On ranking *via* sorting by estimated expected utility

Clément Calauzènes
Criteo AI Lab
Paris, France
c.calauzenes@criteo.com

Nicolas Usunier
Facebook AI Research
Paris, France
usunier@fb.com

Abstract

Ranking tasks are defined through losses that measure trade-offs between different desiderata such as the relevance and the diversity of the items at the top of the list. This paper addresses the question of which of these tasks are asymptotically solved by sorting by decreasing order of expected utility, for some suitable notion of utility, or, equivalently, *when is square loss regression consistent for ranking* via *score-and-sort?* We answer to this question by finding a characterization of ranking losses for which a suitable regression is consistent. This characterization has two strong corollaries. First, whenever there exists a consistent approach based on convex risk minimization, there also is a consistent approach based on regression. Second, when regression is not consistent, there are data distributions for which consistent surrogate approaches necessarily have non-trivial local minima, and for which optimal scoring function are necessarily discontinuous, even when the underlying data distribution is regular. In addition to providing a better understanding of surrogate approaches for ranking, these results illustrate the intrinsic difficulty of solving general ranking problems with the score-and-sort approach.

1 Introduction

The usual approach in learning to rank is to score each item (e.g., a document) given the input (e.g., a search query), and produce the ranking by sorting in decreasing order of scores. This score-and-sort approach follows the probability ranking principle of information retrieval [29], which stipulates that documents should be rank-ordered according to their estimated probability of relevance to the query.

In practice, the definition of a "good" ranking requires more than estimates of relevance. For instance, in scenarios where several users issue the same query "jaguar" but with different intended meanings (e.g., the animal or the car brand), it is desirable to produce diverse rankings where each user finds a relevant document as early as possible. While in the probability ranking principle, items are ranked in decreasing order of their *expected utility* to the users, in sophisticated ranking tasks with a trade-off between relevance and diversity, the definition of a utility per item is not trivial, if not impossible.

In this paper, we study what ranking tasks are solved via sorting by expected utilities, in a general supervised ranking framework that captures different types of ground-truth signal and losses. Since utilities can serve as target values to learn the scoring function through square loss regression, the optimality of sorting by expected utilities is equivalent to the consistency of regression. The main question we address is thus: *When is square loss regression consistent for ranking* via *score-and-sort?*

The consistency of regression for ranking, and more generally the consistency of convex risk minimization for ranking, are still only partly understood. Existing consistency results only apply to gain-based losses such as the Discounted Cumulative Gain and precision and recall at K [12, 28, 6, 19, 14], for which there is an explicit utility function [see e.g., 8, Table 1]. For other losses, only impossibility results are known. Duchi et al. [15], followed by Calauzènes et al. [7] and Ramaswamy et al. [27] proved that convex approaches are, in general, inconsistent with the usual loss when ranking from

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

pairwise preferences, as well as with the Expected Reciprocal Rank [9] and the well-known Average Precision [22], two metrics that are respectively diversity-inducing and diversity-averse [10]. Thus, extending our analysis to general convex risk minimization, two questions remain open: *i) Are there ranking losses for which regression is not consistent, but for which there exists a consistent convex risk minimization approach?* and *ii) When only non-convex surrogate approaches are consistent, is the learning or optimization problem intrinsically more difficult?*

In Section 3, we describe our main result: sorting by expected utilities is optimal if and only if the sublevel sets of the ranking loss are connected, for a suitable notion of connectedness in the space of permutations. This result identifies the fundamental property of ranking losses that is related to the existence of a consistent regression approach. We give an intuitive interpretation of this result in terms of continuity of optimal scoring functions: when regression is not consistent, there necessarily exist data distributions with continuous conditional distributions $x \mapsto P(y|x)$ but such that all optimal scoring functions are discontinuous. On the other hand, when there is a utility function, expected utilities inherit the regularity properties of $x \mapsto P(y|x)$.

In Section 4, we elaborate on the two types of ranking tasks and discuss surrogate approaches. We first answer question *i)* and show that for every ranking loss, whenever a convex risk minimization approach is consistent, then a suitable regression is consistent. Second, we give elements of answer to question *ii)*: when surrogate losses need be non-convex to be consistent, we show that every Lipschitz surrogate loss must have bad local minima.

Our results establish that the general class of convex surrogate losses cannot solve more tasks than plain regression. This clarifies that sophisticated convex approaches for ranking should be justified by better sample complexity more than "better fit" to a specific ranking task, since asymptotically they are either inconsistent or equivalent to a suitable regression. Moreover, the necessary non-global minima of surrogate losses and the discontinuity of optimal scoring are the first formal arguments for the intrinsic difficulties that arise when using score-and-sort for general ranking problems.

2 Preliminaries: Learning to rank and Consistency

2.1 Learning to Rank

Supervised learning to rank. We consider a framework of label or subset ranking [13, 12]. The learner predicts rankings over n items based on input features $x \in \mathcal{X} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ has nonempty interior. Rankings are represented by permutations, and we denote by $\mathfrak{S}_n$ the set of all permutations of $[n] = \{1, \ldots, n\}$. The learner has access to a supervision signal in $\mathcal{Y}$. When the task is fully supervised, $\mathcal{Y} = \mathfrak{S}_n$, but we also allow for weakly supervised settings where a supervision $y \in \mathcal{Y}$ is a vector of relevance judgements for each item, a preference graph, or a partial ranking. Our analysis is agnostic to the type of supervision, we only assume that $\mathcal{Y}$ is finite. The task loss $L : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$ measures the quality of a ranking given the supervision. The goal is to learn, from supervised training data, a ranking function $h : \mathcal{X} \to \mathfrak{S}_n$ with low *task risk* $\mathcal{R}_{L,P}(h) = \mathbb{E}_P\big[L(Y, h(X))\big]$, where the expectation is taken according to the data distribution P. We give later examples of usual task losses and their associated $\mathcal{Y}$ in Table 1 (Section 3).

The score-and-sort approach. A usual approach to learning to rank is to sort the items by decreasing order of learnt scores. Given a vector of scores $s \in \mathbb{R}^n$, $\mathrm{argsort}(s)$ returns *the set* of permutations that are compatible with a decreasing order of score:

$$\mathrm{argsort}(s) = \{\sigma \in \mathfrak{S}_n : \forall k \in [n-1], s_{\sigma(k)} \geq s_{\sigma(k+1)}\}\,.$$

We overload L for a set of rankings $\pi \subseteq \mathfrak{S}_n$, using its average value $L(y, \pi) = \frac{1}{|\pi|}\sum_{\sigma\in\pi} L(y, \sigma)$. Thus, given a *scoring function* f, i.e., a measurable function $f : \mathcal{X} \to \mathbb{R}^n$, the task risk for the score-and-sort approach is $\mathcal{R}_{L,P}(\mathrm{argsort} \circ f) = \mathbb{E}_P\big[L(Y, \mathrm{argsort}(f(X)))\big]$.

Similarly to previous studies on consistency for ranking [e.g., 12, 15, 28, 7, 8, 26], x contains the information about the input and all items. This is the natural setup when ranking class labels in a multiclass/multilabel setting, since in that case there are only input features (e.g., an image). In recommender systems or search engines, this means that we allos the score of an item to depend on the other available items. This setup makes sure that every ranking function can be implemented by the score-and-sort approach. The difficulty of learning to rank comes from the requirement to learn the scoring function from noisy supervision, instead of from a deterministic ground-truth ranking.

2.2 Consistency, Calibration and Utilities

Calibration of surrogate losses and consistency. In the score-and-sort approach, the scoring function is not trained by minimizing $\mathcal{R}_L(\mathrm{argsort}\circ f)$ (or its empirical counterpart) because it is computationally hard in general. Rather, a *surrogate loss* is used, which is a measurable function $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$. Learning algorithms aim at minimizing the *surrogate risk*, defined for a scoring function $f : \mathcal{X} \to \mathbb{R}^n$ as $\mathcal{R}_{\Phi,P}(f) = \mathbb{E}_P\big[\Phi(Y, f(X))\big]$. We analyze the consistency of surrogate risk minimization, which informally states that minimizing $\mathcal{R}_{\Phi,P}(f)$ over f leads to minimizing $\mathcal{R}_{L,P}(\mathrm{argsort}\circ f)$. More formally, an *excess risk bound* between $\mathcal{R}_\Phi$ and $\mathcal{R}_L$ is a continuous function $\delta : \mathbb{R}_+ \to \mathbb{R}_+$ with $\delta(\epsilon) \xrightarrow[\epsilon\to 0]{} 0$ such that, for every distribution P over $\mathcal{X} \times \mathcal{Y}$:

$$\forall f : \mathcal{X} \to \mathbb{R}^n, \quad \mathcal{R}_{L,P}(\mathrm{argsort}\circ f) - \inf_{h:\mathcal{X}\to\mathfrak{S}_n} \mathcal{R}_{L,P}(h) \le \delta\Big(\mathcal{R}_{\Phi,P}(f) - \inf_{g:\mathcal{X}\to\mathbb{R}^n} \mathcal{R}_{\Phi,P}(g)\Big).$$

where we implicitly restrict to measurable functions. Excess risks bounds give distribution-independent guarantees on the convergence of the task risk given the convergence of the surrogate risk to its infimum. The guarantees are asymptotic in nature, because the infimum over all measurable functions of the risks are, in general, approachable only as the number of samples tends to infinity.

Following Steinwart [31], we study consistency through *inner risks* and *calibration*. Let $\mathcal{Q}$ be the set of probability mass functions over $\mathcal{Y}$. Given $q \in \mathcal{Q}$ and $\psi : \mathcal{Y} \to \mathbb{R}$, let $\mathbb{E}_q\big[\psi(Y)\big] = \sum_{y\in\mathcal{Y}} q(y)\psi(y)$. The inner task risk $\ell : \mathcal{Q} \times \mathfrak{S}_n \to \mathbb{R}$ and the inner surrogate risk $\phi : \mathcal{Q} \times \mathbb{R}^n \to \mathbb{R}_+$ are:

$$\ell(q,\sigma) = \mathbb{E}_q\big[L(Y,\sigma)\big] \qquad \phi(q,s) = \mathbb{E}_q\big[\Phi(Y,s)\big] \qquad \underline{\phi}(q) = \inf_{s\in\mathbb{R}^n} \phi(q,s).$$

The terminology "inner risk" is justified by the equality $\mathcal{R}_{L,P}(h) = \mathbb{E}_P\big[\ell\big(P(.|X), h(X)\big)\big]$, which holds for every $h : \mathcal{X} \to \mathfrak{S}_n$ and where $x \mapsto P(.|x)$ is the conditional distribution of Y given $X = x$. Similarly to L, we extend ℓ to subsets $\pi \subseteq \mathfrak{S}_n$ using an average: $\ell(q,\pi) = \frac{1}{|\pi|}\sum_{\sigma\in\pi} \ell(q,\sigma)$..

Our definition of calibration is adapted from [31, Def 2.7] to the case where $\mathcal{Y}$ and $\mathfrak{S}_n$ are finite:

Definition 1. *Given $L : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$, the surrogate loss $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$ is L-*calibrated *if*

$$\forall q \in \mathcal{Q}, \exists \delta > 0, \forall s \in \mathbb{R}^n, \quad \phi(q,s) - \underline{\phi}(q) \le \delta \implies \mathrm{argsort}(s) \subseteq \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} \ell(q,\sigma)\,.$$

Since $\mathcal{Y}$ is finite, using the notation of Def 1, we have (see [31, Th. 2.8] and Appendix B):

Proposition 2. *Φ is L-calibrated if and only if there is an excess risk bound between $\mathcal{R}_\Phi$ and $\mathcal{R}_L$.*

Consistency of regression and utility functions. We first study a problem equivalent to the consistency of square loss regression. In genral, the supervision y does not necessarily contain suitable target scores for the regression, so we use intermediate utility values. Given a task loss $L : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$, we say that $u : \mathcal{Y} \to \mathbb{R}^n$ is a *a utility function* for L if the square loss $\Phi^{\mathrm{sq}}_u(y,s) = (s - u(y))^2$ is L-calibrated. We say that L is *compatible with expected utility* (CEU) if a utility function exists for L. Denoting ϕ^{sq}_u the inner risk of Φ^{sq}_u, we have $\mathrm{argmin}_s\, \phi^{\mathrm{sq}}_u(q,s) = \{\mathbb{E}_q\big[u(Y)\big]\}$, so that

$$L \text{ is CEU} \iff \exists u : \mathcal{Y} \to \mathbb{R}^n, \forall q \in \mathcal{Q},\ \mathrm{argsort}\Big(\mathbb{E}_q\big[u(y)\big]\Big) \subseteq \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} \ell(q,\sigma).$$

The characterization of CEU task losses for ranking is our main technical result, presented in Section 3. We extend the analysis to general surrogate losses in Section 4.

2.3 Ranking losses

An arbitrary function $\mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$ is not necessarily a valid task loss for ranking. The definition of a *ranking loss* below isolates minimal properties that make a function $\mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$ suitable to specify a ranking task. We use τ_{ij} to denote the transposition that swaps items i and j, i.e., for $i' \in [n]$, $\tau_{ij}(i') = i'$ if $i' \notin \{i,j\}$, $\tau_{ij}(i) = j$ and $\tau_{ij}(j) = i$. Notice that given the definition of argsort in (1), lower ranks are better, $\sigma(k)$ is the item at rank k and $\sigma^{-1}(i)$ is the rank of item i.

Definition 3 (Ranking loss)**.** *A task loss $L : \mathcal{Y} \times \mathfrak{S}_n \to \mathbb{R}$ is a* ranking loss *if its inner risk satisfies*

- *Items are equivalent* a priori*: $\forall q \in \mathcal{Q}, \forall i,j \in [n], \exists q' \in \mathcal{Q}$ s.t. $\forall \sigma \in \mathfrak{S}_n, \ell(q,\sigma) = \ell(q',\tau_{ij}\sigma)$.*

- *One distribution over $\mathcal{Y}$ strictly prefers item i over other items:* $\forall i \in [n],\ \exists q_{\text{top}}^{(i)} \in \mathcal{Q}$ *s.t.:*

  1. $\ell(q_{\text{top}}^{(i)}, .)$ *strictly decreases with the rank of i:*
  $$\forall \sigma, \nu \in \mathfrak{S}_n, \left(\sigma^{-1}(i) < \nu^{-1}(i) \Rightarrow \ell(q_{\text{top}}^{(i)}, \sigma) < \ell(q_{\text{top}}^{(i)}, \nu)\right) ,$$
  2. *Ranks of items other than i do not matter:*
  $$\forall \sigma \in \mathfrak{S}_n, \forall j \neq i \in [n], \forall j' \neq i \in [n],\ \ \ell(q_{\text{top}}^{(i)}, \sigma) = \ell(q_{\text{top}}^{(i)}, \tau_{jj'}\sigma) .$$

The first point of the definition is always satisfied in practice since indexes of items do not carry any meaning. In usual ranking tasks, $q_{\text{top}}^{(i)}$ is trivial. For instance, when the supervision is a vector of relevance judgements, $q_{\text{top}}^{(i)}$ is a Dirac on y that gives a high relevance to i and the same lower relevance to all other items. The only practical restriction of this definition is the requirement of *strict* improvement in the definition of $q_{\text{top}}^{(i)}$. It is satisfied by some notions of DCG, the AP and the ERR, but not by task losses for K-subset selection and top-K ranking tasks (in these tasks, ranks of i larger than K lead to the same loss). In the appendix, we extend all our results to a more general case that captures selection and top-k ranking tasks (see Appendix A for the general setup). In the main paper, we focus on ranking only to keep the exposition simple.

For ranking losses, calibration is an "equality of argmins" instead of an inclusion:

Theorem 4 ([7, Th. 2])**.** *Given a ranking loss L, $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$ is L-calibrated if and only if*
$$\forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall \delta \in (0, \delta_0],\ \ \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \ell(q, \sigma) = \bigcup_{s : \phi(q,s) - \underline{\phi}(q) < \delta} \operatorname{argsort}(s)$$

This equality means that for L-calibrated surrogate losses, properties of sublevel sets of the inner risk $isurrogate$ translate into similar properties on ℓ and vice-versa. The next section characterizes the sublevel sets of ranking losses that are CEU, while Section 4 focuses on calibrated surrogate losses.

3 Utility functions and connectedness

This section studies sublevel sets of inner ranking risks in terms of their connectedness:

Definition 5. *A set $\pi \subseteq \mathfrak{S}_n$ is* connected *if there is a connected set $S \subseteq \mathbb{R}^n$ s.t.* $\pi = \bigcup_{s \in S} \operatorname{argsort}(s)$.

The relationship with topological notion of connectedness is given in Appendix A.

In our main result below, we denote by $\operatorname{lev}_\epsilon \ell(q, .)$ the strict ϵ-sublevel set of the excess inner risk at q: $\operatorname{lev}_\epsilon \ell(q, .) = \{\sigma : \ell(q, \sigma) - \min_{\sigma'} \ell(q, \sigma') < \epsilon\}$.

Theorem 6. *For a ranking loss L, the following statements are equivalent:*

(i) *L is CEU,*

(ii) $\forall q, \operatorname{argmin}_\sigma \ell(q, \sigma)$ *is connected.*

(iii) $\forall \epsilon > 0, \forall q, \operatorname{lev}_\epsilon \ell(q, .)$ *is connected,*

Moreover, the function $\tilde{u} : \mathcal{Y} \to \mathbb{R}^n$ defined as: $\forall i \in [n], \tilde{u}_i(y) = -\sum_{\sigma \in \mathfrak{S}_n} \mathbb{1}_{\{\sigma(1)=i\}} L(y, \sigma)$ *is a utility function for L whenever there exists a utility function for L (i.e., whenever L is CEU).*

The proof is given in Appendix C. The theorem has four parts: the existence of the utility function, the connected argmins, and the connected sublevel sets, which in turn give an explicit formula for a utility function for L when one exists. We discuss these four aspects in more detail below.

3.1 Disconnected argmins and discontinuity of optimal scoring functions

While the connectedness of argmins and of sublevel sets is the fundamental concept underlying our result, we first give a more intuitive interpretation of the connectedness of argmins (point *ii)* of Th. 6) in terms of the continuity of optimal predictors. The full proof is in Appendix D.

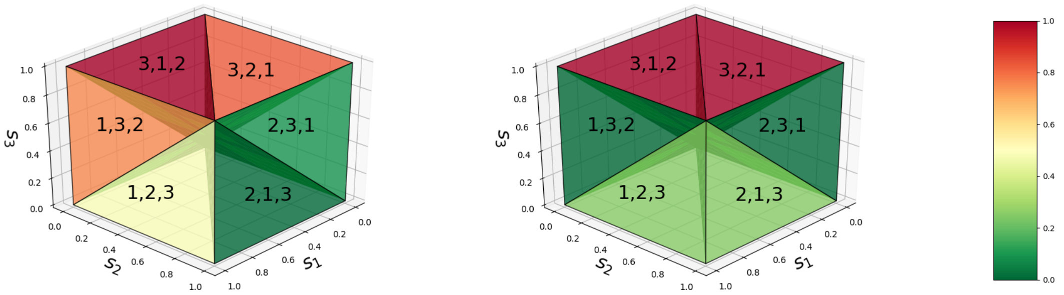

Figure 1: Graphical representation of level sets of the DCG (left) and the ERR (right) for three items. (normalized in [0,1], lower is better). For the DCG, the distribution over $\mathcal{Y}$ is a Dirac on $(1, 2, 0)$. For the ERR it is a mixture of a Dirac on $(1, 1, 0)$ and on $(0, 0, 1)$ with weights $\frac{7}{12}$ and $\frac{5}{12}$. Each axis is the score of one item, and each region is the corresponding argsort. Two regions share an edge (i.e., are adjacent)if the permutations differ only by a transposition of adjacent items. A subset of regions is *connected* if it is possible to stay in that subset while moving from one region to the other using a path of adjacent regions. The sublevel sets for the DCG are all connected, but for the ERR, the sublevel set highlighted in dark green has two connected components.

Corollary 7. *A ranking loss L is CEU if and only if: for every distribution P over $\mathcal{X} \times \mathcal{Y}$ such that $x \mapsto P(.|x)$ is continuous, there is a* continuous *optimal scoring function for $\mathcal{R}_{L,P}$.*

The result follows from the preservation of connectedness by continuous functions. If a ranking loss has a utility function u, an optimal scoring function is $x \mapsto \mathbb{E}_{P(.|x)}[u(Y)]$, which inherits all the regularity properties of $x \mapsto P(.|x)$ since $\mathcal{Y}$ is finite. The value of Cor. 7 is thus to show that *only* CEU losses always have continuous optimal scoring functions when $x \mapsto P(.|x)$ is continuous.

