THE ASYMPTOTICS OF RANKING ALGORITHMS

BY JOHN C. DUCHI[1,2], LESTER MACKEY[1,2] AND MICHAEL I. JORDAN[2]

*University of California, Berkeley, Stanford University
and University of California, Berkeley*

We consider the predictive problem of supervised ranking, where the task is to rank sets of candidate items returned in response to queries. Although there exist statistical procedures that come with guarantees of consistency in this setting, these procedures require that individuals provide a complete ranking of all items, which is rarely feasible in practice. Instead, individuals routinely provide partial preference information, such as pairwise comparisons of items, and more practical approaches to ranking have aimed at modeling this partial preference data directly. As we show, however, such an approach raises serious theoretical challenges. Indeed, we demonstrate that many commonly used surrogate losses for pairwise comparison data do not yield consistency; surprisingly, we show inconsistency even in low-noise settings. With these negative results as motivation, we present a new approach to supervised ranking based on aggregation of partial preferences, and we develop U-statistic-based empirical risk minimization procedures. We present an asymptotic analysis of these new procedures, showing that they yield consistency results that parallel those available for classification. We complement our theoretical results with an experiment studying the new procedures in a large-scale web-ranking task.

1. Introduction. Recent years have seen significant developments in the theory of classification, most notably binary classification, where strong theoretical results are available that quantify rates of convergence and shed light on qualitative aspects of the problem [3, 45]. Extensions to multi-class classification have also been explored, and connections to the theory of regression are increasingly well understood, so that overall a satisfactory theory of supervised machine learning has begun to emerge [40, 44].

In many real-world problems in which labels or responses are available, however, the problem is not merely to classify or predict a real-valued response, but rather to list a set of items in order. The theory of supervised learning cannot be

considered complete until it also provides a treatment of such *ranking* problems. For example, in information retrieval, the goal is to rank a set of documents in order of relevance to a user's search query; in medicine, the object is often to rank drugs in order of probable curative outcomes for a given disease; and in recommendation or advertising systems, the aim is to present a set of products in order of a customer's willingness to purchase or consume. In each example, the intention is to order a set of items in accordance with the preferences of an individual or population. While such problems are often converted to classification problems for simplicity (e.g., a document is classified as "relevant" or not), decision makers frequently require the ranks (e.g., a search engine must display documents in a particular order on the page). Despite its ubiquity, our statistical understanding of ranking falls short of our understanding of classification and regression. Our aim here is to characterize the statistical behavior of computationally tractable inference procedures for ranking under natural data-generating mechanisms.

We consider a general decision-theoretic formulation of the *supervised ranking problem* in which preference data are drawn i.i.d. from an unknown distribution, where each datum consists of a *query*, $Q \in \mathcal{Q}$, and a *preference judgment*, $Y \in \mathcal{Y}$, over a set m of candidate items that are available based on the query Q. The exact nature of the query and preference judgment depend on the ranking context. In the setting of information retrieval, for example, each datum corresponds to a user issuing a natural language query and expressing a preference by selecting or clicking on zero or more of the returned results. The statistical task is to discover a function that provides a query-specific ordering of items that best respects the observed preferences. This query-indexed setting is especially natural for tasks like information retrieval in which a different ranking of webpages is needed for each natural language query.

Following existing literature, we estimate a *scoring function* $f : \mathcal{Q} \to \mathbb{R}^m$, where $f(q)$ assigns a score to each of m candidate items for the query q, and the results are ranked according to their scores [22, 24]. Throughout the paper, we adopt a decision-theoretic perspective and assume that given a query-judgment pair (Q, Y), we evaluate the scoring function f via a loss $L(f(Q), Y)$. The goal is to choose the f minimizing the risk

$$(1) \qquad\qquad R(f) := \mathbb{E}\big[L(f(Q), Y)\big].$$

While minimizing the risk (1) directly is in general intractable, researchers in machine learning and information retrieval have developed surrogate loss functions that yield procedures for selecting f. Unfortunately, as we show, extant procedures fail to solve the ranking problem under reasonable data generating mechanisms. The goal in the remainder of the paper is to explain this failure and to propose a novel solution strategy based on preference aggregation.

Let us begin to elucidate the shortcomings of current approaches to ranking. One main problem lies in their unrealistic assumptions about available data. The

losses proposed and most commonly used for evaluation in the information retrieval literature [27, 30] have a common form, generally referred to as (Normalized) Discounted Cumulative Gain ((N)DCG). The NDCG family requires that the preference judgments Y associated with the datum (Q, Y) be a vector $Y \in \mathbb{R}^m$ of *relevance scores* for the entire set of items; that is, Y_j denotes the real-valued relevance of item j to the query Q. While having complete preference information makes it possible to design procedures that asymptotically minimize NDCG losses (e.g., [12]), in practice such complete preferences are unrealistic: they are expensive to collect and difficult to trust. In biological applications, evaluating the effects of all drugs involved in a study—or all doses—on a single subject is infeasible. In web search, users click on only one or two results: no feedback is available for most items. Even when practical and ethical considerations do not preclude collecting complete preference information from participants in a study, a long line of psychological work highlights the inconsistency with which humans assign numerical values to multiple objects (e.g., [31, 39, 41]).

The inherent practical difficulties that arise in using losses based on relevance scores has led other researchers to propose loss functions that are suitable for *partial preference data* [16, 22, 28]. Such data arise naturally in a number of real-world situations; for example, a patient's prognosis may improve or deteriorate after administration of treatment, competitions and sporting matches provide paired results, and shoppers at a store purchase one item but not others. Moreover, the psychological literature shows that human beings are quite good at performing pairwise distinctions and forming relative judgments (see, e.g., [37] and references therein).

More formally, let $\alpha := f(Q) \in \mathbb{R}^m$ denote the vector of predicted scores for each item associated with query Q. If a preference Y indicates that item i is preferred to j then the natural associated loss is the zero-one loss $L(\alpha, Y) = 1(\alpha_i \leq \alpha_j)$. Minimizing such a loss is well known to be computationally intractable; nonetheless, the classification literature [3, 40, 44, 45] has shown that it is possible to design convex Fisher-consistent surrogate losses for the 0–1 loss in classification settings and has linked Fisher consistency to consistency. By reduction to classification, similar consistency results are possible in certain bipartite or binary ranking scenarios [10]. One might therefore hope to make use of these surrogate losses in the ranking setting to obtain similar guarantees. Unfortunately, however, this hope is not borne out; as we illustrate in Section 3, it is generally computationally intractable to minimize any Fisher-consistent loss for ranking, and even in favorable low-noise cases, convex surrogates that yield Fisher consistency for binary classification fail to be Fisher-consistent for ranking.

We find ourselves at an impasse: existing methods based on practical data-collection strategies do not yield a satisfactory theory, and those methods that do have theoretical justification are not practical. Our approach to this difficulty is to take a new approach to supervised ranking problems in which partial preference data are aggregated before being used for estimation. The point of departure for

this approach is the notion of *rank aggregation* (e.g., [21]), which has a long history in voting [15], social choice theory [2, 11] and statistics [29, 42]. In Section 2, we discuss some of the ways in which partial preference data can be aggregated, and we propose a new family of U-statistic-based loss functions that are computationally tractable. Sections 3 and 4 present a theoretical analysis of procedures based on these loss functions, establishing their consistency. We provide a further discussion of practical rank aggregation strategies in Section 5 and present experimental results in Section 6. Section 7 contains our conclusions, with proofs deferred to appendices.

2. Ranking with rank aggregation. We begin by considering several ways in which partial preference data arise in practice. We then turn to a formal treatment of our aggregation-based strategy for supervised ranking.

1. *Paired comparison data*. Data in which an individual judges one item to be preferred over another in the context of a query are common. Competitions and sporting matches, where each pairwise comparison may be accompanied by a magnitude such as a difference of scores, naturally generate such data. In practice, a single individual will not provide feedback for all possible pairwise comparisons, and we do not assume transitivity among the observed preferences for an individual. Thus, it is natural to model the pairwise preference judgment space $\mathcal{Y}$ as the set of weighted directed graphs on m nodes.

2. *Selection data*. A ubiquitous source of partial preference information is the selection behavior of a user presented with a small set of potentially ordered items. For example, in response to a search query, a web search engine presents an ordered list of webpages and records the URL a user clicks on, and a store records inventory and tracks the items customers purchase. Such selections provide partial information: that a user or customer prefers one item to others presented.

3. *Partial orders*. An individual may also provide preference feedback in terms of a partial ordering over a set of candidates or items. In the context of elections, for example, each preference judgment $Y \in \mathcal{Y}$ specifies a partial order $\prec_Y$ over candidates such that candidate i is preferred to candidate j whenever $i \prec_Y j$. A partial order need not specify a preference between every pair of items.

Using these examples as motivation, we wish to develop a formal treatment of ranking based on aggregation. To provide intuition for the framework presented in the remainder of this section, let us consider a simple aggregation strategy appropriate for the case of paired comparison data. Let each relevance judgment $Y \in \mathcal{Y}$ be a weighted adjacency matrix where the (i, j)th entry expresses a preference for item i over j whenever this entry is nonzero. In this case, a natural aggregation strategy is to average all observed adjacency matrices for a fixed query. Specifically, for a set of adjacency matrices $\{Y_l\}_{l=1}^k$ representing user preferences for a given query, we form the average $(1/k) \sum_{l=1}^k Y_l$. As $k \to \infty$, the average adjacency matrix captures the mean population preferences, and we thereby obtain complete preference information over the m items.

This averaging of partial preferences is one example of a general class of aggregation strategies that form the basis of our theoretical framework. To formalize this notion, we modify the loss formulation slightly and hereafter assume that the loss function L is a mapping $\mathbb{R}^m \times \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ is a problem-specific *structure space*. We further assume the existence of a series of *structure functions*, $s_k : \mathcal{Y}^k \to \mathcal{S}$, that map sets of preference judgments $\{Y_j\}$ into $\mathcal{S}$. The loss L depends on the preference feedback $(Y_1, \ldots, Y_k)$ for a given query only via the structure $s_k(Y_1, \ldots, Y_k)$. In the example of the previous paragraph, $\mathcal{S}$ is the set of $m \times m$ adjacency matrices, and $s_k(Y_1, \ldots, Y_k) = (1/k) \sum_{l=1}^{k} Y_l$. A typical loss for this setting is the pairwise loss [22, 28]

$$L(\alpha, s(Y_1, \ldots, Y_k)) \equiv L(\alpha, A) := \sum_{i<j} A_{ij} \mathbf{1}(\alpha_i \leq \alpha_j) + \sum_{i>j} A_{ij} \mathbf{1}(\alpha_i < \alpha_j),$$

where α is a set of scores and $A = s_k(Y_1, \ldots, Y_k)$ is the average adjacency matrix with entries A_{ij}. In Section 5, we provide other examples of structure functions for different data collection mechanisms and losses. Hereafter, we abbreviate $s_k(Y_1, \ldots, Y_k)$ as $s(Y_1, \ldots, Y_k)$ whenever the input length k is clear from context.

To meaningfully characterize the asymptotics of inference procedures, we make a mild assumption on the limiting behavior of the structure functions.

ASSUMPTION A. Fix a query $Q = q$. Let the sequence $Y_1, Y_2, \ldots$ be drawn i.i.d. conditional on q, and define the random variables $S_k := s(Y_1, \ldots, Y_k)$. If μ_q^k denotes the distribution of S_k, there exists a limiting law μ_q such that

$$\mu_q^k \overset{d}{\to} \mu_q \qquad \text{as } k \to \infty.$$

For example, the averaging structure function satisfies Assumption A so long as $\mathbb{E}[|Y_{ij}| \mid Q] < \infty$ with probability 1. Aside from the requirements of Assumption A, we allow arbitrary aggregation within the structure function.

In addition, our main assumption on the loss function L is as follows:

ASSUMPTION B. The loss function $L : \mathbb{R}^m \times \mathcal{S} \to \mathbb{R}$ is bounded in $[0, 1]$, and, for any fixed vector $\alpha \in \mathbb{R}^m$, $L(\alpha, \cdot)$ is continuous in the topology of $\mathcal{S}$.

With our assumptions on the asymptotics of the structure function s and the loss L in place, we now describe the risk functions that guide our design of inference procedures. We begin with the pointwise conditional risk, which maps predicted scores and a measure μ on $\mathcal{S}$ to $[0, 1]$:

$$(2) \qquad \ell : \mathbb{R}^m \times \mathcal{M}(\mathcal{S}) \to [0, 1] \qquad \text{where } \ell(\alpha, \mu) := \int L(\alpha, s) \, d\mu(s).$$

Here, $\mathcal{M}(\mathcal{S})$ denotes the closure of the subset of probability measures on the set $\mathcal{S}$ for which ℓ is defined. For any query q and $\alpha \in \mathbb{R}^m$, we have $\lim_i \ell(\alpha, \mu_q^i) =$

$\ell(\alpha, \mu_q)$ by the definition of convergence in distribution. This convergence motivates our decision-theoretic approach.

Our goal in ranking is thus to minimize the risk

$$(3) \qquad R(f) := \sum_q p_q \ell(f(q), \mu_q),$$

where p_q denotes the probability that the query $Q = q$ is issued. The risk of the scoring function f can also be obtained in the limit as the number of preference judgments for each query goes to infinity:

$$(4) \qquad R(f) = \lim_k \mathbb{E}[L(f(Q), s(Y_1, \ldots, Y_k))] = \lim_k \sum_q p_q \ell(f(q), \mu_q^k).$$

That the limiting expectation (4) is equal to the risk (3) follows from the definition of weak convergence.

We face two main difficulties in the study of the minimization of the risk (3). The first difficulty is that of *Fisher consistency* mentioned previously: since L may be nonsmooth in the function f and is typically intractable to minimize, when will the minimization of a tractable surrogate lead to the minimization of the loss (3)? We provide a precise formulation of and answer to this question in Section 3. In addition, we demonstrate the inconsistency of many commonly used pairwise ranking surrogates and show that aggregation leads to tractable Fisher consistent inference procedures for both complete and partial data losses.

The second difficulty is that of *consistency*: for a given Fisher consistent surrogate for the risk (3), are there tractable statistical procedures that converge to a minimizer of the risk? Yes: in Section 4, we develop a new family of aggregation losses based on U-statistics of increasing order, showing that uniform laws of large numbers hold for the resulting M-estimators.

3. Fisher consistency of surrogate risk minimization. In this section, we formally define the Fisher consistency of a surrogate loss and give general necessary and sufficient conditions for consistency to hold for losses satisfying Assumption B. To begin, we assume that the space $\mathcal{Q}$ of queries is countable (or finite) and thus bijective with $\mathbb{N}$. Recalling the definition (3) of the risk and the pointwise conditional risk (2), we define the Bayes risk for R as the minimal risk over all measurable functions $f : \mathcal{Q} \to \mathbb{R}^m$:

$$R^* := \inf_f R(f) = \sum_q p_q \inf_{\alpha \in \mathbb{R}^m} \ell(\alpha, \mu_q).$$

The second equality follows because $\mathcal{Q}$ is countable and the infimum is taken over all measurable functions.

Since it is infeasible to minimize the risk (3) directly, we consider a bounded-below surrogate φ to minimize in place of L. For each structure $s \in \mathcal{S}$, we write $\varphi(\cdot, s) : \mathbb{R}^m \to \mathbb{R}_+$, and we assume that for $\alpha \in \mathbb{R}^m$, the function $s \mapsto \varphi(\alpha, s)$ is

continuous with respect to the topology on $\mathcal{S}$. We then define the conditional φ-risk as

$$\ell_\varphi(\alpha, \mu) := \int_{\mathcal{S}} \varphi(\alpha, s) \, d\mu(s) \tag{5}$$

and the asymptotic φ-risk of the function f as

$$R_\varphi(f) := \sum_q p_q \ell_\varphi(f(q), \mu_q), \tag{6}$$

whenever each $\ell_\varphi(f(q), \mu_q)$ exists [otherwise $R_\varphi(f) = +\infty$]. The optimal φ-risk is defined to be $R_\varphi^* := \inf_f R_\varphi(f)$, and throughout we make the assumption that there exist measurable f such that $R_\varphi(f) < +\infty$ so that R_φ^* is finite. The following is our general notion of Fisher consistency.

DEFINITION 1. The surrogate loss φ is *Fisher-consistent* for the loss L if for any $\{p_q\}$ and probability measures $\mu_q \in \mathcal{M}(\mathcal{S})$, the convergence

$$R_\varphi(f_n) \to R_\varphi^* \qquad \text{implies } R(f_n) \to R^*.$$

To achieve more actionable risk bounds and to more accurately compare surrogate risks, we also draw upon a uniform statement of consistency:

DEFINITION 2. The surrogate loss φ is *uniformly Fisher-consistent* for the loss L if for any $\varepsilon > 0$, there exists a $\delta(\varepsilon) > 0$ such that for any $\{p_q\}$ and probability measures $\mu_q \in \mathcal{M}(\mathcal{S})$,

$$R_\varphi(f) < R_\varphi^* + \delta(\varepsilon) \qquad \text{implies } R(f) < R^* + \varepsilon. \tag{7}$$

The bound (7) is equivalent to the assertion that there exists a nondecreasing function ζ such that $\zeta(0) = 0$ and $R(f) - R^* \leq \zeta(R_\varphi(f) - R_\varphi^*)$. Bounds of this form have been completely characterized in the case of binary classification [3], and Steinwart [40] has given necessary and sufficient conditions for uniform Fisher-consistency to hold in general risk minimization problems. We now turn to analyzing conditions under which a surrogate loss φ is Fisher-consistent for ranking.

3.1. *General theory.* The main approach in establishing conditions for the surrogate risk Fisher consistency in Definition 1 is to move from global conditions for Fisher consistency to local, pointwise Fisher consistency. Following the treatment of Steinwart [40], we begin by defining a function measuring the discriminating ability of the surrogate φ:

$$H(\varepsilon) := \inf_{\mu \in \mathcal{M}(\mathcal{S}), \alpha} \left\{ \ell_\varphi(\alpha, \mu) - \inf_{\alpha'} \ell_\varphi(\alpha', \mu) \mid \ell(\alpha, \mu) - \inf_{\alpha'} \ell(\alpha', \mu) \geq \varepsilon \right\}. \tag{8}$$

This function is familiar from work on surrogate risk Fisher consistency in classification [3] and measures surrogate risk suboptimality as a function of risk suboptimality. A reasonable conditional φ-risk will declare a set of scores $\alpha \in \mathbb{R}^m$ suboptimal whenever the conditional risk ℓ declares them suboptimal. This corresponds to $H(\varepsilon) > 0$ whenever $\varepsilon > 0$, and we call any loss satisfying this condition *pointwise consistent*.

From these definitions, we can conclude the following consistency result, which is analogous to the results of [40]. For completeness, we provide a proof in the supplementary material [18].

PROPOSITION 1. *Let $\varphi : \mathbb{R}^m \times \mathcal{S} \to \mathbb{R}_+$ be a bounded-below loss function such that for some f, $R_\varphi(f) < +\infty$. Then φ is pointwise consistent if and only if the uniform Fisher-consistency definition* (7) *holds.*

Proposition 1 makes it clear that pointwise consistency for general measures μ on the set of structures $\mathcal{S}$ is a stronger condition than that of Fisher consistency in Definition 1. In some situations, however, it is possible to connect the weaker surrogate risk Fisher consistency of Definition 1 with uniform Fisher consistency and pointwise consistency. Ranking problems with appropriate choices of the space $\mathcal{S}$ give rise to such connections. Indeed, consider the following:

ASSUMPTION C. The space of possible structures $\mathcal{S}$ is finite, and the loss L is discrete, meaning that it takes on only finitely many values.

