in G} \phi\left(\alpha_i - \alpha_j - h(a_{ij}^G)\right), \qquad (13)$$

where h is continuous and $h(0) = 0$. It is clear from the reduction to binary classification that h must be increasing for the loss in Eq. (13) to be edge-consistent. When ϕ is a decreasing function, this intuitively says that the larger a_{ij} is, the larger α_i should be when compared to α_j. Nonetheless, as we show below, such a loss is inconsistent even in low-noise settings.

Theorem 12. *Let φ be a loss that can be written as*

$$\varphi(\boldsymbol{\alpha}, G) = \sum_{(i \to j) \in G} \phi(\alpha_i - \alpha_j - h(a_{ij}^G))$$

for h continuous and increasing with $h(0) = 0$. Even in the low-noise setting, for ϕ convex and satisfying Assumption A, φ is not edge-consistent.

Proof Assume for the sake of contradiction that φ is edge-consistent. As noted before, $\phi'(0) < 0$, and since ϕ is differentiable almost everywhere (Rockafellar, 1970, Theorem 25.3), ϕ is differentiable at $-c$ for some $c > 0$ in the range of h. Considering the four-graph setting with graphs containing one edge each, $G_1 = (\{1,2,3\}, \{(1 \to 2)\})$, $G_2 = (\{1,2,3\}, \{(2 \to 3)\})$, $G_3 = (\{1,2,3\}, \{(1 \to 3)\})$, and $G_4 = (\{1,2,3\}, \{(3 \to 1)\})$, choose constant edge weights $a_{12}^{G_1} = a_{13}^{G_2} = a_{23}^{G_3} = a_{31}^{G_4} = h^{-1}(c) > 0$, and

set $\boldsymbol{p} = (.25, .01, .5, .24)$. In this setting,

$$W(\boldsymbol{p}, \boldsymbol{\alpha}) = p_{G_1}\tilde{\phi}(\alpha_1 - \alpha_2) + p_{G_2}\tilde{\phi}(\alpha_2 - \alpha_3) + p_{G_3}\tilde{\phi}(\alpha_1 - \alpha_3) + p_{G_4}\tilde{\phi}(\alpha_3 - \alpha_1),$$

for $\tilde{\phi}(x) = \phi(x - c)$. Notably, $\tilde{\phi}$ is convex, satisfies Assumption A, and $\tilde{\phi}'(0) = \phi'(-c) < 0$. Moreover, $a_{13} - a_{31} = h^{-1}(c)(p_{G_3} - p_{G_4}) \geq h^{-1}(c)(p_{G_1} + p_{G_2}) = a_{12} + a_{23} > 0$, so $\widetilde{G}$ is a DAG satisfying the low-noise condition. However, $p_{G_2} < \frac{p_{G_4} p_{G_1}}{p_{G_3} + p_{G_1}}$. Hence, by Lemma 10, $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\}$, a contradiction. □

3. Conditions for Consistency

The prospects for consistent surrogate ranking appear bleak given the results of the previous section. Nevertheless, we demonstrate in this section that there exist surrogate losses that yield consistency under some restrictions on problem noise. We consider a new loss—specifically, a linear loss in which we penalize $(\alpha_j - \alpha_i)$ proportional to the weight a_{ij} in the given graph G. To keep the loss well-behaved and disallow wild fluctuations, we also regularize the $\boldsymbol{\alpha}$ values. That is, our loss takes the form

$$\varphi(\boldsymbol{\alpha}, G) = \sum_{(i \to j) \in G} a_{ij}^G (\alpha_j - \alpha_i) + \nu \sum_i r(\alpha_i). \quad (14)$$

We assume that r is strictly convex and 1-coercive, that is, that r asymptotically grows faster than any linear function. These conditions imply that the loss of Eq. (14) is bounded below. Moreover, we have the basis for consistency:

Theorem 13. *Let the loss take the form of a generalized disagreement error of Eq. (7) and the surrogate loss take the form of Eq. (14) where $\nu > 0$ and r is strictly convex and 1-coercive. If the pair $(\mathcal{G}, \boldsymbol{p})$ induces a difference graph $\widetilde{G}$ that is a DAG, then*

$$W^*(\boldsymbol{p}) < \inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\} \quad \Leftrightarrow$$
$$\sum_j a_{ij} - a_{ji} > \sum_j a_{kj} - a_{jk} \text{ for } i, k \text{ s.t. } a_{ik} > a_{ki}.$$