Sketch of proof. The “only if” direction is straightforward. The full proof of the ”if” direction is deferred to the appendix. The main line is the following: take a ranking loss L such that there is a continuous optimal scoring function for every distribution such that $P(.|x)$ is continuous. Aiming for a contradiction, assume that L is not CEU. Then, by Theorem 6, there is a q such that $\operatorname{argmin}_\sigma \ell(q, \sigma)$ is disconnected. Taking $\mathcal{X} = [0, 1]$ without loss of generality and using a uniform marginal distribution over $\mathcal{X}$, we construct a conditional distribution $P(.|x)$ that continuously goes between two distributions q_0 and q_1, where each one of $\operatorname{argmin}_\sigma \ell(q_0, \sigma)$ and $\operatorname{argmin}_\sigma \ell(q_1, \sigma)$ are included in different connected components of q. In that construction, q itself is used as an intermediate point between q_0 and q_1 to make sure that $x \mapsto P(.|x)$ is continuous. If this distribution over $\mathcal{X} \times \mathcal{Y}$ had a continuous optimal scoring function, then this continuous function would connect two distinct connected components of $\operatorname{argmin}_\sigma \ell(q, \sigma)$, which is a contradiction. □

This interpretation of Theorem 6 in terms of continuous optimal predictors indicates that CEU and non-CEU losses lead to learning problems of different intrinsic difficulty. We elaborate more on this dichotomy between CEU and non-CEU ranking losses in the next subsection and in Section 4.

3.2 Disconnected sublevel sets and bad local minima

We now interpret the connectedness of sublevel sets in terms of local minima of the ranking loss. A permutation is a local minimum if exchanging two adjacent items only increases the loss:

Definition 8. *Given a distribution $q \in \mathcal{Q}$ and a loss L, a ranking $\sigma \in \mathfrak{S}_n$ is a* local minimum *if for any $r \in [n-1]$, $\ell(q, \sigma) \leq \ell(q, \sigma\tau_{r,r+1})$.*

The relationship with the connectedness of sublevel sets is more apparent with this characterization of connectedness in $\mathfrak{S}_n$ (the proof is in Appendix C):

Proposition 9. *A set $\pi \subseteq \mathfrak{S}_n$ is connected if and only if for every $\sigma, \nu \in \pi$ there is path between σ and ν in π using transpositions of adjacent items, i.e. $\exists \sigma_0 = \sigma, \sigma_1, ..., \sigma_M = \nu$ where $\forall m, \exists k$: $\sigma_{m+1} = \sigma_m \tau_{kk+1}$ and $\forall m, \sigma_m \in \pi$.*

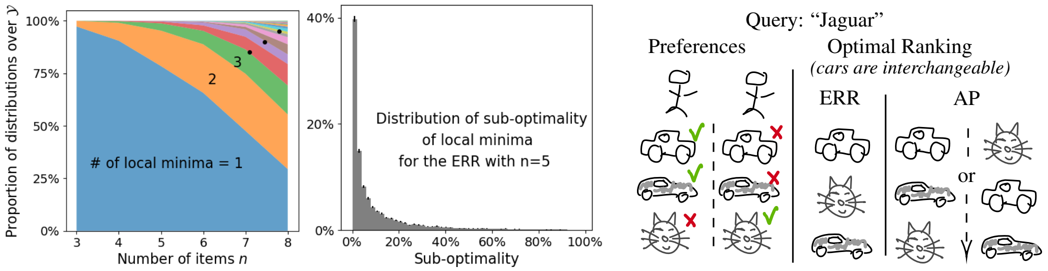

Figure 2: *(left)* Proportions of distributions $q \in \mathcal{Q}$ (y-axis) vs number of local minima (colors) vs number of items (x-axis) for the ERR. For $n = 8$, nearly 70% of distributions have several local minima. *(middle)* distribution of sub-optimality of the sub-optimal local minima for the ERR for $n = 4$. The sub-optimality of a local minimum is $\frac{\text{value}-\min}{\max - \min}$ in $\%$. 0.25% of local minima are more than 90% sub-optimal. *(right)* Illustration of optimal rankings for the ERR (diversity-inducing) and the AP (diversity-averse), for the fictional search engine scenario with the ambiguous query "jaguar".

That is, in a connected set π, we can go from one ranking to another by iteratively swapping adjacent items (as in the bubble sort algorithm), while staying in π. A gaphical representation of connected and disconnected sets is given in Figure 1.

When L is CEU, then by point *iii)* of Th. 6, all sublevel sets are conected, which means that for any ranking in a sublevel set, we can find a path using transpositions of adjacent items to an optimal ranking that never increases the loss. Thus, for CEU losses, all local minima are global minima.

Conversely, when L is not CEU, then for some distributions q, the inner ranking risk ℓ has disconnected sublevel sets, leading to bad local minima. We show in Appendix E that the set $\{q \in \mathcal{Q} : q \text{ has at least one non-global local minimum}\}$ has non-zero Lebesgue measure for all such L. In order to give concrete numbers, we consider the Expected Reciprocal Rank (ERR) [9], which is not CEU as we see later in Section 3.3. By uniformly sampling $q \in \mathcal{Q}$, we empirically estimated proportions of distributions q vs number of local minima for $\ell(q, .)$, for different numbers of items n. The results are plotted in Fig. 2 *(left)*. The probability that $\ell(q, .)$ has several local minima increases rapidly with n, which is expected since when n is small a larger proportion of permutations are adjacent to each other. For $n = 8$, we found that the inner risk has up to 23 local minima, with several local minima for nearly 70% of distributions. Fig. 2 *(middle)* displays the sub-optimality of these local minima, showing that 25% of them are more than 10% sub-optimal. Appendix E contains more details and presents additional results for the Average Precision that are qualitatively similar.

The absence of local minima in L when L is CEU echoes the absence of local minima of the square loss Φ_u^{sq} that is L-calibrated. The result is far from trivial though, because calibration is only a property of minimizers of the inner ranking risk. The strength of our result is to understand that connectedness of argmins actually implies connectedness of all sublevel sets.

3.3 Utilities, diverse rankings and disconnected argmins

We now discuss Theorem 6 in light of prior works. First, it is well-known [12, 6, 28] that gain-based metrics of the form $\text{DCG}_{w,u}(y, \sigma) = -\sum_{k=1}^{n} w_k u_{\sigma(k)}(y)$, which include the Discounted Cumulative gain or precision/recall at K have u as a utility function. We provide example formulas for well-known ranking losses in Table 1, more examples can be found in [see e.g., 8, Table 1]. We show in Appendix F that u is equal to $\tilde{u}$ given in Th. 6 up to an affine transformation.

Conversely, it is also known that the ERR and the AP are not CEU, because of their non-neutral behavior with respect to diverse rankings [6, 7, 26]. To illustrate the effect of diversity on connectedness of the argmins, we use the examples of Calauzènes et al. [7, Table 2]. Let us consider a fictional search engine scenario depicted in Figure 2 (right), in which the ambiguous query "jaguar" is interpreted differently by two annotators (e.g., as the animal, or as the car brand) leading to noise in the supervision. Optimal rankings for a diversity-inducing loss (e.g., the ERR) alternate items relevant for each interpretation, while for a diversity-averse loss (e.g., the AP), they cluster items relevant to the same interpretation. In both cases, we cannot find a path between optimal rankings by swapping adjacent items without leaving the argmin, so the argmin is not connected.

$\mathcal{Y}$	Loss	Formula	$u_i(y)$
$y \in \{0,1\}^n$	Prec@K	$-\sum_{k=1}^{K} \frac{y_{\sigma(k)}}{K}$	y_i
	AP	$\frac{-1}{\|y\|_1} \sum_{i:y_i=1} \mathrm{Prec@}\sigma^{-1}(\mathrm{i})(\mathrm{y}, \sigma)$	$\times$
	AUC	$\sum_{\substack{i:y_i=1 \\ j:y_j=0}} \frac{\mathbb{1}_{\{\sigma^{-1}(i)>\sigma^{-1}(j)\}}}{\|y\|_1(n-\|y\|_1)}$	$\frac{y_i}{\|y\|_1(n-\|y\|_1)}$
$y \in \{0,\dots,p\}^n$	DCG@K	$-\sum_{k=1}^{K} \frac{2^{y_{\sigma(k)}}-1}{\log(1+k)}$	$2^{y_i} - 1$
	NDCG@K	$\frac{-\mathrm{DCG@K}(y,\sigma)}{\max_{\nu \in \mathfrak{S}_n} \mathrm{DCG@K}(y,\nu)}$	$\frac{2^{y_i}-1}{\max_{\nu \in \mathfrak{S}_n} \mathrm{DCG@K}(y,\nu)}$
	ERR	$-\sum_{k=1}^{n} \frac{R_{\sigma(k)}}{k} \prod_{r=1}^{k-1} (1-R_{\sigma(r)})\,, \qquad R_i = \frac{2^i-1}{2^p}$	$\times$
$y \in \mathrm{DAG}_n$	PD	$\sum_{i \to j \in y} \mathbb{1}_{\{\sigma^{-1}(i)>\sigma^{-1}(j)\}}$	$\times$
$y \in \mathfrak{S}_n$	Spearman	$\frac{6}{n(n^2-1)} \sum_{i=1}^{n} (\sigma^{-1}(i) - y^{-1}(i))^2 - 1$	$n - y^{-1}(i)$

Table 1: Example of ranking losses with their utilities, if any. We give examples with different types of supervision, including DAG_n, which is the set of directed acyclic graphs used in the computation of the pairwise disagreement (PD) studied by Duchi et al. [15].

The works above studied convex approaches for ranking. As we explain next, compared to these works, the value of Th. 6 is to show that only CEU ranking losses have a convex, calibrated surrogate.

4 Calibrated Surrogate Losses

When a ranking loss L has a utility function u, the square loss $\Phi_u^{\mathrm{sq}} : y, s \mapsto (s - u(y))^2$ is L-calibrated. Many other convex surrogate losses are applicable when a utility function is available: losses based on a general Bregman divergence D of the form $y, s \mapsto D(\tilde{u}(y), \psi(s))$ where ψ is a link function [28], pointwise losses such as $y, s \mapsto \sum_i \tilde{u}_i(y) \log(1 + e^{-s_i}) + \lambda \sum_i \log(1 + e_i^s)$ where $\lambda > 0$ is a hyperparameter, as well as pairwise losses such as $y, s \mapsto \sum_{i,j} \tilde{u}_i(y) \log(1 + e^{s_j - s_i})$ [8]. These losses are *convex*, in the sense that $\Phi : \mathcal{Y} \times \mathbb{R}^n\mathbb{R}_+$ is convex when every $y \in \mathcal{Y}$, the function $s \mapsto \Phi(y, s)$ is convex. Also, for all of them, the minimizers of the inner risk are equal to expected utilities up to a strictly monotonic transform. These losses are well understood, and come with numerous guarantees, such as explicit excess risk bounds for gain-based losses [12, 28, 8]. There are also fast rates of estimation of the scoring function when $x \mapsto P(.|x)$ is smooth [see e.g., 4].

Our main result above, Theorem 6,implies that whenever there exists a convex, calibrated loss, then these utility-based surrogate losses are calibrated (if given the suitable utility function):

Theorem 10. *Given a ranking loss L, there is an L-calibrated* convex *loss if and only if L is CEU.*

Proof of Theorem 10. The reverse implication is straightforward: if L admits a utility u, then $y, s \mapsto (s - u(y))^2$ is convex and L-calibrated. For the direct implication, let $q \in \mathcal{Q}$. Since $\phi(q, .)$ is convex, its sublevel sets are all convex. Thus, if Φ is L-calibrated, then $\mathrm{argmin}_\sigma\, \ell(q, \sigma)$ is generated by a convex (and thus connected) set of scores by Theorem 4. By the definition of connectedness (Def. 5) $\mathrm{argmin}_\sigma\, \ell(q, \sigma)$ is connected. This holds for all q, so L admits a utility function by Theorem 6. □

This result establishes the fundamental role of utilities and regression in learning to rank with convex risk minimization: for non-CEU ranking losses, only non-convex surrogate losses are calibrated. For instance, it implies that sophisticated convex approaches such as structural SVMs for ranking [see e.g., 21, 39] do not asymptotically solve more tasks than regression.

4.1 The case of convex surrogate losses

Th. 10 closes the question of the calibration of convex losses for ranking. The question was initiated by Duchi et al. [15, Section 2.1] in the context of learning from pairwise preferences, and then extended to general ranking and classification losses [7, 26].

Calauzènes et al. [7] proved that the AP, the ERR and the pairwise disagreement cannot have calibrated, convex surrogate losses because $\operatorname{argmin}_\sigma \ell(q, ., \sigma)$ is not connected for specific choices of q. Their result is essentially *i)* $\implies$ *ii)* of our Theorem 6. By proving the reverse implication, we show that *the general class of convex surrogate losses is not more general than square loss regression.* Notice that *i)* $\implies$ *ii)* is an immediate consequence of Th. 4 (also [7, Theorem 2]), while proving the reverse implication is technically much more challenging.

Ramaswamy and Agarwal [26] defined the concept of *convex calibration dimension*. On the one hand, their approach is more general and they study the minimum number of degrees of freedom that are required for a convex approach to be calibrated. However, they use an unstructured inference procedure, which in general does not correspond to sorting and may be computationally hard. Thus, they do not study when convex losses are calibrated with a score-and-sort approach.

4.2 When no convex surrogate loss is calibrated

For non-CEU ranking losses, Th. 10 shows that L-calibrated surrogate losses are necessarily non-convex. Additionally, we explained in Sections 3.1 and 3.2 that non-CEU ranking losses have a complex landscape with local minima and discontinuous optimal scoring functions. We now describe analogous undesirable properties for their calibrated (non-convex) surrogate losses.

To illustrate the claims of this section, we perform simulations using a non-convex surrogate loss defined by smoothing the task loss, similarly to [33, 35, 20]. We follow the idea of the probit loss for classification [18], which is a Gaussian smoothing of the 0/1 loss. Here, we use a Gumbel kernel κ instead of a Gaussian kernel, because the resulting smoothed loss has a closed form formulation using the Plackett-Luce model [38], and is always L-calibrated (a proof is given in Appendix G).

Proposition 11. *For any ranking loss L, the following surrogate loss Φ_L^{NC} is L-calibrated.*

$$\Phi_L^{\text{NC}} : y, s \mapsto \int_{\mathbb{R}^n} L(y, \operatorname{argsort}(u - s))\kappa(u)\mathrm{d}u = \sum_{\sigma \in \mathfrak{S}_n} L(y, \sigma) \prod_{r=1}^{n} \frac{e^{s_{\sigma(r)}}}{\sum_{k \geq r} e^{s_{\sigma(k)}}}$$

Existence of sub-optimal local minima. Th. 12 below is the counterpart of Section 3.2 for surrogate losses. In essence, it shows that when L is not CEU, then bounded, Lipschitz surrogate losses that are L-calibrated have bad local minima. In order to deal with surrogate losses like Φ_L^{NC} where critical points are at infinity, we use the following generalization of local minimum. Let $\operatorname{lev}_\epsilon \phi(q, .)$ denote the strict ϵ-sublevel set of the excess inner (surrogate) risk: $\operatorname{lev}_\epsilon \phi(q, .) = \{s \in \mathbb{R}^n : \phi(q, s) - \underline{\phi}(q) < \epsilon\}$. Given $\epsilon > 0$, a *local valley* is a connected component of $\operatorname{lev}_\epsilon \phi(q, .)$. A *bad local valley* is a local valley C such that $\inf_{s \in C} \phi(q, s) > \underline{\phi}(q)$ [23, Def. 4.1].

Theorem 12. *Let L be a non-CEU ranking loss. Let $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$ such that $s \mapsto \Phi(y, s)$ is bounded and Lipschitz for every $y \in \mathcal{Y}$. If Φ is L-calibrated, then the set $\{q \in \mathcal{Q} :$ $\phi(q, .)$ has bad local valleys$\}$ has non-zero Lebesgue measure.*

To give more precise figures on these local valleys, Fig. 3 *(left and middle)* shows the distribution of sub-optimalities of bad local valleys for the surrogate Φ_L^{NC} applied to the ERR and the AP. These local valleys are found by gradient descent, starting from a relatively large random initialization to analyze the global loss surface. The distributions q are uniformly sampled over $\mathcal{Q}$, rejecting the distributions q where $\ell(q, .)$ does not have any local minima, as for Fig. 2 *(middle)*. Fig. 3 *(right)* show the proportions or runs on these distributions that end up stuck in a bad local valley when using an initialization close to 0 (which empirically was best to avoid bad local valleys). As for the ranking losses, we see that the surrogate losses also have “bad local minima”, but also that gradient descent algorithms might be stuck in them. Details of these experiments are given in Appendix H.

(Bad) approximations of optimal scoring functions by Lipschitz functions. In Section 3.1, we showed that optimal scoring functions for non-CEU ranking losses are discontinuous for some

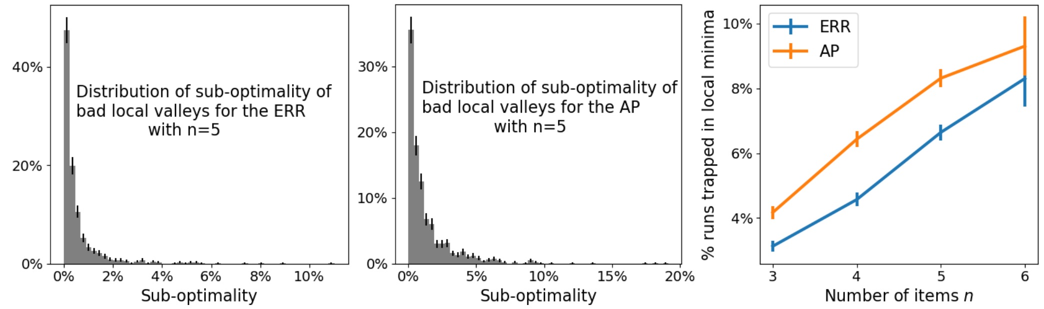

Figure 3: *Left and middle* Distributions of sub-optimalities of local valleys for the surrogate Φ_L^{NC} applied to the ERR and the AP on random distributions where the ERR/the AP have bad local minima. *(right)* percentage of optimization runs based on gradient descent that end up stuck in local minima.

distributions where $x \mapsto P(.|x)$ is continuous. Prop. 13 below is a stronger analogous statement for surrogate losses when $x \mapsto P(.|x)$ is Lipschitz. The proof can be found in Appendix H.

Proposition 13. *Let L be a non-CEU ranking loss. Let $\Phi : \mathcal{Y} \times \mathbb{R}^n \to [0, B_\Phi]$ be an L-calibrated loss such that $\Phi(y, .)$ is β_Φ-Lipschitz for all $y \in \mathcal{Y}$.*

Then, there exists constants $c, c' > 0$ and a probability measure P over $[0,1] \times \mathcal{Y}$ where $x \mapsto P(.|x)$ is Lipschitz, such that for all $\beta \geq 0$:

$$\inf_{\substack{f:[0,1]\to\mathbb{R}^n \\ f:\beta-\mathit{Lipschitz}}} \mathcal{R}_{\Phi,P}(f) - \inf_{g:\mathcal{X}\to\mathbb{R}^n} \mathcal{R}_{\Phi,P}(g) \geq \min\big(c', \frac{c}{8B_\Phi + \beta_\Phi \beta}\big).$$

Notice that for CEU ranking losses with utility function u, denoting β_P the Lipschitz coefficient of $x \mapsto P(.|x)$, the optimal scoring function is $\|u\|_\infty \beta_P$-Lipschitz. In contrast, for non-CEU losses, the lower bound above shows that for some distributions, the Lipschitz constant of the scoring function β needs to grow to infinity to minimize a Lipschitz, calibrated surrogate loss.

5 Dicussion and conclusion

For supervised ranking with the score-and-sort approach, learning the scoring function through regression is consistent for all ranking tasks for which a convex risk minimization approach is consistent. When regression is not consistent, surrogate risks have bad local minima and their minimizers cannot be approximated by regular functions even when $P(.|x)$ is regular. These results demonstrate the fundamental role of regression among convex methods for ranking. They also cast light on the undesirable properties of the score-and-sort approach for non-CEU ranking losses.