Binary and multiclass classification provide examples of settings in which Assumption C is appropriate, since the set of structures $\mathcal{S}$ is the set of class labels, and L is usually a version of the 0–1 loss. We also sometimes make a weaker version of Assumption C:

ASSUMPTION C′. The (topological) space of possible structures $\mathcal{S}$ is compact, and for some $d \in \mathbb{N}$ there exists a partition $\mathcal{A}_1, \ldots, \mathcal{A}_d$ of $\mathbb{R}^m$ such that for any $s \in \mathcal{S}$,

$$L(\alpha, s) = L(\alpha', s) \qquad \text{whenever } \alpha, \alpha' \in \mathcal{A}_i.$$

Assumption C′ may be more natural in ranking settings than Assumption C. The compactness assumption holds, for example, if $\mathcal{S} \subset \mathbb{R}^m$ is closed and bounded, such as in our pairwise aggregation example in Section 2. Losses L that depend only on the relative order of the coordinate values of $\alpha \in \mathbb{R}^m$—common in ranking problems—provide a collection of examples for which the partitioning condition holds.

Under Assumption C or C′, we can provide a definition of local consistency that is often more user-friendly than pointwise consistency (8):

DEFINITION 3. Let φ be a bounded-below surrogate loss such that $\varphi(\cdot, s)$ is continuous for all $s \in \mathcal{S}$. The function φ is *structure-consistent* with respect to the loss L if for all $\mu \in \mathcal{M}(\mathcal{S})$,

$$\ell_\varphi^*(\mu) := \inf_\alpha \ell_\varphi(\alpha, \mu) < \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) \mid \alpha \notin \operatorname*{argmin}_{\alpha'} \ell(\alpha', \mu) \right\}.$$

Definition 3 describes the set of loss functions φ satisfying the intuitively desirable property that the surrogate φ cannot be minimized if the scores $\alpha \in \mathbb{R}^m$ are restricted to not minimize the loss L. As we see presently, Definition 3 captures exactly what it means for a surrogate loss φ to be Fisher-consistent when one of Assumptions C or C' holds. Moreover, the set of Fisher-consistent surrogates coincides with the set of uniformly Fisher-consistent surrogates in this case. The following theorem formally states this result; we give a proof in the supplementary material [18].

THEOREM 1. *Let* $\varphi : \mathbb{R}^m \times \mathcal{S} \to \mathbb{R}_+$ *satisfy* $R_\varphi(f) < +\infty$ *for some measurable* f. *If Assumption* C *holds, then*:

(a) *If* φ *is structure consistent* (*Definition* 3), *then* φ *is uniformly Fisher-consistent for the loss* L (*Definition* 2).

(b) *If* φ *is Fisher-consistent for the loss* L (*Definition* 1), *then* φ *is structure consistent*.

If the function $\varphi(\cdot, s)$ *is convex for* $s \in \mathcal{S}$, *and for* $\mu \in \mathcal{M}(\mathcal{S})$ *the conditional risk* $\ell_\varphi(\alpha, \mu) \to \infty$ *as* $\|\alpha\| \to \infty$, *then Assumption* C' *implies* (a) *and* (b).

Theorem 1 shows that as long as Assumption C holds, pointwise consistency, structure consistency, and both uniform and nonuniform surrogate loss consistency coincide. These four also coincide under the weaker Assumption C' so long as the surrogate is 0-coercive, which is not restrictive in practice. As a final note, we recall a result due to Steinwart [40], which gives general necessary and sufficient conditions for the consistency in Definition 1 to hold, using a weaker version of the suboptimality function (8) that depends on μ:

$$(9) \quad H(\varepsilon, \mu) := \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) - \inf_{\alpha'} \ell_\varphi(\alpha', \mu) \mid \ell(\alpha, \mu) - \inf_{\alpha'} \ell(\alpha', \mu) \geq \varepsilon \right\}.$$

PROPOSITION 2 (Steinwart [40], Theorems 2.8 and 3.3). *The suboptimality function* (9) *satisfies* $H(\varepsilon, \mu_q) > 0$ *for any* $\varepsilon > 0$ *and* μ_q *with* $q \in \mathcal{Q}$ *and* $p_q > 0$ *if and only if* φ *is Fisher-consistent for the loss* L (*Definition* 1).

As a corollary of this result, any structure-consistent surrogate loss φ (in the sense of Definition 3) is Fisher-consistent for the loss L whenever the conditional risk $\ell(\alpha, \mu)$ has finite range, so that $\alpha \notin \operatorname{argmin}_{\alpha'} \ell(\alpha', \mu) \neq \varnothing$ implies the existence of an $\varepsilon > 0$ such that $\ell(\alpha, \mu) - \inf_{\alpha'} \ell(\alpha, \mu) \geq \varepsilon$.

3.2. *The difficulty of Fisher consistency for ranking.* We now turn to the question of whether there exist structure-consistent ranking losses. In a preliminary version of this work [19], we focused on the practical setting of learning from pairwise preference data and demonstrated that many popular ranking surrogates are inconsistent for standard pairwise ranking losses. We review and generalize our main inconsistency results here, noting that while the losses considered use pairwise preferences, they perform no aggregation. Their theoretically poor performance provides motivation for the aggregation strategies proposed in this work; we explore the connections in Section 5 (focusing on pairwise losses in Section 5.3). We provide proofs of our inconsistency results in the supplementary material [18].

To place ourselves in the general structural setting of the paper, we consider the structure function $s(Y_1, \ldots, Y_k) = Y_1$ which performs no aggregation for all k, and we let Y denote the weighted adjacency matrix of a directed acyclic graph (DAG) G, so that Y_{ij} is the weight of the directed edge $(i \to j)$ in the graph G. We consider a pairwise loss that imposes a separate penalty for each misordered pair of results:

$$(10) \qquad L(\alpha, Y) = \sum_{i < j} Y_{ij} 1(\alpha_i \leq \alpha_j) + \sum_{i > j} Y_{ij} 1(\alpha_i < \alpha_j),$$

where we distinguish the cases $i < j$ and $i > j$ to avoid doubly penalizing $1(\alpha_i = \alpha_j)$. When pairwise preference judgments are available, use of such losses is common. Indeed, this loss generalizes the disagreement error described by Dekel et al. [16] and is similar to losses used by Joachims [28]. If we define $Y_{ij}^\mu := \int Y_{ij} \, d\mu(Y)$, then

$$(11) \qquad \ell(\alpha, \mu) = \sum_{i < j} Y_{ij}^\mu 1(\alpha_i \leq \alpha_j) + \sum_{i > j} Y_{ij}^\mu 1(\alpha_i < \alpha_j).$$

We assume that the number of nodes in any graph G (or, equivalently, the number of results returned by any query) is bounded by a finite constant m. Hence, the conditional risk (11) has a finite range; if there are a finite number of preference labels Y or the set of weights is compact, Assumptions C or C′ are satisfied, whence Theorem 1 applies.

3.2.1. *General inconsistency.* Let the set P denote the complexity class of problems solvable in polynomial time and *NP* denote the class of nondeterministic polynomial time problems (see, e.g., [26]). Our first inconsistency result (see also [19], Lemma 7) is that unless $P = NP$ (a widely doubted proposition), any loss that is tractable to minimize cannot be a Fisher-consistent surrogate for the loss (10) and its associated risk.

PROPOSITION 3. *Finding an α minimizing ℓ is NP-hard.*

In particular, most convex functions are minimizable to an accuracy of ε in time polynomial in the dimension of the problem times a multiple of $\log \frac{1}{\varepsilon}$, known as poly-logarithmic time [4]. Since any α minimizing $\ell_\varphi(\alpha, \mu)$ must minimize $\ell(\alpha, \mu)$ for a Fisher-consistent surrogate φ, and $\ell(\cdot, \mu)$ has a finite range (so that optimizing ℓ_φ to a fixed ε accuracy is sufficient), convex surrogate losses are inconsistent for the pairwise loss (10) unless $P = NP$.

3.2.2. *Low-noise inconsistency.* We now turn to showing that, surprisingly, many common convex surrogates are inconsistent even in low-noise settings in which it is easy to find an α minimizing $\ell(\alpha, \mu)$. (Weaker versions of the results in this section appeared in our preliminary paper [19].) Inspecting the loss definition (10), a natural choice for a surrogate loss is one of the form [16, 22, 24]

$$(12) \qquad \varphi(\alpha, Y) = \sum_{i,j} h(Y_{ij}) \phi(\alpha_i - \alpha_j),$$

where $\phi \geq 0$ is a convex function, and h is a some function of the penalties Y_{ij}. This surrogate implicitly uses the structure function $s(Y_1, \ldots, Y_k) = Y_1$ and performs no preference aggregation. The conditional surrogate risk is thus $\ell_\varphi(\alpha, \mu) = \sum_{i \neq j} h_{ij} \phi(\alpha_i - \alpha_j)$, where $h_{ij} := \int h(Y_{ij}) d\mu(Y)$. Surrogates of the form (12) are convenient in margin-based binary classification, where the complete description by Bartlett, Jordan and McAuliffe [3] shows ϕ is Fisher-consistent if and only if it is differentiable at 0 with $\phi'(0) < 0$.

We now precisely define our low-noise setting. For any measure μ on a space $\mathcal{Y}$ of adjacency matrices, let the directed graph G_μ be the *difference graph*, that is, the graph with edge weights $\max\{Y_{ij}^\mu - Y_{ji}^\mu, 0\}$ on edges $(i \to j)$, where $Y_{ij}^\mu = \int Y_{ij} d\mu(Y)$. Then we say that the edge $(i \to j) \notin G_\mu$ if $Y_{ij}^\mu \leq Y_{ji}^\mu$ (see Figure 1). We define the following low-noise condition based on self-reinforcement of edges in the difference graph.

DEFINITION 4. The measure μ on a set Y of adjacency matrices is *low-noise* when the corresponding difference graph G_μ satisfies the following reverse triangle inequality: whenever there is an edge $(i \to j)$ and an edge $(j \to k)$ in G_μ, the weight $Y_{ik}^\mu - Y_{ki}^\mu$ on the edge $(i \to k)$ is greater than or equal to the path weight $Y_{ij}^\mu - Y_{ji}^\mu + Y_{jk}^\mu - Y_{kj}^\mu$ on the path $(i \to j \to k)$.
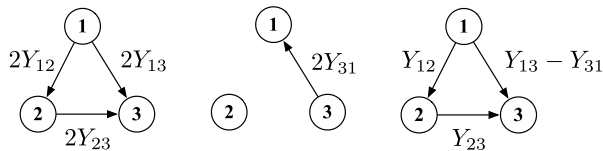

FIG. 1. *The two leftmost DAGs occur with probability $\frac{1}{2}$, yielding the difference graph G_μ at right, assuming $Y_{23} > Y_{32}$.*

If μ satisfies Definition 4, its difference graph G_μ is a DAG. Indeed, the definition ensures that all global preference information in G_μ (the sum of weights along any path) conforms with and reinforces local preference information (the weight on a single edge). Hence, we would expect any reasonable ranking method to be consistent in this setting. Nevertheless, typical pairwise surrogate losses are inconsistent in this low-noise setting (see also the weaker Theorem 11 in our preliminary work [19]):

THEOREM 2. *Let φ be a loss of the form* (12) *and assume $h(0) = 0$. If ϕ is convex, then even in the low-noise setting of Definition 4 the loss φ is not structure-consistent.*

Given the difficulties we encounter using losses of the form (12), it is reasonable to consider a reformulation of the surrogate. A natural alternative is a margin-based loss, which encodes a desire to separate ranking scores by large margins dependent on the preferences in a graph. Similar losses have been proposed, for example, by [38]. The next result shows that convex margin-based losses are also inconsistent, even in low-noise settings. (See also the weaker Theorem 12 of our preliminary work [19].)

THEOREM 3. *Let $h : \mathbb{R} \to \mathbb{R}$ and φ be a loss of the form*

$$(13) \qquad \varphi(\alpha, Y) = \sum_{i,j:Y_{ij}>0} \phi\big(\alpha_i - \alpha_j - h(Y_{ij})\big).$$

If ϕ is convex, then even in the low-noise setting of Definition 4 the loss φ is not structure-consistent.

3.3. *Achieving Fisher consistency.* Although Section 3.2 suggests an inherent difficulty in the development of tractable losses for ranking, tractable Fisher consistency is in fact achievable if one has access to *complete* preference data. We review a few of the known results here, showing how they follow from the Fisher consistency guarantees in Section 3.1, and derive some new Fisher consistency guarantees for the complete data setting (we defer all proofs to the supplementary material [18]). These results may appear to be of limited practical value, since complete preference judgments are typically unavailable or untrustworthy, but, as we show in Sections 4 and 5, they can be combined with aggregation strategies to yield procedures that are both practical and come with consistency guarantees.

We first define the normalized discounted cumulative gain (NDCG) family of complete data losses. Such losses are common in applications like web search, since they penalize ranking errors at the top of a ranked list more heavily than errors farther down the list. Let $s \in \mathcal{S} \subseteq \mathbb{R}^m$ be a vector of relevance scores and $\alpha \in \mathbb{R}^m$ be a vector of predicted scores. Define π_α to be the permutation associated with α, so that $\pi_\alpha(j)$ is the rank of item j in the ordering induced by α. Following

Ravikumar et al. [35], a general class of NDCG loss functions can be defined as follows:

$$(14) \quad L(\alpha, s) = 1 - \frac{1}{Z(s)} \sum_{j=1}^{m} \frac{G(s_j)}{F(\pi_\alpha(j))}, \qquad Z(s) = \max_{\alpha'} \sum_{j=1}^{m} \frac{G(s_j)}{F(\pi_{\alpha'}(j))},$$

where G and F are functions monotonically increasing in their arguments. By inspection, $L \in [0, 1]$, and we remark that the standard NDCG criterion [27] uses $G(s_j) = 2^{s_j} - 1$ and $F(j) = \log(1 + j)$. The "precision at k" loss [30] can also be written in the form (14), where $G(s_j) = s_j$ (assuming that $s_j \geq 0$) and $F(j) = 1$ for $j \leq k$ and $F(j) = +\infty$ otherwise, which measures the relevance of the top k items given by the vector α. This form generalizes standard forms of precision, which assume $s_j \in \{0, 1\}$.

To analyze the consistency of surrogate losses for the NDCG family (14), we first compute the loss $\ell(\alpha, \mu)$ and then state a corollary to Proposition 2. Observe that for any $\mu \in \mathcal{M}(\mathcal{S})$,

$$\ell(\alpha, \mu) = 1 - \sum_{j=1}^{m} \frac{1}{F(\pi_\alpha(j))} \int \frac{G(s_j)}{Z(s)} d\mu(s).$$

Since the function F is increasing in its argument, minimizing $\ell(\alpha, \mu)$ corresponds to choosing any vector α whose values α_j obey the same order as the m points $\int G(s_j)/Z(s) d\mu(s)$. In particular, the range of ℓ is finite for any μ since it depends only on the permutation induced by α, so we have Corollary 1.

COROLLARY 1. *Define the set*

$$(15) \quad A(\mu) = \left\{ \alpha \in \mathbb{R}^m \mid \alpha_j > \alpha_l \text{ when } \int \frac{G(s_j)}{Z(s)} d\mu(s) > \int \frac{G(s_l)}{Z(s)} d\mu(s) \right\}.$$

A surrogate loss φ is Fisher-consistent for the NDCG family (14) if and only if for all $\mu \in \mathcal{M}(\mathcal{S})$,

$$\inf_{\alpha} \left\{ \ell_\varphi(\alpha, \mu) - \inf_{\alpha'} \ell_\varphi(\alpha', \mu) \mid \alpha \notin A(\mu) \right\} > 0.$$

Corollary 1 recovers the main flavor of the consistency results in the papers of Ravikumar et al. [35] and Buffoni et al. [6]. The surrogate φ is consistent if and only if it preserves the order of the integrated terms $\int G(s_j)/Z(s) d\mu(s)$: any sequence α_n tending to the infimum of $\ell_\varphi(\alpha, \mu)$ must satisfy $\alpha_n \in A(\mu)$ for large enough n. Zhang [44] presents several examples of such losses; as a corollary to his Theorem 5 (also noted by [6]), the loss

$$\varphi(\alpha, s) := \sum_{j=1}^{m} \frac{G(s_j)}{Z(s)} \sum_{l=1}^{m} \phi(\alpha_l - \alpha_j)$$

is convex and structure-consistent (in the sense of Definition 3) whenever $\phi : \mathbb{R} \to \mathbb{R}_+$ is nonincreasing, differentiable and satisfies $\phi'(0) < 0$. The papers [6, 35] contain more examples and a deeper study of NDCG losses. To extend Corollary 1 to a uniform result, we note that if $G(s_j) > 0$ for all j and $\mathcal{S}$ is compact, then φ is 0-coercive over the set $\{\alpha_1 = 0\}$,[3] whence Theorem 1 implies that structure consistency coincides with uniform consistency.

Another family of loss functions is based on a cascade model of user behavior [8]. These losses model dependency among items or results by assuming that a user scans an ordered list of results from top to bottom and selects the first satisfactory result. Here, satisfaction is determined independently at each position. Let $\pi_\alpha^{-1}(i)$ denote the index of item that $\alpha \in \mathbb{R}^m$ assigns to rank i. The form of such expected reciprocal rank (ERR) losses is

$$(16) \qquad L(\alpha, s) = 1 - \sum_{i=1}^{m} \frac{1}{F(i)} G(s_{\pi_\alpha^{-1}(i)}) \prod_{j=1}^{i-1} \big(1 - G(s_{\pi_\alpha^{-1}(j)})\big),$$

where $G : \mathbb{R} \to [0, 1]$ is a nondecreasing function that indicates the prior probability that a result with score s_j is selected, and $F : \mathbb{N} \to [1, \infty)$ is an increasing function that more heavily weights the first items. The ERR family also satisfies $L \in [0, 1]$, and empirically correlates well with user satisfaction in ranking tasks [8].

Computing the expected conditional risk $\ell(\alpha, \mu)$ for general $\mu \in \mathcal{M}(\mathcal{S})$ is difficult, but we can compute it when μ is a product measure over $s_1, \ldots, s_m$. Indeed, in this case, we have

$$\ell(\alpha, \mu) = 1 - \sum_{i=1}^{m} \frac{1}{F(i)} \int G(s_{\pi_\alpha^{-1}(i)}) \prod_{j=1}^{i-1} \big(1 - G(s_{\pi_\alpha^{-1}(j)})\big) \, d\mu(s)$$

$$= 1 - \sum_{i=1}^{m} \frac{1}{F(i)} \mathbb{E}_\mu\big[G(s_{\pi_\alpha^{-1}(i)})\big] \prod_{j=1}^{i-1} \big(1 - \mathbb{E}_\mu\big[G(s_{\pi_\alpha^{-1}(j)})\big]\big).$$

When one believes that the values $G(s_i)$ represent the a priori relevance of the result i, this independence assumption is not unreasonable, and indeed, in Section 5 we provide examples in which it holds. Regardless, we see that $\ell(\alpha, \mu)$ depends only on the permutation π_α, and we can compute the minimizers of the conditional risk for the ERR family (16) using the following lemma, with proof provided in the supplementary material [18].