Proof We first note that $\widetilde{G}$ is a DAG if and only if

$$A(\boldsymbol{p}) = \{\boldsymbol{\alpha} : \alpha_i > \alpha_j \text{ for } i < j \text{ with } a_{ij} > a_{ji}, \; \alpha_i \geq \alpha_j \text{ for } i > j \text{ with } a_{ij} > a_{ji}\}.$$

(For a proof see Lemma 16, though essentially all we do is write out $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$.) We have that

$$W(\boldsymbol{p}, \boldsymbol{\alpha}) = \sum_i \Big(\alpha_i \sum_j (a_{ji} - a_{ij}) + \nu r(\alpha_i)\Big).$$

Standard subgradient calculus gives that at optimum,

$$r'(\alpha_i) = \frac{\sum_j a_{ij} - a_{ji}}{\nu}.$$

Since r is strictly convex, r' is a strictly increasing set-valued map with increasing inverse $s(g) = \{\alpha : g \in \partial r(\alpha)\}$. Optimality is therefore attained uniquely at

$$\alpha_i^* = s\left(\frac{\sum_j a_{ij} - a_{ji}}{\nu}\right). \quad (15)$$

Note that for any i, k, $\alpha_i^* > \alpha_k^*$ if and only if $s\left(\frac{\sum_j a_{ij} - a_{ji}}{\nu}\right) > s\left(\frac{\sum_j a_{kj} - a_{jk}}{\nu}\right)$, which in turn occurs if and only if $\frac{\sum_j a_{ij} - a_{ji}}{\nu} > \frac{\sum_j a_{kj} - a_{jk}}{\nu}$. Hence, the optimal $\boldsymbol{\alpha}^*$ of Eq. (15) is in $A(\boldsymbol{p})$ if and only if

$$\frac{\sum_j a_{ij} - a_{ji}}{\nu} > \frac{\sum_j a_{kj} - a_{jk}}{\nu} \quad \text{when } a_{ik} > a_{ki}. \quad (16)$$

Thus, $W^*(\boldsymbol{p}) = \inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\}$ whenever Eq. (16) is violated. On the other hand, suppose Eq. (16) is satisfied. Then for all $\boldsymbol{\alpha}$ satisfying

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty < \min_{\{i,k : a_{ik} > a_{ki}\}} \frac{1}{2}(\alpha_i^* - \alpha_k^*),$$

we have $\boldsymbol{\alpha} \in A(\boldsymbol{p})$, and $\inf_{\boldsymbol{\alpha}} \{W(\boldsymbol{p}, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \notin A(\boldsymbol{p})\} > W^*(\boldsymbol{p})$ since $\boldsymbol{\alpha}^*$ is the unique global minimum. □

We now prove a simple lemma showing that low-noise settings satisfy the conditions of Theorem 13.

Lemma 14. *If $(\mathcal{G}, \boldsymbol{p})$ is low noise, then for the associated difference graph $\widetilde{G}$, whenever $a_{ik} > a_{ki}$,*

$$\sum_j a_{ij} - a_{ji} > \sum_j a_{kj} - a_{jk}.$$

Proof Fix (i, k) with $a_{ik} > a_{ki}$. There are two cases for a third node j: either $a_{ij} - a_{ji} > 0$ or $a_{ij} - a_{ji} \leq 0$. In the first case, there is an edge $(i \to j) \in \widetilde{G}$. If $(k \to j) \in \widetilde{G}$, the low-noise condition implies $a_{ij} - a_{ji} \geq a_{kj} - a_{jk} + a_{ik} - a_{ki} > a_{kj} - a_{jk}$. Otherwise, $a_{kj} - a_{jk} \leq 0 < a_{ij} - a_{ji}$. In the other case, $a_{ij} - a_{ji} \leq 0$. If the inequality is strict, then $(j \to i) \in \widetilde{G}$, so the low-noise condition implies that $a_{ji} - a_{ij} < a_{ji} - a_{ij} + a_{ik} - a_{ki} \leq a_{jk} - a_{kj}$, or $a_{kj} - a_{jk} < a_{ij} - a_{ji}$. Otherwise, $a_{ij} = a_{ji}$, and the low-noise condition guarantees that $(j \to k) \notin \widetilde{G}$, so $a_{kj} - a_{jk} \leq 0 = a_{ij} - a_{ji}$.