For tasks with non-CEU ranking losses, one possible avenue is to develop efficient direct loss minimization approaches, such as approximations of Φ_L^{NC} above or as proposed by Song et al. [30]. Another direction is to find alternatives to score-and-sort. Ramaswamy and Agarwal [26] developped a more general approach which allows for consistent convex approaches, but in higher dimension than the number of items. Excess risk bounds follow from the work on stuctured prediction [11, 25]. The drawback of these approaches is that inference might be NP-hard [15, 26]. In multilabel classification, in some rare cases, efficient inference procedures have been found [36], but not yet in ranking. A third direction is to relax the requirement of asymptotic optimality. By choosing a sensible but efficient inference procedure and making additional assumptions on the data distribution. Chapelle et al. [10] followed this approach for divesifying search results. For now, the theoretical properties of for such approaches have not been investigated. A possible starting point would be to build on the recent work on excess risk bounds for non-calibrated losses [32].

Broader Impact

The framework of ranking we studied plays a fundamental role in information retrieval and information filtering systems. While these systems play an undeniable positive role in society because they increase the efficiency of information access, they also shape the landscape of what their users get to know. This lead to questions regarding equal representation in search engines result [34], how they represent social groups [17], and whether they exacerbate filter bubbles or act as echo chambers [3]. As the most recent trends in learning to rank involve randomized experiments and online learning [16, 1] to improve quality, such user experiments also need be carried out with care [5].

Creating a strong theory of learning to rank is important to address the challenges of information access systems. For instance, methods to produce diverse rankings might be part of the solution, and for now the theory of machine learning for diverse rankings is scarse. This paper partly fills this gap.

References

[1] A. Agarwal, K. Takatsu, I. Zaitsev, and T. Joachims. A general framework for counterfactual learning-to-rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5–14, 2019.

[2] F. G. Arenas. Alexandroff spaces. 1999.

[3] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[4] A. R. Barron. *Complexity Regularization with Application to Artificial Neural Networks*, pages 561–576. Springer Netherlands, Dordrecht, 1991.

[5] S. Bird, S. Barocas, K. Crawford, F. Diaz, and H. Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[6] D. Buffoni, C. Calauzènes, P. Gallinari, and N. Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 825–832, 2011.

[7] C. Calauzènes, N. Usunier, and P. Gallinari. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems 25*, pages 197–205. 2012.

[8] C. Calauzènes, N. Usunier, and P. Gallinari. Calibration and regret bounds for order-preserving surrogate losses in learning to rank. *Machine learning*, 93(2-3):227–260, 2013.

[9] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630, 2009.

[10] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[11] C. Ciliberto, A. Rudi, and L. Rosasco. A consistent regularization approach for structured prediction. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4419–4427, 2016.

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

[13] O. Dekel, Y. Singer, and C. D. Manning. Log-linear models for label ranking. In *Advances in neural information processing systems*, pages 497–504, 2004.

[14] K. Dembczynski, W. Kotlowski, and E. Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012.

[15] J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 327–334, 2010.

[16] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017.

[17] M. Kay, C. Matuszek, and S. A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM, 2015.

[18] J. Keshet and D. A. McAllester. Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing Systems 24*, pages 2205–2212. 2011.

[19] W. Kotłowski, K. Dembczyński, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1113–1120, 2011.

[20] J.-W. Kuo, P.-J. Cheng, and H.-M. Wang. Learning to rank from bayesian decision inference. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 827–836, 2009.

[21] Q. Le and A. Smola. Direct optimization of ranking measures. *arXiv preprint arXiv:0704.3359*, 2007.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[23] Q. Nguyen. On connected sublevel sets in deep learning. *arXiv preprint arXiv:1901.07417*, 2019.

[24] E. A. Ok. *Real Analysis with Economic Applications*. Number mathecon1 in Online economics textbooks. SUNY-Oswego, Department of Economics, January 2004.

[25] A. Osokin, F. Bach, and S. Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pages 302–313, 2017.

[26] H. G. Ramaswamy and S. Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

[27] H. G. Ramaswamy, S. Agarwal, and A. Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2013.

[28] P. Ravikumar, A. Tewari, and E. Yang. On ndcg consistency of listwise ranking methods. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 618–626, 2011.

[29] S. E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4): 294–304, 1977.

[30] Y. Song, A. Schwing, R. Urtasun, et al. Training deep neural networks via direct loss minimization. In *International Conference on Machine Learning*, pages 2169–2177, 2016.

[31] I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

[32] K. Struminsky, S. Lacoste-Julien, and A. Osokin. Quantifying learning guarantees for convex but inconsistent surrogates. In *Advances in Neural Information Processing Systems*, pages 669–677, 2018.

[33] M. Taylor, J. Guiver, S. Robertson, and T. Minka. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86, 2008.

[34] L. Vaughan and Y. Zhang. Equal representation by search engines? a comparison of websites across countries and domains. *Journal of computer-mediated communication*, 12(3):888–909, 2007.

[35] M. N. Volkovs and R. S. Zemel. Boltzrank: learning to maximize expected ranking gain. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1089–1096, 2009.

[36] W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, and E. Hüllermeier. On the bayes-optimality of f-measure maximizers. *Journal of Machine Learning Research*, 15:3333–3388, 2014.

[37] R. Wijsman. Continuity of the Bayes risk. *The Annals of Mathematical Statistics*, 41(3): 1083–1085, 1970.

[38] J. I. Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2): 109–144, 1977.

[39] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278, 2007.

A General Setting

Disclaimer: As mentioned in the body of the paper, we address, in this Appendix, a setting more general than ranking. We pay attention to often reinterpret the results with the ranking setting described in the paper to ease the reader's task.

In supervised learning, a learner has access to input features $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and wants to predict output variables $z \in \mathcal{Z}$ (finite). To do so, she can learn a prediction function $h : \mathcal{X} \to \mathcal{Z}$, belonging to the space of measurable functions from $\mathcal{X}$ to $\mathcal{Z}$ denoted $\mathcal{H}$, using some feedback $y \in \mathcal{Y}$ (finite) and minimizing a task loss $L : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ in expectation over some joint distribution P on the observable variables:

$$\inf_{h \in \mathcal{H}} \mathcal{R}_{L,P}(h) \qquad \text{where } \mathcal{R}_{L,P}(h) = \mathbb{E}_P \left[L(Y, h(X)) \right]$$

Note that the infimum exists because L is bounded (because $\mathcal{Z}$ and $\mathcal{Y}$ being finite). We denote by $\|L\|_\infty = \max_{y,z} |L(y, z)|$.

Prediction space: $\mathcal{Z}$ vs $\mathfrak{S}_n$. Intuitively, when predictions are ranking $\sigma \in \mathfrak{S}_n$, they cannot encode *indifference* between items (the ordering is strict). However, all the results still holds when items can be indifferent (the ordering is a weak order). Further, moving from total orders to weak orders allows to extends the range of applications. For instance, multiclass classification can be seen as predicting a weak order amongst those that strictly prefers one item (label) to all the other items, that are indifferent one from each other. Top-k ranking and subset selection can be handled similarly. The prediction space $\mathcal{Z}$ allows to handle the additional formalism (detailed below) required to handle this generalization.

Two notes on notation. For $n \in \mathbb{N}$, we denote $[n]$ the set of integers going from 1 to n. The indicator function is denoted $\mathbb{1}_{[.]}$. For $s \in \mathbb{R}^n$ and $\varepsilon > 0$, $\mathcal{B}_2(s, \varepsilon) = \{s' \in \mathbb{R}^n, \|s' - s\|_2 < \varepsilon\}$. Similarly, $\mathcal{B}_\infty(s, \varepsilon)$ denotes the open ball of infinity norm. For $q \in \mathcal{Q}$,the $\|.\|_1$ ball in $\mathcal{Q}$ is denoted by $\mathcal{B}_1(q, \varepsilon) = \{q' \in \mathcal{Q} : \|q - q'\|_1 < \varepsilon\}$. Random variable are uppercase version of their realisation counterparts. Given a function $g : \mathcal{U} \to \mathbb{R}$ bounded below and some $\varepsilon \geq 0$, we define the strict ε-optimal (relative) sub-level set of g,

$$\operatorname{lev}_\varepsilon g = \{u \in \mathcal{U} : g(u) - \inf_{u' \in \mathcal{U}} g(u') < \varepsilon\} .$$

Further, we denote $\operatorname{argmin} g = \operatorname{argmin}_{u \in \mathcal{U}} g(u)$.

A.1 Score-and-sort for weak orders ($\mathcal{Z}$)

Scores and argsort We consider the case of ranking or selection tasks solved by sorting according to predicted scores. The basic operation we consider, called argsort, is the set-value function that associates to a vector of scores given to items the set of total orders compatible with the scores:

$$\begin{aligned} \operatorname{argsort} : \mathbb{R}^n &\twoheadrightarrow \mathfrak{S}_n \\ s &\mapsto \{\sigma \in \mathfrak{S}_n : \forall k \in [n-1], s_{\sigma(k)} \geq s_{\sigma(k+1)}\} . \end{aligned} \tag{1}$$

Note that we consider argsort as a set-valued function (and not as a function with values in $2^{\mathfrak{S}_n}$), so that follows the usual convention, for any subset $S \in \mathbb{R}^n$, the notation $\operatorname{argsort}(S)$ is defined as:

$$\forall S \subseteq \mathbb{R}^n, \operatorname{argsort}(S) = \bigcup_{s \in S} \operatorname{argsort}(s) .$$

Transpositions We make heavy use of transpositions: given i, j in $[n]$, the transposition of i and j is the permutation τ_{ij} such that $\tau_{ij}(i') = i'$ if $i' \notin \{i, j\}$, $\tau_{ij}(j) = i$ and $\tau_{ij}(i) = j$. Given σ, we denote by $\tau_{ij}\sigma$ the composition of τ_{ij} and σ.

Given $z \subseteq \mathfrak{S}_n$, we denote by $\tau_{ij} z = \{\tau_{ij}\sigma : \sigma \in z\}$.

Notice the following difference between a subset $z \subseteq \mathfrak{S}_n$ and a permutation $\sigma \in \mathfrak{S}_n$ when it comes to transpositions: the sequential application of transpositions to a permutation commute and it is an associative operation (because they are all members of the symmetric group). That is, if $\sigma \in \mathfrak{S}_n$ we have $\tau_{ij}\tau_{i'j'}\sigma = (\tau_{ij}\tau_{i'j'})\sigma = \tau_{i'j'}\tau_{ij}\sigma$. This is not the case in general when applying sequences of transpositions in $\mathcal{Z}$. For instance let $z = \{(123), (213)\}$. Then $\tau_{12}z = z$, so $\tau_{13}\tau_{12}z = \tau_{13}z = \{(321), (231)\}$, while $\tau_{12}\tau_{13}z = \tau_{12}\{(321), (231)\} = \{(312), (132)\}$. Thus, the composition of a permutation and an element of $\mathcal{Z}$ is not defined in general, it is only defined for transpositions, and these should only be read as usual function compositions.

The decision space The final prediction space corresponds to a selection or partial ranking task, generally speaking, a weak order. The decision space $\mathcal{Z}$ should satisfy the following properties:

(A1). *The decision space $\mathcal{Z} \subseteq 2^{\mathfrak{S}_n}$ satisfies*

(Total preorders) $\forall z \in \mathcal{Z}, \exists s \in \mathbb{R}^n : z = \operatorname{argsort}(s) \subseteq \mathfrak{S}_n$.

(Partition of $\mathfrak{S}_n$) $\bigcup_{z\in\mathcal{Z}} z = \mathfrak{S}_n$ *and* $\forall z, w \in \mathcal{Z}, (z \neq w) \Rightarrow (z \cap w = \emptyset)$.

(Closed under transpositions) $\forall i, j, \forall z \in \mathcal{Z}, \tau_{ij}z \in \mathcal{Z}$.

Since $\mathcal{Z}$ partitions $\mathfrak{S}_n$, elements of $\mathcal{Z}$ can be seen as equivalence classes on rankings in $\mathfrak{S}_n$, and we can define the function Λ as the quotient map which assigns a permutation to its representative member of $\mathcal{Z}$:

$$\begin{aligned} \Lambda : \mathfrak{S}_n &\to \mathcal{Z} \\ \sigma &\mapsto z \ \text{ s.t. } \ \sigma \in z \,, \end{aligned}$$

The prediction function is then the composition $\Lambda \circ \operatorname{argsort}$:

$$\forall s\mathbb{R}^n, \ \ \operatorname{pred}(s) = \Lambda \circ \operatorname{argsort}(s) = \{\Lambda(\sigma) : \sigma \in \operatorname{argsort}(s)\} \,.$$

A straightforward consequence of Assumption (A1) that we make use of when discussing connectedness later is that Λ is equivariant by transposition[1]:

$$\forall \sigma, \ \ \forall i, j, \ \ \Lambda(\tau_{ij}\sigma) = \tau_{ij}\Lambda(\sigma) \,.$$

Notation for orders and preorder Given $\sigma \in \mathfrak{S}_n$, and $i, j \in [n]$, we use the notation

$$i \succ_\sigma j \ \ \Leftrightarrow \ \ \sigma^{-1}(i) < \sigma^{-1}(j) \qquad \text{"}\sigma \text{ prefers } i \text{ to } j\text{"}.$$

This is coherent with an ordering in decreasing order of scores in (1), where $\sigma^{-1}(i)$ is the rank of item i and lower ranks are associated with higher scores (i.e., rank 1 is best), the usual convention in learning to rank.

We use the following notation for the preorder induced by z on $[n]$: $\forall z \in \mathcal{Z}, \forall i, j \in [n]$:

- $i \succeq_z j \quad \Leftrightarrow \quad \exists \sigma \in z, i \succ_\sigma j$,
- $i \succ_z j \quad \Leftrightarrow \quad \forall \sigma \in z, i \succ_\sigma j$,
- $\succeq_z$ induces *indifference classes*: $i \sim_z j \quad \Leftrightarrow \quad \exists \sigma, \sigma' \in z, i \succ_\sigma j$ and $j \succ_{\sigma'} i$.

Given the properties of $\mathcal{Z}$ (Assumption (A1)), the following properties are straightforward to prove: $\forall z \in \mathcal{Z}, \forall s \in \mathbb{R}^n$ s.t. $z = \operatorname{argsort}(s), \forall i, j \in [n]$:

$$\begin{aligned} &a) i \succeq_z j \Leftrightarrow s_i \geq s_j \quad b) i \succ_z j \Leftrightarrow s_i > s_j \quad c) i \sim_z j \Leftrightarrow s_i = s_j \\ &d) i \sim_z j \Leftrightarrow \tau_{ij}z = z \end{aligned} \tag{2}$$

a), b), c) above simply mean that our definition of $\succeq_z$ is coherent with the preorder induced by the scores. d) is a direct consequence of c). Another noticeable implication of (A1) is that the number of indifference classes is the same for all z. This can be seen by combining d) above together with the fact that z is both a partition and stable by transposition.

[1] proof: $\tau_{ij}\Lambda(\sigma) \in \mathcal{Z}$ because $\mathcal{Z}$ is closed by transpositions. By the definition of $\tau_{ij}z$ for $z \in \mathcal{Z}$, $\tau_{ij}\Lambda(\sigma)$ must contain $\tau_{ij}\sigma$. Since $\mathcal{Z}$ is a partition of $\mathfrak{S}_n$, there is a unique element of $\mathcal{Z}$ that contains $\tau_{ij}\sigma$, which is, by definition of Λ, $\Lambda(\tau_{ij}\sigma)$.

A.2 Assumption on task losses

Let $\mathcal{Q}$ be the set of probability mass functions over $\mathcal{Y}$. For any $q \in \mathcal{Q}$, we define the *inner risk* for the task loss L as

$$\forall z \in \mathcal{Z}, \ell(q, z) = \sum_{y \in \mathcal{Y}} q(y) L(y, z) \qquad \text{and} \qquad \underline{\ell}(q) = \min_{z \in \mathcal{Z}} \ell(q, z)$$

We now detail our assumptions that subsumes the *ranking task loss* definition (Def. 3).

(A2) (Items are equivalent *a priori*). $\forall q \in \mathcal{Q}, \forall i, j \in [n], \exists q' \in \mathcal{Q}$ *s.t.* $\forall z \in \mathcal{Z}, \ell(q, z) = \ell(q', \tau_{ij} z)$.

For the next assumption, for any z and any item i, let us denote the rank of i in the preorder z as:

$$z^{-1}(i) = |\{j \in [n] : j \succ_z i\}| + 1.$$

The $+1$ makes sure this definition of rank matches the definition for permutations (first rank is 1 rather than 0).

(A3) (One distribution over $\mathcal{Y}$ strictly prefers item i over other items)**.** $\forall i \in [n], \exists q^{(i)}_{\text{top}} \in \mathcal{Q}$ *such that*

1. $\forall z, w \in \mathcal{Z}, \left(z^{-1}(i) < w^{-1}(i) \Rightarrow \ell(q^{(i)}_{\text{top}}, z) < \ell(q^{(i)}_{\text{top}}, w)\right)$
2. $\forall z \in \mathcal{Z}, \forall k \neq i, \forall l \neq i, \ell(q^{(i)}_{\text{top}}, z) = \ell(q^{(i)}_{\text{top}}, \tau_{kl} z)$

The following lemma is a straightforward consequence of the three assumptions above:

Lemma 14. *Under assumptions (A1), (A2) and (A3), we have:*

i) $\forall i \in [n], \exists q^{(i)}_{\text{bottom}} \in \mathcal{Q}$ *such that*

  (a) $\forall z, w \in \mathcal{Z}, \left(z^{-1}(i) < w^{-1}(i) \Rightarrow \ell(q^{(i)}_{\text{bottom}}, z) > \ell(q^{(i)}_{\text{bottom}}, w)\right)$

  (b) $\forall z \in \mathcal{Z}, \forall k \neq i, \forall l \neq i, \ell(q^{(i)}_{\text{bottom}}, z) = \ell(q^{(i)}_{\text{bottom}}, \tau_{kl} z)$

ii) $\forall z \in \mathcal{Z}, \exists q^{(z)} \in \mathcal{Q}$ *such that* $\operatorname{argmin} \ell(q^{(z)}, .) = \{z\}$.

Proof. Notice that without loss of generality, using (A2), we can assume $\forall z, \ell(q^{(i)}_{\text{top}}, z) = \ell(q^{(j)}_{\text{top}}, \tau_{ij} z)$

Then, define

$$q^{(i)}_{\text{bottom}} = \frac{1}{n-1} \sum_{j \neq i} q^{(j)}_{\text{top}}.$$

$q^{(i)}_{\text{bottom}} \in \mathcal{Q}$ by convexity of the simplex. It is then easy to check that $q^{(i)}_{\text{bottom}}$ satisfies the two desired conditions.

For $q^{(z)}$, define $\alpha_i = n - z^{-1}(i)$ and $\tilde{\alpha}_i = \alpha_i / \sum_j \alpha_j$. Then, define

$$q^{(z)} = \sum_i \tilde{\alpha}_i q^{(i)}_{\text{top}} \,.$$

$q^{(z)} \in \mathcal{Q}$ by convexity of the simplex and it is easy to check that $\operatorname{argmin} \ell(q^{(z)}, .) = \{z\}$. □

An immediate consequence of point *ii)* of Lemma 14 above, is the following straightforward result, which we use throughout the proof of the main results to reduce all cases to cases where argmins of the task loss are unique:

Lemma 15 (Tie-breaking)**.** *Under (A1), (A2) and (A3),* $\forall q \in \mathcal{Q}, \forall z \in \operatorname{argmin} \ell(q, .), \ \exists q' \in \mathcal{Q}$ *s.t.*

1. $\operatorname{argmin} \ell(q', .) = \{z\}$,
2. $\forall z_1, z_2 \in \mathcal{Z}, \ell(q, z_1) > \ell(q, z_2) \Rightarrow \ell(q', z_1) > \ell(q', z_2)$

Proof. Let $q \in \mathcal{Q}$ and $z \in \operatorname{argmin} \ell(q, .)$. By point *ii)* of Lemma 14, we know $\exists q^{(z)} \in \mathcal{Q}, \operatorname{argmin} \ell(q^{(z)}, .) = \{z\}$. We denote $q_\alpha = (1-\alpha)q + \alpha q^{(z)}$.