LEMMA 1. *Let $p_i = \mathbb{E}_\mu[G(s_i)]$. The permutation π minimizing $\ell(\alpha, \mu)$ is in decreasing order of the p_i.*

---

[3]The loss is invariant to linear shifts by the ones vector $\mathbb{1}$, so we may arbitrarily set a value for α_1.

Lemma 1 shows that an order-preserving property is necessary and sufficient for the Fisher-consistency of a surrogate φ for the ERR family (16), as it was for the NDCG family (14). To see this, we apply a variant of Corollary 1 where $A(\mu)$ as defined in equation (15) is replaced with the set

$$A(\mu) = \left\{ \alpha \in \mathbb{R}^m \mid \alpha_j > \alpha_l \text{ whenever } \int G(s_j) \, d\mu(s) > \int G(s_l) \, d\mu(s) \right\}.$$

Theorem 5 of [44] implies that $\varphi(\alpha, s) = \sum_{j=1}^m G(s_j) \sum_{l=1}^m \phi(\alpha_l - \alpha_j)$ is a consistent surrogate when ϕ is convex, differentiable and nonincreasing with $\phi'(0) < 0$. Theorem 1 also yields an equivalence between structure and uniform consistency under suitable conditions on $\mathcal{S}$.

Before concluding this section, we make a final remark, which has bearing on the aggregation strategies we discuss in Section 5. We have assumed that the structure spaces $\mathcal{S}$ for the NDCG (14) and ERR (16) loss families consist of real-valued relevance scores. This is certainly not necessary. In some situations, it may be more beneficial to think of $s \in \mathcal{S}$ as simply an ordered list of the results or as a directed acyclic graph over $\{1, \ldots, m\}$. We can then apply a transformation $r : \mathcal{S} \to \mathbb{R}^m$ to get relevance scores, using $r(s)$ in place of s in the losses (14) and (16). This has the advantage of causing $\mathcal{S}$ to be finite, so Theorem 1 applies, and there exists a nondecreasing function ζ with $\zeta(0) = 0$ such that for any distribution and any measurable f,

$$R(f) - R^* \leq \zeta\big(R_\varphi(f) - R_\varphi^*\big).$$

4. Uniform laws and asymptotic consistency. In Section 3, we gave examples of losses based on readily available pairwise data but for which Fisher-consistent tractable surrogates do not exist. The existence of Fisher-consistent tractable surrogates for other forms of data, as in Section 3.3, suggests that aggregation of pairwise and partial data into more complete data structures, such as lists or scores, makes the problem easier. However, it is not obvious how to design statistical procedures based on aggregation. In this section, we formally define a class of suitable estimators that permit us to take advantage of the weak convergence of Assumption A and show that uniform laws of large numbers hold for our surrogate losses. This means that we can indeed asymptotically minimize the risk (3) as desired.

Our aim is to develop an empirical analogue of the population surrogate risk (6) that converges uniformly to the population risk under minimal assumptions on the loss φ and structure function s. Given a dataset $\{(Q_i, Y_i)\}_{i=1}^n$ with $(Q_i, Y_i) \in \mathcal{Q} \times \mathcal{Y}$, we begin by defining, for each query q, the batch of data belonging to the query, $\mathcal{B}(q) = \{i \in \{1, \ldots, n\} \mid Q_i = q\}$, and the empirical count of the number of items in the batch, $\widehat{n}_q = |\mathcal{B}(q)|$. As a first attempt at developing an empirical objective, we might consider an empirical surrogate risk based on complete aggregation over

the batch of data belonging to each query:

$$
(17) \qquad \frac{1}{n} \sum_q \widehat{n}_q \varphi\big(f(q), s(\{Y_{i_j} \mid i_j \in \mathcal{B}(q)\})\big).
$$

While we would expect this risk to converge uniformly when φ is a sufficiently smooth function of its structure argument, the analysis of the complete aggregation risk (17) requires overly detailed knowledge of the surrogate φ and the structure function s.

To develop a more broadly applicable statistical procedure, we instead consider an empirical surrogate based on U-statistics. By trading off the nearness of an order-k U-statistic to an i.i.d. sample and the nearness of the limiting structure distribution μ_q to a structure $s(Y_1, \ldots, Y_k)$ aggregated over k draws, we can obtain consistency under mild assumptions on φ and s. More specifically, for each query q, we consider the surrogate loss

$$
(18) \qquad \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \varphi\big(f(q), s(Y_{i_1}, \ldots, Y_{i_k})\big).
$$

When $\widehat{n}_q < k$, we adopt the convention $\binom{\widehat{n}_q}{k} = 1$, and the above sum becomes the single term $\varphi(f(q), s(\{Y_{i_j} \mid i_j \in \mathcal{B}(q)\}))$ as in the expression (17). Hence, our U-statistic loss recovers the complete aggregation loss (17) when $k = \infty$.

An alternative formulation to loss (18) might consist of $\lceil |\mathcal{B}(q)|/k \rceil$ aggregation terms per query, with each query-preference pair appearing in a single term. However, the instability of such a strategy is high: a change in the ordering of the data or a substitution of queries could have a large effect on the final estimator. The U-statistic (18) grants robustness to such perturbations in the data. Moreover, by choosing the right rate of increase of the aggregation order k as a function of n, we obtain consistent procedures for a broad class of surrogates φ and structures s.

We associate with the surrogate loss (18) a surrogate empirical risk that weights each query by its empirical probability of appearance:

$$
(19) \qquad \widehat{R}_{\varphi,n}(f) := \frac{1}{n} \sum_q \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \varphi\big(f(q), s(Y_{i_1}, \ldots, Y_{i_k})\big).
$$

Let $\mathbb{P}_n$ denote the probability distribution of the queries given that the dataset is of size n. Then by iteration of expectation and Fubini's theorem, the surrogate risk (19) is an unbiased estimate of the population quantity

$$
(20) \quad R_{\varphi,n}(f) := \sum_q \left[\sum_{l=1}^n l\mathbb{P}_n(\widehat{n}_q = l) \mathbb{E}\big[\varphi(f(Q), s(Y_1, \ldots, Y_{l \wedge k})) \,\big|\, Q = q\big] \right].
$$

It remains to establish a uniform law of large numbers guaranteeing the convergence of the empirical risk (19) to the target population risk (6). Under suitable

conditions such as those of Section 3, this ensures the asymptotic consistency of computationally tractable statistical procedures. Hereafter, we assume that we have a nondecreasing sequence of function classes $\mathcal{F}_n$, where any $f \in \mathcal{F}_n$ is a scoring function for queries, mapping $f : \mathcal{Q} \to \mathbb{R}^m$ and giving scores to the (at most m) results for each query $q \in \mathcal{Q}$. Our goal is to give sufficient conditions for the convergence in probability

$$(21) \qquad \sup_{f \in \mathcal{F}_n} |\widehat{R}_{\varphi,n}(f) - R_\varphi(f)| \xrightarrow{p} 0 \qquad \text{as } n \to \infty.$$

While we do not provide fully general conditions under which the convergence (21) occurs, we provide representative, checkable conditions sufficient for convergence. At a high level, to establish (21), we control the uniform difference between the expectations $R_{\varphi,n}(f)$ and $R_\varphi(f)$ and bound the distance between the empirical risk $\widehat{R}_{\varphi,n}$ and its expectation $R_{\varphi,n}$ via covering number arguments. We now specify assumptions under which our results hold, deferring all proofs to the supplementary material [18].

Without loss of generality, we assume that p_q, the true probability of seeing the query q, is nonincreasing in the query index q. First, we describe the tails of the query distribution:

ASSUMPTION D. There exist constants $\beta > 0$ and $K_1 > 0$ such that $p_q \leq K_1 q^{-\beta-1}$ for all q. That is, $p_q = \mathcal{O}(q^{-\beta-1})$.

Infinite sets of queries $\mathcal{Q}$ are reasonable, since search engines, for example, receive a large volume of entirely new queries each day. Our arguments also apply when $\mathcal{Q}$ is finite, in which case we can take $\beta \uparrow \infty$.

Our second main assumption concerns the behavior of the surrogate loss φ over the function class $\mathcal{F}_n$, which we assume is contained in a normed space with norm $\|\cdot\|$.

ASSUMPTION E (Bounded Lipschitz losses). The surrogate loss function φ is bounded and Lipschitz continuous over $\mathcal{F}_n$: for any $s \in \mathcal{S}$, any $f, f_1, f_2 \in \mathcal{F}_n$, and any $q \in \mathcal{Q}$, there exist constants B_n and $L_n < \infty$ such that

$$0 \leq \varphi(f(q), s) \leq B_n$$

and

$$\left|\varphi(f_1(q), s) - \varphi(f_2(q), s)\right| \leq L_n \|f_1 - f_2\|.$$

This assumption is satisfied whenever $\varphi(\cdot, s)$ is convex and $\mathcal{F}_n$ is compact [and contained in the interior of the domain of $\varphi(\cdot, s)$] [25]. Our final assumption gives control over the sizes of the function classes $\mathcal{F}_n$ as measured by their covering numbers. (The ε-covering number of $\mathcal{F}$ is the smallest N for which there are f^i, $i \leq N$, such that $\min_i \|f^i - f\| \leq \varepsilon$ for any $f \in \mathcal{F}$.)

ASSUMPTION F. For all $\varepsilon > 0$, $\mathcal{F}_n$ has ε-covering number $N(\varepsilon, n) < \infty$.

With these assumptions in place, we give a few representative conditions that enable us to guarantee uniform convergence (21). Roughly, these conditions control the interaction between the size of the function classes $\mathcal{F}_n$ and the order k of aggregation used with n data points. To that end, we let the aggregation order k_n grow with n. In stating the conditions, we make use of the shorthand $\mathbb{E}_q[\varphi(f(q), s(Y_{1:k}))]$ for $\mathbb{E}[\varphi(f(Q), s(Y_1, \ldots, Y_k)) \mid Q = q]$.

CONDITION I. There exist a $\rho > 0$ and constant C such that for all $q \in \mathcal{Q}$, $n \in \mathbb{N}$, $k \in \mathbb{N}$, and $f \in \mathcal{F}_n$,

$$\left| \mathbb{E}_q[\varphi(f(q), s(Y_1, \ldots, Y_k))] - \lim_{k'} \mathbb{E}_q[\varphi(f(q), s(Y_1, \ldots, Y_{k'}))] \right| \leq C B_n k^{-\rho}.$$

Additionally, the sequences B_n and k_n satisfy $B_n = o(k_n^\rho)$.

This condition is not unreasonable; when φ and s are suitably continuous, we expect $\rho \geq \frac{1}{2}$. We also consider an alternative covering number condition.

CONDITION I'. Sequences $\{\varepsilon_n\} \subset \mathbb{R}_+$ and $\{k_n\} \subset \mathbb{N}$ and an ε_n-cover $\mathcal{F}_n^1, \ldots,$ $\mathcal{F}_n^{N(\varepsilon_n, n)}$ of $\mathcal{F}_n$ can be chosen such that

$$\max_{i \in \{1, \ldots, N(\varepsilon_n, n)\}} \inf_{f \in \mathcal{F}_n^i} \left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_1, \ldots, Y_{k_n}))] \right| + 2L_n \varepsilon_n \to 0.$$

Condition I' is weaker than Condition I, since it does not require uniform convergence over $q \in \mathcal{Q}$. If the function class $\mathcal{F}$ is fixed for all n, then the weak convergence of $s(Y_1, \ldots, Y_k)$ as in Assumption A guarantees Condition I', since $N(\varepsilon, n) = N(\varepsilon, n') < \infty$, and we may take ε arbitrarily small. We require one additional condition, which relates the growth of k_n, B_n, and the function classes $\mathcal{F}_n$ more directly.

CONDITION II. The sequences k_n and B_n satisfy $k_n B_n^{(1+\beta)/\beta} = o(n)$. Additionally, for any fixed $\varepsilon > 0$, the sequences satisfy

$$k_n B_n \left[\log N\left(\frac{\varepsilon}{4L_n}, n \right) \right]^{1/2} = o(\sqrt{n}).$$

By inspection, Condition II is satisfied for any $k_n = o(\sqrt{n})$ if the function classes $\mathcal{F}_n$ are fixed for all n. Similarly, if for all $k \geq k_0$, $s(Y_1, \ldots, Y_k) = s(Y_1, \ldots, Y_{k_0})$, so s depends only on its first k_0 arguments, Condition II holds whenever $\max\{B_n^{(1+\beta)/\beta}, B_n^2 \log N(\varepsilon/4L_n, n)\} = o(n)$. If the function classes $\mathcal{F}_n$ consist of linear functionals represented by vectors $\theta \in \mathbb{R}^{d_n}$ in a ball of some finite

radius, then $\log N(\varepsilon, n) \approx d_n \log \varepsilon^{-1}$, which means that Condition II roughly requires $k_n \sqrt{d_n/n} \to 0$ as $n \to \infty$. Modulo the factor k_n, this condition is familiar from its necessity in the convergence of parametric statistical problems.

The conditions in place, we come to our main result on the convergence of our U-statistic-based empirical loss minimization procedures.

THEOREM 4. *Assume Condition* I *or* I′ *and additionally assume the growth Condition* II. *Under Assumptions* D, E *and* F,

$$\sup_{f \in \mathcal{F}_n} \left| \widehat{R}_{\varphi,n}(f) - R_\varphi(f) \right| \xrightarrow{p} 0 \qquad \text{as } n \to \infty.$$

We remark in passing that if Condition II holds, with the change that the $o(\sqrt{n})$ bound is replaced by $O(n^\rho)$ for some $\rho < \frac{1}{2}$, the conclusion of Theorem 4 can be strengthened to both convergence almost surely and in expectation.

By inspection, Theorem 4 provides our desired convergence guarantee (21). By combining the Fisher-consistent loss families outlined in Section 3.3 with the consistency guarantees provided by Theorem 4, it is thus possible to design statistical procedures that are both computationally tractable—minimizing only convex risks—and asymptotically consistent.

5. Rank aggregation strategies. In this section, we give several examples of practical strategies for aggregating disparate user preferences under our framework. Motivated by the statistical advantages of complete preference data highlighted in Section 3.3, we first present strategies for constructing complete vectors of relevance scores from pairwise preference data. We then discuss a model for the selection or "click" data that arises in web search and information retrieval and show that maximum likelihood estimation under this model allows for consistent ranking. We conclude this section with a brief overview of structured aggregation strategies.

5.1. *Recovering scores from pairwise preferences.* Here we treat partial preference observations as noisy evidence of an underlying complete ranking and attempt to achieve consistency with respect to a complete preference data loss. We consider three methods that take as input pairwise preferences and output a relevance score vector $s \in \mathbb{R}^m$. Such procedures fit naturally into our ranking-with-aggregation framework: the results in Section 3.3 and Section 4 show that a Fisher-consistent loss is consistent for the limiting distribution of the scores s produced by the aggregation procedure. Thus, it is the responsibility of the statistician—the designer of an aggregation procedure—to determine whether the scores accurately reflect the judgments of the population. We present our first example in some detail to show how aggregation of pairwise judgments can lead to consistency in our framework and follow with brief descriptions of alternate aggregation strategies.

For an introduction to the design of aggregation strategies for pairwise data, see Tsukida and Gupta [43] as well as the book by David [14].

Thurstone–Mosteller least squares and skew-symmetric scoring. The first aggregation strategy constructs a relevance score vector s in two phases. First, it aggregates a sequence of observed preference judgments $Y_i \in \mathcal{Y}$, provided in any form, into a skew-symmetric matrix $A \in \mathbb{R}^{m \times m}$ satisfying $A = -A^\top$. Each entry A_{ij} encodes the extent to which item i is preferred to item j. Given such a skew-symmetric matrix, Thurstone and Mosteller [32] recommend deriving a score vector s such that $s_i - s_j \approx A_{ij}$. In practice, one may not observe preference information for every pair of results, so we define a masking matrix $\Omega \in \{0, 1\}^{m \times m}$ with $\Omega = \Omega^\top$, $\Omega_{ii} = 1$, and $\Omega_{ij} = 1$ if and only if preference information has been observed for the pair $i \neq j$. Letting $\circ$ denote the Hadamard product, a natural objective for selecting scores (e.g., [23]) is the least squares objective

$$(22) \quad \underset{x \,:\, x^\top \mathbb{1} = 0}{\text{minimize}} \quad \frac{1}{4} \sum_{i,j} \Omega_{ij}\big(A_{ij} - (x_i - x_j)\big)^2 = \frac{1}{4}\big\|\Omega \circ \big(A - (\mathbb{1}x^\top - x\mathbb{1}^\top)\big)\big\|_{\mathrm{Fr}}^2.$$

The gradient of the objective (22) is

$$D_\Omega x - (\Omega \circ A)\mathbb{1} - \Omega x \qquad \text{where } D_\Omega := \mathrm{diag}(\Omega\mathbb{1}).$$

Setting $s = (D_\Omega - \Omega)^\dagger(\Omega \circ A)\mathbb{1}$ yields the solution to the minimization problem (22), since $D_\Omega - \Omega$ is an unnormalized graph Laplacian matrix [9], and therefore $\mathbb{1}^\top s = \mathbb{1}^\top (D_\Omega - \Omega)^\dagger(\Omega \circ A)\mathbb{1} = 0$.

If $\Omega = \mathbb{1}\mathbb{1}^\top$, so that all pairwise preferences are observed, then the eigenvalue decomposition of $D_\Omega - \Omega = mI - \mathbb{1}\mathbb{1}^\top$ can be computed explicitly as $V\Sigma V^\top$, where V is any orthonormal matrix whose first column is $1/\sqrt{m}$, and Σ is a diagonal matrix with entries 0 (once) and m repeated $m - 1$ times. Thus, letting s_A and s_B denote solutions to the minimization problem (22) with different skew-symmetric matrices A and B and noting that $A\mathbb{1} \perp \mathbb{1}$ since $\mathbb{1}^\top A\mathbb{1} = 0$, we have the Lipschitz continuity of the solutions s in A:

$$\|s_A - s_B\|_2^2 = \big\|(mI - \mathbb{1}\mathbb{1}^\top)^\dagger(A - B)\mathbb{1}\big\|_2^2 = \frac{1}{m^2}\big\|(A - B)\mathbb{1}\big\|_2^2 \leq \frac{1}{m}\|\!|A - B|\!\|_2^2.$$

Similarly, when Ω is fixed, the score structure s is likewise Lipschitz in A for any norm $\|\!|\cdot|\!\|$ on skew-symmetric matrices.

A variety of procedures are available for aggregating pairwise comparison data $Y_i \in \mathcal{Y}$ into a skew-symmetric matrix A. One example, the Bradley–Terry–Luce (BTL) model [5], is based upon empirical log-odds ratios. Specifically, assume that $Y_i \in \mathcal{Y}$ are pairwise comparisons of the form $j \succ l$, meaning item j is preferred to item l. Then we can set

$$A_{jl} = \log \frac{\widehat{\mathbb{P}}(j \succ l) + c}{\widehat{\mathbb{P}}(j \prec l) + c} \qquad \text{for observed pairs } j, l,$$

where $\widehat{\mathbb{P}}$ denotes the empirical distribution over $\{Y_1, \ldots, Y_k\}$ and $c > 0$ is a smoothing parameter.