The inequality in the statement of the lemma is strict, because $a_{ik} - a_{ki} > 0 = a_{kk} - a_{kk}$. □

The converse of the lemma is, in general, false. Combining the above lemma with Theorem 13, we have

Corollary 15. *The linear loss of Eq. (14) is edge-consistent if $(\mathcal{G}, \boldsymbol{p})$ is low noise for all query-result pairs.*

With the above corollary, we have a consistent loss: the value-regularized linear loss is edge (and hence asymptotically) consistent in low-noise settings. It is not difficult to see that the value regularization from r is necessary; if r is not in the objective in Eq. (14), then $\varphi(\cdot, G)$ can be sent to $-\infty$ with $\boldsymbol{\alpha} \notin A(\boldsymbol{p})$.

3.1. Relationship to prior work

One of the main results on consistency to date is due to Cossock & Zhang (2008), who work in a setting in which each item x_i to be ranked is associated with a score y_i. In this setting we can show that the resulting graphs $(\mathcal{G}, \boldsymbol{p})$ satisfy our low-noise condition. Indeed, for every observed pair of results and scores $\langle (x_i, y_i), (x_j, y_j) \rangle$, we can construct a graph $G = (\{x_i, x_j\}, \{(i \to j)\})$ and set $a_{ij}^G = y_i - y_j$. Then in the limit, we have $a_{ij} = \bar{y}_i - \bar{y}_j$, where $\bar{y}_i$ is the true score of item x_i, and clearly $a_{ik} = a_{ij} + a_{jk}$ so that $\widetilde{G}$ satisfies the low-noise condition.

Another related line of work is due to Xia et al. (2008), who introduce a notion of *order-preserving probability spaces*. These are inherently different from our work, which we show by considering graphs on nodes 1, 2, and 3. First, consider a low-noise setting in which the difference graph $\widetilde{G}$ consists of edges $(1 \to 2)$ and $(1 \to 3)$. Our losses are indifferent to whether we order result 2 ahead of or behind 3, and this cannot be captured by an order-preserving probability space.

Conversely, consider an order-preserving probability space over the three nodes, where the data we receive consists of full orderings of $1, 2, 3$. To translate this into our framework, we must convert each of these orderings into a DAG G with associated edge weights. We assume that the weight on each edge is only a function of the distance between the entries in the ordering. Suppose we observe two orderings $\pi = \{1 \to 2 \to 3\}$ and $\widehat{\pi} = \{2 \to 3 \to 1\}$ with probabilities $p_\pi > p_{\widehat{\pi}} \geq 0$ and $p_\pi + p_{\widehat{\pi}} = 1$, which is an order preserving probability space (see Def. 3 and Theorem 5 in Xia et al., 2008). If we assume w.l.o.g. that any adjacent pair in the list has edge weight equal to one and that pairs of distance equal to two in the list have edge weight w_2, then there is no way to set w_2 so that the resulting $(\mathcal{G}, \boldsymbol{p})$ satisfies the low-noise condition in Def. 8. The associated difference graph $\widetilde{G}$ will have edge weights $a_{12} = p_\pi - w_2 p_{\widehat{\pi}}$, $a_{13} = w_2 p_\pi - p_{\widehat{\pi}}$, and $a_{23} = p_\pi + p_{\widehat{\pi}}$. To satisfy the low-noise condition in Def. 8 and have the ordering π minimize the true loss, we must have $w_2 p_\pi - p_{\widehat{\pi}} = a_{13} \geq a_{12} + a_{23} = p_\pi - w_2 p_{\widehat{\pi}} + p_\pi + p_{\widehat{\pi}}$ so that $w_2 p_\pi + w_2 p_{\widehat{\pi}} \geq 2p_\pi + 2p_{\widehat{\pi}}$. That is, $w_2 \geq 2$. On the other hand, we must have $a_{12} > 0$ so that $p_\pi > w_2 p_{\widehat{\pi}}$ or $w_2 < p_\pi / p_{\widehat{\pi}}$; taking $p_\pi \downarrow .5$ and $p_{\widehat{\pi}} \uparrow .5$, we have $w_2 \leq 1$. Thus no construction that assigns a fixed weight to edges associated with permutations can transform an order-preserving probability space into graphs satisfying the low-noise conditions here. Nonetheless, our general consistency result, Theorem 6, implicitly handles order-preserving probability spaces, which assume that graphs G contain all results and the loss $L(\boldsymbol{f}(\boldsymbol{x}, q), G) = 0$ if $\boldsymbol{f}$ agrees with G on all orderings and is 1 otherwise.