1. As z is the only element of $\mathcal{Z}$ optimal for both q and $q^{(z)}$, we have that $\forall \alpha \in (0, 1], \operatorname{argmin} \ell(p_\alpha, .) = \{z\}$.
2. Denoting $\epsilon = \min_{z_1, z_2 : \ell(q, z_1) \neq \ell(q, z_2)} |\ell(q, z_1) - \ell(q, z_2)|$, Let $0 < \alpha < \frac{\epsilon}{\epsilon + \max_{z'}(\ell(q^{(z)}, z') - \underline{\ell}(q^{(z)}))}$. Then, for any z_1, z_2 such that $\ell(q, z_1) > \ell(q, z_2)$:
$$\begin{aligned}\ell(q_\alpha, z_1) - \ell(q_\alpha, z_2) &= (1-\alpha)(\ell(q, z_1) - \ell(q, z_2)) + \alpha(\ell(q^{(z)}, z_1) - \ell(q^{(z)}, z_2)) \\ &\geq (1-\alpha)\epsilon - \alpha \max_{z' \in \mathcal{Z}} (\ell(q^{(z)}, z') - \underline{\ell}(q^{(z)})) > 0 \quad \text{for our choice of } \alpha.\end{aligned}$$

Thus, for our choice of α, q_α satisfies both conditions. □

A.3 Notation

For conciseness, in the following, we may skip the sets that universal quantifiers range over when default values apply and are unambiguous:

- i, j always denote items in $[n]$, if the range is not specified, $\forall i$, $\forall j$ or $\forall i, j$ respectively mean $\forall i \in [n]$, $\forall j \in [n]$ and $\forall i, j \in [n]$
- k, m denote ranks, also in $[n]$. We use the same shorthands as above
- $\mathfrak{S}_n$ denotes the set of permutations of $[n]$. σ, μ, σ' denote permutations. $\forall \sigma$ means $\forall \sigma \in \mathfrak{S}_n$, same for μ and σ'.
- z, w, z' denote possible predictions, as above, $\forall z$ (without specific range for z) means $\forall z \in \mathcal{Z}$
- q, q' are members of $\mathcal{Q}$, i.e., $\forall q$ is a shorthand for $\forall q \in \mathcal{Q}$.
- s, s' are vectors of scores that belong to $\mathbb{R}^n$, $\forall s$ means $\forall s \in \mathbb{R}^n$.

A.4 Connectedness in $\mathfrak{S}_n$ and $\mathcal{Z}$

Definition 16 (Connectedness in $\mathfrak{S}_n$)**.** *A subset $\pi \subseteq \mathfrak{S}_n$ is* connected *in $\mathfrak{S}_n$ if there is a connected subset $S \subseteq \mathbb{R}^n$ such that $\pi = \operatorname{argsort}(S)$.*

Definition 17 (Connectedness in $\mathcal{Z}$)**.** *A subset $\zeta \subseteq \mathcal{Z}$ is* connected *in $\mathcal{Z}$ if there is a connected subset $S \subseteq \mathbb{R}^n$ such that $\zeta = \operatorname{pred}(S)$.*

Remark 18 (Link with topological connectedness)**.** *These notions of connectedness are not* stricto sensu *corresponding to topological connectedness on $\mathfrak{S}_n$ and $\mathcal{Z}$. To derive the topological connectedness that corresponds to these definitions, it is necessary to see* argsort *as a function valued in $2^{\mathfrak{S}_n}$ rather than a set-valued function. When doing so, the image of $\mathbb{R}^n$ by* argsort *is exactly the set of weak orders, on which it is possible to put the specialization topology, where the specialization corresponds to the inclusion relation. It is also sometimes referred to as the Alexandroff topology[2]. This topology makes* argsort *continuous (as a function valued in $2^{\mathfrak{S}_n}$) and open, which explains the use of* argsort *and* pred *to have a suitable (but not topological) definition of connectedness on $\mathfrak{S}_n$ and $\mathcal{Z}$.*

Definition 19 (u.h.c.)**.** *Let $\mathcal{U}$ and $\mathcal{V}$ be two topological spaces. A set-valued function $g : \mathcal{U} \twoheadrightarrow \mathcal{V}$ is* upper hemicontinuous *(u.h.c.) if for any non-empty open set $O \subseteq \mathcal{V}$, the upper inverse of O by g, $\operatorname{uinv}[g](O) = \{u \in \mathcal{U} : g(u) \subseteq O\}$ is open.*

Proposition 20. *Given a discrete space $\mathcal{V}$ (the topology is its whole power set), some $m \in \mathbb{N}$, a set-valued function $g : \mathbb{R}^m \twoheadrightarrow \mathcal{V}$ is u.h.c. if and only if*
$$\forall s \in \mathbb{R}^m, \exists \epsilon > 0, g(\mathcal{B}_\infty(s, \epsilon)) = g(s)$$

Proof. Direct implication. Let $s \in \mathbb{R}^n$. Because $\mathcal{V}$ is discrete, $g(s)$ is open. Because g is u.h.c. and $g(s)$ is open, $\operatorname{uinv}[g](g(s))$ is open in $\mathbb{R}^n$. As it contains s itself, $\exists \epsilon > 0, \mathcal{B}_\infty(s, \epsilon) \subseteq \operatorname{uinv}[g](g(s))$. Hence $\exists \epsilon > 0, g(\mathcal{B}_\infty(s, \epsilon)) \subseteq g(s)$. The equality comes from s being itself in the ball.

Reverse implication. Let $O \subseteq \mathcal{V}$ non-empty. For any $s \in \operatorname{uinv}[g](O), \exists \epsilon_s > 0, g(\mathcal{B}_\infty(s, \epsilon_s)) = g(s)$. Thus $\mathcal{B}_\infty(s, \epsilon_s) \subseteq \operatorname{uinv}[g](O)$. Finally, $\operatorname{uinv}[g](O) = \bigcup_{s \in \operatorname{uinv}[g](O)} \mathcal{B}_\infty(s, \epsilon_s)$ which is open. □

Some basic properties of connectedness are given below. These are useful in general to understand basic properties of $\operatorname{argsort}$ and pred. Point *vii)* allows to focus on connected subsets S that are open, which is convenient since open subsets or $\mathbb{R}^n$ are connected if and only if they are path-connected, and path-connectedness is easier to manipulate in our case.

Lemma 21.

i) $\forall \sigma, \{s : \operatorname{argsort}(s) = \{\sigma\}\}$ *is open and connected.*

ii) $\forall s, \{s' : \ \operatorname{argsort}(s') = \operatorname{argsort}(s)\}$ *is connected.*

iii) $\forall s, \{s' : \ \operatorname{argsort}(s') \subseteq \operatorname{argsort}(s)\}$ *is connected.*

iv) $\forall s, \{s' : \operatorname{pred}(s) = \operatorname{pred}(s')\}$ *is connected.*

v) $\operatorname{argsort}$ *is u.h.c. with respect to the discrete topology on* $\mathfrak{S}_n$*, i.e.*

$$\forall s, \exists \epsilon > 0, \quad \operatorname{argsort}(\mathcal{B}_\infty(s, \epsilon)) = \operatorname{argsort}(s). \tag{3}$$

vi) $\forall s, \forall \epsilon > 0, \forall \sigma \in \operatorname{argsort}(s), \quad \exists s' : \|s - s'\|_\infty < \epsilon$ *and* $\operatorname{argsort}(s') = \{\sigma\}$.

vii) $\pi \subseteq \mathfrak{S}_n$ *is connected in* $\mathfrak{S}_n$ *if and only if there is an* open *connected subset* $S \subseteq \mathbb{R}^n$ *such that* $\pi = \operatorname{argsort}(S)$*. Similarly,* $\zeta \subseteq \mathcal{Z}$ *is connected if and only if there is an open subset* S *such that* $\zeta = \operatorname{pred}(S)$.

Proof.

i) The set $\{s : \operatorname{argsort}(s) = \{\sigma\}\}$ is open and convex (and thus connected) as the intersection of $n-1$ open half spaces $\{s_{\sigma(k)} > s_{\sigma(k+1)}\}$ for $k = 1 \ldots n-1$.

ii) As for point *i)* above, $\{s' : \ \operatorname{argsort}(s') = \operatorname{argsort}(s)\}$ is an intersection of open half-spaces and hyperplanes: use the half-space $\{s' : (s_i - s_j)(s'_i - s'_j) > 0\}$ for i, j such that $s_i \neq s_j$ and the hyperplane $\{s' : s'_i = s'_j\}$ for i, j such that $s_i = s_j$. The intersection is convex and thus connected.

iii) It is a consequence of points *v)* and *vi)*, which are proved below, and point *ii)* above. Fix s, denote by $S = \{s' : \ \operatorname{argsort}(s') \subseteq \operatorname{argsort}(s)\}$ and take an arbitrary $s' \in S$. Denote by $\mathcal{B}_\infty(s, \epsilon) = \{s' : \|s - s'\|_\infty < \epsilon\}$. Fix $\epsilon > 0$ to satisfy (3) for both s and s', i.e., $\operatorname{argsort}(\mathcal{B}_\infty(s, \epsilon)) = \operatorname{argsort}(s)$ and $\operatorname{argsort}(\mathcal{B}_\infty(s', \epsilon)) = \operatorname{argsort}(s')$. let $\sigma \in \operatorname{argsort}(\mathcal{B}_\infty(s', \epsilon)) = \operatorname{argsort}(s')$. Then

$$\mathcal{B}_\infty(s, \epsilon) \cup \mathcal{B}_\infty(s', \epsilon) \cup \{s'' : \operatorname{argsort}(s'') = \{\sigma\}\}$$

is connected and included in S. S is thus a union of connected subsets (of the form above) with non-empty intersection $\{s\}$, and is thus connected.

iv) Fix s and let s_0 such that $\operatorname{argsort}(s_0) = \operatorname{pred}(s)$ (s_0 exists by definition of pred). The set $\{s' : \operatorname{pred}(s) = \operatorname{pred}(s')\}$ is then equal to $\{s' : \operatorname{argsort}(s) \subseteq \operatorname{argsort}(s_0)\}$, which is connected by point *iii)* above.

v) To show inclusion, take $\epsilon < \min\{s_i - s_j : i, j \text{ such that } s_i > s_j\}$. The ranking induced by any $s' \in \mathcal{B}_\infty(s, \epsilon)$ is necessarily compatible with a ranking induced by s. Equality holds since s is in the open ball.

vi) Take ϵ as above, take $s'_i = s_i - \frac{\epsilon\sigma^{-1}(i)}{n}$ (using $-\sigma^{-1}(i)$ because lower ranks are better). by the choice of ϵ, the relative ordering of s'_i and s'_j is the same as s_i and s_j when $s_i \neq s_j$, and the ordering is exactly the one induced by σ on equivalence classes of ties in s.

vii) Let $\pi = \operatorname{argsort}(S)$ for some connected $S \subseteq \mathbb{R}^n$. For $s \in S$, let $\epsilon_s > 0$ such that $\operatorname{argsort}(\mathcal{B}_\infty(s, \epsilon_s)) = \operatorname{argsort}(s)$. Then $\pi = \operatorname{argsort}(\bigcup_{s \in S} \mathcal{B}_\infty(s, \epsilon_s))$, and $\bigcup_{s \in S} \mathcal{B}_\infty(s, \epsilon_s)$ is open (as a union of open balls) and connected (as a union of connected sets with a connected set that intersects all of them). The argument is the same for connectedness in $\mathcal{Z}$.

□

The following lemma makes sure the definitions of connectedness in $\mathcal{Z}$ and $\mathfrak{S}_n$ are coherent.

Lemma 22. *$\zeta \subseteq \mathcal{Z}$ is connected in $\mathcal{Z}$ if and only if $\bigcup_{z\in\zeta} z$ is connected in $\mathfrak{S}_n$.*

Proof. First point: The if direction holds because $\zeta = \Lambda(\bigcup_{z\in\zeta} z)$, thus if $\bigcup_{z\in\zeta} z = \mathrm{argsort}(S)$ for some connected S, then $\zeta = \Lambda \circ \mathrm{argsort}(S) = \mathrm{pred}(S)$. For the only if direction, let us assume $\zeta \subseteq \mathcal{Z}$ is connected, i.e. there is a connected $S \subseteq \mathbb{R}^n$ such that $\zeta = \mathrm{pred}(S)$. Then

$$\begin{aligned}\bigcup_{z\in\zeta} z = \bigcup_{s\in S} \mathrm{pred}(s) &= \bigcup_{s\in S} \mathrm{argsort}(\{s' : \mathrm{argsort}(s') \subseteq \mathrm{pred}(s)\}) \\ &= \mathrm{argsort}(\underbrace{\bigcup_{s\in S}\{s' : \mathrm{argsort}(s') \subseteq \mathrm{pred}(s)\})}_{\tilde{S}}\end{aligned}$$

Thee set $\tilde{S}$ is the union of connected sets (by Lemma 21 point *iii)*), and there is a connected set (S) that intersects all members of the union. Thus, it is connected. □

Remark 23. *An analog statement from $\mathfrak{S}_n$ to $\mathcal{Z}$ is false in general: there exists n and pred satisfying all assumptions such that there is $\pi \subseteq \mathfrak{S}_n$ not connected but $\Lambda(\pi)$ is connected.*

To see this, we anticipate a bit on the idea of paths of adjacent items (Lemma 36 below). Consider a top-2 selection task with n=3. Let $\pi = \{2 \succ 1 \succ 3, 3 \succ 1 \succ 2\}$. π is not connected because there is no path between the two permutations. However, $\Lambda(\pi) = \{1 \sim 2 \succ 3, 1 \sim 3 \succ 2\}$, which is connected because there is a path (transposition of 2 and 3).

B Calibration and Consistency

Disclaimer: As mentioned in the body of the paper, we address, in these Appendices, a setting more general than ranking. This setting is described in Section A. We restate first the statement of the paper (on ranking), then state the more general statement (on weak orders) and prove the latter.

Subsection B.1 describes the definition and basic notions of calibration and uniform calibration,

Subsection B.2 provides the proof that calibration is not just an inclusion of argmins, but rather an equality, stated in Theorem 4 (also [7, Th. 2]),

Subsection B.3 contains the proof that uniform calibration is equivalent calibration under our assumptions as mentioned in Proposition 2.

B.1 Calibration and Consistency

Following the body of the paper, a surrogate loss $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$ is a measurable function that aims to be minimized. Similarly to the task loss, we can define the outer and inner risks, given a function $f : \mathcal{X} \to \mathbb{R}^n$ belonging to the space of measurable function from $\mathcal{X}$ to $\mathbb{R}^n$, denoted $\mathcal{F}$, a distribution P over $\mathcal{X} \times \mathcal{Y}$, $s \in \mathbb{R}^n$ and a distribution $q \in \mathcal{Q}$,

$$\mathcal{R}_{\Phi,P}(f) = \mathbb{E}_P[\Phi(Y, f(X))]\,, \qquad \phi(q,s) = \mathbb{E}_q[\Phi(Y,s)]\,, \qquad \underline{\phi}(q) = \inf_{s'\in\mathbb{R}^n} \phi(q,s')\,.$$

We can state the definition of calibration in $\mathcal{Z}$,

Definition 24 (calibration)**.** *Φ is L-calibrated if and only if*

$$\forall q \in \mathcal{Q}, \exists \delta > 0,\ \mathrm{pred}(\mathrm{lev}_\delta \phi(q,.)) \subseteq \mathrm{argmin}\, \ell(q,.)$$

And the one of uniform calibration,

Definition 25 (uniform calibration)**.** *Φ is L-uniformly calibrated if and only if*

$$\forall \varepsilon > 0, \exists \delta > 0, \forall q \in \mathcal{Q},\ \mathrm{pred}(\mathrm{lev}_\delta \phi(q,.)) \subseteq \mathrm{lev}_\varepsilon \ell(q,.)$$

Indeed, uniform calibration leads to the existence of an excess risk bound allowing to control the task risk by the surrogate risk.

Definition 26 (excess risk bound)**.** *A continuous function* $\delta : \mathbb{R}_+ \to \mathbb{R}_+$ *is an* excess risk bound *if*

$$\forall P, \forall f \in \mathcal{F}, \quad \mathcal{R}_{L,P}(\text{pred} \circ f) - \inf_{h \in \mathcal{H}} \mathcal{R}_{L,P}(h) \leq \delta \left(\mathcal{R}_{\Phi,P}(f) - \inf_{g \in \mathcal{F}} \mathcal{R}_{\Phi,P}(g) \right)$$

and

$$\delta(\epsilon) \xrightarrow[\epsilon \to 0]{} 0$$

Proposition 27. Φ *is uniformly L-calibrated if and only if there exists an excess risk bound.*

Proof. This result is directly derived from Steinwart [31] Th. 2.13 (direct implication) and Th. 2.17 (reverse implication). □

B.2 Proof of characterization of Calibration in Theorem 4 (equality of optimal predictions)

In this section, for the sake of completness, we provide a full proof for the generalization of Theorem 4 under our assumptions. First, lets remind the exact statement of the paper in the ranking setting.

Theorem 4 ([7, Th. 2])**.** *Given a ranking loss L,* $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}_+$ *is L-calibrated if and only if*

$$\forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall \delta \in (0, \delta_0], \quad \operatorname*{argmin}_{\sigma \in \mathfrak{S}_n} \ell(q, \sigma) = \bigcup_{s : \phi(q,s) - \underline{\phi}(q) < \delta} \operatorname{argsort}(s)$$

Then, the version of the same theorem in $\mathcal{Z}$,

Theorem 28. *Assuming (A1), (A2) and (A3),* Φ *is L-calibrated if and only if*

$$\forall q \in \mathcal{Q}, \exists \delta_0 {>} 0, \forall 0 {<} \delta {<} \delta_0, \text{pred}(\text{lev}_\delta \phi(q, .)) = \operatorname{argmin} \ell(q, .)$$

Before providing to the proof of Theorem 28, we state a general result of upper hemicontinuity that underlies many of the proofs in these appendices. It is a simplified version of the Berge Maximum Theorem [24]. One application is that $\operatorname{argmin} \ell(q, .)$ is u.h.c. as a function of q.

Lemma 29. *Given a function* $g : \mathcal{Q} \times \mathcal{Z} \to \mathbb{R}$*, continuous in its first argument, the set-valued function*

$$\begin{cases} \mathcal{Q} & \twoheadrightarrow \mathcal{Z} \\ q & \mapsto \operatorname*{argmin}_{z \in \mathcal{Z}} g(q, z) \end{cases} \qquad \textit{is u.h.c.}$$

Proof. We use the characterization of u.h.c. from Prop.20 as $\mathcal{Q} \subseteq \mathbb{R}^{|\mathcal{Y}|}$. Let $q \in \mathcal{Q}$ be a distribution. We denote $\mathcal{A}_q = \operatorname{argmin}_{z \in \mathcal{Z}} g(q, z)$ and

$$\epsilon_q = \min_{z \in \mathcal{A}_q, w \in \mathcal{Z} \setminus \mathcal{A}_q} |g(q, z) - g(q, w)|$$

For some $\delta > 0$, and for any $\tilde{q} \in \mathcal{B}_\infty(q, \delta)$, we use the following decomposition,

$$g(\tilde{q}, w) - g(\tilde{q}, z) = \underbrace{g(\tilde{q}, w) - g(q, w)}_{\substack{\exists \delta_1 > 0,\ \_ < \frac{\epsilon_q}{4} \\ \text{(continuity)}}} + \underbrace{g(q, w) - g(q, z)}_{> \epsilon_q} + \underbrace{g(q, z) - g(\tilde{q}, z)}_{\substack{\exists \delta_2 > 0,\ \_ < \frac{\epsilon_q}{4} \\ \text{(continuity)}}}$$

Thanks to the continuity of g in its first argument, we have that

$$\exists \delta_q > 0, \forall \tilde{q} \in \mathcal{B}_\infty(q, \delta), g(\tilde{q}, w) - g(\tilde{q}, z) > \frac{\epsilon_q}{2} \, .$$

Thus,

$$\exists \delta_q > 0, \bigcup_{\tilde{q} \in \mathcal{B}_\infty(q, \delta)} \operatorname*{argmin}_{z \in \mathcal{Z}} g(\tilde{q}, z) \subseteq \operatorname*{argmin}_{z \in \mathcal{Z}} g(q, z)$$

The equality comes from the fact q itself is in the ball. □

We can now proceed to the proof of Theorem 28.

Proof of Theorem 28.