Since the proposed structure s is a continuous function of the skew-symmetric matrix A, the limiting distribution μ is a point mass whenever A converges almost surely, as it does in the BTL model. If aggregation is carried out using only a finite number of preferences rather than letting k approach ∞ with n, then μ converges to a nondegenerate distribution. Theorem 1 grants uniform consistency since the score space $\mathcal{S}$ is finite.

Borda count and budgeted aggregation. The Borda count [15] provides a computationally efficient method for computing scores from election results. In a general election setting, the procedure counts the number of times that a particular item was rated as the best, second best, and so on. Given a skew-symmetric matrix A representing the outcomes of elections, the Borda count assigns the scores $s = A\mathbb{1}$. As above, a skew-symmetric matrix A can be constructed from input preferences $\{Y_1, \ldots, Y_k\}$, and the choice of this first-level aggregation can greatly affect the resulting rankings. Ammar and Shah [1] suggest that if one has limited computational budget and only pairwise preference information then one should assign to item j the score

$$s_j = \frac{1}{m-1} \sum_{l \neq j} \widehat{\mathbb{P}}(j \succ l),$$

which estimates the probability of winning an election against an opponent chosen uniformly. This is equivalent to the Borda count when we choose $A_{jl} = \widehat{\mathbb{P}}(j \succ l) - \widehat{\mathbb{P}}(j \prec l)$ as the entries in the skew-symmetric aggregate A.

Principal eigenvector method. Saaty [36] describes the principal eigenvector method, which begins by forming a reciprocal matrix $A \in \mathbb{R}^{m \times m}$, with positive entries $A_{ij} = (A_{ji})^{-1}$, from pairwise comparison judgments. Here A_{ij} encodes a multiplicative preference for item i over item j; the idea is that ratios preserve preference strength [36]. To generate A, one may use, for example, smoothed empirical ratios $A_{jl} = \frac{\widehat{\mathbb{P}}(j \succ l) + c}{\widehat{\mathbb{P}}(j \prec l) + c}$. Saaty recommends finding a vector s so that $s_i/s_j \approx A_{ij}$, suggesting using the Perron vector of the matrix, that is, the first eigenvector of A.

5.2. *Cascade models for selection data.* Cascade models [8, 13] explain the behavior of a user presented with an ordered list of items, for example from a web search. In a cascade model, a user considers results in the presented order and selects the first to satisfy him or her. The model assumes the result l satisfies a user with probability p_l, independently of previous items in the list. It is natural to express a variety of ranking losses, including the expected reciprocal rank (ERR) family (16), as expected disutility under a cascade model, but computation and optimization of these losses require knowledge of the satisfaction probabilities p_l. When the satisfaction probabilities are unknown, Chapelle et al. [8] recommend plugging in those values p_l that maximize the likelihood of observed click data.

Here we show that risk consistency for the ERR family is straightforward to characterize when scores are estimated via maximum likelihood.

To this end, fix a query q, and let each affiliated preference judgment Y_i consist of a triple (m_i, π_i, c_i), where m_i is the number of results presented to the user, π_i is the order of the presented results, which maps positions $\{1, \ldots, m_i\}$ to the full result set $\{1, \ldots, m\}$, and $c_i \in \{1, \ldots, m_i + 1\}$ is the position clicked on by the user ($m_i + 1$ if the user chooses nothing). The likelihood g of an i.i.d. sequence $\{Y_1, \ldots, Y_k\}$ under a cascade model p is

$$g(p, \{Y_1, \ldots, Y_k\}) = \prod_{i=1}^{k} p_{\pi_i(c_i)}^{1(c_i \leq m_i)} \prod_{j=1}^{c_i - 1} (1 - p_{\pi_i(j)}),$$

and the maximum likelihood estimator of the satisfaction probabilities has the closed form

$$\widehat{p}_l(Y_1, \ldots, Y_k) = \frac{\sum_{i=1}^{k} 1(\pi_i(c_i) = l)}{\sum_{i=1}^{k} \sum_{j=1}^{c_i} 1(\pi_i(j) = l)}, \qquad l = 1, \ldots, m.$$

To incorporate this maximum likelihood aggregation procedure into our framework, we define the structure function s to be the vector

$$s(Y_1, \ldots, Y_k) := \widehat{p}(Y_1, \ldots, Y_k) \in \mathbb{R}^m$$

of maximum likelihood probabilities, and we take as our loss L any member of the ERR family (16). The strong law of large numbers implies the a.s. convergence of $\widehat{p}$ to a vector $p \in [0, 1]^m$, so that the limiting law $\mu_q(\{p\}) = 1$. Since μ_q is a product measure over $[0, 1]^m$, Lemma 1 implies that any α inducing the same ordering over results as p minimizes the conditional ERR risk $\ell(\alpha, \mu)$. By application of Theorems 1 (or Proposition 2) and 4, it is possible to asymptotically minimize the expected reciprocal rank by aggregation.

5.3. *Structured aggregation.* Our framework can leverage aggregation procedures (see, e.g., [21]) that map input preferences into representations of combinatorial objects. Consider the setting of Section 3.2, in which each observed preference judgment Y is the weighted adjacency matrix of a directed acyclic graph, our loss of interest L is the edgewise indicator loss (10), and our candidate surrogate losses have the form (18). Theorems 2 and 3 establish that risk consistency is not generally attainable when $s(Y_1, \ldots, Y_k) = Y_1$. In certain cases, aggregation can recover consistency. Indeed, define

$$s(Y_1, \ldots, Y_k) := \frac{1}{k} \sum_{i=1}^{k} Y_i,$$

the average of the input adjacency matrices. For an i.i.d. sequence $Y_1, Y_2, \ldots$ associated with a given query q, we have $s(Y_1, \ldots, Y_n) \overset{\text{a.s.}}{\to} \mathbb{E}(Y \mid Q = q)$ by the strong

law of large numbers, and hence the asymptotic surrogate risk

$$R_\varphi(f) = \sum_q p_q \int \varphi(f(q), s) \, d\mu_q(s) = \sum_q p_q \varphi(f(q), \mathbb{E}(Y \mid Q = q)).$$

Recalling the conditional pairwise risk (11), we can rewrite the risk as

$$R(f) = \sum_q p_q \left[\sum_{i<j} Y_{ij}^{\mu_q} 1(f_i(q) \le f_j(q)) + \sum_{i>j} Y_{ij}^{\mu_q} 1(f_i(q) < f_j(q)) \right]$$

$$= \sum_q p_q \sum_{i>j} \mathbb{E}[Y_{ij} \mid Q = q]$$

$$+ \sum_q p_q \sum_{i<j} \mathbb{E}[Y_{ij} - Y_{ji} \mid Q = q] 1(f_i(q) \le f_j(q)).$$

The discussion immediately following Proposition 2 shows that any consistent surrogate φ must be bounded away from its minimum for $\alpha \notin \text{argmin}_{\alpha'} \ell_\varphi(\alpha', \mu)$. Since the limiting distribution μ is a point mass at some adjacency matrix s for each q, a surrogate loss φ is consistent if and only if

$$\inf_\alpha \left\{ \varphi(\alpha, s) - \inf_{\alpha'} \varphi(\alpha', s) \mid \alpha \notin \underset{\alpha'}{\text{argmin}} \, L(\alpha', s) \right\} > 0.$$

In the important special case when the difference graph G_μ associated with $\mathbb{E}[Y \mid Q = q]$ is a DAG for each query q (recall Section 3.2.2), structure consistency is obtained if for each $\alpha^* \in \text{argmin}_\alpha \varphi(\alpha, s)$, $\text{sign}(\alpha_i^* - \alpha_j^*) = \text{sign}(s_{ij} - s_{ji})$ for each pair of results i, j. As an example, in this setting

$$(23) \qquad\qquad \varphi(\alpha, s) := \sum_{i,j} [s_{ij} - s_{ji}]_+ \phi(\alpha_i - \alpha_j)$$

is consistent when ϕ is nonincreasing, convex, and has derivative $\phi'(0) < 0$.

The Fisher-consistent loss (23) is similar to the inconsistent losses (12) considered in Section 3.2, but the coefficients adjoining each $\phi(\alpha_i - \alpha_j)$ summand exhibit a key difference. While the inconsistent losses employ coefficients based solely on the average $i \to j$ weight s_{ij}, the consistent loss coefficients are nonlinear functions of the edge weight differences $s_{ij} - s_{ji}$: they are precisely the edge weights of the difference graph G_μ introduced Section 3.2.2. Since at least one of the two coefficients $[s_{ij} - s_{ji}]_+$ and $[s_{ji} - s_{ij}]_+$ is always zero, the loss (23) penalizes misordering either edge $i \to j$ or $j \to i$. This contrasts with the inconsistent surrogates of Section 3.2, which simultaneously associate nonzero convex losses with opposing edges $i \to j$ and $j \to i$. Note also that our argument for the consistency of the loss (23) does not require Definition 4's low-noise assumption: consistency holds under the weaker condition that, on average, a population's preferences are acyclic.

6. Experimental study and implementation. In this section, we describe strategies for solving the convex programs that emerge from our aggregation approach to ranking and demonstrate the empirical utility of our proposed procedures. We begin with a broad description of implementation strategies and end with a presentation of specific experiments.

6.1. *Minimizing the empirical risk.* At first glance, the empirical risk (19) appears difficult to minimize, since the number of terms grows exponentially in the level of aggregation k. Fortunately, we may leverage techniques from the stochastic optimization literature [17, 33] to minimize the risk (19) in time linear in k and independent of n. Let us consider minimizing a function of the form

$$(24) \qquad \widehat{R}_N(f) := \frac{1}{N} \sum_{i=1}^{N} \varphi(f, s^i) + \Phi(f),$$

where $\{s^i\}_{i=1}^N$ is some collection of data, $\varphi(\cdot, s)$ is convex in its first argument, and Φ is a convex regularizing function (possibly zero).

Duchi and Singer [17], using ideas similar to those of Nemirovski et al. [33], develop a specialized stochastic gradient descent method for minimizing composite objectives of the form (24). Such methods maintain a parameter f^t, which is assumed to live in convex subset $\mathcal{F}$ of a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and iteratively update f^t as follows. At iteration t, an index $i_t \in [N]$ is chosen uniformly at random and the gradient $\nabla_f \varphi(f^t, s^{i_t})$ is computed at f^t. The parameter f is then updated via

$$(25) \qquad f^{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \langle f, \nabla \varphi(f^t, s^{i_t}) \rangle + \Phi(f) + \frac{1}{2\eta_t} \| f - f^t \|^2 \right\},$$

where $\eta_t > 0$ is an iteration-dependent stepsize and $\| \cdot \|$ denotes the Hilbert norm. The convergence guarantees of the update (25) are well understood [17, 20, 33]. Define $\overline{f}^T = (1/T) \sum_{t=1}^T f^t$ to be the average parameter after T iterations. If the function $\widehat{R}_N$ is strongly convex—meaning it has at least quadratic curvature—the step-size choice $\eta_t \propto 1/t$ gives

$$\mathbb{E}[\widehat{R}_N(\overline{f}^T)] - \inf_{f \in \mathcal{F}} \widehat{R}_N(f) = \mathcal{O}\left(\frac{1}{T}\right),$$

where the expectation is taken with respect to the indices i_t chosen during each iteration of the algorithm. In the convex case (without assuming any stronger properties than convexity), the step-size choice $\eta_t \propto 1/\sqrt{t}$ yields

$$\mathbb{E}[\widehat{R}_N(\overline{f}^T)] - \inf_{f \in \mathcal{F}} \widehat{R}_N(f) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

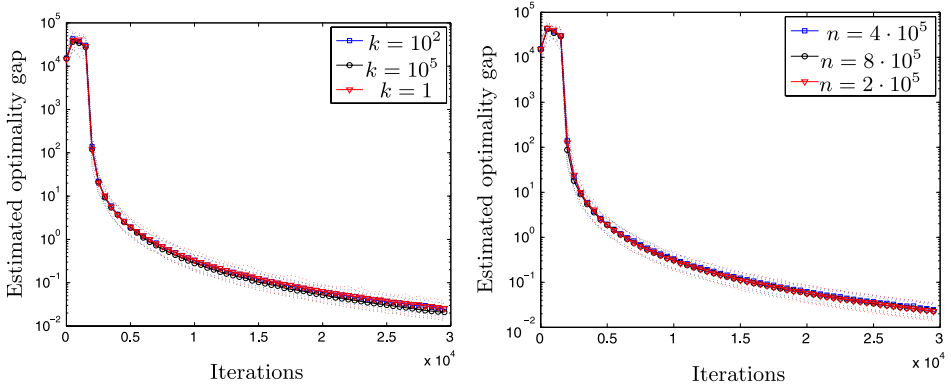These guarantees also hold with high probability [20, 33].

FIG. 2.   *Timing experiments for different values of k and n when applying the method* (25). *The horizontal axes are the number of stochastic gradient iterations*; *the vertical axes are the estimated optimality gap for the empirical surrogate risk. Left*: *varying amount of aggregation k, fixed* $n = 4 \cdot 10^5$. *Right*: *varying total number of samples n, fixed* $k = 10^2$.

Neither of the convergence rates $1/T$ or $1/\sqrt{T}$ depends on the number of terms N in the stochastic objective (24). As a consequence, we can apply the composite stochastic gradient method (25) directly to the empirical risk (19): we sample a query q with probability $\widehat{n}_q/n$, after which we uniformly sample one of the $\binom{\widehat{n}_q}{k}$ collections $\{i_1, \ldots, i_k\}$ of k indices associated with query q, and we then perform the gradient update (25) using the gradient sample $\nabla \varphi(f^t, s(Y_{i_1}, \ldots, Y_{i_k}))$. This stochastic gradient scheme means that we can minimize the empirical risk in a number of iterations independent of both n and k; the run-time behavior of the method scales independently of n and depends on k only so much as computing an instantaneous gradient $\nabla \varphi(f, s(Y_1, \ldots, Y_k))$ increases with k.

In Figure 2, we show empirical evidence that the stochastic method (25) works as described. In particular, we minimize the empirical U-statistic-based risk (19) with the loss (28) we employ in our experiments in the next section. In each plot in Figure 2, we give an estimated optimality gap, $\widehat{R}_{\varphi,n}(f^t) - \inf_{f \in \mathcal{F}} \widehat{R}_{\varphi,n}(f)$, as a function of t, the number of iterations. As in the section to follow, $\mathcal{F}$ consists of linear functionals parameterized by a vector $\theta \in \mathbb{R}^d$ with $d = 136$. To estimate $\inf_{f \in \mathcal{F}}$, we perform 100,000 updates of the procedure (25), then estimate $\inf_{f \in \mathcal{F}} \widehat{R}_{\varphi,n}(f)$ using the output predictor $\widehat{f}$ evaluated on an additional (independent) 50,000 samples (the number of terms in the true objective is too large to evaluate). To estimate the risk $\widehat{R}_{\varphi,n}(f^t)$, we use a moving average of the previous 100 sampled losses $\varphi(f^\tau, s^{i_\tau})$ for $\tau \in \{t - 99, \ldots, t\}$, which is an unbiased estimate of an upper bound on the empirical risk $\widehat{R}_{\varphi,n}(f^t)$ (see, e.g., [7]). We perform the experiment 20 times and plot averages as well as 90% confidence intervals. As predicted by our theoretical results, the number of iterations to attain a particular accuracy is essentially independent of n and k; all the plots lie on one another.

6.2. *Experimental evaluation.* To perform our experimental evaluation, we use a subset of the Microsoft Learning to Rank Web10K dataset [34], which consists of 10,000 web searches (queries) issued to the Microsoft Bing search engine, a set of approximately 100 potential results for each query, and a relevance score $r \in \mathbb{R}$ associated with each query/result pair. A query/result pair is represented by a $d = 136$-dimensional feature vector of standard document-retrieval features.

To understand the benefits of aggregation and consistency in the presence of partial preference data, we generate pairwise data from the observed query/result pairs, so that we know the true asymptotic generating distribution. We adopt a loss L from the NDCG-family (14) and compare three surrogate losses: a Fisher-consistent regression surrogate based on aggregation, an inconsistent but commonly used pairwise logistic loss [16], and a Fisher-consistent loss that requires access to complete preference data [35]. Recalling the NDCG score (14) of a prediction vector $\alpha \in \mathbb{R}^m$ for scores $s \in \mathbb{R}^m$ (where π_α is the permutation induced by α), we have the loss

$$L(\alpha, s) = 1 - \frac{1}{Z(s)} \sum_{j=1}^{m} \frac{G(s_j)}{F(\pi_\alpha(j))},$$

where $Z(s)$ is the normalizing value for the NDCG score, and $F(\cdot)$ and $G(\cdot)$ are increasing functions.

Given a set of queries q and relevance scores $r_i \in \mathbb{R}$, we generate n pairwise preference observations according to a Bradley–Terry–Luce (BTL) model [5]. That is, for each observation, we choose a query q uniformly at random and then select a uniformly random pair (i, j) of results to compare. The pair is ordered as $i \succ j$ (item i is preferred to j) with probability p_{ij}, and $j \succ i$ with probability $1 - p_{ij} = p_{ji}$, where

$$(26) \qquad p_{ij} = \frac{\exp(r_i - r_j)}{1 + \exp(r_i - r_j)}$$

for r_i and r_j the respective relevances of results i and j under query q.

We define our structure functions s_k as score vectors in $\mathbb{R}^m$, where given a set of k preference pairs, the score for item i is

$$s_k(i) = \frac{1}{m - 1} \sum_{j \neq i} \log \frac{\widehat{\mathbb{P}}(j \prec i)}{\widehat{\mathbb{P}}(j \succ i)},$$

the average empirical log-odds of result i being preferred to any other result. Under the BTL model (26), as $k \to \infty$ the structural score converges for each $i \in [m]$ to

$$(27) \qquad s(i) = \frac{1}{m - 1} \sum_{j \neq i} \left[\log\left(1 + \exp(r_i - r_j)\right) - \log\left(1 + \exp(r_j - r_i)\right) \right].$$

In our setting, we may thus evaluate the asymptotic NDCG risk of a scoring function f by computing the asymptotic scores (27). In addition, Corollary 1 shows that if all minimizers of a loss obey the ordering of the values

$$\int_{\mathcal{S}} \frac{G(s(j))}{Z(s)} \, d\mu(s), \qquad j \in \{1, \ldots, m\}$$

then the loss is Fisher-consistent. A well-known example [12, 35] of such a loss is the least-squares loss, where the regression labels are $G(s(j))/Z(s)$:

$$(28) \qquad \varphi(\alpha, s) = \frac{1}{2m} \sum_{j=1}^{m} \left(\alpha_j - \frac{G(s(j))}{Z(s)} \right)^2.$$

We compare the least-squares aggregation loss with a pairwise logistic loss natural for the pairwise data generated according to the BTL model (26). Specifically, given a data pair with $i \succ j$, the logistic surrogate loss is

$$(29) \qquad \varphi(\alpha, i \succ j) = \log\big(1 + \exp(\alpha_j - \alpha_i)\big),$$

which is equivalent or similar to previous losses used for pairwise data in the ranking literature [16, 28]. For completeness, we also compare with a Fisher-consistent surrogate that requires access to complete preference information in the form of the asymptotic structure scores (27). Following Ravikumar et al. [35], we obtain such a surrogate by granting the regression loss (28) direct access to the asymptotic structure scores. Note that such a construction would be infeasible in any true pairwise data setting.