4. Experiments

While the focus of this work is a theoretical investigation of consistency, we have also conducted experiments that study the value-regularized linear loss our analysis suggests. We perform experiments on a collaborative filtering task in which the goal is to recommend movies to a user based on the user's and other users' movie ratings. We use one of the MovieLens datasets (GroupLens Lab, 2008), which consists of 100,000 ratings, on a scale of 1 to 5, for 1682 different movies by 943 users. In this case, our "query" is a user u, and the set of possible results consists of all 1682 movies. We learn a linear model so that $f_i(\boldsymbol{x}, u) = w^T d(x_i, u)$, where d is a mapping from movie x_i and user u to a feature vector. We use features that have proven successful in settings such as the Netflix challenge, including the age of the movie, its genre(s), the average rating of the user for other movies in the same genre(s), the average rating of the movie, and ratings given to the movie by users similar to and dissimilar from u in rating of other movies.

To create pairs to train our models, we randomly sample pairs (x_i, x_j) of movies rated by a user. Each sampled pair of rated movies then gets a per-user weight a_{ij}^u that we set to be the difference in their ratings. As discussed in Sec. 3.1, this guarantees that $(\mathcal{G}, \boldsymbol{p})$ is low noise. We sample across users to get n samples total. We then learn the weight vector w using one of three methods: the value-regularized linear method in this paper, a pairwise hinge loss (Herbrich et al., 2000), and a pairwise logistic loss (Dekel et al., 2004). Specifically, the surrogates are

$$\sum_{i,j,u} a_{ij}^u w^T (d(x_j, u) - d(x_i, u)) + \theta \sum_{i,u} (w^T d(x_i, u))^2$$

$$\sum_{i,j,u} a_{ij}^u \left[1 - w^T (d(x_i, u) - d(x_j, u))\right]_+$$

$$\sum_{i,j,u} a_{ij}^u \log\left(1 + e^{w^T (d(x_j, u) - d(x_i, u))}\right),$$

where $[z]_+ = \max\{z, 0\}$. We set $\theta = 10^{-4}$ (it needed simply be a small number), and also added ℓ_2-

Train pairs	Hinge	Logistic	Linear
20000	.478 (.008)	.479 (.010)	**.465** (.006)
40000	.477 (.008)	.478 (.010)	**.464** (.006)
80000	.480 (.007)	.478 (.009)	**.462** (.005)
120000	.477 (.008)	.477 (.009)	**.463** (.006)
160000	.474 (.007)	.474 (.007)	**.461** (.004)

Table 1. Test losses for different surrogate losses.

regularization in the form of $\lambda\|w\|^2$ to each problem. We cross-validated λ separately for each loss.

We partitioned the data into five subsets, and, in each of 15 experiments, we used one subset for validation, one for testing, and three for training. In every experiment, we subsampled 40,000 rated movie pairs from the test set for final evaluation. Once we had learned a vector w for each of the three methods, we computed its average generalized pairwise loss (Eq. (7)). We show the results in Table 1. The leftmost column contains the number of pairs that were subsampled for training, and the remaining columns show the average pairwise loss on the test set for each of the methods (with standard error in parentheses). Each number is the mean of 15 independent training runs, and bold denotes the lowest loss. It is interesting to note that the linear loss always achieves the lowest test loss averaged across all tests. In fact, it achieved the lowest test loss of all three methods in all but one of our experimental runs. (We use these three losses to focus exclusively on learning in a pairwise setting—Cossock & Zhang (2008) learn using relevance scores, while Xia et al. (2008) require full ordered lists of results as training data rather than pairs.) Finally, we note that there is a closed form for the minimizer of the linear loss, which makes it computationally attractive.