(Φ, pred) is L-calibrated

$$\Leftrightarrow \forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall 0 < \delta < \delta_0, \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) \subseteq \mathrm{argmin}\, \ell(q, .) \quad \text{(Definition 25)}$$

We now focus on showing

$$\begin{gathered}\forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall 0 < \delta \ < \delta_0, \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) \subseteq \mathrm{argmin}\, \ell(q, .) \\ \Leftrightarrow \\ \forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall 0 < \delta \ < \delta_0, \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) = \mathrm{argmin}\, \ell(q, .)\end{gathered}$$

Or otherwise stated, that

$$\forall q \in \mathcal{Q}, \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) \subseteq \mathrm{argmin}\, \ell(q, .) \ \Leftrightarrow \ \forall q \in \mathcal{Q}, \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) = \mathrm{argmin}\, \ell(q, .) \tag{4}$$

First, for any $q \in \mathcal{Q}$, after defining $\tilde{\phi}(q, z) = \inf_{s \in \mathrm{pred}^{-1}(z)} \phi(q, s)$, we have that

$$\bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) = \mathrm{argmin}\, \tilde{\phi}(q, .)\,.$$

Now, thanks to Wijsman [37, Th. 2], we know that $\tilde{\phi}$ is continuous with respect to its first argument, then Lemma 29 gives us that $\mathrm{argmin}\, \tilde{\phi}(q, .)$ is u.h.c. and thanks to Prop. 20, we have

$$\exists \delta > 0, \bigcup_{\tilde{q} \in B_\delta(q)} \mathrm{argmin}\, \tilde{\phi}(\tilde{q}, .) = \mathrm{argmin}\, \tilde{\phi}(q, .) = \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .))\,. \tag{5}$$

Let's take $z \in \mathrm{argmin}\, \ell(q, .)$. By Lemma 14 *ii)*, $\exists q^z \in \mathcal{Q}, \mathrm{argmin}\, \ell(q^z, .) = \{z\}$ and we can define $\tilde{q}^\alpha = (1 - \alpha)q + \alpha q^z$ for some $\alpha \in (0, 1)$. Because $\mathrm{argmin}\, \ell(q, .) \cap \mathrm{argmin}\, \ell(q^z, .) = \{z\}$, then $\mathrm{argmin}\, \ell(\tilde{q}^\alpha, .) = \{z\}$ (z is the only element optimal for both components of the mixture). Finally, we prove (4):

$$\begin{aligned}&\forall q \ \in \mathcal{Q}, \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) \subseteq \mathrm{argmin}\, \ell(q, .) \\ &\quad \Rightarrow \forall q \in \mathcal{Q}, \forall z \in \mathrm{argmin}\, \ell(q, .), \forall \alpha \in (0, 1), \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(\tilde{q}^\alpha, .)) = \{z\} && \text{(previous line applied at } \tilde{q}^\alpha\text{)} \\ &\quad \Rightarrow \forall q \in \mathcal{Q}, \forall z \in \mathrm{argmin}\, \ell(q, .), \forall \alpha \in (0, 1), \mathrm{argmin}\, \tilde{\phi}(\tilde{q}^\alpha, .) = \{z\} && \text{(from (5))} \\ &\quad \Rightarrow \forall q \in \mathcal{Q}, \forall z \in \mathrm{argmin}\, \ell(q, .), \exists \alpha \in (0, 1), \{z\} = \mathrm{argmin}\, \tilde{\phi}(\tilde{q}^\alpha, .) \subseteq \mathrm{argmin}\, \tilde{\phi}(q, .) && \text{(from (5))} \\ &\quad \Rightarrow \forall q \in \mathcal{Q}, \forall z \in \mathrm{argmin}\, \ell(q, .), \{z\} \subseteq \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .)) && \text{(from (5) again)} \\ &\quad \Rightarrow \forall q \in \mathcal{Q}, \mathrm{argmin}\, \ell(q, .) \subseteq \bigcap_{\delta > 0} \mathrm{pred}(\mathrm{lev}_\delta \phi(q, .))\end{aligned}$$

□

B.3 Equivalence between calibration and uniform calibration

In this section, we prove the equivalence between calibration and the existence of an excess risk bound stated in Proposition 2. First, we remind the exact statement of the paper in the ranking setting.

Proposition 2. *Φ is L-calibrated if and only if there is an excess risk bound between $\mathcal{R}_\Phi$ and $\mathcal{R}_L$.*

Since we already established in Proposition 27 that uniform calibration is equivalent to the existence of an excess risk bound, we now show the equivalence between calibration and uniform calibration in our general setting:

Theorem 30. *Assuming (A1), Φ is L-calibrated if and only if Φ is L-uniformly calibrated.*

Proof. If Φ is L-uniformly calibrated, then it is L-calibrated, so we just have to prove the only if direction.

Using the notation $\tilde{\phi}(q,z) = \inf_{s \in \mathrm{pred}^{-1}(z)} \phi(q,s) \in \mathbb{R}$ for $z \in \mathcal{Z}$, let us notice that all the following functions are continuous on $\mathcal{Q}$ as $\mathcal{Y}$ is finite [37, Theorem 2]:

$$\begin{aligned} &q \mapsto \tilde{\phi}(q,z) && q \mapsto \underline{\phi}(q) = \inf_{s \in \mathbb{R}^n} \phi(q,s) = \min_{z \in \mathcal{Z}} \tilde{\phi}(q,z) \\ &q \mapsto \ell(q,z) && q \mapsto \underline{\ell}(q) = \min_{z \in \mathcal{Z}} \ell(q,z)\,. \end{aligned}$$

Consequently, for every $z \in \mathcal{Z}$, both functions

$$g_z : q \mapsto \ell(q,z) - \underline{\ell}(q) \text{ and } \qquad h_z : q \mapsto \tilde{\phi}(q,z) - \underline{\phi}(q)$$

are continuous on $\mathcal{Q}$. Also notice that since both $\mathcal{Y}$ and $\mathcal{Z}$ are finite, both ℓ and $\tilde{\phi}$ are bounded on their domain. More formally, denoting $B_\ell = \max_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}} L(y,z) - \min_{z \in \mathcal{Z}} \min_{y \in \mathcal{Y}} L(y,z)$, we have $g_z \in [0, B_\ell]$.

Let us now fix $\epsilon > 0$ and denote

$$\tilde{\Delta}(z,\epsilon) = \big\{ q \in \mathcal{Q} : g_z(q) \geq \epsilon \big\} = g_z^{-1}([\epsilon, B_\ell])\,.$$

Since g_z is continuous from a compact set ($\mathcal{Q}$) to $\mathbb{R}$, it is a proper map, which means that preimages of compact subsets of $\mathbb{R}$ are compact. In particular, it implies that $\tilde{\Delta}(z,\epsilon)$ is compact. Since h_z is also continuous on $\mathcal{Q}$, it implies that is reaches its minimum on $\tilde{\Delta}(z,\epsilon)$. Let us then denote:

$$\delta(\epsilon) = \min_{z \in \mathcal{Z}} \min_{q \in \tilde{\Delta}(z,\epsilon)} h_z(q)\,.$$

Then, since Φ is L-calibrated, $g_z(q) > \epsilon \Rightarrow h_z(q) > 0$ for all z, q and thus $\delta(\epsilon) > 0$.

Thus, for all $\epsilon > 0$, $q \in \mathcal{Q}$ and $s \in \mathbb{R}^n$, we have:

$$\begin{aligned} & & &\phi(q,s) - \underline{\phi}(q) < \delta(\epsilon) & \\ \Rightarrow & & \forall z \in \mathrm{pred}(s),\ \ &\tilde{\phi}(q,z) - \underline{\phi}(q) < \delta(\epsilon) & \text{(by def. of } \tilde{\phi}) \\ \Rightarrow & & \forall z \in \mathrm{pred}(s),\ \ &\ell(q,z) - \underline{\ell}(q) < \epsilon & \text{(by Def of } \delta(\epsilon)) \\ \Rightarrow & & &\ell(q,s) - \underline{\ell}(q) < \epsilon & \end{aligned}$$

which means that Φ is L-uniformly calibrated. □

C Proof of the main result

Disclaimer: As mentioned in the body of the paper, we address, in these Appendices, a setting more general than ranking. This setting is described in Section A. We restate first the statement of the paper (on ranking), then state the more general statement (on weak orders) and prove the latter.

The objective of this section is to prove the following Theorem 6. First we restate the theorem in the ranking setting.

Theorem 6. *For a ranking loss L, the following statements are equivalent:*

(i) L *is CEU,*

(ii) $\forall q, \operatorname{argmin}_\sigma \ell(q,\sigma)$ *is connected.*

(iii) $\forall \epsilon > 0, \forall q, \mathrm{lev}_\epsilon \ell(q,.)$ *is connected,*

Moreover, the function $\tilde{u} : \mathcal{Y} \to \mathbb{R}^n$ *defined as:* $\forall i \in [n], \tilde{u}_i(y) = -\sum_{\sigma \in \mathfrak{S}_n} \mathbb{1}_{\{\sigma(1)=i\}} L(y,\sigma)$ *is a utility function for* L *whenever there exists a utility function for* L *(i.e., whenever* L *is CEU).*

Before stating its generalization in $\mathcal{Z}$, we need to note the definition of CEU task loss extends in a strait-forward way to $\mathcal{Z}$. We state it for the sake of completeness.

Definition 31 (CEU)**.** *A task loss* L *is* compatible with expected utility *(CEU) if there exists a function* $u : \mathcal{Y} \to \mathbb{R}^n$ *such that* $\Phi_u^{\mathrm{sq}} : y, s \mapsto (u(y) - s)^2$ *is* L*-calibrated.*

Theorem 32. *Under assumptions (A1), (A2) and (A3), for a task loss L, the following statements are equivalent:*

(i) L *is CEU,*

(ii) $\forall q, \operatorname{argmin} \ell(q, .)$ *is connected.*

(iii) $\forall \epsilon > 0, \forall q, \operatorname{lev}_\epsilon \ell(q, .)$ *is connected,*

Moreover, the function $\tilde{u} : \mathcal{Y} \to \mathbb{R}^n$ *defined as:* $\forall i \in [n], \tilde{u}_i(y) = -\sum_{z \in \mathcal{Z} : z^{-1}(i) = 1} L(y, z)$ *is a utility function for* L *whenever there exists a utility function for* L *(i.e., whenever* L *is CEU).*

We now proceed with the proof that this whole section is dedicated to. The two main intermediary results of this proof are Corollary 41 and Theorem 42.

In Appendix C.1, we elaborate on more properties of connectedness in $\mathfrak{S}_n$ and $\mathcal{Z}$. In particular, we prove Lemma 36 which is the generalized version of Proposition 9 in the main paper regarding the characterization of connectedness with paths of adjacent items. This proposition is a key technical result that is fundamental to the results of Appendices D, E and H.

In Appendix C.2, we give technical lemmas that lead to the main steps in the full proof. The main result of this section, Lemma 38 allows to define sequences of distributions in cQ for which the argmin matches a path of adjacent transposition between permutation. This is the key component to extend properties of argmins to properties of the entire loss.

In Appendix C.3, we put the tools together, and first prove the strict monotonicity of loss with connected argmins (Corollary 41) , which means that preferences between items as defined by optimal z are reflected by preferences even for non-optimal z'. This, in turn, allow us to prove the existence of the utility function and its analytical formula 42. The final proof at the end of that section makes the link with connected sublevel sets.

C.1 Properties of connected sets in $\mathfrak{S}_n$ and $\mathcal{Z}$

The main technical property of connectedness is the existence of paths with transpositions of adjacent items:

Definition 33 (Adjacent items)**.** $\forall \sigma \in \mathfrak{S}_n$*, items* i *and* j *are* adjacent[2] *in* σ *if* $|\sigma^{-1}(i) - \sigma^{-1}(j)| \leq 1$.

$\forall z \in \mathcal{Z}$*, items* i *and* j *are* adjacent *in* z *if* $\exists \sigma \in z$ *such that* i *and* j *are adjacent in* σ.

Definition 34 (paths of adjacent items)**.**
Path in $\mathfrak{S}_n$: *Let* $\pi \subseteq \mathfrak{S}_n$*, and* $\sigma, \sigma' \in \pi$*. A* path of adjacent transpositions of length $M \in \mathbb{N}$ *from* σ *to* σ' *in* π *is a sequence* $(\sigma_0, ..., \sigma_M) \in \pi^M$ *such that:*

- $\sigma_0 = \sigma$ *and* $\sigma_M = \sigma'$
- $\forall m \in [M]$*, there is* $(i_m, i'_m) \in [n]^2$ *such that* i_m *and* i'_m *are adjacent in* σ_{m-1} *and* $\sigma_m = \tau_{i_m i'_m} \sigma_{m-1}$.

Path in $\mathcal{Z}$: *Let* $\zeta \subseteq \mathcal{Z}$*, and* $z, z' \in \zeta$*. A* path of adjacent transpositions of length $M \in \mathbb{N}$ *from* z *to* z' *in* ζ *is a sequence* $(z_0, ..., z_M) \in \zeta^M$ *such that:*

- $z_0 = z$ *and* $z_M = z'$
- $\forall m \in [M]$*, there is* $(i_m, i'_m) \in [n]^2$ *such that* i_m *and* i'_m *are adjacent in* z_{m-1} *and* $z_m = \tau_{i_m i'_m} z_{m-1}$.

We say that there is a path from σ and σ' in $\pi \subseteq \mathfrak{S}_n$ if there is $M \in \mathbb{N}$ such that there is a path of adjacent transpositions of length M from σ to σ' in π. A similar terminology is used in $\mathcal{Z}$.

[2] In that definition, we consider an item to be adjacent to itself. We could prevent that without altering anything in the proofs.

Paths of adjacent items between two permutations can be found by bubble sort. We describe a naive version of the algorithm and then explain the important property that we use later in the proofs.

Algorithm 1: (Naive) Bubble Sort

```
input  : $n$ and $\sigma, \sigma' \in \mathfrak{S}_n$
output : a sequence of transpositions $\tau^{(1)}, \dots, \tau^{(t)}$ such that $\sigma = \tau^{(t)} \dots \tau^{(1)}\sigma'$
$k = n$;
$\nu = \sigma'$;
$t = 1$;
for $m \leftarrow 1$ to $n - 1$ do
    for $k \leftarrow 1$ to $m - i$ do
        if $\sigma^{-1}(\nu(k)) > \sigma^{-1}(\nu(k+1))$ then
            $\tau^{(t)} = \tau_{\nu(k)\nu(k+1)}$;
            $\nu = \tau^{(t)}\nu$;
            $t = t + 1$;
        end
    end
end
```

Remark 35. *Note the* monotonicity *property of the bubble sort. We use it later in the proofs. Given two items $i \neq j \in [n]$, if $i \succ_\sigma j$ and $i \prec_{\sigma'} j$, then either τ_{ij} or τ_{ji} appears exactly once in the output sequence $\tau^{(1)}, \dots, \tau^{(t)}$ of bubble sort. Otherwise, neither τ_{ij} nor τ_{ji} appears in the output sequence of bubble sort.*

The next lemma is the general version of Proposition 9 in the main paper.

Lemma 36 (connectedness and paths in $\mathfrak{S}_n$)**.** *$\pi \subseteq \mathfrak{S}_n$ is connected in $\mathfrak{S}_n$ if and only if for every $\sigma, \sigma' \in \pi$, there is a path from σ to σ' in π.*

As a corollary, $\zeta \subseteq \mathcal{Z}$ is connected in $\mathcal{Z}$ if and only if for every $z, z' \in \zeta$ there is a path between z and z' in ζ.

Proof. **"if" direction** Let σ and σ' two permutations such that there is a path between them in π. Denote $(\sigma_m)_{m=0}^{M}$ the sequence of permutations along that path ($\sigma_0 = \sigma$ and $\sigma_M = \sigma'$). We show that there is a connected set S such that $\{\sigma_0, ..., \sigma_M\} = \operatorname{argsort}(S)$.

The first step is to notice that if σ_m and σ_{m+1} are two permutations that are equal up to an adjacent transposition, then there is $s^{(m)}$ such that $\operatorname{argsort}(s^{(m)}) = \{\sigma_m, \sigma_{m+1}\}$ (use $s_i^{(m)} = -(\sigma_m^{-1}(i) + \sigma_{m+1}^{-1}(i))$). We thus have:

$$\{\sigma_0, ..., \sigma_M\} = \operatorname{argsort}(\bigcup_{m=0..M-1} \underbrace{\{s : \operatorname{argsort}(s) \subseteq \operatorname{argsort}(s^{(m)})\}}_{S_m})$$

Each S_m is connected by Lemma 21 point *iii)*, and $S_m \cap S_{m+1} = \{s : \operatorname{argsort}(s) = \sigma_{m+1}\}$ is open and connected. Thus $\bigcup_m S_m$ is connected. Finally, π is the union of all these sets for all possible pairs $\sigma, \sigma' \in \pi$. The union itself is connected

"only if" direction We prove the result in two steps. First, we prove that if $\pi = \operatorname{argsort}(s)$ for some $s \in \mathbb{R}^n$, then any two σ, σ' are connected by a path. We then go to the more general case $\pi = \operatorname{argsort}(S)$ with $S \subseteq \mathbb{R}^n$.

Case 1: $\exists s \in \mathbb{R}^n, \pi = \operatorname{argsort}(s)$. Let $\sigma, \sigma' \in \pi$. By definition of argsort, $i \succ_\sigma j$ and $j \succ_{\sigma'} i$ is only possible for i, j such that $s_i = s_j$. Apply `bubblesort`(target=σ, input=σ'). Bubble sort gives a path $(\sigma_m)_{m=0}^{M}$ between σ and σ'; applying recursively the above remark to $i \succ_\sigma j$ and $j \succ_{\sigma_m} i$, the adjacent transpositions only exchange items i, j that are tied in s. Thus, by the definition of argsort, $\forall m \in \{0, ..., M\}$, $\sigma_m \in \operatorname{argsort}(s) = \pi$.

Step 2: general case: $\pi = \operatorname{argsort}(S)$ for connected $S \subseteq \mathbb{R}^n$ Using Lemma 21 (point *vii)*) we can assume without loss of generality that S is open and thus path-connected. Let $\sigma, \sigma' \in \pi$. Let $\gamma : [0, 1] \to S$ be a continuous function such that $\sigma \in \operatorname{argsort}(\gamma(0))$ and $\sigma' \in \operatorname{argsort}(\gamma(1))$; such a γ exists by the assumption $\pi = \operatorname{argsort}(S)$ for path-connected S. By definition of γ, we have

$\text{argsort}(\gamma([0,1])) \subseteq \pi$. Now consider the undirected graph $G = (\pi, E)$ with set of nodes π and edges E, where $(\nu, \nu') \in E$ if $\exists t \in [0,1]$ such that $\{\nu, \nu'\} \subseteq \text{argsort}(\gamma(t))$. We now prove that there is a path in G (in the usual sense of paths in a graph) between the nodes corresponding to σ and σ'. To that end, first notice that by Lemma 21 point *v)*, $\text{argsort}(\gamma(t-\epsilon)) \subseteq \text{argsort}(\gamma(t))$ for small enough ϵ and $t \in (0,1]$, since γ is continuous (a similar statements holds for $t+\epsilon$ instead of $t-\epsilon$). This means that all permutations in consecutive values of $t \mapsto \text{argsort}(\gamma(t))$ are neighbors in G. Now let us consider

$$t_0 = \sup\{t \in [0,1] : \forall \nu \in \text{argsort}(\gamma(t)), \nu \text{ is connected to } \sigma \text{ in } G\}.$$

t_0 is well defined since all permutations of $\gamma(0)$ are connected to σ in G. First notice that since argsort is u.h.c. (Lemma 21 point *v)*), the sup above is actually a max. We now prove that it implies that all permutations in $\text{argsort}(\gamma(t))$ are connected to σ in G. Aiming for a contradiction, assume $t_0 < 1$. Then, by the remark above, there is $\epsilon > 0$ such that $\text{argsort}(\gamma(t_0+\epsilon)) \subseteq \text{argsort}(\gamma(t_0))$. Then, all permutations of $\text{argsort}(\gamma(t_0+\epsilon))$ are connected in G to all permutations of $\text{argsort}(\gamma(t_0))$, and are thus connected to σ in G. This contradicts the definition of t_0.

Thus, all permutations in $\gamma(1)$ are connected in G to all permutations in $\gamma(0)$. To finish the proof, let us remind that by the Step 1 above, for all t, all permutations within $\text{argsort}(\gamma(t))$ are connected by a path (of adjacent transpositions) in $\mathfrak{S}_n$. Thus, any two permutations connected in G are also connected by a path of adjacent transpositions. Thus, σ and σ' are connected by a path of adjacent transpositions. □

C.2 Main property of Losses with connected argmins

To clarify the results of the section, we state the connectedness of the argmins of the losses as an assumption:

(A4). $\forall q \in \mathcal{Q}, \text{argmin}\,\ell(q,.)$ *is connected in* $\mathcal{Z}$.

The proof of the final result is the combination of the following observation, which is a general property of preorders, together with the lemma that follows.