Having described our sampling procedure, aggregation strategy, and loss functions, we now describe our model. We let x_i^q denote the feature vector for the ith result from query q, and we model the scoring function $f(q)_i = \langle \theta, x_i^q \rangle$ for a vector $\theta \in \mathbb{R}^d$. For the regression loss (28), we minimize the U-statistic-based empirical risk (19) over a variety of orders k, while for the pairwise logistic loss (29), we minimize the empirical risk over all pairs sampled according to the BTL model (26). We regularize our estimates by adding $\Phi(\theta) = (\lambda/2)\|\theta\|_2^2$ to the objective minimized, and we use the specialized stochastic method (25) to minimize the empirical risk.

Our goals in the experiments are to understand the behavior of the empirical risk minimizer as the order k of the aggregating statistic is varied and to evaluate the extent to which aggregation improves the estimated scoring function. A secondary concern is to verify that the method is insensitive to the amount λ of regularization performed on θ. We run each experiment 50 times and report confidence intervals based on those 50 experiments.

Let $\theta_{n,k}^{\text{reg}}$ denote the estimate of θ obtained from minimizing the empirical risk (19) with the regression loss (28) on n samples with aggregation order k, let θ_n^{log} denote the estimate of θ obtained from minimizing the empirical pairwise logistic loss (29), and let θ^{full} denote the estimate of θ obtained from minimizing the
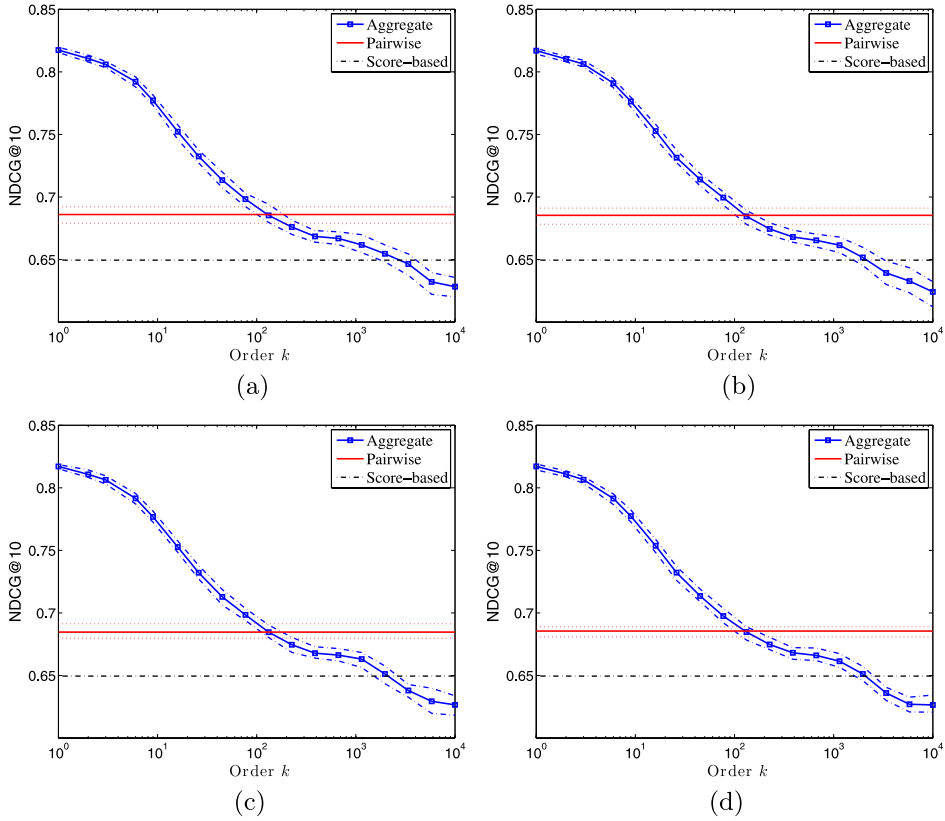
FIG. 3. *NDCG risk and 95% confidence intervals for θ estimated using the logistic pairwise loss* (29) *and the U-statistic empirical risk with φ chosen to be regression loss* (28). *The horizontal axis of each plot is the order k of the aggregation in the U-statistic* (19), *the vertical axis is the NDCG risk, and each plot corresponds to a different number n of samples.* (a) $n = 2 \cdot 10^5$; (b) $n = 4 \cdot 10^5$; (c) $n = 8 \cdot 10^5$; (d) $n = 1.6 \cdot 10^6$.

empirical risk with surrogate loss (28) using the asymptotic structure scores (27) directly. Then each plot of Figure 3 displays the risk $R(\theta_{n,k}^{\mathrm{reg}})$ as a function of the aggregation order k, using $R(\theta_n^{\log})$ and $R(\theta^{\mathrm{full}})$ as references. The four plots in the figure correspond to different numbers n of data pairs.

Broadly, the four plots in Figure 3 match our theoretical results. Consistently across the plots, we see that for small k, it appears there is not sufficient aggregation in the regression-loss-based empirical risk, and for such small k the pairwise logistic loss is better. However, as the order of aggregation k grows, the risk performance of $\theta_{n,k}^{\mathrm{reg}}$ improves. In addition, with larger sample sizes n, the difference between the risk of $\theta_n^{\log}$ and $\theta_{n,k}^{\mathrm{reg}}$ becomes more pronounced. The second salient feature of the plots is a moderate flattening of the risk $R(\theta_{n,k}^{\mathrm{reg}})$ and widening of the confidence interval for large values of k. This seems consistent with the estima-
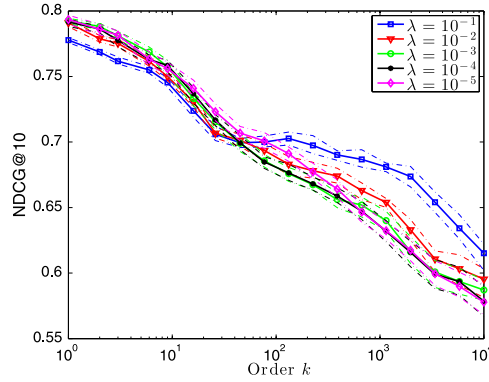
FIG. 4. *NDCG risk and 95% confidence intervals for θ estimated using the U-statistic empirical risk* (19) *with φ chosen as the regression loss* (28) *under various choices of the regularization parameter,* λ.

tion error guarantees in the theoretical results in Lemmas 7 and 10 in the appendices, where the order k being large has an evidently detrimental effect. Interestingly, however, large values of k still yield significant improvements over $R(\theta_n^{\log})$. For very large k, the improved performance of $\theta_{n,k}^{\text{reg}}$ over θ^{full} is a consequence of sampling artifacts and the fact that we use a finite dimensional representation. [By using sufficiently many dimensions d, the estimator θ^{full} attains zero risk by matching the asymptotic scores (27) directly.]

Figure 4 displays the risk $R(\theta_{n,k}^{\text{reg}})$ for $n = 800{,}000$ pairs, $k = 100$, and multiple values of the regularization multiplier λ on $\|\theta\|_2^2$. The results, which are consistent across many choices of n, suggest that minimization of the aggregated empirical risk (19) is robust to the choice of regularization multiplier.

7. Conclusions. In this paper, we demonstrated both the difficulty and the feasibility of designing consistent, practicable procedures for ranking. By giving necessary and sufficient conditions for the Fisher consistency of ranking algorithms, we proved that many natural ranking procedures based on surrogate losses are inconsistent, even in low-noise settings. To address this inconsistency while accommodating the incomplete nature of typical ranking data, we proposed a new family of surrogate losses, based on U-statistics, that aggregate disparate partial preferences. We showed how our losses can fruitfully leverage any well behaved rank aggregation procedure and demonstrated their empirical benefits over more standard surrogates in a series of ranking experiments.

Our work thus takes a step toward bringing the consistency literature for ranking in line with that for classification, and we anticipate several directions of further development. First, it would be interesting to formulate low-noise conditions under which faster rates of convergence are possible for ranking risk minimization (see, e.g., the work of [10], which focuses on the minimization of a single pairwise loss).

Additionally, it may be interesting to study structure functions s that yield nonpoint distributions μ as the number of arguments k grows to infinity. For example, would scaling the Thurstone–Mosteller least-squares solutions (22) by $\sqrt{k}$—to achieve asymptotic normality—induce greater robustness in the empirical minimizer of the U-statistic risk (19)? Finally, exploring tractable formulations of other supervised learning problems in which label data is naturally incomplete could be fruitful.

SUPPLEMENTARY MATERIAL

Proofs of results (DOI: 10.1214/13-AOS1142SUPP; .pdf). The supplementary material contains proofs of our results.

REFERENCES

[1] AMMAR, A. and SHAH, D. (2011). Ranking: Compare, don't score. In *The 49th Allerton Conference on Communication, Control, and Computing*. IEEE, Washington, DC.

[2] ARROW, K. J. (1951). *Social Choice and Individual Values. Cowles Commission Monograph* **12**. Wiley, New York, NY. MR0039976

[3] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[4] BEN-TAL, A. and NEMIROVSKI, A. (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, Philadelphia, PA. MR1857264

[5] BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925

[6] BUFFONI, D., CALAUZENES, C., GALLINARI, P. and USUNIER, N. (2011). Learning scoring functions with order-preserving losses and standardized supervision. In *Proceedings of the 28th International Conference on Machine Learning* 825–832. Omnipress, Madison, WI.

[7] CESA-BIANCHI, N., CONCONI, A. and GENTILE, C. (2002). On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems* 14 359–366. MIT Press, Cambridge, MA.

[8] CHAPELLE, O., METZLER, D., ZHANG, Y. and GRINSPAN, P. (2009). Expected reciprocal rank for graded relevance. In *Conference on Information and Knowledge Management*. ACM, New York.

[9] CHUNG, F. R. K. (1997). *Spectral Graph Theory. CBMS Regional Conference Series in Mathematics* **92**. Conference Board of the Mathematical Sciences, Washington, DC. MR1421568

[10] CLÉMENÇON, S., LUGOSI, G. and VAYATIS, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.* **36** 844–874. MR2396817

[11] CONDORCET, N. (1785). Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix. Paris.

[12] COSSOCK, D. and ZHANG, T. (2008). Statistical analysis of Bayes optimal subset ranking. *IEEE Trans. Inform. Theory* **54** 5140–5154. MR2589888

[13] CRASWELL, N., ZOETER, O., TAYLOR, M. J. and RAMSEY, B. (2008). An experimental comparison of click position-bias models. In *Web Search and Data Mining (WSDM)* 87–94. ACM, New York.

[14] DAVID, H. A. (1969). *The Method of Paired Comparisons*. Charles Griffin & Company, London.

[15] DE BORDA, J. C. (1781). Memoire sur les Elections au Scrutin. Histoire de l'Academie Royale des Sciences, Paris.

[16] DEKEL, O., MANNING, C. and SINGER, Y. (2004). Log-linear models for label ranking. In *Advances in Neural Information Processing Systems* **16**.

[17] DUCHI, J. and SINGER, Y. (2009). Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.* **10** 2899–2934. MR2579916

[18] DUCHI, J. C., MACKEY, L. and JORDAN, M. I. (2013). Supplement to "The asymptotics of ranking algorithms." DOI:10.1214/13-AOS1142SUPP.

[19] DUCHI, J. C., MACKEY, L. and JORDAN, M. I. (2010). On the consistency of ranking algorithms. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML-10*) (J. Fürnkranz and T. Joachims, eds.) 327–334. Omnipress, Madison, WI.

[20] DUCHI, J. C., SHALEV-SHWARTZ, S., SINGER, Y. and TEWARI, A. (2010). Composite objective mirror descent. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*.

[21] DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank aggregation methods for the web. In *Proceedings of the Tenth International Conference on World Wide Web* (*WWW*10) 613–622. ACM, New York.

[22] FREUND, Y., IYER, R., SCHAPIRE, R. E. and SINGER, Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.* **4** 933–969. MR2125342

[23] GULLIKSEN, H. (1956). A least squares method for paired comparisons with incomplete data. *Psychometrika* **21** 125–134.

[24] HERBRICH, R., GRAEPEL, T. and OBERMAYER, K. (2000). Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA.

[25] HIRIART-URRUTY, J. and LEMARÉCHAL, C. (1996). *Convex Analysis and Minimization Algorithms I & II*. Springer, New York.

[26] HOPCROFT, J. E. and ULLMAN, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA. MR0645539

[27] JÄRVELIN, K. and KEKÄLÄINEN, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* **20** 422–446.

[28] JOACHIMS, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*. ACM, New York.

[29] MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267

[30] MANNING, C., RAGHAVAN, P. and SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge.

[31] MILLER, G. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63** 81–97.

[32] MOSTELLER, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16** 3–9.

[33] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041

[34] QIN, T., LIU, T. Y., DING, W., XU, J. and LI, H. (2012). Microsoft learning to rank datasets. Available at http://research.microsoft.com/en-us/projects/mslr/.

[35] RAVIKUMAR, P., TEWARI, A. and YANG, E. (2011). On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings* **15** 618–626. Society for Artificial Intelligence and Statistics.

[36] SAATY, T. L. (2003). Decision-making with the AHP: Why is the principal eigenvector necessary. *European J. Oper. Res*. **145** 85–91. MR1947158

[37] SAATY, T. L. (2008). Relative measurement and its generalization in decision making. Why pairwise comparisons are central in mathematics for the measurement of intangible factors. The analytic hierarchy/network process. *Rev. R. Acad. Cienc. Exactas FíS. Nat. Ser. A Math. RACSAM* **102** 251–318. MR2479460

[38] SHASHUA, A. and LEVIN, A. (2002). Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems* **15**.

[39] SHIFFRIN, R. M. and NOSOFSKY, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review* **101** 357–361.

[40] STEINWART, I. (2007). How to compare different loss functions and their risks. *Constr. Approx*. **26** 225–287. MR2327600

[41] STEWART, N., BROWN, G. and CHATER, N. (2005). Absolute identification by relative judgment. *Psychological Review* **112** 881–911.

[42] THURSTONE, L. L. (1927). A law of comparative judgment. *Psychological Review* **34** 273–286.

[43] TSUKIDA, K. and GUPTA, M. R. (2011). How to analyze paired comparison data. Technical Report UWEETR-2011-0004, Univ. Washington, Dept. Electrical Engineering.

[44] ZHANG, T. (2004). Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res*. **5** 1225–1251. MR2248016

[45] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist*. **32** 56–85. MR2051001

J. C. DUCHI
M. I. JORDAN
DEPARTMENTS OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA
USA
E-MAIL: jduchi@cs.berkeley.edu
         jordan@stat.berkeley.edu

L. MACKEY
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
USA
E-MAIL: lmackey@stanford.edu

SUPPLEMENTARY MATERIAL: THE ASYMPTOTICS OF RANKING ALGORITHMS

By John C. Duchi , Lester Mackey and Michael I. Jordan

University of California, Berkeley and Stanford University

In this supplement, we provide the proofs of our main results, their corollaries, and supporting lemmas.

8. Proofs of Consistency Results.

8.1. *Proof of Proposition 1.* We begin by recalling the definition of the convex conjugate h^* of a function h and its biconjugate h^{**}

$$h^*(y) := \sup_x \{yx - h(x)\} \quad \text{and} \quad h^{**}(x) := \sup_y \{xy - h^*(y)\},$$

noting that $h^{**}(x) \leq h(x)$ and that h^{**} is closed convex [4]. The following lemma is useful for the proof of Proposition 1.

Lemma 2. *Let* $h : [0,1] \to \mathbb{R}_+ \cup \{\infty\}$ *be non-decreasing with* $h(0) = 0$ *and* $h(x) > 0$ *for* $x > 0$. *Then* $h^{**}(x) > 0$ *for* $x > 0$.

Proof. For the sake of contradiction assume that there is some $z \in (0,1)$ with $h^{**}(z) = 0$. Since h^{**} is convex with $h^{**}(0) = 0$, we have $h^{**}([0,z]) = \{0\}$, and in particular $h^{**}(z/2) = 0$. By assumption $h(z/2) = b > 0$, whence $h(1) \geq b > 0$. Thus we can define the closed convex function

$$g(x) = \begin{cases} 0 & \text{if } x \leq z/2 \\ \frac{b}{1-z/2}\,(x - z/2) & \text{otherwise} \end{cases}$$

and have $g(x) \leq h(x)$ since $h(1) \geq h(z/2) = b > 0$. It is clear, however, that $g(z) > 0 = h^{**}(z)$, which contradicts the fact [4] that h^{**} is the greatest closed convex minorant of h. □

Proof of Proposition 1 The proof is analogous to proofs of similar results due to Zhang [11], Tewari and Bartlett [9], and Steinwart [8]. First, note that the function H is non-decreasing in its argument and satisfies

$H(0) = 0$ and $H(1 + \epsilon) = \infty$ for all $\epsilon > 0$. Jensen's inequality implies that

$$
\begin{aligned}
H^{**}(R(f) - R^*) = H^{**} &\left(\sum_q p_q \left[\ell(f(q), \mu_q) - \inf_\alpha \ell(\alpha, \mu_q) \right] \right) \\
&\leq \sum_q p_q H^{**} \left(\ell(f(q), \mu_q) - \inf_\alpha \ell(\alpha, \mu_q) \right) \\
&\leq \sum_q p_q H \left(\ell(f(q), \mu_q) - \inf_\alpha \ell(\alpha, \mu_q) \right) \\
&\leq \sum_q p_q \left(\ell_\varphi(f(q), \mu_q) - \inf_\alpha \ell_\varphi(\alpha, \mu_q) \right) \\
= R_\varphi(f) &- R_\varphi^*,
\end{aligned}
$$

(1)

the final inequality following by the definition (8) of H. Lemma 2 implies that $H^{**}(\epsilon) > 0$ for $\epsilon > 0$, and since H^{**} is closed convex and non-negative, it is continuous from the right at 0 so long as $H^{**}(\epsilon) < \infty$ for some $\epsilon > 0$.

Now, assume that φ is pointwise consistent. Let f_n be a sequence of measurable functions such that $R_\varphi(f_n) \to R_\varphi^*$. For shorthand, define $h(\epsilon) = H^{**}(\epsilon)$. If $H^{**}(\epsilon) = \infty$ for all $\epsilon > 0$, then $H(\epsilon) \geq H^{**}(\epsilon) = \infty$ for all $\epsilon > 0$. Then the definition (8) and the finiteness $R_\varphi(f) < \infty$ imply that $\ell(f(q), \mu) - \inf_{\alpha'} \ell(\alpha', \mu) = 0$ for all $\mu \in \mathcal{M}(\mathcal{S})$, and hence $\delta(\epsilon) = \infty$ in Definition 2.

Thus we can assume that $H(\epsilon) < \infty$ for some $\epsilon > 0$, and let $h'_-(\epsilon)$ denote the right-derivative of h at the point ϵ for $\epsilon \in \operatorname{int} \operatorname{dom} h$. This interior is non-empty by assumption, and the right-derivative exists and is positive since h is closed convex and H is non-decreasing with $H(x) > H(0)$ for $x > 0$ [4]. Now fix some $\epsilon > 0$, and let $b = h'_-(\epsilon/2) > 0$. Since h is convex and b is a subgradient of h, we have $h(\epsilon) \geq h(\epsilon/2) + b\epsilon/2 > b\epsilon/2$, so that if $h(\delta) < b\epsilon/2$, then $\delta < \epsilon$. Applying the bound (1), we can take $N \in \mathbb{N}$ such that $n \geq N$ implies $R_\varphi(f_n) - R_\varphi^* < b\epsilon/2$, and as such $H^{**}(R(f_n) - R^*) < b\epsilon/2$ so that $R(f_n) - R^* < \epsilon$.