5. Discussion

In this paper we have presented results on both the difficulty and the feasibility of surrogate loss consistency for ranking. We have presented the negative result that many natural candidates for surrogate ranking are not consistent in general or even under low-noise restrictions, and we have presented a class of surrogate losses that achieve consistency under reasonable noise restrictions. We have also demonstrated the potential usefulness of the new loss functions in practice. This work thus takes a step toward bringing the consistency literature for ranking in line with that for classification. A natural next step in this agenda is to establish rates for ranking algorithms; we believe that our analysis can be extended to the analysis of rates. Finally, given the difficulty of achieving consistency using surrogate losses that decompose along edges, it may be beneficial to explore non-decomposable losses.

Acknowledgments

JD and LM were supported by NDSEG fellowships. We thank the reviewers for their helpful feedback.

References

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

Ben-Tal, A. and Nemirovski, A. *Lectures on Modern Convex Optimization*. SIAM, 2001.

Cossock, D. and Zhang, T. Statistical analysis of Bayes optimal subset ranking. *IEEE Transaction on Information Theory*, 16:1274–1286, 2008.

Dekel, O., Manning, C., and Singer, Y. Log-linear models for label ranking. In *Advances in Neural Information Processing Systems 16*, 2004.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. Efficient boosting algorithms for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

GroupLens Lab. MovieLens dataset, 2008. URL http://www.grouplens.org/taxonomy/term/14.

Herbrich, R., Graepel, T., and Obermayer, K. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*. MIT Press, 2000.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Karp, R. M. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972.

Rockafellar, R.T. *Convex Analysis*. Princeton University Press, 1970.

Rudin, C. 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, 2009.

Shashua, A. and Levin, A. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems 15*, 2002.

Tewari, A. and Bartlett, P. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–10025, 2007.

Xia, F., Liu, T. Y., Wang, J., Zhang, W., and Li, H. Listwise approach to learning to rank – theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

A. Auxiliary Proofs

Proof of Lemma 1 The proof of this lemma is analogous to Zhang's Theorem 24. Jensen's inequality implies

$$\begin{aligned}
&\zeta\left(\mathbb{E}(\ell(\boldsymbol{p},\boldsymbol{f})-\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p},\boldsymbol{\alpha}))\right)\\
&\leq \mathbb{E}\zeta(\ell(\boldsymbol{p},\boldsymbol{f})-\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p},\boldsymbol{\alpha}))\\
&\leq \mathbb{E}_{\boldsymbol{X},Q}H(\ell(\boldsymbol{p},\boldsymbol{f})-\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p},\boldsymbol{\alpha}))\\
&\leq \mathbb{E}_{\boldsymbol{X},Q}\left(\sum_G p_G\varphi(\boldsymbol{f}(\boldsymbol{X},Q),G)-\inf_{\boldsymbol{\alpha}'}\sum_G p_G\varphi(\boldsymbol{\alpha}',G)\right)\\
&= R_\varphi(\boldsymbol{f})-R^*_\varphi.
\end{aligned}$$

The second to last line is a consequence of the fact that $H(\ell(\boldsymbol{p},\boldsymbol{\alpha})-\inf_{\boldsymbol{\alpha}'}\ell(\boldsymbol{p},\boldsymbol{\alpha}'))\leq \sum_G p_G\varphi(\boldsymbol{\alpha},G)-\inf_{\boldsymbol{\alpha}'}\sum_G p_G\varphi(\boldsymbol{\alpha}',G)$ for $\boldsymbol{p}\in\Delta$ and any $\boldsymbol{\alpha}$. □

Proof of Lemma 3 The proof of this lemma is entirely similar to the proofs of lemma 16 from (Tewari & Bartlett, 2007) and lemma 27 from (Zhang, 2004), but we include it for completeness.

Let $\boldsymbol{p}^{(n)}$ be a sequence converging to $\boldsymbol{p}$ and B_r be a closed ball of radius r in $\mathbb{R}^{|\mathcal{G}|}$. Since $\sum_G p_G^{(n)}\varphi(\boldsymbol{\alpha},G)\to\sum_G\varphi(\boldsymbol{\alpha},G)$ uniformly in $\boldsymbol{\alpha}\in B_r$ (we have equicontinuity),

$$\inf_{\boldsymbol{\alpha}\in B_r}\sum_{G\in\mathcal{G}}p_G^{(n)}\varphi(\boldsymbol{\alpha},G)\to\inf_{\boldsymbol{\alpha}\in B_r}\sum_{G\in\mathcal{G}}p_G\varphi(\boldsymbol{\alpha},G).$$