The general observation is the following: Given two preorders z and w that have the same (strict) preferences between two items i and j, then we can find a path of adjacent transpositions between them that never transposes i and j. Thus the strict preferences between i and j are kept constant along the path.

Lemma 37. *$\forall z, w \in \mathcal{Z}, \forall i,j \in [n]$ s.t. $i \succ_z j$ and $i \succ_w j$, $\exists M \in \mathbb{N}, \exists (i_m, i'_m)_{m=0}^M \in [n]^{2M}$ s.t. denoting $z_0 = z$ and $z_m = \tau_{i_m, i'_m} z_{m-1}$, we have*

(i) $\forall m \in [M], i_q, i'_m$ *are adjacent in* z_{m-1},

(ii) $z_M = w$,

(iii) $\forall m \in [M], \{k_m, k'_m\} \neq \{i,j\}$.

Proof. For two permutations, the existence of such a path is given by bubble sort: since bubble sort is monotonic (see Remark 35), it never exchanges two items that are in an ordering compatible with the target weak order. The result in $\mathcal{Z}$ follows by compositing with Λ the path between $\sigma \in z$ and $\sigma' \in z'$. □

The next lemma is the main technical step in the proof of our final result, and transposes the previous lemma in sequences of argmins for different distributions, when the argmins are always connected. In a similar way to the path between preorders above, the lemma considers a distribution for which the argmin of the loss is a single preorder z, and two fixed items i, j. Then, it shows that for any two adjacent items $\{k,l\} \in z$, different from $\{i,j\}$, we can find a distribution for which the argmin is $\tau_{kl} z$, which at the same time preserves strict preferences (by the inner risk) between z' and $\tau_{ij} z'$:

Lemma 38 (Adjacent transpositions of argmin)**.** *Under assumptions (A1), (A2) and (A3) and (A4): $\forall q, z$ s.t. $\text{argmin}\,\ell(q,.) = \{z\}$, $\forall i,j,k,l \in [n]$ s.t. $\{i,j\} \neq \{k,l\}$ and (k,l) are adjacent in z, we have: $\exists q' \in \mathcal{Q}$ s.t.*

1. $\text{argmin}\,\ell(q',.) = \{\tau_{kl} z\}$,

2. $\forall z', \ell(q, z') > \ell(q, \tau_{ij} z') \Rightarrow \ell(q', z') > \ell(q', \tau_{ij} z')$

Proof. Note that the result trivially holds with $q' = q$ when $k \sim_z l$, since in that case $\tau_{kl} z = z$ (by (2) d)). We thus focus on the case $k \not\sim_z l$. Without loss of generality, we assume $l \succ_z k$.

First case: $k \neq i$ **and** $k \neq j$**.** Consider $q_\alpha = (1-\alpha) q + \alpha q_{\rm top}^{(k)}$ and let $\alpha_0 = \inf\{\alpha : \{z\} \neq \operatorname{argmin} \ell(q_\alpha, .)\}$. Notice that α_0 is well defined since all $z' \in \operatorname{argmin} \ell(q_1, .)$ must have ${z'}^{-1}(k) = 1$ by definition of $q_{\rm top}^{(k)}$, while $z^{-1}(k) > z^{-1}(l)$ by our assumption $l \succ_z k$. Moreover, since $\operatorname{argmin} \ell(., .)$ is uhc, the infimum is a minimum, and thus $\alpha_0 > 0$ since $\operatorname{argmin} \ell(q_0, .) = \{z\}$.

Let $q' = q_{\alpha_0}$. Since $\operatorname{argmin} \ell(., .)$ is uhc, we have that $\exists \epsilon > 0$ such that $\operatorname{argmin} \ell(\mathcal{B}_\infty(q', \epsilon), .) = \operatorname{argmin} \ell(q', .)$. Taking one such ϵ with the additional requirement $\epsilon < \alpha_0$, we have:

1. $z \in \operatorname{argmin} \ell(q', .)$, because $\{z\} = \operatorname{argmin} \ell(q_{\alpha_0 - \epsilon}, .) \subseteq \operatorname{argmin} \ell(q', .)$. The first equality comes from the definition of α_0 and the inclusion comes from the choice of ϵ.
2. $\operatorname{argmin} \ell(q', .) \neq \{z\}$, since the infimum in the definiton of α_0 is a minimum.

Let $z' \in \operatorname{argmin} \ell(q', .)$ with $z' \neq z$. Since $\operatorname{argmin} \ell(q', .)$ is connected by (A4), Lemma 36 shows that there is a path of adjacent transpositions between z and z' in $\operatorname{argmin} \ell(q', .)$. In particular, there exists two items k', l' adjacent in z such that $\tau_{k',l'} z \in \operatorname{argmin} \ell(q', .)$ and $\tau_{k',l'} z \neq z$. The result follows from the three additional remarks:

1. We necessarily have $k' = k$ or $l' = k$.

   Indeed, aiming for a contradiction, assume $k \notin \{k', l'\}$. By the definition of $q_{\rm top}^{(k)}$, we have $\ell(q_{\rm top}^{(k)}, \tau_{k'l'} z) = \ell(q_{\rm top}^{(k)}, z)$, and thus

   $$\begin{aligned} \ell(q', \tau_{k'l'} z) &= (1-\alpha_0) \ell(q, \tau_{k'l'} z) + \alpha_0 \ell(q_{\rm top}^{(k)}, z) && (6) \\ &> (1-\alpha_0) \ell(q, z) + \alpha_0 \ell(q_{\rm top}^{(k)}, z) && \text{because } \tau_{k'l'} z \notin \operatorname{argmin} \ell(q, .) \\ &= \ell(q', z) \end{aligned}$$

   which contradicts $\tau_{k'l'} z \in \operatorname{argmin} \ell(q', .)$.

2. w.l.o.g., let $k' = k$. Then $\ell(q', \tau_{kl'} z) = \ell(q', \tau_{kl} z)$ and thus $\tau_{kl} z \in \operatorname{argmin} \ell(q', .)$.

   Indeed, since $\tau_{k'l'} z \in \operatorname{argmin} \ell(q', .)$, this means $\ell(q_{\rm top}^{(k)}, \tau_{kl'} z) < \ell(q_{\rm top}^{(k)}, z)$ and thus $z^{-1}(l') < z^{-1}(k)$ by the definition of $q_{\rm top}^{(k)}$. Since k and l' are adjacent in z, and the rank of l' is strictly smaller than that of l, we have $z^{-1}(l') = z^{-1}(l)$ and thus $l \sim_z l'$. This implies $\ell(q', \tau_{kl'} z) = \ell(q', \tau_{kl} z)$.

3. $\forall z'', \ell(q', z'') - \ell(q', \tau_{ij} z'') = (1-\alpha_0) \big(\ell(q, z'') - \ell(q, \tau_{ij} z'') \big)$.

   This follows from the same calculation as (6), using the definition of $q_{\rm top}^{(k)}$ and $k \notin \{i, j\}$. This remark implies that strict inequalities between $\ell(q, z'')$ and $\ell(q, \tau_{ij} z'')$ are preserved in q'.

The final result immediately follows, using the tie breaking lemma (Lemma 15) to find some other distribution for which $\tau_{k'l} z$ is the unique element of the argmin.

Second case: $k \in \{i, j\}$**.** Then, by assumption we have $l \neq i$ and $l \neq j$. In that case, we follow similar steps, using $q_{\rm bottom}^{(l)}$ instead of $q_{\rm top}^{(k)}$. □

C.3 Strict monotonicity for losses with connected argmin

This section provides the final proof. The proof of the main result (coonected argmins are equivalent to the existence of a utility function) is based on the argument of *strict monotocity* of task losses with connected argmins (Lemma 40 and Corollary 41). In essence, the strict monotonicity extends to the entire task loss the property of its argmin, when these are connected: whenever item i is preferred to item j in the argmin, it is always preferred. The strict monotocity allows to directly prove the existence of a utility function (Theorem 42). The utility function itself proves the calibration of the square loss, which in turn implies the connectedness of argmins using the characterization of calibration based on equality of argmins (Theorem 28). The strict monotonicity also readily implies connectedness of all sublevel sets, since it shows we can find a path between any permutation and an optimal permutation that never increases the task loss.

Before going to the proofs of strict monotonicity, we give the following alternative to the tie-breaking lemma (Lemma 15)[3]:

Lemma 39. *Under assumptions (A1), (A2) and (A3):*

Let $q \in \mathcal{Q}$ and i, j, z' such that $i \succ_{z'} j$. If $\ell(q, z') = \ell(q, \tau_{ij} z')$, then $\exists q' \in \mathcal{Q}$, such that $\operatorname{argmin} \ell(q', .) \subseteq \operatorname{argmin} \ell(q, .)$ and $\ell(q', z') > \ell(q', \tau_{ij} z')$.

Proof. Let $\epsilon = \min_{z_1, z_2 : \ell(q, z_1) \neq \ell(q, z_2)} |\ell(q, z_1) - \ell(q, z_2)|$, take α such that $0 < \alpha < \frac{\epsilon}{\epsilon + \max_{z_0} \ell(q, z_0)}$, and define $q' = (1 - \alpha) q + \alpha q^{(\tau_{ij} z')}$ (as defined in Lemma 14). The choice of α makes sure that any strict preferences between any z_1 and z_2 by $\ell(q, .)$ are preserved in $\ell(q', .)$. Only ties in $\ell(q, .)$ can be changed in strict inequalities in $\ell(q', .)$. Thus $\operatorname{argmin} \ell(q', .) \subseteq \operatorname{argmin} \ell(q, .)$. Moreover, since $\ell(q^{(\tau_{ij} z')}, \tau_{ij} z') < \ell(q^{(\tau_{ij} z')}, , z')$ by definition of $q^{(\tau_{ij} z')}$, we have $\ell(q, z') > \ell(q, \tau_{ij} z')$, which is the desired result. □

We decribe the *strict monotonicity* property of losses with connected argmins (A4) in the next results.

Lemma 40 (Strict monotonicity, base case)**.** *Under assumptions (A1), (A2) and (A3) and (A4), we have:*

$\forall q, z$ s.t. $\operatorname{argmin} \ell(q, .) = \{z\}$: $\forall i, j$ such that $i \succ_z j$, $\forall z'$ such that $i \succ_{z'} j$, $\ell(q, z') < \ell(q, \tau_{ij} z')$.

Proof. Let i, j such that $i \succ_z j$ and z' such that $i \succ_{z'} j$. By Lemma 37, there is a path of adjacent transpositions between z and z' that never swaps i and j. Let us denote by $z_0, ..., z_M$ such a path (with $z_0 = z$ and $z_M = z'$). Since we never swap i and j, denoting $k_m, k_{m'}$ the adjacent items of z_{m-1} that are swapped between z_{m-1} and z_m, we have $\{i, j\} \neq \{k_m, k'_m\}$. We can thus apply Lemma 38, and find $q_0 = q, q_1, ..., q_M$ such that:

1. $\forall m \in [M], \ \operatorname{argmin} \ell(q_m, .) = \{z_m\}$
2. $\forall m \in [M], \quad \forall z''$ such that $\ell(q_{m-1}, z'') > \ell(q_{m-1}, \tau_{ij} z'')$, we have $\ell(q_m, z'') > \ell(q_m, \tau_{ij} z'')$

An immediate consequence is that

$$\ell(q, z') \leq \ell(q, \tau_{ij} z').$$

Indeed, aiming for a contradiction, assume that $\ell(q, z') > \ell(q, \tau_{ij} z')$. Then by point 2. above, the sign of the difference would be kept along the path, all the way to q_M, i.e. we should have $\ell(q_M, z') > \ell(q_M, \tau_{ij} z')$. But $z' \in \operatorname{argmin} \ell(q_M, .)$, so this impossible.

Now, by Lemma 39, if we had $\ell(q, z') = \ell(q, \tau_{ij} z')$, we could find q'' with $\operatorname{argmin} \ell(q'', .) = \{z\}$ and $\ell(q'', z') > \ell(q'', \tau_{ij} z')$, which is impossible as we just stated. Thus, we have $\ell(q, z') < \ell(q, \tau_{ij} z')$. □

The lemma above extends to all possible q, not only those for which the argmin is a single element, in the following way:

Corollary 41 (Strict monotonicity)**.** *Under the assumptions of Lemma 40, let $q \in \mathcal{Q}$ and $i, j \in [n]$. There are three cases:*

[3] Note that even though tie breaking lemmas are purely technical, they are very important because connectedness is materialized by ties. There are two ways to break ties: the tie breaking lemma chooses the argmin, but cannot guarantee that other ties outside the argmin are suitably broken. On the other hand, Lemma 39 breaks a specific tie, but cannot choose precisely the resulting argmin.

i) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succ_z j$*, then* $\forall z' : i \succ_z j, \ell(q,z') < \ell(q, \tau_{i,j}z')$*;*

ii) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succeq_z j$ *and* $\exists z_0 \in \operatorname{argmin} \ell(q,.), i \succ_{z_0} j$*, then:*
$\forall z' : i \succ_{z'} j, \ell(q,z') \leq \ell(q, \tau_{i,j}z')$*.*
(Notice in that case $\ell(q,z_0) < \ell(q,\tau_{i,j}z_0)$*.)*

iii) if $\exists z_0, z_1 \in \operatorname{argmin} \ell(q,.)$ *such that* $i \succ_{z_0} j$ *and* $j \succ_{z_1} i$*, then* $\forall z' : \ell(q,z') = \ell(q,\tau_{i,j}z')$*.*

Proof.

i) Let q such that $\forall z \in \operatorname{argmin} \ell(q,.), i \succ_z j$. Aiming for a contradiction, assume $\ell(q,z') > \ell(q,\tau_{i,j}z')$. Let $z \in \operatorname{argmin} \ell(q,.)$. By the tie-breaking lemma (Lemma 15), we could find q' with $\operatorname{argmin} \ell(q',.) = \{z\}$ and $\ell(q',z') > \ell(q',\tau_{i,j}z')$, but since $i \succ_z j$ this contradicts the base case (Lemma 40). We thus have $\ell(q,z') \leq \ell(q,\tau_{i,j}z')$. However, by Lemma 39, $\ell(q,z') = \ell(q,\tau_{i,j}z')$ would also yield a contradiction. We thus have $\ell(q,z') < \ell(q,\tau_{i,j}z')$, which is the desired result.

ii) $\forall z \in \operatorname{argmin} \ell(q,.), i \succeq_z j$ and $\exists z_0 \in \operatorname{argmin} \ell(q,.), i \succ_{z_0} j$. With the same arguments as above, taking $z \in \operatorname{argmin} \ell(q,.)$ such that $i \succ_z j$ (which exists by assumption) when using the tie-breaking lemma, we obtain $\forall z', \ell(q,z') \leq \ell(q,\tau_{i,j}z')$.

iii) Let q such that $\exists z_0, z_1 \in \operatorname{argmin} \ell(q,.)$ such that $i \succ_{z_0} j$ and $j \succ_{z_1} i$. If there is z' such that $i \succ_z j, \ell(q,z') > \ell(q,\tau_{i,j}z')$ then we can use the tie-breaking lemma with z_0 and that would contradicts the base case Lemma 40. Likewise, a z' such that $i \succ_z j, \ell(q,z') < \ell(q,\tau_{i,j}z')$ would contradict the base case after applying the tie breaking lemma with z_1. Thus, equality must hold for all z'.

□

The following result is a direct consequence of the strict monotonicity, and is the main result of the paper: there is a utility function wuch that the expected utility gives an optimal scoring function:

Theorem 42. *Let* $\tilde{u} : \mathcal{Y} \to \mathbb{R}^n$ *defined as:* $\tilde{u}_i(y) = -\sum_{z \in \mathcal{Z}: z^{-1}(i)=1} L(y,z)$*.*

For $q \in \mathcal{Q}$*, denote the expected utility by* $\tilde{U}(q) = \sum_{y \in \mathcal{Y}} q_y \tilde{u}(y)$*.*

Under assumptions (A1), (A2) and (A3) and (A4), we have:

$$\operatorname{argmin} \ell(q,.) = \operatorname{pred}(\tilde{U}(q))$$

Obviously, utilities are cardinal, in the sense that any affine transformation of u is also a utility.

Proof. Let $q \in \mathcal{Q}$ and $i, j \in [n]$. From the definition of u and the linearity (w.r.t. q) of $\tilde{U}$, we have:

$$\begin{aligned}\tilde{U}_i(q) - \tilde{U}_j(q) &= \sum_{z \in \mathcal{Z}: z^{-1}(j)=1} \ell(q,z) - \sum_{z \in \mathcal{Z}: z^{-1}(j)=1} \ell(q,z) \\ &= \sum_{z \in \mathcal{Z}: z^{-1}(i)=1} \big(\ell(q,\tau_{ij}z) - \ell(q,z)\big)\,.\end{aligned}$$

because $\{\tau_{ij}z \in \mathcal{Z} : z^{-1}(j) = 1\} = \{z \in \mathcal{Z} : z^{-1}(i) = 1\}$. Thus, coming back to the three cases of Lemma 41, we immediately have:

i) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succ_z j$, then $\tilde{U}_i(q) > \tilde{U}_j(q)$;

ii) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succeq_z j$ and $\exists z_0 \in \operatorname{argmin} \ell(q,.), i \succ_{z_0} j$, then $\tilde{U}_i(q) > \tilde{U}_j(q)$;
(The strict inequality comes from $\ell(q,z_0) < \ell(q,\tau_{i,j}z_0)$.)

iii) if $\exists z_0, z_1 \in \operatorname{argmin} \ell(q,.)$ such that $i \succ_{z_0} j$ and $j \succ_{z_1} i$, then $\tilde{U}_i(q) = \tilde{U}_j(q)$

The second and main step of the theorem is to show $\operatorname{argsort}(\tilde{U}(q)) \subseteq \bigcup_{z \in \operatorname{argmin} \ell(q,.)} z$. The main arguments consist in the following disjunction of cases. Denoting $\bigcup \operatorname{argmin} \ell(q,.) = \bigcup_{z \in \operatorname{argmin} \ell(q,.)} z$, we have:

a) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succ_z j$, then $\forall \sigma \in \bigcup \operatorname{argmin} \ell(q,.), i \succ_\sigma j$;

b) if $\forall z \in \operatorname{argmin} \ell(q,.), i \succeq_z j$ and $\exists z_0 \in \operatorname{argmin} \ell(q,.), i \succ_{z_0} j$, then $\forall \sigma \in \bigcup \operatorname{argmin} \ell(q,.)$ we either have $i \succ_\sigma j$ or $\tau_{i,j}\sigma \in \bigcup \operatorname{argmin} \ell(q,.)$.

c) if $\exists z_0, z_1 \in \operatorname{argmin} \ell(q,.)$ such that $i \succ_{z_0} j$ and $j \succ_{z_1} i$, then $\forall \sigma \in \bigcup \operatorname{argmin} \ell(q,.)$, we have $\tau_{ij}\sigma \in \bigcup \operatorname{argmin} \ell(q,.)$ (because $\ell(q, \Lambda(\sigma)) = \ell(q, \Lambda(\tau_{ij}\sigma))$ by *iii)* of Corollary 41).

d) if $\forall z \in \operatorname{argmin} \ell(q,.), i \sim_z j$, then $\forall \sigma \bigcup \operatorname{argmin} \ell(q,.)$, we have $\tau_{ij}\sigma \in \bigcup \operatorname{argmin} \ell(q,.)$ (by definition of $\sim_z$).

The critical implication of points *a-d)* above is that for all $\sigma \in \bigcup \operatorname{argmin} \ell(q,.)$, if $\tilde{U}_i(q) > \tilde{U}_i(q)$ then σ either gives a relative ordering of i, j properly, or $\tau_{ij}\sigma \in \bigcup \operatorname{argmin} \ell(q,.)$. If $\tilde{U}_i(q) = \tilde{U}_i(q)$, which happens only in cases *c)* and *d)* above, then $\{\sigma, \tau_{ij}\sigma\} \subseteq \bigcup \operatorname{argmin} \ell(q,.)$. Successive applications of this remark allows us to construct $\sigma \in \operatorname{argsort}(\tilde{U}(q)) \cap \bigcup_{z \in \operatorname{argmin} \ell(q,.)}$: start from an arbitrary $\sigma \in \bigcup_{z \in \operatorname{argmin} \ell(q,.)}$, and take any $\sigma' \in \operatorname{argsort}(\tilde{U}(q))$. Apply bubble sort starting from σ with target σ'. By the remark above, every step of bubble sort stays in $\bigcup \operatorname{argmin} \ell(q,.)$, which proves $\sigma' \in \bigcup \operatorname{argmin} \ell(q,.)$, and thus $\operatorname{argsort}(\tilde{U}(q)) \subset \bigcup \operatorname{argmin} \ell(q,.)$.