Now we turn to the converse, where we assume that φ is uniformly consistent. Let $\{p_q\}$ be supported completely on a single q. Then

$$
R(f) - R^* = \ell(f(q), \mu_q) - \inf_{\alpha'} \ell(\alpha', \mu_q),
$$

and

$$
R_\varphi(f) - R_\varphi^* = \ell_\varphi(f(q), \mu_q) - \inf_{\alpha'} \ell_\varphi(\alpha', \mu_q).
$$

Fix an $\epsilon > 0$, and let $\alpha = f(q) \in \mathbb{R}^m$. By the contrapositive of Definition 2,

we know that there exists $\delta(\epsilon) > 0$ such that

$$R(f) - R^* = \ell(\alpha, \mu_q) - \inf_{\alpha'} \ell(\alpha', \mu_q) \geq \epsilon \quad \text{implies}$$

$$R_\varphi(f) - R_\varphi^* = \ell_\varphi(\alpha, \mu_q) - \inf_{\alpha'} \ell_\varphi(\alpha', \mu_q) \geq \delta(\epsilon),$$

which holds for any measurable f (and hence any $\alpha \in \mathbb{R}^m$). Since the measure μ_q was arbitrary, we see that the function $H(\epsilon)$ defined by the infimum (8) satisfies $H(\epsilon) \geq \delta(\epsilon) > 0$ for all $\epsilon > 0$. $\qquad \square$

8.2. *Proof of Theorem 1.* We prove Theorem 1 in two parts: first under Assumption C and second under Assumption C' under the additional condition of 0-coercivity of $\ell_\varphi(\cdot, \mu)$ (recall a function $g : \mathbb{R}^m \to \mathbb{R}$ is called *k-coercive* [4] if $g(x)/\|x\|^k \to +\infty$ as $\|x\| \to \infty$).

The proof of Theorem 1 using Assumption C requires three lemmas that make it quite straightforward. We begin by recalling the definition (3) of the optimal conditional surrogate loss $\ell_\varphi^*(\mu) := \inf_\alpha \ell_\varphi(\alpha, \mu)$. We have the following three results, each assuming the conditions of Theorem 1 with Assumption C.

LEMMA 3 (Zhang [11], Lemma 27; Tewari and Bartlett [9], Lemma 16). *The function ℓ_φ^* is continuous in the measure μ.*

LEMMA 4. *Let φ be structure-consistent and $\{\alpha_n\} \subset \mathbb{R}^m$ be a sequence of vectors. Then $\ell_\varphi(\alpha_n, \mu) \to \ell_\varphi^*(\mu)$ implies that $\ell(\alpha_n, \mu) \to \inf_\alpha \ell(\alpha, \mu)$, and for large enough n, the vector $\alpha_n \in \operatorname{argmin}_\alpha \ell(\alpha, \mu)$.*

PROOF. Suppose for contraposition that the sequence of vectors $\{\alpha_n\}$ has $\ell(\alpha_n, \mu) \not\to \inf_\alpha \ell(\alpha, \mu)$. Then there is a subsequence n_j of α_n and some $\epsilon > 0$ such that $\ell(\alpha_{n_j}, \mu) \geq \inf_\alpha \ell(\alpha, \mu) + \epsilon$ for all j. Thus we have $\alpha_{n_j} \notin \operatorname{argmin}_\alpha \ell(\alpha, \mu)$ for any j, and Definition 3 implies that

$$\ell_\varphi(\alpha_{n_j}, \mu) \geq \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) \mid \alpha \notin \operatorname{argmin}_\alpha \ell(\alpha, \mu) \right\} > \ell_\varphi^*(\mu).$$

As a consequence, we have $\ell_\varphi(\alpha_{n_j}, \mu)$ uniformly bounded from $\ell_\varphi^*(\mu)$, which is the contrapositive of the lemma's first claim. Thus, whenever $\ell_\varphi(\alpha_n, \mu) \to \ell_\varphi^*(\mu)$, we have $\ell(\alpha_n, \mu) \to \inf_\alpha \ell(\alpha, \mu)$, and since the range of L (and hence ℓ by Assumption C) is finite by assumption, we must have $\alpha_n \in \operatorname{argmin}_\alpha \ell(\alpha, \mu)$ for sufficiently large n. $\qquad \square$

LEMMA 5.   *Let φ be structure-consistent. Then the suboptimality function (8) satisfies $H(\epsilon) > 0$ for all $\epsilon > 0$.*

PROOF. An equivalent statement to the result that $H(\epsilon) > 0$ for all $\epsilon > 0$ is that for any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that for any $\mu \in \mathcal{M}(\mathcal{S})$,

$$\ell(\alpha, \mu) - \inf_{\alpha'} \ell(\alpha', \mu) \geq \epsilon \quad \text{implies} \quad \ell_\varphi(\alpha, \mu) - \ell_\varphi^*(\mu) \geq \delta(\epsilon).$$

We give a proof by contradiction using Lemmas 3 and 4. Assume that there is a sequence of pairs $(\alpha_n, \mu^n) \in \mathbb{R}^m \times \mathcal{M}(\mathcal{S})$ with

$$\ell(\alpha_n, \mu^n) - \inf_{\alpha'} \ell(\alpha', \mu^n) \geq \epsilon \quad \text{and} \quad \ell_\varphi(\alpha_n, \mu^n) - \ell_\varphi^*(\mu^n) \to 0.$$

By the compactness of $\mathcal{M}(\mathcal{S})$ (recall $\mathcal{S}$ is finite), we can take a subsequence n_j such that $\mu^{n_j} \to \mu$ for some $\mu \in \mathcal{M}(\mathcal{S})$. Then Lemma 3 gives $\ell_\varphi^*(\mu^{n_j}) \to \ell_\varphi^*(\mu)$; this coupled with the fact that $\ell_\varphi(\alpha_n, \mu^n) - \ell_\varphi^*(\mu^n) \to 0$ implies

$$(2) \qquad\qquad \lim_j \ell_\varphi(\alpha_{n_j}, \mu^{n_j}) = \lim_j \ell_\varphi^*(\mu^{n_j}) = \ell_\varphi^*(\mu).$$

If we can show that $\lim_j \ell_\varphi(\alpha_{n_j}, \mu) = \ell_\varphi^*(\mu)$, then we are done. Indeed, if this is the case then Lemma 4 implies that $\alpha_{n_j} \in \arg\min_\alpha \ell(\alpha, \mu)$ eventually, whence $\ell(\alpha_{n_j}, \mu) - \inf_{\alpha'} \ell(\alpha', \mu) = 0$. Then the continuity of $\ell(\alpha, \mu)$ in μ contradicts $\ell(\alpha_n, \mu^n) - \inf_{\alpha'} \ell(\alpha', \mu^n) \geq \epsilon$ for all n.

To the end of showing $\lim_j \ell_\varphi(\alpha_{n_j}, \mu) = \ell_\varphi^*(\mu)$, fix an arbitrary $\epsilon' > 0$ and recalling that $\mathcal{S}$ is finite, let $\mathcal{S}_+$ denote the set of $s \in \mathcal{S}$ such that $\mu(s) > 0$. Since $\mu^{n_j} \to \mu$, there must be some $B \in \mathbb{R}$ and $J \in \mathbb{N}$ such that $\varphi(\alpha_{n_j}, s) \leq B$ for all $j \geq J$ and $s \in \mathcal{S}_+$. Now, choose $J' \geq J$ so that $j \geq J'$ implies $\mu(s) \leq \mu^{n_j}(s) + \epsilon'/B|\mathcal{S}|$. Then

$$\int_{\mathcal{S}_+} \varphi(\alpha_{n_j}, s) d\mu(s) \leq \int_{\mathcal{S}_+} \varphi(\alpha_{n_j}, s) d\mu^{n_j}(s) + B|\mathcal{S}| \cdot \frac{\epsilon'}{B|\mathcal{S}|}$$

for $j \geq J'$, and thus

$$\begin{aligned}
\limsup_j \ell_\varphi(\alpha_{n_j}, \mu) &= \limsup_j \int_{\mathcal{S}_+} \varphi(\alpha_{n_j}, s) d\mu(s) \\
&\leq \limsup_j \int_{\mathcal{S}_+} \varphi(\alpha_{n_j}, s) d\mu^{n_j}(s) + \epsilon' \\
&\leq \limsup_j \int_{\mathcal{S}} \varphi(\alpha_{n_j}, s) d\mu^{n_j}(s) + \epsilon' = \ell_\varphi^*(\mu) + \epsilon'
\end{aligned}$$

by using the limit (2). Since ϵ' was arbitrary, we see that $\lim_j \ell_\varphi(\alpha_{n_j}, \mu) = \ell_\varphi^*(\mu)$ as desired.   □

Proof of Theorem 1 under Assumption C. Part (a) of the theorem follows immediately from Lemma 5, which implies that for any $\epsilon > 0$, $H(\epsilon) > 0$, whence we can apply Proposition 1.

For part (b), we focus on the case when $\{p_q\}$ is supported completely on a single q and prove the result by contradiction. Assume that φ is not structure consistent, so that there exists a $\mu \in \mathcal{M}(\mathcal{S})$ satisfying

$$\ell_\varphi^*(\mu) = \inf_\alpha \ell_\varphi(\alpha, \mu) = \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) \mid \alpha \notin \operatorname*{argmin}_{\alpha'} \ell(\alpha', \mu) \right\}.$$

There exists a sequence $\alpha_n \notin \operatorname{argmin}_\alpha \ell(\alpha, \mu)$ with $\ell_\varphi(\alpha_n, \mu) \to \ell_\varphi^*(\mu)$. Define the sequence of functions f_n with $f_n(q) = \alpha_n$, and let $\mu_q = \mu$. In this case, we have $R_\varphi(f) = \ell_\varphi(f(q), \mu)$, but by the finiteness of the range of $\ell(\cdot, \mu)$, there is some $\epsilon > 0$ such that $\ell(\alpha_n, \mu) \geq \inf_\alpha \ell(\alpha, \mu) + \epsilon$ for all n. Thus we see that although $R_\varphi(f_n) \to R_\varphi^*$, we have

$$R(f_n) = \ell(f_n(q), \mu) = \ell(\alpha_n, \mu) \not\to R^* = \inf_\alpha \ell(\alpha, \mu).$$

Thus we fail to have consistency as desired. $\square$

Proof of Theorem 1 under Assumption C' and coercivity. Now we turn to proving the result of Theorem 1 when Assumption C' holds, but we also assume that for any $\mu \in \mathcal{M}(\mathcal{S})$, the function $\alpha \mapsto \ell_\varphi(\alpha, \mu)$ is 0-coercive. We begin by noting that since $\mathcal{S}$ is compact, Prohorov's Theorem implies the collection of probability measures defined on $\mathcal{S}$ is compact with respect to the topology of weak convergence [2, Chapter 5]. We let $\rightsquigarrow$ denote weak convergence (or convergence in distribution) of measures. To prove the result, all we require is an analogue of Lemma 5.

Before proceeding, we recall a few results from variational analysis, referring to the book of Rockafellar and Wets [7] for a deeper exploration of these issues. For notational convenience, for a function $g : \mathbb{R}^m \to \mathbb{R}$ we define the approximate minimal set

$$\epsilon\text{-argmin}\, g := \left\{ x \in \mathbb{R}^m : g(x) \leq \inf_{x \in \operatorname{dom} g} g(x) + \epsilon \right\}.$$

PROPOSITION 4. *Let $g_n : \mathbb{R}^m \to \mathbb{R}$ be a sequence of closed convex functions converging pointwise to a 0-coercive function g. Then g is closed convex. Moreover, if $\epsilon_n > 0$ is a sequence with $\epsilon_n \to 0$, then any sequence $x_n \in \epsilon_n$-argmin g_n has a cluster point and*

$$\limsup_n g(x_n) \leq \inf_x g(x).$$

PROOF. This is a recapitulation of results due to Rockafellar and Wets [7] and Hiriart-Urruty and Lemaréchal [4]. We begin by noting that since the g_n converge pointwise, the function g is convex [4, Thoerem IV.3.1.5]. Pointwise convergence of convex g_n implies convergence of their epigraphs in the sense of set convergence [7, Theorem 7.17], and hence g is closed convex [7, Proposition 7.4(a)]. Since g is coercive (by assumption), Theorem 7.33 of [7] implies that there exists a bounded set $C \subset \mathbb{R}^m$ such that $\epsilon_n\text{-argmin}\, g_n \subset C$ for all suitably large n. The result also states that for any $\epsilon > 0$, eventually $\epsilon_n\text{-argmin}\, g_n \subset \epsilon\text{-argmin}\, g$, and the sequence of non-empty sets $\epsilon_n\text{-argmin}\, g_n$ has all its cluster points in $\text{argmin}\, g$. $\square$

We now give our promised analogue of Lemma 5.

LEMMA 6.   *Let φ be structure consistent and assume that for each $\mu \in \mathcal{M}(\mathcal{S})$, the function $\alpha \mapsto \ell_\varphi(\alpha, \mu)$ is 0-coercive. Then the suboptimality function (8) satisfies $H(\epsilon) > 0$ for all $\epsilon > 0$.*

PROOF. The proof follows that of Lemma 5, but we require the somewhat stronger tools of Proposition 4, since $\mathcal{S}$ is no longer necessarily finite.

Assume for the sake of contradiction that there exists $\epsilon > 0$ and a sequence of pairs $(\alpha_n, \mu^n) \in \mathbb{R}^m \times \mathcal{M}(\mathcal{S})$ satisfying

$$\ell(\alpha_n, \mu^n) - \inf_{\alpha'} \ell(\alpha', \mu^n) \geq \epsilon \quad \text{and} \quad \ell_\varphi(\alpha_n, \mu^n) - \ell_\varphi^*(\mu^n) \to 0.$$

Since $\mathcal{M}(\mathcal{S})$ is compact with respect to weak convergence (Prohorov's theorem), we can take a subsequence n_j such that $\mu^{n_j} \rightsquigarrow \mu$ for some $\mu \in \mathcal{M}(\mathcal{S})$.

By weak-convergence, we have the pointwise convergence

$$\lim_j \ell_\varphi(\alpha, \mu^{n_j}) = \ell_\varphi(\alpha, \mu) \ \text{ for all } \alpha \in \mathbb{R}^m.$$

Of course, the funtions $\ell_\varphi(\cdot, \mu^n)$ and $\ell_\varphi(\cdot, \mu)$ are convex, and the latter is 0-coercive, whence Proposition 4 applies. We find in particular

$$\limsup_j \ell_\varphi(\alpha_{n_j}, \mu) \leq \inf_\alpha \ell_\varphi(\alpha, \mu) = \ell_\varphi^*(\mu).$$

Lemma 4 holds under Assumption C', and the proof is, *mutatis mutandis*, the same. Thus, Applying Lemma 4, we find that $\alpha_{n_j} \in \text{argmin}_\alpha \ell(\alpha, \mu)$ eventually, whence $\ell(\alpha_{n_j}, \mu) - \inf_{\alpha'} \ell(\alpha', \mu) = 0$. Since $\mu \mapsto \ell(\alpha, \mu)$ is continuous, this contradicts the assumption that $\ell(\alpha_n, \mu^n) - \inf_{\alpha'} \ell(\alpha', \mu^n) \geq \epsilon$ for all n, which is our desired contradiction. So $H(\epsilon) > 0$. $\square$

Finalizing the proof.   The remainder of the proof is completely identical to that under Assumption C, but now instead of Lemma 5 we apply Lemma 6.

$\square$

8.3. *Proof of Lemma 1.*   Consider the function

$$g(p) = \sum_{i=1}^{m} \frac{1}{F(i)} p_i \prod_{j=1}^{i-1} (1 - p_j),$$

and assume that $p_i < p_{i+1}$ for some index i. Let $p' = p$ except that the values of p_i and p_{i+1} are swapped. Then

$$\begin{aligned}
g(p) - g(p') &= \frac{1}{F(i)} p_i \prod_{j=1}^{i-1} (1 - p_j) + \frac{1}{F(i+1)} p_{i+1} \prod_{j=1}^{i} (1 - p_j) \\
&\quad - \frac{1}{F(i)} p_{i+1} \prod_{j=1}^{i-1} (1 - p_j) - \frac{1}{F(i+1)} p_i (1 - p_{i+1}) \prod_{j=1}^{i-1} (1 - p_i) \\
&= \frac{1}{F(i)} (p_i - p_{i+1}) \prod_{j=1}^{i-1} (1 - p_j) \\
&\quad + \frac{1}{F(i+1)} \left(p_{i+1}(1 - p_i) - p_i(1 - p_{i+1}) \right) \prod_{j=1}^{i-1} (1 - p_j) \\
&= (p_i - p_{i+1}) \left[\frac{1}{F(i)} - \frac{1}{F(i+1)} \right] \prod_{j=1}^{i-1} (1 - p_j) < 0
\end{aligned}$$

since F is an increasing function. Thus swapping p_i and p_{i+1} increases the objective $g(p)$, proving the lemma.

9. Proofs of Uniform Laws.   In this section, we provide the proof of Theorem 4. The proof has two parts, the first controlling the expectation (20), the second controlling the convergence of the empirical risk $\widehat{R}_{\varphi,n}$ to its expectation $R_{\varphi,n}$.

9.1. *Expectation.*   We begin by studying the expectation (20). Our goal is to understand the rate at which we have convergence

(3) $$R_{\varphi,n}(f) \to \sum_{q} p_q \mathbb{E}_q \left[\varphi(f(q), s(Y_1, \ldots, Y_k)) \mid Q = q \right].$$

If the convergence (3) occurs sufficiently quickly, then we can allow k to increase to infinity, capturing the asymptotic surrogate risk (6). The following

lemma, whose proof we provide in Appendix 9.3, gives sufficient conditions for such a convergence result to hold.

LEMMA 7. *Let Assumptions D and E hold for the sequence of function classes $\mathcal{F}_n$. There exists a constant $C(K_1, \beta) < \infty$, such that*

$$\sup_{f \in \mathcal{F}_n} \left| R_{\varphi,n}(f) - \sum_q p_q \mathbb{E}\left[\varphi(f(Q), s(Y_1, \ldots, Y_k)) \mid Q = q\right] \right|$$
$$\leq C(K_1, \beta) B_n n^{-\frac{\beta}{1+\beta}} \left(k + \log k + 2 \log n\right)^{\frac{\beta}{1+\beta}}.$$

Lemma 7 is suggestive of Condition II and its requirements on B_n and k_n. We now show how Conditions I or I' and Condition II imply that $\sup_{f \in \mathcal{F}_n} |R_{\varphi,n}(f) - R_\varphi(f)| \to 0$. First, we show how Condition I implies I'. Indeed, since $\sum_q p_q = 1$, Condition I implies that for any $f \in \mathcal{F}_n$

$$\left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right|$$
$$\leq \sum_q p_q \left| \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] - \lim_{k'} \mathbb{E}_q[\varphi(f(q), s(Y_{1:k'}))] \right| \leq C B_n k_n^{-\rho},$$

which tends to 0 whenever $B_n = o(k_n^\rho)$. Then for any $\epsilon_n > 0$ we obtain

$$\max_{i \in [N(\epsilon_n, n)]} \inf_{f \in \mathcal{F}_n^i} \left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right| \leq C B_n k_n^{-\rho} \to 0,$$

and we may assume $\epsilon_n L_n \to 0$ in Condition I'. This established, the following lemma captures the convergence of expectations (see Appendix 9.4).