We then have $W^*(\boldsymbol{p}^{(n)})\leq\inf_{\boldsymbol{\alpha}\in B_r}\sum_G p_G^{(n)}\varphi(\boldsymbol{\alpha},G)\to\inf_{\boldsymbol{\alpha}\in B_r}\sum_{G\in\mathcal{G}}p_G\varphi(\boldsymbol{\alpha},G)$, and hence (as the infimum is bounded below) we can let $r\to\infty$ and have

$$\limsup_n W^*(\boldsymbol{p}^{(n)})\leq\inf_{\boldsymbol{\alpha}}\sum_{G\in\mathcal{G}}p_G\varphi(\boldsymbol{\alpha},G)=W^*(\boldsymbol{p}).$$

To bound the limit infimum, assume without loss of generality that there is some subset $\mathcal{G}'\subset\mathcal{G}$ so that $p_G=0$ for all $G\notin\mathcal{G}'$. Define

$$\bar{W}(\boldsymbol{p},\boldsymbol{\alpha})=\sum_{G\in\mathcal{G}'}p_G\varphi(\boldsymbol{\alpha},G)\quad\text{and}\quad\bar{W}^*(\boldsymbol{p})=\inf_{\boldsymbol{\alpha}}\bar{W}(\boldsymbol{p},\boldsymbol{\alpha}).$$

Note that $\bar{W}^*(\boldsymbol{p}^{(n)})$ is eventually bounded so that each sequence of surrogate loss terms $p_G^{(n)}\varphi(\cdot,G)$ is also eventually bounded. Thus we have the desired uniform convergence, and

$$\liminf_n W^*(\boldsymbol{p}^{(n)})\geq\liminf_n\bar{W}^*(\boldsymbol{p}^{(n)})=W^*(\boldsymbol{p}).$$

This completes the proof that $W^*(\boldsymbol{p}^{(n)})\to W^*(\boldsymbol{p})$. □

Proof of Lemma 4 Suppose that $\ell(\boldsymbol{p},\boldsymbol{\alpha}^{(n)})\not\to\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p},\boldsymbol{\alpha})$. Then there is a subsequence $\boldsymbol{\alpha}^{(n_j)}$ and $\varepsilon>0$ such that $\ell(\boldsymbol{p},\boldsymbol{\alpha}^{(n_j)})\geq\ell(\boldsymbol{p},\boldsymbol{\alpha})+\varepsilon$. This in turn, since there is a finite set of orderings of the entries in $\boldsymbol{\alpha}$, implies that $\boldsymbol{\alpha}^{(n_j)}\notin A(\boldsymbol{p})$. But by the definition of edge-consistency,

$$W(\boldsymbol{p},\boldsymbol{\alpha}^{(n_j)})>\inf_{\boldsymbol{\alpha}}\{W(\boldsymbol{p},\boldsymbol{\alpha}):\boldsymbol{\alpha}\notin A(\boldsymbol{p})\}>W^*(\boldsymbol{p}).$$

The $W(\boldsymbol{p},\boldsymbol{\alpha}^{(n_j)})$ are thus bounded uniformly away from $W^*(\boldsymbol{p})$, a contradiction. □

Proof of Lemma 5 We prove this by contradiction and using the continuity result in Lemma 3. Assume that the statement does not hold, so that there is a sequence $(\boldsymbol{p}^{(n)},\boldsymbol{\alpha}^{(n)})$ with $\ell(\boldsymbol{p}^{(n)},\boldsymbol{\alpha}^{(n)})-\inf_{\boldsymbol{\alpha}'}\ell(\boldsymbol{p}^{(n)},\boldsymbol{\alpha}')\geq\varepsilon$ but $W(\boldsymbol{p}^{(n)},\boldsymbol{\alpha}^{(n)})-W^*(\boldsymbol{p}^{(n)})\to 0$. Because $\Delta_{|\mathcal{G}|}$ is compact, we choose a convergent subsequence n_j so that $\boldsymbol{p}^{(n_j)}\to\boldsymbol{p}$ for some $\boldsymbol{p}\in\Delta_{|\mathcal{G}|}$. By Lemma 3 and our assumption that $W(\boldsymbol{p}^{(n_j)},\boldsymbol{\alpha}^{(n_j)})-W^*(\boldsymbol{p}^{(n_j)})\to 0$, $W(\boldsymbol{p}^{(n_j)},\boldsymbol{\alpha}^{(n_j)})\to W^*(\boldsymbol{p})$. Similar to the proof of the previous lemma, we assume that there is a set $\mathcal{G}'\subset\mathcal{G}$ with $p_G>0$ for $G\in\mathcal{G}'$. We have