We thus proved $\forall q \in \mathcal{Q}, \operatorname{pred}(\tilde{U}q) \subset \operatorname{argmin} \ell(q,.)$. Equality is proved since this inclusion proves the square loss $y, s \mapsto (s - \tilde{u}(z))^2$ is L-calibrated, so equality follows from equality of argmins (Theorem 28).

□

Now, we have all the tools to wrap up the proof of Theorem 32.

Proof of Theorem 32.
(i) ⇒ (ii). As L is CEU, there exists u such that Φ_u^{sq} is L-calibrated. Thanks to Th. 28, we know

$$\forall q \in \mathcal{Q}, \exists \delta_0 > 0, \forall 0 < \delta < \delta_0, \operatorname{pred}(\operatorname{lev}_\delta \phi_u^{\text{sq}}(q,.)) = \operatorname{argmin} \ell(q,.)$$

As Φ_u^{sq} is convex in s and thus $\operatorname{lev}_\delta \phi_u^{\text{sq}}(q,.)$ is connected, then $\forall q \in \mathcal{Q},\ \operatorname{argmin} \ell(q,.)$ is connected.

(ii) ⇒ (i). This is exactly Th. 42.

(ii) ⇒ (iii). By Cor. 41, we have the property we referred to as *strict monotonicity* of the loss L. We use Lemma 36 to prove the connectedness by using path of adjacent items. Let us take $q \in \mathcal{Q}$, $\varepsilon > 0$ and $z \in \operatorname{lev}_\varepsilon \ell(q,.)$ and $\sigma \in z$. We know there exists $s \in \mathbb{R}^n$ such that $\operatorname{argmin} \ell(q,.) = \operatorname{pred}(s)$ (from (ii) ⇒ (i)), thus we can find $\nu \in \mathfrak{S}_n$ such that $\operatorname{pred}(\nu) \in \operatorname{argmin} \ell(q,.)$ and for any i, j such that $\tau_{ij} \in$ `bubblesort`(target=ν, input=σ), we have $\forall z' \in \operatorname{argmin} \ell(q,.), i \prec_z j$. Indeed, if there were $\tau_{ij} \in$ `bubblesort`(target=ν, input=σ) such that $\exists z' \operatorname{argmin} \ell(q,.), i \succeq_z j$ because $\operatorname{argmin} \ell(q,.) = \operatorname{pred}(s)$ then $\tau_{ij}\nu \in \operatorname{argsort}(s)$ and we could choose $\tau_{ij}\nu$ instead of ν.

Denoting $(\tau^{(m)})_{m=1}^M$ the output of `bubblesort`(target=ν, input=σ), we just prove $\tau^{(1)} z \in \operatorname{lev}_\varepsilon \ell(q,.)$, and then finish by induction.
Case 1: $i \sim_z j$. In this case $\tau_{ij} z = z \in \operatorname{lev}_\varepsilon \ell(q,.)$.
Case 2: $i \succ_z j$ By Cor. 41 *(i)*, $\ell(q, \tau_{ij} z) > \ell(q, z)$. Hence $\tau_{ij} z \in \operatorname{lev}_\varepsilon \ell(q,.)$. Then, by induction, for any $m \in [M]$ we have $\tau^{(m)} \ldots \tau^{(1)} z \in \operatorname{lev}_\varepsilon \ell(q,.)$.
Finally, any $z \in \operatorname{lev}_\varepsilon \ell(q,.)$ is connected to an element of $\operatorname{argmin} \ell(q,.)$ which is itself a connected set. Thus, $\operatorname{lev}_\varepsilon \ell(q,.)$ is connected.

(iii) ⇒ (ii). Because $\forall q \in \mathcal{Q}, \exists \varepsilon > 0, \operatorname{lev}_\varepsilon \ell(q,.) = \operatorname{argmin} \ell(q,.)$. □

D Proof of Corollary 7

Disclaimer: As mentioned in the body of the paper, we address, in these Appendices, a setting more general than ranking. This setting is described in Section A. We restate first the statement of the paper (on ranking), then state the more general statement (on weak orders) and prove the latter.

This section provides the proof of Corollary 7. First, we remind the exact statement of the paper in the ranking setting.

Corollary 7. *A ranking loss L is CEU if and only if: for every distribution P over $\mathcal{X} \times \mathcal{Y}$ such that $x \mapsto P(.|x)$ is continuous, there is a* continuous *optimal scoring function for $\mathcal{R}_{L,P}$.*

Then, the version of the same statement in $\mathcal{Z}$,

Corollary 43. *Under (A1), (A2) and (A3), the task loss L is CEU if and only if for every distribution P over $\mathcal{X} \times \mathcal{Y}$ such that $x \mapsto P(.|x)$ is continuous, there is a* continuous *optimal scoring function for $\mathcal{R}_{L,P}$ – i.e. a function $f : \mathcal{X} \to \mathbb{R}^n$ continuous such that $\forall x \in \mathcal{X}$, $\operatorname{pred}(f(x)) \subseteq \operatorname{argmin} \ell(P(.|x), .)$.*

The proof of that corollary, as well as all proofs regarding local minima when there are disconnected argmins rely on the following path between probability distributions $\bar{q}$. The proof is straightforward.

Lemma 44. *Let q_0 such that $|\operatorname{argmin} \ell(q_0, .)| > 1$. Let $z, z' \in \operatorname{argmin} \ell(q_0, .)$ with $z \neq z'$. Define the following path between probability distributions in $\mathcal{Q}$:*

$$\forall \alpha \in [0,1], \quad \bar{q}(\alpha) = \begin{cases} (1-2\alpha)q^{(z)} + 2\alpha q_0 & \text{if } \alpha \in [0, \frac{1}{2}] \\ (2-2\alpha)q_0 + (2\alpha-1)q^{(z')} & \text{if } \alpha \in [\frac{1}{2}, 1] \end{cases}.$$

Then, let

$$\epsilon_0 = \min_{q' \in \{q_0, q^{(z)}, q^{(z')}\}} \min_{z'' \notin \operatorname{argmin} \ell(q',.)} \big(\ell(q', z'') - \underline{\ell}(q')\big)$$

Then, $\epsilon > 0$ and we have:

$$\forall \alpha \in [0,1], \forall z'' \in \operatorname{lev}_{\epsilon_0} \ell(\bar{q}(\alpha), .), z'' \in \operatorname{argmin} \ell(q_0, .).$$

Proof. We prove the case $\alpha \in [0, \frac{1}{2}]$, the other case is similar. Let $\alpha \in [0, \frac{1}{2}]$. We first notice $z \in \operatorname{argmin} \ell(\bar{q}(\alpha), .)$. Thus, by developping $\bar{q}(\alpha)$ we have, for $z'' \neq z$:

$$\ell(\bar{q}(\alpha), z'') - \underline{\ell}(\bar{q}(\alpha)) = (1-2\alpha)\big(\ell(q^{(z)}, z'') - \ell(q^{(z)}, z)\big) + 2\alpha\big(\ell(q_0, z'') - \ell(q_0, z)\big) \geq \epsilon. \qquad \square$$

Proof of Corollary 43. The direct implication is straightforward: if L is CEU, we can choose $f : x \mapsto \mathbb{E}_{P(.|x)}[u(Y)]$ as optimal scoring function by Theorem 32, which is continuous whenever $x \mapsto P(.|x)$ is continuous.

For the reverse implication, let $q_0 \in \mathcal{Q}$. We have to prove that $\operatorname{argmin} \ell(q_0, .)$ is connected. Notice that if $|\operatorname{argmin} \ell(q_0, .)| = 1$ it is connected, so we focus on the case $|\operatorname{argmin} \ell(q_0, .)| > 1$. Let $z, z' \in \operatorname{argmin} \ell(q_0, .)$ with $z \neq z'$. Let $\mathcal{X} = [0,1]$, and take the distribution P over $\mathcal{X} \times \mathcal{Y}$ such that the marginal distribution over $\mathcal{X}$ is uniform and $x \mapsto P(.|x)$ is the path $\bar{q}$ constructed as in Lemma 44. If there is a continuous optimal scoring function $f : [0,1] \to \mathbb{R}^n$, it means that $\forall \alpha \in [0,1]$, $\operatorname{pred}(f(\alpha)) \in \operatorname{argmin} \ell(\bar{q}(\alpha), .)$ and thus $\operatorname{pred}(f([0,1])) \subseteq \operatorname{argmin} \ell(q_0, .)$ by Lemma 44.

Notice that f is continuous and thus preserves connectedness, so $\operatorname{pred}(f([0,1]))$ is connceted in $\mathcal{Z}$ by definition. By the characterization of connectedness in $\mathcal{Z}$ through paths of adjacent transpositions (Theorem 36), there is a path between z and z' in $\operatorname{argmin} \ell(q_0, .)$. The construction above can be repeated for every $z, z' \in \operatorname{argmin} \ell(q_0, .)$, which implies that $\operatorname{argmin} \ell(q_0, .)$ is connected. $\square$

E Local Minima of the Task Loss

E.1 Proof of Theorem 46

Definition 8. *Given a distribution $q \in \mathcal{Q}$ and a loss L, a ranking $\sigma \in \mathfrak{S}_n$ is a* local minimum *if for any $r \in [n-1]$, $\ell(q, \sigma) \leq \ell(q, \sigma\tau_{r,r+1})$.*

The definition of local minima for ranking losses 8 is extended to $\mathcal{Z}$ in a straightforward manner:

Definition 45. *Given a loss L and $q \in \mathcal{Q}$, a prediction $z \in \mathcal{Z}$ is* a local minimum *if, for any pair of adjacent items (i, j), we have $\ell(q, z) \leq \ell(q, \tau_{ij} z)$.*

Theorem 46. *Under (A1), (A2) and (A3), if a task loss L is not CEU, then the subset of distribution for which ℓ has non-global local minima has non-zero measure.*

Proof. We use the following property of the excess inner risk, where we denote $\|L\|_\infty = \max_{y,z} |L(y, z)|$. The calculation is similar to the one later in Lemma 48 for surrogate losses.

$$\forall q, q', \forall z, \;\; \big|\ell(q', z) - \underline{\ell}(q') - \big(\ell(q, z) - \underline{\ell}(q)\big)\big| \leq 2\|q - q'\|_1 \|L\|_\infty.$$

Let us take q_0 such that $\operatorname{argmin} \ell(q_0, .)$ is not connected, and $z \in \operatorname{argmin} \ell(q_0, .)$.

Let $\bar{q}(\alpha) = (1 - \alpha)q_0 + \alpha q^{(z)}$. We define the *gap* with respect to q_0 of $q \in \mathcal{Q}$ as:

$$G(q) = \min_{z' \notin \operatorname{argmin} \ell(q_0,.)} \big(\ell(q, z') - \underline{\ell}(q)\big).$$

We use $\epsilon_0 = G(q_0)$.

The important aspect of this gap is that if z and z' are both in $\operatorname{argmin} \ell(q_0, .)$ but not connected in $\operatorname{argmin} \ell(q_0, .)$, this disconnectedness remains in sublevel sets of $\ell(q, .)$ for other distributions q: If, for some $\epsilon > 0$, z and z' are not connected in $\operatorname{lev}_\epsilon \ell(q, .)$ with $\epsilon \leq G(q)$, then z and z' are in different connected components of $\operatorname{lev}_\epsilon \ell(q, .)$. We prove the existence of bad local minima by showing that there are suboptimal connected components smaller than the gap.

Let $\alpha_0 \in (0, 1]$. We then have $\{z\} = \operatorname{argmin} \ell(\bar{q}(\alpha_0), .)$. Moreover, let $\tilde{\epsilon} = \min_{z' \neq z}(\ell(q^{(z)}, z') - \underline{\ell}(q^{(z)}))$. Notice that $\tilde{\epsilon} > 0$ by definition of $q^{(z)}$. We have[4]:

$$G(\bar{q}(\alpha_0)) \geq \epsilon_0 - 4\alpha_0 \|L\|_\infty,$$
$$\forall z' \in \operatorname{argmin} \ell(q_0, .), z' \neq z, \quad 4\alpha_0\|L\|_\infty \geq \ell(\bar{q}(\alpha_0), z') - \underline{\ell}(\bar{q}(\alpha_0)) \geq \alpha_0 \tilde{\epsilon}.$$

And thus, given $\eta > 0$, for every q' such that $\|q' - \bar{q}(\alpha_0)\|_1 \leq \eta$, we have

$$G(q') \geq \epsilon_0 - (4\alpha_0 + 2\eta)\|L\|_\infty,$$

and $\forall z' \in \operatorname{argmin} \ell(q_0, .), z' \neq z$:

$$(4\alpha_0 + 2\eta)\|L\|_\infty \geq \ell(q', z') - \underline{\ell}(q') \geq \alpha_0 \tilde{\epsilon} - 2\eta \|L\|_\infty.$$

Let $\alpha_0 \in (0, 1)$ and $\eta \in (0, 1)$ such that

$$0 < \alpha_0 \tilde{\epsilon} - 2\eta\|L\|_\infty < (4\alpha_0 + 2\eta)\|L\|_\infty < \epsilon_0 - (4\alpha_0 + 2\eta)\|L\|_\infty.$$

These exist since $\epsilon_0 > 0$. Then, for any q' such that $\|\bar{q}(\alpha_0) - q'\| < \eta$, for any $z' \in \operatorname{argmin} \ell(q_0, .)$ such that z' is not in the same connected component as z, with such values of α_0 and η, z' is in a connected component of $\operatorname{lev}_{\epsilon_0 - (4\alpha_0 + 2\eta)\|L\|_\infty} \ell(q', .)$ that is suboptimal (because $\alpha_0 \tilde{\epsilon} - 2\eta\|L\|_\infty > 0$) and disconnected from z (becuase $G(q') \geq \epsilon_0 - (4\alpha_0 + 2\eta)\|L\|_\infty$). Thus, all such q' have a bad local minimum. The result follows, since the measure of the $\|.\|_1$-ball of radius η is non-zero. □

F Proof of Utility Computation on Generalized DCG

The expression of the utility u from Theorem 6 may not be efficient to compute a priori. As it happens, for many evaluation metrics, the expression of u does simplify for numerous common tasks losses.

[4] The last equality comes from noticing $\underline{\ell}(\bar{q}(\alpha_0)) = \ell(\bar{q}(\alpha_0), z)$ and decomposing

$$\ell(\bar{q}(\alpha_0), z') - \underline{\ell}(\bar{q}(\alpha_0)) = \alpha_0\big(\ell(q^{(z)}, z') - \ell(q^{(z)}, z)\big) + (1 - \alpha_0)\big(\ell(q_0, z') - \ell(q_0, z))\big).$$

where the term in q_0 vanishes because both z and z' are optimal for q_0.

Proposition 47. *We assume here a ranking task loss L of the form $L(y,\sigma) = \mathrm{DCG}_{w,u}(y,\sigma) = -\sum_{k=1}^{n} w_k u_{\sigma(k)}(y) - b(y)$ where $\forall i \in [n], u_i$ is increasing and w is decreasing. Then, there exists $B : \mathcal{Y} \to \mathbb{R}_*^+$ and $A > 0$ such that, $\forall i \in [n], \forall y \in \mathcal{Y}$,*

$$\tilde{u}_i(y) = B(y) + A u_i(y)$$

Proof. By definition, we have,

$$\begin{aligned}
u_i(y) &= -\sum_{\sigma \in \mathfrak{S}_n} \mathbb{1}_{[\sigma(1)=i]} L(y,\sigma) \\
&= \sum_{\sigma \in \mathfrak{S}_n} \mathbb{1}_{[\sigma(1)=i]} \left(b(y) + \sum_{r=1}^{n} u_{\sigma(r)}(y) w_r \right) \\
&= b(y)(n-1)! + \sum_{\sigma \in \mathfrak{S}_n} \sum_{r=1}^{n} \mathbb{1}_{[\sigma(1)=i]} u_{\sigma(r)}(y) w_r \\
&= b(y)(n-1)! + u_i(y) w_1 (n-1)! + \sum_{\sigma \in \mathfrak{S}_n} \sum_{r=2}^{n} \mathbb{1}_{[\sigma(1)=i]} u_{\sigma(r)}(y) w_r \\
&= b(y)(n-1)! + u_i(y) w_1 (n-1)! + \sum_{\sigma \in \mathfrak{S}_n} \sum_{r=2}^{n} \sum_{k \neq i} \mathbb{1}_{[\sigma(r)=k]} \mathbb{1}_{[\sigma(1)=i]} u_k(y) w_r \\
&= b(y)(n-1)! + u_i(y) w_1 (n-1)! + \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r \sum_{\sigma \in \mathfrak{S}_n} \mathbb{1}_{[\sigma(r)=k]} \mathbb{1}_{[\sigma(1)=i]} \\
&= b(y)(n-1)! + u_i(y) w_1 (n-1)! + \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r \mathbb{1}_{[k \neq i]} (n-2)! \\
&= b(y)(n-1)! + u_i(y) w_1 (n-1)! - \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r \mathbb{1}_{[k=i]} (n-2)! + \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r \Big(\mathbb{1}_{[k \neq i]} + \mathbb{1}_{[k=i]} \Big) (n-2)! \\
&= b(y)(n-1)! + \sum_{r=2}^{n} \Big(u_i(y) w_1 - u_i(y) w_r \Big) (n-2)! + \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r (n-2)! \\
&= \underbrace{b(y)(n-1)! + \sum_{k=1}^{n} \sum_{r=2}^{n} u_k(y) w_r (n-2)!}_{B(y)} + u_i(y) \underbrace{\sum_{r=2}^{n} \Big(w_1 - w_r \Big) (n-2)!}_{A} \\
&= B(y) + A u_i(y)
\end{aligned}$$

□

G Gumbel Smoothing is a Calibrated Surrogate Loss

Disclaimer: While this section presents a way to build, for any L, a non-convex surrogate loss L-calibrated, it is by no means one to use in practice. Its interest is mostly theoretical as its computational complexity is generally prohibitive.

Disclaimer: For this section, we keep the specific case of the ranking where $\mathcal{Z} = \mathfrak{S}_n$.