LEMMA 8. *Let Condition I' and $k_n B_n^{(1+\beta)/\beta} = o(n)$ hold. Under Assumptions D and E, as $n \to \infty$*

$$\sup_{f \in \mathcal{F}_n} |R_{\varphi,n}(f) - R_\varphi(f)| \to 0.$$

9.2. *High probability guarantees.* We now turn to demonstrating high-probability convergence of the empirical risk (19) to its expectation. Our approach makes use of the smoothness Assumption E along with a covering number argument [e.g., 10]. We use the bounded differences inequality [6] to demonstrate the convergence of the empirical risk (19) to its expectation. We begin by viewing $\widehat{R}_{\varphi,n}(f)$ as a random function of the n query preference evaluation pairs $(Q_1, Y_1), \ldots, (Q_n, Y_n)$; define

$$F\left((Q_1, Y_1), \ldots, (Q_n, Y_n)\right) := \widehat{R}_{\varphi,n}(f).$$

We have the following lemma, proved in Appendix 9.5.

LEMMA 9.   *Assume* $|\varphi(f(q), s(Y_1, \ldots, Y_m))| \leq B$ *for all* $f \in \mathcal{F}$. *Then*

$$\left| F((Q_1, Y_1), \ldots, (Q_n, Y_n)) - F((Q_1, Y_1), \ldots, (Q'_j, Y'_j), \ldots, (Q_n, Y_n)) \right| \leq \frac{4kB}{n}$$

for all j *and* (Q'_j, Y'_j).

As a consequence of Lemma 9, the bounded differences inequality [6] implies that since $|\varphi(f(q), s(Y_1, \ldots, Y_k))| \leq B_n$ for any fixed $f \in \mathcal{F}_n$,

$$(4) \qquad \mathbb{P}\left(\left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \geq \epsilon \right) \leq 2 \exp\left(-\frac{n\epsilon^2}{8k^2 B_n^2} \right).$$

Using Assumption E, we can now apply our covering number argument to get a uniform bound on the deviations $\widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f)$. We obtain

LEMMA 10.   *Let Assumptions* E *and* F *and Condition* II *hold. Then*

$$\sup_{f \in \mathcal{F}_n} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \xrightarrow{p} 0.$$

See Appendix 9.6 for the proof of the lemma, noting that the bound (4) guarantees almost sure and expected convergence in the conclusion of the lemma under the conditions in the remark following Theorem 4.

To complete the proof of Theorem 4, note that Lemma 8 guarantees the uniform convergence of the difference of expectations $R_{\varphi,n}(f) - R_\varphi(f)$; applying the triangle inequality in Lemma 10 completes the argument.

9.3. *Proof of Lemma 7.* We begin by fixing some q_0 and splitting the summation (20) into two terms:

$$R_{\varphi,n}(f) = \sum_{q \leq q_0} \left[\sum_{m=1}^{n} m \mathbb{P}_n(\widehat{n}_q = m) \mathbb{E}\left[\varphi\left(f(Q), s(Y_1, \ldots, Y_{m \wedge k}) \right) \mid Q = q \right] \right]$$

$$+ \sum_{q > q_0} \left[\sum_{m=1}^{n} m \mathbb{P}_n(\widehat{n}_q = m) \mathbb{E}\left[\varphi\left(f(Q), s(Y_1, \ldots, Y_{m \wedge k}) \right) \mid Q = q \right] \right],$$

the sum over queries $q \leq q_0$ and the sum over $q > q_0$. We control each sum in turn, using a Chernoff bound to control the first and the fact that $\sum_{q > q_0} p_q \to 0$ quickly enough as $q_0 \uparrow \infty$ (due to the power law assumption D) for the second. We also use the shorthand notation $\mathbb{E}_q[\varphi(f(q), s(Y_{1:k}))]$ for the quantity $\mathbb{E}[\varphi(f(Q), s(Y_1, \ldots, Y_k)) \mid Q = q]$ to keep the arguments neater.

We begin by studying the expectation of the more probable queries, for which we require a version of the Chernoff bound:

LEMMA 11 (Multiplicative Chernoff bound). *Let X_i be i.i.d. Bernoulli random variables with $\mathbb{P}(X_i = 1) = p$. For any $\delta \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i < (1 - \delta)np\right) \leq \exp\left(-\frac{np\delta^2}{2}\right).$$

As an immediate consequence of the Chernoff bound, we see that for any query q and empirical count $\widehat{n}_q$, we have

$$\mathbb{P}_n(\widehat{n}_q < k) = \mathbb{P}_n\left(\widehat{n}_q < \frac{k}{np_q}np_q\right) = \mathbb{P}_n\left[\widehat{n}_q < \left(1 - \left(1 - \frac{k}{np_q}\right)\right)np_q\right]$$

$$\leq \exp\left(-\frac{np_q}{2}\left(1 - \frac{k}{np_q}\right)^2\right) = \exp\left(-\frac{np_q}{2} + k - \frac{k^2}{2np_q}\right).$$

For any $\epsilon \in (0, 1)$, to have $\mathbb{E}\left[\widehat{n}_q 1\left(\widehat{n}_q < k\right)\right] \leq \epsilon np_q$, it is sufficient that

$$(5) \qquad\qquad n > \frac{2}{p_q}\left[k + \log k + \log\frac{1}{\epsilon}\right].$$

Indeed, for such n we have $np_q \geq 1$ and

$$\mathbb{E}\left[\widehat{n}_q 1\left(\widehat{n}_q < k\right)\right] = \sum_{m=1}^{k-1} m\mathbb{P}_n(\widehat{n}_q = m) \leq k\mathbb{P}_n(\widehat{n}_q < k) \leq k\exp\left(-\frac{np_q}{2} + k\right)$$

$$\leq k\exp\left(-\log k + \log\epsilon\right) = \epsilon \leq \epsilon np_q.$$

For fixed $\epsilon > 0$ and $n \in \mathbb{N}$, we let $q_0 = q_0(\epsilon, n)$ for

$$(6) \qquad\qquad q_0(\epsilon, n) := \max\left\{q : \frac{2}{p_q}\left[k + \log k + \log\frac{1}{\epsilon}\right] < n\right\}.$$

Then for all $q \leq q_0$, since $\sum_{m=1}^{n} m\mathbb{P}_n(\widehat{n}_q = m) = np_q$ and $k\mathbb{P}_n(\widehat{n}_q < k) \leq \epsilon$, we immediately see that

$$(1 - \epsilon)np_q\mathbb{E}_q\left[\varphi(f(q), s(Y_1, \ldots, Y_k))\right]$$

$$\leq \sum_{m=k}^{n} m\mathbb{P}_n(\widehat{n}_q = m)\mathbb{E}_q\left[\varphi(f(q), s(Y_1, \ldots, Y_k))\right]$$

$$\leq \sum_{m=1}^{n} m\mathbb{P}_n(\widehat{n}_q = m)\mathbb{E}_q\left[\varphi(f(q), s(Y_1, \ldots, Y_{m\wedge k}))\right]$$

and additionally that

$$\sum_{m=1}^{k-1} m\mathbb{P}(\widehat{n}_q = m)\mathbb{E}_q\left[\varphi(f(q), s(Y_{1:m}))\right] + \sum_{m=k}^{n} m\mathbb{P}(\widehat{n}_q = m)\mathbb{E}_q\left[\varphi(f(q), s(Y_{1:k}))\right]$$
$$\leq \epsilon B_n + np_q\mathbb{E}_q\left[\varphi(f(q), s(Y_1, \dots, Y_k))\right]$$

by applying Assumption E on the boundedness of φ. We thus see that

$$(1-\epsilon)\sum_{q \leq q_0} p_q\mathbb{E}_q\left[\varphi(f(q), s(Y_{1:k}))\right] \leq R_{\varphi,n}(f)$$

$$(7) \qquad \leq \sum_{q \leq q_0} p_q\mathbb{E}_q\left[\varphi(f(q), s(Y_{1:k}))\right] + \frac{q_0(\epsilon, n)\epsilon B_n}{n} + B_n\sum_{q > q_0} p_q$$

for all $f \in \mathcal{F}_n$. The sandwich inequality (7) suggests that we should have our desired convergence statement (3) so long as the bound B_n does not grow too quickly.

Now we use Assumption D to complete the proof by bounding q_0 and $\sum_{q > q_0} p_q$. Recalling the definition (6) of q_0, we note that for $q > q_0$,

$$p_q \leq \min\left\{K_1 q^{-\beta-1}, \frac{2(k + \log k + \log \epsilon^{-1})}{n}\right\}.$$

Let

$$q_0^* := \left(\frac{K_1 n}{2(k + \log k + \log \epsilon^{-1})}\right)^{\frac{1}{\beta+1}}$$

denote the q solving $K_1 q^{-\beta-1} = 2(k + \log k + \log \epsilon^{-1})/n$. Then

$$\sum_{q > q_0} p_q \leq \int_1^{q_0^*} \frac{2(k + \log k + \log \epsilon^{-1})}{n} dq + \int_{q_0^*}^{\infty} K_1 q^{-\beta-1} dq$$

$$\leq 2q_0^* \frac{k + \log k + \log \frac{1}{\epsilon}}{n} + \frac{K_1}{\beta} q_0^{*-\beta} \leq C\left(\frac{k + \log k + \log \frac{1}{\epsilon}}{n}\right)^{\frac{\beta}{1+\beta}},$$

where C is a constant dependent on K_1 and β. Lastly, we bound $q_0(\epsilon, n)$. Since $p_q \leq K_1 q^{-\beta-1}$, the inequality (5) can be satisfied only if

$$q \leq \left(\frac{K_1 n}{2(k + \log k + \log \epsilon^{-1})}\right)^{\frac{1}{1+\beta}}.$$

Choosing ϵ to be a function of n via $\epsilon = 1/n^2$, we can use Assumption E and the sandwich inequality (7) to find that there is constant C'—dependent

on K_1 and β—such that

$$\sum_q p_q \mathbb{E}_q \left[\varphi(f(q), s(Y_{1:k})) \right] - C'B_n \left(n^{-2} + \left(\frac{k + \log k + 2\log n}{n} \right)^{\frac{\beta}{1+\beta}} \right)$$

$$(8) \quad \leq R_{\varphi,n}(f)$$

$$\leq \sum_q p_q \mathbb{E}_q \left[\varphi(f(q), s(Y_{1:k})) \right] + C'B_n n^{\frac{-2-3\beta}{1+\beta}} (k + \log k + 2\log n)^{-\frac{1}{1+\beta}}$$

$$+ C'B_n n^{\frac{-\beta}{1+\beta}} (k + \log k + 2\log n)^{\frac{\beta}{1+\beta}}.$$

The two-sided bound (8) implies the lemma.

9.4. *Proof of Lemma 8.* Fix $\epsilon > 0$ and let $f_n^i \in \mathcal{F}_n^i$ be an arbitrary member of each of the sets $\mathcal{F}_n^i$. Then

$$\sup_{f \in \mathcal{F}_n} \left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right|$$

$$= \max_{i \in [N(\epsilon,n)]} \sup_{f \in \mathcal{F}_n^i} \left| \sum_q p_q \ell_\varphi(f(q), \mu_q) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right|$$

$$\leq \max_{i \in [N(\epsilon,n)]} \left| R_\varphi(f_n^i) - \sum_q p_q \mathbb{E}_q[\varphi(f_n^i(q), s(Y_{1:k_n}))] \right|$$

$$+ \max_{i \in [N(\epsilon,n)]} \sup_{f \in \mathcal{F}_n^i} \left| \sum_q p_q \left[\ell_\varphi(f_n^i(q), \mu_q) - \ell_\varphi(f(q), \mu_q) \right] \right|$$

$$+ \max_{i \in [N(\epsilon,n)]} \sup_{f \in \mathcal{F}_n^i} \left| \sum_q p_q \left(\mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] - \mathbb{E}_q[\varphi(f_n^i(q), s(Y_{1:k_n}))] \right) \right|$$

by the triangle inequality. Applying Assumption E and that $\mathcal{F}_n^i$ has radius ϵ, the final two terms are each bounded by $L_n \epsilon$, which implies the bound

$$(9) \quad \max_{i \in [N(\epsilon,n)]} \left| R_\varphi(f_n^i) - \sum_q p_q \mathbb{E}_q[\varphi(f_n^i(q), s(Y_{1:k_n}))] \right| + 2L_n \epsilon.$$

Noting that f_n^i was arbitrary, we can strengthen the bound (9) to

$$\sup_{f \in \mathcal{F}_n} \left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right|$$

$$\leq \max_{i \in [N(\epsilon,n)]} \inf_{f \in \mathcal{F}_n^i} \left| R_\varphi(f) - \sum_q p_q \mathbb{E}_q[\varphi(f(q), s(Y_{1:k_n}))] \right| + 2L_n \epsilon.$$

Applying Lemma 7 completes the proof.

9.5. *Proof of Lemma 9.* Without loss of generality, let $j = 1$, and fix the query judgment pairs $(Q_1, Y_1) = (q, Y)$ and $(Q_1', Y_1') = (q', Y')$ with $Q_i' := Q_i$ for all $i \in \{2, \ldots, n\}$. Recall the notation $\mathcal{B}(q)$ from Section 4, and for any $r \in \mathcal{Q}$, let $\mathcal{B}'(r) := \{i \in [n] \mid Q_i' = r\}$ and define $\widehat{n}_r' = |\mathcal{B}'(r)|$. In addition, let $Y_{i_1:i_k}$ denote $Y_{i_1}, \ldots, Y_{i_k}$.

We focus first on the case in which $q \neq q'$. Note that by definition (19)

$$nF((q, Y), (Q_2, Y_2), \ldots, (Q_n, Y_n)) - nF((q', Y'), (Q_2, Y_2), \ldots, (Q_n, Y_n))$$

(10a)
$$= \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \varphi(f(q), s(Y_{i_1:i_k})) - (\widehat{n}_q - 1)\binom{\widehat{n}_q - 1}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}'(q)}} \varphi(f(q), s(Y_{i_1:i_k}))$$

(10b)
$$+ \widehat{n}_{q'} \binom{\widehat{n}_{q'}}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q')}} \varphi(f(q'), s(Y_{i_1:i_k})) - (\widehat{n}_{q'} + 1)\binom{\widehat{n}_{q'} + 1}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}'(q')}} \varphi(f(q'), s(Y_{i_1:i_k}')).$$

We first bound the term (10a). When $\widehat{n}_q \leq k$, this term becomes

$$\widehat{n}_q \varphi(f(q), s(Y_{i_1:i_{\widehat{n}_q}})) - (\widehat{n}_q - 1)\varphi(f(q), s(Y_{i_2:i_{\widehat{n}_q}}))$$

where $i_1, \ldots, i_{\widehat{n}_q}$ denote the distinct elements of $\mathcal{B}(q)$. By assumption, each of the terms φ is bounded by B, so this difference is at most $2\widehat{n}_q B \leq 2kB$.

When $\widehat{n}_q > k$, we note that the term (10a) equals

(11a)
$$\left[\widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} - (\widehat{n}_q - 1)\binom{\widehat{n}_q - 1}{k}^{-1} \right] \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}'(q)}} \varphi(f(q), s(Y_{i_1:i_k}))$$

(11b)
$$+ \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_2 < \cdots < i_k, \\ i_j \in \mathcal{B}'(q)}} \varphi(f(q), s(Y, Y_{i_2:i_k})).$$

Since there are $\binom{\widehat{n}_q - 1}{k}$ terms in the summation (11a) and $\binom{\widehat{n}_q - 1}{k-1}$ terms in the summation (11b), and for any $m, k \in \mathbb{N}$ with $m > k$,

$$m\binom{m}{k}^{-1} - (m-1)\binom{m-1}{k}^{-1} = \frac{k!(m-k)!}{(m-1)!} - (m-1)\frac{(m-1-k)!k!}{(m-1)!}$$
$$= (1-k)\binom{k}{m-1}^{-1},$$

we use the fact that $|\varphi| \le B$ to conclude that

$$\left| \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \varphi(f(q), s(Y_{i_1:i_k})) \right.$$
$$\left. - (\widehat{n}_q - 1) \binom{\widehat{n}_q - 1}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}'(q)}} \varphi(f(q), s(Y_{i_1:i_k})) \right|$$
$$\le |k-1|B + \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \binom{\widehat{n}_q - 1}{k-1} B = |k-1|B + \widehat{n}_q \frac{k}{\widehat{n}_q} B \le 2kB.$$

Hence, the term (10a) can be bounded by $2kB$ for any choice of $\widehat{n}_q$. Bounding the term (10b) requires an entirely similar combinatorial argument and again yields the bound $2kB$ for a total bound of $4kB$.

All that remains is to control the difference of the functions F when $q = q'$. When $\widehat{n}_q \le k$, this difference is given by

$$(\widehat{n}_q/n)|\varphi(f(q), s(Y_{i_1:i_{\widehat{n}_q}})) - \varphi(f(q), s(Y'_{i_1:i_{\widehat{n}_q}}))| \le 2kB/n.$$

When $\widehat{n}_q \ge k$, we have

$$nF((q, Y), (Q_2, Y_2), \ldots, (Q_n, Y_n)) - nF((q, Y'), (Q_2, Y_2), \ldots, (Q_n, Y_n))$$
$$= \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \left[\varphi(f(q), s(Y_{i_1:i_k})) - \varphi(f(q), s(Y'_{i_1:i_k})) \right]$$
$$(12) \quad = \widehat{n}_q \binom{\widehat{n}_q}{k}^{-1} \sum_{\substack{1 = i_1 < \cdots < i_k, \\ i_j \in \mathcal{B}(q)}} \left[\varphi(f(q), s(Y, Y_{i_2:i_k})) - \varphi(f(q), s(Y', Y_{i_2:i_k})) \right].$$

There are $\binom{\widehat{n}_q - 1}{k-1}$ terms in the final summation, and since

$$m \binom{m}{k}^{-1} \binom{m-1}{k-1} = m \frac{k!(m-k)!}{m!} \cdot \frac{(m-1)!}{(k-1)!(m-k)!} = k,$$

the difference (12) is bounded by $2kB$.