$$\begin{aligned}
\limsup_{n_j}\sum_{G\in\mathcal{G}}p_G\varphi(\boldsymbol{\alpha}^{(n_j)},G)&=\limsup_{n_j}\sum_{G\in\mathcal{G}'}p_G^{(n_j)}\varphi(\boldsymbol{\alpha}^{(n_j)},G)\\
&\leq\lim_{n_j}\sum_{G\in\mathcal{G}}p_G^{(n_j)}\varphi(\boldsymbol{\alpha}^{(n_j)},G)=W^*(\boldsymbol{p}).
\end{aligned}$$

The above proves that $W(\boldsymbol{p},\boldsymbol{\alpha}^{(n_j)})\to W^*(\boldsymbol{p})$, and Lemma 4 implies that $\ell(\boldsymbol{p},\boldsymbol{\alpha}^{(n_j)})\to\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p},\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}^{(n_j)}\in A(\boldsymbol{p})$ eventually. The continuity of ℓ in $\boldsymbol{p}$ and the fact that $\boldsymbol{p}^{(n_j)}\to\boldsymbol{p}$, however, contradicts $\ell(\boldsymbol{p}^{(n)},\boldsymbol{\alpha}^{(n)})-\inf_{\boldsymbol{\alpha}}\ell(\boldsymbol{p}^{(n)},\boldsymbol{\alpha})\geq\varepsilon$. □

Proof of Lemma 7 As stated earlier, this is a straightforward consequence of the fact that the feedback arc set problem (Karp, 1972) is NP-complete. In the feedback arc set problem, we are given a directed graph $G=(V,E)$ and integer k and need to determine whether there is a subset $E'\subseteq E$ with $|E'|\leq k$ such that E' contains at least one edge from every directed cycle in G, or, equivalently, that $G'=(V,E\setminus E')$ is a DAG.

Now consider the problem of deciding whether there exists a $\boldsymbol{\alpha}$ with $\ell(\boldsymbol{p},\boldsymbol{\alpha})\leq k$, and let $\overline{G}$ be the graph over all the nodes in the graphs in $\mathcal{G}$ with average edge weights $a_{ij}=\sum_{G\in\mathcal{G}}a_{ij}^G p_G$. Since $\boldsymbol{\alpha}$ induces an order of the nodes in this expected graph $\overline{G}$, this is equivalent

to finding an ordering of the nodes $i_1, \ldots, i_n$ (denoted $i_1 \prec i_2 \prec \cdots \prec i_n$) in $\overline{G}$ such that the sum of the back edges is less than k, i.e.,

$$\sum_{i \succeq j} a_{ij} \leq k.$$

Further, it is clear that removing all the back edges (edges $(i \to j)$ in the expected graph $\overline{G}$ such that $i \succeq j$ in the order) leaves a DAG. Now given a graph $G = (V, E)$, we can construct the expected graph $\overline{G}$ directly from G with weights $a_{ij} = 1$ if $(i \to j) \in E$ and 0 otherwise (to be pedantic, set $p_{ij} = 1/|E|$ and let the ij^{th} graph in the set of possible graphs $\mathcal{G}$ be simply the edge $(i \to j)$ with weight $a_{ij} = |E|$). Then it is clear that there is a $\boldsymbol{\alpha}$ such that $\ell(\boldsymbol{p}, \boldsymbol{\alpha}) \leq k$ if and only if there is a feedback arc set E' with $|E'| \leq k$. □