9.6. *Proof of Lemma 10.* Let $\epsilon' > 0$ be arbitrary. Use Assumption F and partition $\mathcal{F}_n$ into $N = N(\epsilon', n) < \infty$ subsets $\mathcal{F}_n^1, \ldots, \mathcal{F}_n^N$, and fix $f^i \in \mathcal{F}_n^i$ so that for any $f \in \mathcal{F}_n^i$ we have $\|f - f^i\| \le \epsilon'$. Then for any index $i \in$

$\{1, \ldots, N(\epsilon', n)\}$, we see that

$$\sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right|$$

$$= \sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - \widehat{R}_{\varphi,n}(f^i) + \widehat{R}_{\varphi,n}(f^i) - R_{\varphi,n}(f^i) + R_{\varphi,n}(f^i) - R_{\varphi,n}(f) \right|$$

$$\leq \left| \widehat{R}_{\varphi,n}(f^i) - R_{\varphi,n}(f^i) \right| + \sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - \widehat{R}_{\varphi,n}(f^i) \right| + \sup_{f \in \mathcal{F}_n^i} \left| R_{\varphi,n}(f) - R_{\varphi,n}(f^i) \right|.$$

Assumption E guarantees that $\left| \widehat{R}_{\varphi,n}(f) - \widehat{R}_{\varphi,n}(f^i) \right| \leq L_n \left\| f - f^i \right\|$, so

$$\sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - \widehat{R}_{\varphi,n}(f^i) \right| \leq L_n \epsilon'.$$

The same argument applies to the difference $R_{\varphi,n}(f) - R_{\varphi,n}(f^i)$. Thus we use the fact that the classes $\mathcal{F}_n^i$ partition $\mathcal{F}_n$, that is, $\cup_{i=1}^N \mathcal{F}_n^i \supseteq \mathcal{F}_n$, and apply a union bound and the triangle inequality to find that

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}_n} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \geq \epsilon \right)$$

$$= \mathbb{P} \left(\max_{i \in \{1, \ldots, N(\epsilon', n)\}} \sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \geq \epsilon \right)$$

$$\leq \sum_{i=1}^{N(\epsilon', n)} \mathbb{P} \left(\sup_{f \in \mathcal{F}_n^i} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \geq \epsilon \right)$$

$$\leq \sum_{i=1}^{N(\epsilon', n)} \mathbb{P} \left(\left| \widehat{R}_{\varphi,n}(f^i) - R_{\varphi,n}(f^i) \right| \geq \epsilon - 2 L_n \epsilon' \right).$$

Choosing $\epsilon' = \epsilon/(4 L_n)$, we can apply the concentration inequality (4) to conclude that

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}_n} \left| \widehat{R}_{\varphi,n}(f) - R_{\varphi,n}(f) \right| \geq \epsilon \right) \leq 2 \exp \left[\log N \left(\frac{\epsilon}{4 L_n}, n \right) - \frac{n \epsilon^2}{32 k^2 B_n^2} \right].$$

This implies the statement of the lemma.

10. Proofs of Inconsistency Results. The proofs in this section are essentially present in the preliminary version [3] of this work. We present and extend them here for the convenience of the reader.

10.1. *Proof of Proposition 3.* This proposition is a consequence of the fact that the feedback arc set problem [5] is *NP*-complete. In the feedback arc set problem, we are given a directed graph $G = (V, E)$ and an 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 (equivalently, whether $G' = (V, E \setminus E')$ is a directed acyclic graph (DAG)).

Now consider the problem of deciding whether there exists an α with $\ell(\alpha, \mu) \leq k$, and let G_μ denote the graph over the nodes associated with adjacency matrices Y, where G_μ has edge weights equal to the average $Y_{ij}^\mu = \int Y_{ij} d\mu(Y)$. Since α induces an order of the nodes in this "expected" graph G_μ, 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 G_μ such that the sum of the back edges is less than k,

$$\sum_{i \succeq j} Y_{ij}^\mu \leq k.$$

Removing all the back edges (edges $(i \to j)$ in the expected graph G_μ such that $i \succeq j$ in the original ordering) leaves a DAG. Now, given a graph $G = (V, E)$, we can construct the expected graph G_μ directly from G with weights $Y_{ij}^\mu = 1$ if $(i \to j) \in E$ and 0 otherwise (set the probability that edge $(i \to j)$ appears to be $1/|E|$ and set the associated ijth adjacency matrix Y to be $Y = 0$ except that $Y_{ij} = |E|$). Then there is an α such that $\ell(\alpha, \mu) \leq k$ if and only if there is a feedback arc set E' with $|E'| \leq k$.

10.2. *Proof of Theorem 2.* Before proceeding to the proof of Theorem 2 proper, we make a few remarks to reduce the space of functions ϕ that we must consider; we also argue that h must be continuous and strictly increasing.

First, if the surrogate φ from the definition (12) is structure-consistent, it must be *classification calibrated*: ϕ must be differentiable at 0 with $\phi'(0) < 0$. This is a consequence of Bartlett, Jordan, and McAuliffe's analysis of binary classification and the correspondence between binary classification and pairwise ranking [1]. The same reasoning implies that we must have $h \geq 0$ on $\mathbb{R}_+$ and h strictly increasing. We now argue that h must be continuous. Assume for the sake of contradiction that h is discontinuous at some $c > 0$, and without loss of generality assume that h is not right continuous at c. Then there exists an $\epsilon > 0$ such that $h(c + \delta) \geq h(c) + \epsilon$ for all $\delta > 0$. Fix δ and γ such that

$$\gamma < \frac{\epsilon}{4h(c) + 2\epsilon} \quad \text{and} \quad \delta < \frac{4\gamma}{1 - 2\gamma} c.$$

Consider the case in which only two items are to be ranked, 1 and 2, and let the preference $Y_{12} = c + \delta$ appear with probability $\frac{1}{2} - \gamma$ and $Y_{21} = c + \delta$ appear with probability $\frac{1}{2} + \gamma$ and $Y_{12} = c$. In this case,

$$(1-2\gamma)\delta < 4\gamma c \ \text{ or } \ \left(\frac{1}{2} - \gamma\right)\delta - \gamma c < \gamma c, \ \text{ so } \ \left(\frac{1}{2} - \gamma\right)(c+\delta) < \left(\frac{1}{2} + \gamma\right)c.$$

Any correct score assignment must thus rank item 1 before item 2, satisfying $\alpha_1 > \alpha_2$. On the other hand, the condition on γ, ϵ, and $h(c)$ implies $2\gamma h(c) < \left(\frac{1}{2} - \gamma\right)\epsilon$, so

$$\left(\frac{1}{2} + \gamma\right) h(c) < \left(\frac{1}{2} - \gamma\right)(h(c) + \epsilon) < \left(\frac{1}{2} - \gamma\right) h(c + \delta)$$

by assumption on the discontinuity of h. In particular, any minimizer of $(\frac{1}{2}+\gamma)h(c)\phi(\alpha_1-\alpha_2)+(\frac{1}{2}-\gamma)h(c+\delta)\phi(\alpha_2-\alpha_1)$ will in this case set $\alpha_2 > \alpha_1$ since ϕ must be classification calibrated. In this case, the discontinuity of h forces mis-ordering, and hence h must be continuous.

The desired properties of h established, consider the recession function [4] of ϕ,

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

For any ϕ bounded below with $\phi'(0) < 0$, we have that $\phi'_\infty(1) \geq 0$. Throughout this analysis, we require the weaker assumption that ϕ decreases more slowly in the positive direction than the negative:

$$(13) \qquad\qquad \phi'_\infty(1) \geq 0 \quad \text{or} \quad \phi'_\infty(-1) = \infty.$$

Consider two DAGs on nodes 1, 2, and 3 that induce only the four penalty values Y_{12}, Y_{13}, Y_{23}, and Y_{31} (recall Figure 1). In this case, if $Y_{13} > Y_{31}$, any α minimizing $\ell(\alpha, \mu)$ must satisfy $\alpha_1 > \alpha_2 > \alpha_3$. 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 $\partial\phi(x)$ and recall the definition $h_{ij} = \int h(Y_{ij})d\mu$. The subgradient conditions for optimality of

$$(14) \ \ \ell_\varphi(\alpha, \mu) = h_{12}\phi(\alpha_1-\alpha_2)+h_{13}\phi(\alpha_1-\alpha_3)+h_{23}\phi(\alpha_2-\alpha_3)+h_{31}\phi(\alpha_3-\alpha_1)$$

are that

$$0 = h_{12}\phi'(\alpha_1 - \alpha_2) + h_{13}\phi'(\alpha_1 - \alpha_3) - h_{31}\phi'(\alpha_3 - \alpha_1)$$
$$(15) \qquad 0 = -h_{12}\phi'(\alpha_1 - \alpha_2) + h_{23}\phi'(\alpha_2 - \alpha_3).$$

We begin by showing that whenever the condition (13) holds for ϕ, there is a finite minimizer of $\ell_\varphi(\alpha, \mu)$ as defined by (14). The lemma is technical, so we provide its proof in Appendix 10.4.

LEMMA 12. *Let the condition (13) hold and $\phi'(0) < 0$. There is a constant $C < \infty$ and a vector α^* minimizing $\ell_\varphi(\alpha, \mu)$ with $\|\alpha^*\|_\infty \le C$.*

The next lemma provides the essential lever we use to prove the theorem; we provide its proof in Appendix 10.5.

LEMMA 13 (Inconsistency of convex losses). *Suppose that $Y_{13} > Y_{31} > 0$, $Y_{12} > 0$, $Y_{23} > 0$. Let*

$$\ell(\alpha, \mu) = Y_{12}1(\alpha_1 \le \alpha_2) + Y_{13}1(\alpha_1 \le \alpha_3) + Y_{23}1(\alpha_2 \le \alpha_3) + Y_{31}1(\alpha_3 < \alpha_1)$$

and $\ell_\varphi(\alpha, \mu)$ be defined as in Eq. (14). For convex ϕ with $\phi'(0) < 0$, we have

$$\ell_\varphi^*(\mu) = \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) \mid \alpha \notin \operatorname*{argmin}_{\alpha'} \ell(\alpha', \mu) \right\}$$

whenever either of the following conditions is satisfied:

CONDITION 1: $h_{23} < \dfrac{h_{31}h_{12}}{h_{13} + h_{12}}$ *or*   CONDITION 2: $h_{12} < \dfrac{h_{31}h_{23}}{h_{13} + h_{23}}$.

Lemma 13 allows us to construct scenarios under which arbitrary pairwise surrogate losses with convex ϕ are inconsistent. Indeed, assume for the sake of contradiction that φ is structure-consistent. Recall that for ϕ convex, $\phi'(0) < 0$ by classification calibration. We will construct graphs G_1 and G_2 (with associated adjacency matrices Y^{G_1} and Y^{G_2}) so that the resulting expected loss satisfies CONDITION 1 of Lemma 13 while additionally satisfying Definition 4's low-noise condition.

Consider the following two graphs:

$$G_1 = (\{1, 2, 3\}, \{(1 \to 2), (1 \to 3)\}), \quad G_2 = (\{1, 2, 3\}, \{(2 \to 3), (3 \to 1)\}).$$

Fix any weights $Y_{12}^{G_1}, Y_{13}^{G_1}, Y_{31}^{G_2}$ with $Y_{13}^{G_1} > Y_{12}^{G_1} > 0$ and $Y_{13}^{G_1} > Y_{31}^{G_2} > 0$, and let μ place half its mass on G_1 (or Y^{G_1}) and half its mass on G_2. 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 $h_{ij} = \frac{1}{2}h(Y_{ij}^{G_1}) + \frac{1}{2}h(Y_{ij}^{G_2})$ as in the definition of the surrogate risk ℓ_φ. Take $Y_{23}^{G_2} = \min\{\varepsilon, (Y_{13}^{G_1} - Y_{12}^{G_1})/2\}$. Then we have

$$h_{23} = \frac{1}{2}h(Y_{23}^{G_2}) \le h(\varepsilon)/2 < \frac{h_{31}h_{12}}{h_{13} + h_{12}}.$$

Hence CONDITION 1 of Lemma 13 is satisfied, so φ is not edge-consistent. Moreover, the fact that

$$Y_{23}^{G_2} \le \frac{Y_{13}^{G_1} - Y_{12}^{G_1}}{2} < Y_{13}^{G_1} - Y_{12}^{G_1}$$

implies that the expected difference graph G_μ satisfies Definition 4.

10.3. *Proof of Theorem 3.*   Assume for the sake of contradiction that φ is structure-consistent. As in the proof of Theorem 2, the function h must be strictly increasing. Now let $c \in \mathbb{R}$ belong to the image $h((0, \infty))$ and define the shifted function $\widetilde{\phi}(x) = \phi(x - c)$. By the reduction to the binary case, $\widetilde{\phi}$ must be classification calibrated, whence $\widetilde{\phi}$ must be differentiable at 0 and satisfy $\widetilde{\phi}'(0) < 0$. Using a technique similar to that in our proof of Theorem 2, we now construct a setting of four graphs and provide probabilities of appearance satisfying the conditions in Lemma 13 and Definition 4.

Consider the following four graphs on nodes $\{1, 2, 3\}$, each with one edge:

$$G_1 = (\{1, 2, 3\}, \{(1 \to 2)\}), \quad G_2 = (\{1, 2, 3\}, \{(2 \to 3)\}),$$
$$G_3 = (\{1, 2, 3\}, \{(1 \to 3)\}), \quad G_4 = (\{1, 2, 3\}, \{(3 \to 1)\}).$$

Choose constant edge weights $Y_{12}^{G_1} = Y_{13}^{G_2} = Y_{23}^{G_3} = Y_{31}^{G_4} = h^{-1}(c) > 0$ (h is increasing and thus invertible), and set the probabilities of appearance to be $\mu \equiv (.25, .01, .5, .24)$. Recalling $\widetilde{\phi}$, we have

$$\ell_\varphi(\alpha, \mu) = \mu(G_1)\widetilde{\phi}(\alpha_1 - \alpha_2) + \mu(G_2)\widetilde{\phi}(\alpha_2 - \alpha_3)$$
$$+ \mu(G_3)\widetilde{\phi}(\alpha_1 - \alpha_3) + \mu(G_4)\widetilde{\phi}(\alpha_3 - \alpha_1).$$

Notably, $\widetilde{\phi}$ is convex and satisfies the recession condition (13). Moreover

$$Y_{13} - Y_{31} = h^{-1}(c)(\mu(G_3) - \mu(G_4)) \geq h^{-1}(c)\left(\mu(G_1) + \mu(G_2)\right) = Y_{12} + Y_{23} > 0,$$

so G_μ is a DAG satisfying the low-noise condition. However, the probabilities μ satisfiy

$$\mu(G_2) < \frac{\mu(G_4)\mu(G_1)}{\mu(G_3) + \mu(G_1)}.$$

Hence, by Lemma 13, we have the contradiction that

$$\ell_\varphi^*(\mu) = \inf_\alpha \left\{ \ell_\varphi(\alpha, \mu) \mid \alpha \notin \operatorname*{argmin}_{\alpha'} \ell(\alpha, \mu) \right\}.$$

10.4. *Proof of Lemma 12.*   Let $\{\alpha^{(n)}\}_{n=1}^\infty \subset \mathbb{R}^3$ be a sequence satisfying $\ell_\varphi(\alpha^{(n)}, \mu) \to \ell_\varphi^*(\mu)$. Suppose for the sake of contradiction that $\limsup_n (\alpha_i^{(n)} - \alpha_j^{(n)}) = \infty$ for some i, j. The convexity of ϕ coupled with $\phi'(0) < 0$ imply that $\limsup_n \phi(\alpha_j^{(n)} - \alpha_i^{(n)}) = \infty$; the recession condition (13) (that $\phi'_\infty(1) \geq 0$ or $\phi'_\infty(-1) = \infty$) guarantees $\limsup_n \ell_\varphi(\alpha^{(n)}, \mu) = \infty$ whenever $i > j$ or $i = 1$ and $j = 3$. We thus have two remaining cases:

(a) $\limsup_n (\alpha_1^{(n)} - \alpha_2^{(n)}) = \infty$ or   (b) $\limsup_n (\alpha_2^{(n)} - \alpha_3^{(n)}) = \infty$.

In case (a), we note that $\alpha_1^{(n)} - \alpha_2^{(n)} = \alpha_1^{(n)} - \alpha_3^{(n)} + \alpha_3^{(n)} - \alpha_2^{(n)}$, but there must exist a constant C such that $|\alpha_1^{(n)} - \alpha_3^{(n)}| \le C$ for all n by our earlier argument. This would imply that $\limsup_n (\alpha_3^{(n)} - \alpha_2^{(n)}) = \infty$, a contradiction. Similarly, for case (b), we may use $\alpha_2^{(n)} - \alpha_3^{(n)} = \alpha_2^{(n)} - \alpha_1^{(n)} + \alpha_1^{(n)} - \alpha_3^{(n)}$ and our ealier argument that $|\alpha_1^{(n)} - \alpha_3^{(n)}| \le C$ for all n to see that case (b) would require $\limsup_n (\alpha_2^{(n)} - \alpha_1^{(n)}) = \infty$, another contradiction.

As a consequence, there must be some $C < \infty$ with $|\alpha_i^{(n)} - \alpha_j^{(n)}| \le C$ for all i, j, n. The conditional surrogate risk $\ell_\varphi(\alpha, \mu)$ is shift invariant with respect to α, so without loss of generality we may assume $\alpha_3^{(n)} = 0$, and thus $|\alpha_i^{(n)}| \le C$. Convex functions are continuous on compact domains [4, Chapter IV.3], and thus some α with $\|\alpha\|_\infty \le C$ attains the infimum

$$\inf_{\|\alpha\|_\infty \le C} \ell_\varphi(\alpha, \mu) = \ell_\varphi^*(\alpha, \mu).$$

10.5. *Proof of Lemma 13.* Lemma 12 shows that the optimal $\ell_\varphi^*(\mu)$ is attained by some finite α. Thus, we fix an α^* satisfying the optimality condition (15) and let $\delta_{ij} = \alpha_i^* - \alpha_j^*$ and $g_{ij} = \phi'(\delta_{ij})$ for $i \ne j$. We make use of the monotonicity of subgradients, that is, $\delta_{ij} > \delta_{kl}$ implies $g_{ij} \ge g_{kl}$ [e.g., 4, Chapter VI]. By condition (15),

$$(16a) \qquad g_{13} - g_{12} = \frac{h_{31}}{h_{13}} g_{31} - \left(1 + \frac{h_{12}}{h_{13}}\right) g_{12}$$

$$(16b) \qquad g_{13} - g_{23} = \frac{h_{31}}{h_{13}} g_{31} - \left(1 + \frac{h_{23}}{h_{13}}\right) g_{23}.$$

Suppose for the sake of contradiction that $\alpha^* \in \operatorname{argmin}_\alpha \ell(\alpha, \mu)$. 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 $g_{13} \ge g_{12}$. If $g_{12} \ge 0$, we thus have that $g_{13} \ge 0$ and by (15), $g_{31} \ge 0$. This is a contradiction since $\delta_{31} < 0$ gives $g_{31} \le \phi'(0) < 0$. Hence, $g_{12} < 0$. By identical reasoning, we also have that $g_{23} < 0$.

Now, $\delta_{23} > 0 > \delta_{31}$ implies that $g_{23} \ge g_{31}$, which combined with the equality (16a) and the fact that $g_{23} = (h_{12}/h_{23}) g_{12}$ (by the first-order optimality equation (15)) gives

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

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 $g_{13} \ge g_{12}$, 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 equality (16b) yields

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

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.

References.

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

[2] P. Billingsley. *Convergence of Probability Measures*. Wiley, Second edition, 1999.

[3] J. C. Duchi, L. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[4] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, New York, 1996.

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

[6] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

[7] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

[8] I. Steinwart. How to compare different loss functions. *Constructive Approximation*, 26:225–287, 2007.

[9] A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[10] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

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

DEPARTMENTS OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
E-MAIL: jduchi@cs.berkeley.edu
         jordan@stat.berkeley.edu

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
E-MAIL: lmackey@stanford.edu