Proof of Lemma 9 Let $(\boldsymbol{\alpha}^{(n)})_{n=1}^{\infty}$ be a sequence with $\boldsymbol{\alpha}^{(n)} \in A(\boldsymbol{p}) = \{\boldsymbol{\alpha} : \alpha_1 > \alpha_2 > \alpha_3\}$ such that $W(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) \to W^*(\boldsymbol{p})$. Now suppose that $\limsup_n(\alpha_i^{(n)} - \alpha_j^{(n)}) = \infty$ for some $i < j$. Since $\boldsymbol{\alpha}^{(n)} \in A(\boldsymbol{p})$, this implies $\limsup_n(\alpha_1^{(n)} - \alpha_3^{(n)}) = \infty$. But ϕ is convex with $\phi'(0) < 0$ and so is unbounded above, so certainly $\limsup_n \phi(\alpha_3^{(n)} - \alpha_1^{(n)}) = \infty$. The assumption that $\phi'_\infty(1) = 0$ or $\phi'_\infty(-1) = -\infty$ then implies that $\limsup_n W(\boldsymbol{p}, \boldsymbol{\alpha}^{(n)}) = \infty$. This contradiction gives that there must be some C with $|\alpha_i^{(n)} - \alpha_j^{(n)}| \leq C$ for all i, j, n. $W(\boldsymbol{p}, \boldsymbol{\alpha})$ is shift invariant with respect to $\boldsymbol{\alpha}$, so without loss of generality we can let $\alpha_3^{(n)} = 0$, and thus $|\alpha_i^{(n)}| \leq C$. Convex functions are continuous on compact domains (Rockafellar, 1970, Chapter 10), and $\inf_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_\infty \leq C} W(\boldsymbol{p}, \boldsymbol{\alpha}) = W^*(\boldsymbol{p})$, which proves the lemma. □

Lemma 16. *Let $\widetilde{G}$ be the difference graph for the pair $(\mathcal{G}, \boldsymbol{p})$. $\widetilde{G}$ is a DAG if and only if*

$$A(\boldsymbol{p}) = \{\boldsymbol{\alpha} : \alpha_i > \alpha_j \text{ for } i < j \text{ with } a_{ij} > a_{ji},\\ \alpha_i \geq \alpha_j \text{ for } i > j \text{ with } a_{ij} > a_{ji}\}.$$

Proof We first prove that if $\widetilde{G}$ is a DAG, then $A(\boldsymbol{p})$ is the set above. Assume without loss of generality that the nodes in $\widetilde{G}$ are topologically sorted in the order $1, 2, \ldots, m$, so that $a_{ij} \geq a_{ji}$ for $i < j$. Then we can write

$$\begin{aligned}\ell(\boldsymbol{p}, \boldsymbol{\alpha}) &= \sum_{i<j} a_{ij} 1_{(\alpha_i \leq \alpha_j)} + \sum_{i>j} a_{ij} 1_{(\alpha_i < \alpha_j)} \\ &= \sum_{i=1}^{m} \sum_{j=i+1}^{m} a_{ij} 1_{(\alpha_i \leq \alpha_j)} + a_{ji} 1_{(\alpha_j < \alpha_i)} \\ &= \sum_{i=1}^{m} \sum_{j=i+1}^{m} (a_{ij} - a_{ji}) 1_{(\alpha_i \leq \alpha_j)} + a_{ji}.\end{aligned}$$

Clearly, if $a_{ij} > a_{ji}$ for some $i < j$, then any minimizing $\boldsymbol{\alpha}$ must satisfy $\alpha_i > \alpha_j$. If $a_{ij} - a_{ji} = 0$ for some $j > i$, then the relative order of α_i and α_j does not affect $\ell(\boldsymbol{p}, \boldsymbol{\alpha})$.

Now suppose that $\widetilde{G}$ is not a DAG but that $A(\boldsymbol{p})$ takes the form described in the statement of the lemma. In this case, there are (at least) two nodes i and j with a path going from i to j and from j to i in $\widetilde{G}$. Let the nodes on the path from i to j be $l_1, \ldots, l_k$ and from j to i be $l'_1, \ldots, l'_{k'}$. By assumption, we must have that for any $\boldsymbol{\alpha} \in A(\boldsymbol{p})$,

$$\alpha_i \geq \alpha_{l_1} \geq \cdots \geq \alpha_{l_k} \geq \alpha_j \geq \alpha_{l'_1} \geq \cdots \geq \alpha_{l'_{k'}} \geq \alpha_i.$$

One of the inequalities must be strict because we will have some l or $l' > i$, which is clearly a contradiction. □