We examine here a convolution with a well-behaved kernel to smooth $L(., \mathrm{argsort}(.))$. We obtain the following surrogate loss,

$$\Phi_L^{\mathrm{NC}} : y, s \mapsto \int_{\mathbb{R}^n} L(y, \mathrm{argsort}(u - s)) \kappa(u) \mathrm{d}u$$

The particular kernel κ we choose here is a Gumbel density as it can easily be reformulated in the ranking space using a Plackett-Luce model. Indeed, for any $s \in \mathbb{R}^n$ and $u \sim \mathrm{Gumbel}(0,1)^n$, we have that $\operatorname{argsort}(s+u)$ follows a Plackett-Luce distribution[38], allowing to express Φ_L^{NC} in closed-form (with prohibitive computational cost for large values of n):

$$\begin{aligned}
\Phi_L^{\text{NC}}(y,s) &= \int_{\mathbb{R}^n} L(y, \operatorname{argsort}(s-u))\kappa(u)\mathrm{d}u \\
&= \int_{\mathbb{R}^n} L(y, \operatorname{argsort}(s-u))k(u)\mathrm{d}u \\
&= \sum_{\sigma\in\mathfrak{S}_n} L(y,\sigma) \underbrace{\int_{\mathbb{R}^n} \mathbb{1}_{[\operatorname{argsort}(s-u)=\sigma]}k(u)\mathrm{d}u}_{\mathbb{P}(\sigma|e^s) \text{ of a Plackett-Luce model}} \\
&= \sum_{\sigma\in\mathfrak{S}_n} L(y,\sigma) \prod_{r=1}^{n} \frac{e^{s_{\sigma(r)}}}{\sum_{k\geq r} e^{s_{\sigma(k)}}}
\end{aligned}$$

Proposition 11. *For any ranking loss L, the following surrogate loss Φ_L^{NC} is L-calibrated.*

$$\Phi_L^{\text{NC}} : y, s \mapsto \int_{\mathbb{R}^n} L(y, \operatorname{argsort}(u-s))\kappa(u)\mathrm{d}u = \sum_{\sigma\in\mathfrak{S}_n} L(y,\sigma) \prod_{r=1}^{n} \frac{e^{s_{\sigma(r)}}}{\sum_{k\geq r} e^{s_{\sigma(k)}}}$$

Proof. Considering the surrogate $\Phi = \Phi_L^{\text{NC}}$, let $q \in \mathcal{P}$, to show the calibration, we show that

1. $\underline{\phi}(q) = \underline{\ell}(q)$.
2. $\forall s \in \mathbb{R}^n$ such that $\exists \nu \in \operatorname{argsort}(s), \ell(q,\nu) > \underline{\ell}(q)$ we have $\phi(q,s) \geq \frac{n!-1}{n!}\underline{\ell}(q) + \frac{1}{n!}\ell(q,\nu) > \underline{\ell}(q)$

Let's prove 1. first. We can choose $\sigma \in \operatorname{argmin} \ell(q,.)$ and $s \in \mathbb{R}^n$ such that $\operatorname{argsort}(s) = \{\sigma\}$. Taking $\alpha > 0$, we have $\mathbb{P}(\sigma|e^{\alpha s}) \xrightarrow[\alpha\to\infty]{} 1$, meaning that $\phi(q, \alpha s) \xrightarrow[\alpha\to\infty]{} \ell(q,\sigma) = \underline{\ell}(q)$.
Let's prove 2. now and consider $s \in \mathbb{R}^n, \exists \nu \in \operatorname{argsort}(s), \ell(q,\nu) > \underline{\ell}(q)$. Because $\nu \in \operatorname{argsort}(s)$ it is an event of maximal probability in the Plackett-Luce model $\mathbb{P}(.|e^s)$. Hence $\mathbb{P}(\nu|e^s) > \frac{1}{n!}$.

$$\begin{aligned}
\phi(q,s) &= \sum_{\sigma\in\mathfrak{S}_n} \ell(q,\sigma)\mathbb{P}(\sigma|e^s) \\
&\geq \frac{1}{n!}\ell(q,\nu) + \left(\mathbb{P}(\nu|e^s) - \frac{1}{n!}\right)\ell(q,\nu) + \sum_{\sigma\neq\nu} \ell(q,\sigma)\mathbb{P}(\sigma|e^s) \\
&\geq \frac{1}{n!}\ell(q,\nu) + \left(\mathbb{P}(\nu|e^s) - \frac{1}{n!}\right)\underline{\ell}(q) + \sum_{\sigma\neq\nu} \underline{\ell}(q)\mathbb{P}(\sigma|e^s) \\
&\geq \frac{1}{n!}\ell(q,\nu) + \frac{n!-1}{n!}\underline{\ell}(q)
\end{aligned}$$

□

H Local Minima of Surrogate Loss

H.1 Simulations to Analyze Surrogate Bad Local Valleys

We describe here the simulations ran to analyze the loss surface of the Φ_L^{NC}. We remind its definition:

$$\Phi_L^{\text{NC}} : y, s \mapsto \int_{\mathbb{R}^n} L(y, \operatorname{argsort}(u-s))\kappa(u)\mathrm{d}u = \sum_{\sigma\in\mathfrak{S}_n} L(y,\sigma) \prod_{r=1}^{n} \frac{e^{s_{\sigma(r)}}}{\sum_{k\geq r} e^{s_{\sigma(k)}}}.$$

Let us first make a few remarks on Φ_L^{NC} to explain the intention and the design of the experiments. First, about the smoothing. It is common, when smoothing a function with a convolution, to have the strength of the smoothing controlled by the bandwidth of the convolution kernel (e.g., variance of a Gaussian kernel). This is unnecessary here: the function $s \mapsto L(y, \operatorname{argsort}(s))$ is invariant by re-scaling of the scores s (contrarily to the convolution kernel). Thus, the loss L is more smoothed towards vectors s of small norm (up to removing a translation) and almost not smoothed when the norm of s goes towards infinity.

Second, because the task loss is more smoothed towards scores of low norm, initializing optimization algorithms close to 0 (all items have the same score) proves to be an empirically good heuristic.

We performed two experiments, similarly to the simulation for the task losses described in Section 3. Given n items, we ran simulations where each simulation consists in sampling a distribution q then, for each q, we randomly choose a starting point s_0 for the optimization and run a gradient descent on Φ_L^{NC} with a basic line search to handle the poor conditioning of Φ_L^{NC}. To handle bad valleys where the infimum is not reached while avoiding numerical issues, we project the optimization to stay inside a $\|.\|_2$ ball of radius 100. This is sufficient to ensure that the local minimum reached in the $\|.\|_2$ ball generates the same ranking σ as the scores at infinity. We then compute the sub-optimality of this local valley as $\eta = \frac{\ell(q,\sigma)-\min \ell(q,.)}{\max \ell(q,.)-\min \ell(q,.)}$.

Simulation 3: Distribution of sub-optimality of bad local valleys. Here, q is sampled uniformly on $\mathcal{Q}$ and s_0 is sampled from a $\mathcal{N}(0, \mathrm{Id}_n)$, where Id_n is the $n \times x$ identity matrix. Figure 3 *(left and middle)* (reproduced in Figure 4) illustrates the results of the distribution of η given that $\eta > 0$. In these plots, the suboptimality of *bad* local valleys are considered in the distribution.

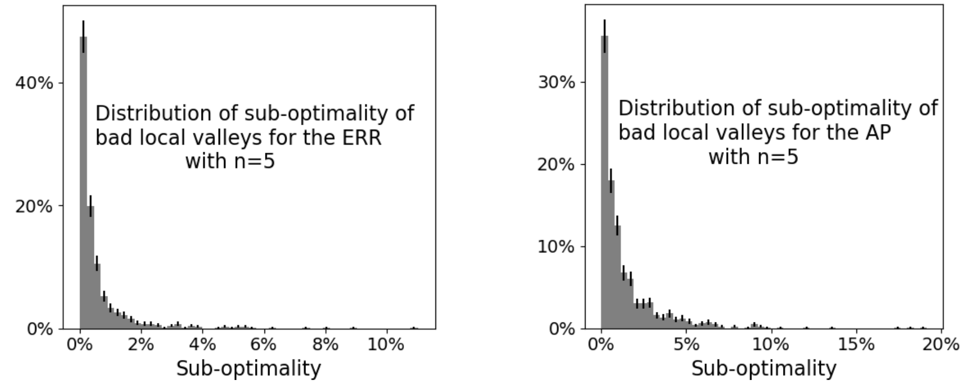

Figure 4: Distribution of sub-optimality of bad local valleys. Left: ERR. Right:AP.

Simulation 4: Percentage of optimization runs stuck in bad local valleys. Here, we want to answer the question: *How often can we expect the optimization of the surrogate to be stuck in a bad local valley when the task loss has several local minima?* To answer this, q is sampled uniformly amongst distributions for which $\ell(q, .)$ has at least two local minima, and s_0 is set to 0 as it proved an empirically good intialization. Figure 3 (right) (reproduced here as Figure 5) illustrates the percentage of runs that are stuck in local minima as a function of the number of items n for the ERR and the AP. For $n = 3$ 5-10% of the runs are stuck in local minima, and is growing with the number of items n until $n = 6$. Moreover, the bad local valleys of Φ_L^{NC} found, actually corresponds to rankings that are bad local minima for the task loss itself. This suggests that these local minima of the surrogate loss reflect an intrinsic difficulty of the optimization for calibrated surrogate losses rather than an artifact of Φ_L^{NC}. We thus conjecture that optimization is difficult for other surrogate losses as well.

H.2 Preliminaries for the study of surrogate losses

The two subsequent sections study properties of surrogate losses calibrated with non-CEU losses. We summarize here basic results that the two sections use.

We first summarize the assumptions we use:

(A5). *The four following statements hold:*

i) (A1), (A2) and (A3) hold,

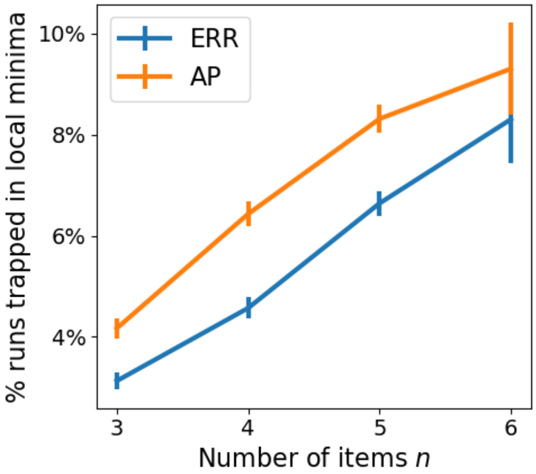

Figure 5: Percentage of optimizations runs on Φ_L^{NC} that end up stuck in a bad local valley over distribution for which the task loss as several local minima.

ii) $\exists q_0$ *such that* $\operatorname{argmin} \ell(q_0, .)$ *is not connected,*

iii) $\Phi : \mathcal{Y} \times \mathbb{R}^n \to \mathbb{R}$ *is such that* $\forall y \in \mathcal{Y}$*,* $s \mapsto \Phi(y, s)$ *is* β_Φ*-Lipschitz and uniformly bounded by* B_Φ*,*

iv) Φ *is* L*-calibrated.*

The assumption above implies Lipschitzness of the excess inner risk $\phi(q, s) - \underline{\phi}(q)$ with respect to both $q \in \mathcal{Q}$ and $s \in \mathbb{R}^n$:

Lemma 48. *If* Φ *is bounded and Lipschitz as defined in Assumption (A5) iii), then:*

$$\forall q, q', \forall s, s', \quad |\phi(q, s) - \underline{\phi}(q) - \phi(q', s') - \underline{\phi}(q')| \leq 2\|q - q'\|_1 B_\Phi + \beta_\Phi \|s - s'\|_2$$

Proof. First consider $A = \phi(q, s) - \underline{\phi}(q) - \phi(q', s') - \underline{\phi}(q')$. Let $\epsilon > 0$ and s^* such that $\underline{\phi}(q) \geq \phi(q, s^*) - \epsilon$. We have

$$\begin{aligned} A &\leq \phi(q, s) - \phi(q, s^*) + \epsilon - \phi(q', s) + \phi(q', s^*) \\ &\leq \|q - q'\|_1 \max_y |\Phi(y, s) - \Phi(y, s^*)| + \phi(q', s) - \phi(q', s) + \epsilon \\ &\leq 2\|q - q'\|_1 B_\Phi + \beta_\Phi \|s - s'\|_2 + \epsilon\,, \end{aligned}$$

where the first inequality holds since $\phi(q', s^*) \geq \underline{\phi}(q')$ and the second from Hölder's inequality. Following the same steps, we find the same upper bound for $\phi(q', s') - \underline{\phi}(q') - \phi(q, s) - \underline{\phi}(q)$. The final result is obtained in the limit $\epsilon \to 0$. □

The following result give the main implications of calibration of Φ. It are extended version of Lemma 44 for surrogate losses instead of task losses. It gives basic properties of segments joining a distribution q_0 with a disconnected argmin, with distributions of type $q^{(z)}$ where $z \in \operatorname{argmin} \ell(q_0, ., .)$. The lemma has several parts that are used in different proofs.

Lemma 49. *Under Assumption (A5), let* $q_0 \in \mathcal{Q}$ *such that* $\operatorname{argmin} \ell(q_0, .)$ *is disconnected.*

i) *Let* $z, z' \in \operatorname{argmin} \ell(q_0, .)$ *such that* z *and* z' *are not in the same connected components of* $\operatorname{argmin} \ell(q_0, .)$*. Let* $\bar{q}$ *be defined as in Lemma 44. Then* $\exists \delta_0$ *such that for every continuous function* $f : [0, 1] \to \mathbb{R}^n$ *such that* $\{z, z'\} \subseteq \operatorname{pred}(f([0, 1]))$*, we have:*

$$\max_{\alpha \in [0,1]} \big(\phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))\big) \geq \delta.$$

ii) *Let* $\bar{q} : [0, 1] \to \mathcal{Q}$ *defined by:*

$$\forall \alpha \in [0, 1], \quad \bar{q}(\alpha) = (1 - \alpha) q_0 + \alpha q^{(z)}.$$

Let $G_0 = \min\limits_{z \notin \text{argmin}_{q_0}} \min\limits_{q \in \{q_0, q^{(z)}\}} \big(\ell(q,z) - \underline{\ell}(q)\big)$*. Then* $G_0 > 0$*.*

Moreover, let $\eta_0 \in (0, \frac{G_0}{2\|L\|_\infty})$*. Then* $\exists \delta_0 > 0$ *such that:*

$$\forall \alpha \in [0,1], \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta), \quad \forall s \in \text{lev}_{\delta_0} \phi(q,.), \text{pred}(s) \subseteq \text{argmin}\, \ell(q_0,.).$$

iii) For any $q \in \mathcal{Q}$ *such that* $\text{argmin}\, \ell(q,.) = \{z\}$*, there is* $\delta_0 > 0$ *and* $\eta_0 > 0$ *such that:*

$$\forall s \in \text{lev}_{\delta_0} \phi(q,.), \forall q' \in \mathcal{B}_1(q, \eta_0), \ \ \text{pred}(s) = \{z\}.$$

Proof. Point *i)*. Using uniform calibration (Theorem 30, let ϵ_0 be defined as in Lemma 44, and let δ_0 such that $\forall q \in \mathcal{Q}, s \in \text{lev}_{\delta_0} \ell(q,.), \text{pred}(s) \subseteq \text{lev}_{\epsilon_0} \ell(q,.)$. Combining with Lemma 44, we have:

$$\forall \alpha \in [0,1], \forall s \in \text{lev}_{\delta_0} \phi(\bar{q}(\alpha),.), \text{pred}(s) \subseteq \text{argmin}\, \ell(q_0,.). \tag{7}$$

Since connectedness is preserved by continuous functions, any continuous function $f : [0,1] \to \mathbb{R}^n$ such that $\{z, z'\} \subseteq \text{pred}(f([0,1]))$ must have $\text{pred}(f([0,1])) \not\subseteq \text{argmin}\, \ell(q_0,.)$. Using (7), this means $\exists \alpha, f(\alpha) \notin \text{lev}_{\delta_0} \phi(\bar{q}(\alpha),.)$, which is equivalent to the desired result.

Point *ii)*. With the same arguments as Lemma 44, for every $\alpha \in [0,1]$, if $z \in \text{lev}_{G_0} \ell(\bar{q}(\alpha),.)$ then $z \in \text{argmin}\, \ell(q_0,.)$.

Now, let us take $\eta \in (0, \frac{G_0}{2\|L\|_\infty})$. For every $\alpha \in [0,1]$, the Lipschitzness of $q \mapsto \ell(q,z) - \underline{\ell}(q)$, which follows from similar arguments as Lemma 48, gives:

$$\forall \alpha \in [0,1], \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta), \forall z \notin \text{argmin}\, \ell(q_0,.), \ \ \ell(q,z) - \underline{\ell}(q) \geq G_0 - 2\eta\|L\|_\infty > 0 \tag{8}$$

By uniform calibration (Theorem 30), there is $\delta_0 > 0$ such that $\forall \delta \in (0, \delta_0), \forall q \in \mathcal{Q}, \forall s \in \text{lev}_\delta \phi(q,.), \text{pred}(s) \subseteq \text{lev}_{G_0 - 2\eta\|L\|_\infty} \ell(q,.)$. With this choice of δ_0, using (8) gives the result.

Point *iii)*. The proof is exactly the same a before: Let $\epsilon = \min_{z' \neq z}(\ell(q,z') - \underline{\ell}(z))$. We have $\epsilon > 0$ by the assumption $\text{argmin}\, \ell(q,.) = \{z\}$. Let $\eta_0 \in (0, \frac{\epsilon}{2\|L\|_\infty})$. We have

$$\forall q' \in \mathcal{B}_1(q, \eta), \min_{z' \neq z}(\ell(q', z') - \underline{\ell}(z)) \geq \epsilon - 2\eta\|L\|_\infty > 0.$$

And thus $\forall q' \in \mathcal{B}_1(q, \eta), \{z\} = \text{argmin}\, \ell(q,.)'$. By uniform calibration (Th. 30), using $\delta_0 > 0$ such that $\forall q' \in \mathcal{Q}, \forall s \in \text{lev}_{\delta_0} \phi(q',.), s \in \text{lev}_{\epsilon - 2\eta\|L\|_\infty} \ell(q',.)$ gives the result. □

H.3 Lower bound on approximation error of surrogate loss by Lipschitz functions

The proposition below is a stronger version of Proposition 13. It makes it precise that the conditional distribution over $\mathcal{Y}$ is infinitely many times differentiable with bounded derivatives at any order (denoted $x \mapsto P(.|x) \in \mathcal{W}^\infty([0,1]^\mathcal{Q})$), and that there is a single point in $[0,1]$ on which the argmin of ℓ is not a singleton (and thus disconnected) (condition $|\{\alpha \in [0,1] : |\text{argmin}\, \ell(p_{Y|X=\alpha}|,.)| > 1\}| = 1$). This makes sure that the distibution is not too peculiar.

Proposition 50. *Under Assumption (A5), there is a probability measure* P *over* $[0,1] \times \mathcal{Y}$ *where the marginal distribution over* $[0,1]$*, and* $x \mapsto P(.|x) \in \mathcal{W}^\infty([0,1]^\mathcal{Q})$ *such that* $|\{\alpha \in [0,1] : |\text{argmin}\, \ell(p_{Y|X=\alpha}|,.)| > 1\}| = 1$*, and constants* $c, c' > 0$*, such that for all* $\beta \geq 0$

$$\inf_{\substack{f:[0,1]\to\mathbb{R}^n \\ f:\beta-\textit{Lipschitz}}} \mathcal{R}_{\Phi,P}(f) - \inf_{g:\mathcal{X}\to\mathbb{R}^n} \mathcal{R}_{\Phi,P}(g) \geq \min\big(c', \frac{c}{8B_\Phi + \beta_\Phi \beta}\big).$$

Proof. Let q_0 such that $\text{argmin}\, \ell(q_0,.)$ is disconnected and let $z, z' \in \text{argmin}\, \ell(q_0,.)$ such that z and z' belong to two different connected components of $\text{argmin}\, \ell(q_0,.)$. Let $\bar{q}(\alpha)$ defined as in Lemma 44. Clearly, $\bar{q} \in \mathcal{W}^\infty([0,1]^{|\mathcal{Y}|})$, i.e., $\bar{q}$ has bounded derivatives of any order. Also there is only one value of α ($\alpha = \frac{1}{2}$) for which $\text{argmin}\, \ell(\bar{q}(\alpha),.)$ is not a singleton.

Let f be β-Lipschitz. We consider two cases:

- If $\{z, z'\} \subset \mathrm{pred}(f([0,1]))$:

  First, we show that $\alpha \mapsto \phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))$ is Lipschitz. Notice that $\|\bar{q}(\alpha) - \bar{q}(\alpha')\|_1 \le 2|\alpha - \alpha'|$ when α and α' are in the same help-segment of $[0,1]$. When they are not, say $\alpha \in [0, \frac{1}{2})$ and $\alpha' \in [\frac{1}{2}, 1]$, we have

  $$\begin{aligned} \|\bar{q}(\alpha) - \bar{q}(\alpha')\|_1 &\le \|(2 - \alpha - \alpha')q_0\|_1 + 2\left\|(\frac{1}{2} - \alpha)q^{(z)}\right\|_1 + 2\left\|(\frac{1}{2} - \alpha')q^{(z')}\right\|_1 \\ &\le 4\big(|\frac{1}{2} - \alpha| + |\frac{1}{2} - \alpha'|\big) = 4(\alpha' - \alpha). \end{aligned}$$

  We thus have that $\alpha \mapsto \phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))$ is $8B_\Phi + \beta_\Phi\beta$-Lipschitz by Lemma 48.

  Moreover, by point *i)* of Lemma 49, choose $\delta > 0$ such that $\max_{\alpha\in[0,1]} \phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha)) \ge \delta$. The minimum of the integral $\int_{[0,1]} \phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))d\alpha$ is attained for a trapezoidal or triangle shape, with the maximum attained at a bound (0 or 1), of length $\min(1, \frac{\delta}{8B_\Phi+\beta_\Phi\beta})$ (because the domain is of length 1). We can conclude

  $$\mathcal{R}_{\Phi,P}(f) - \inf_{g:\mathcal{X}\to\mathbb{R}^n} \mathcal{R}_{\Phi,P}(g) = \int_{[0,1]} \big(\phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))\big)d\alpha \ge \frac{1}{2}\min(\delta, \frac{\delta^2}{8B_\Phi + \beta_\Phi\beta})$$

- If $\{z, z'\} \not\subseteq \mathrm{pred}(f([0,1]))$:

  Assume for instance $z \notin \mathrm{pred}(f([0,1]))$. Let $\epsilon = \min_{z''\neq z} \ell(q^{(z)}, z'') - \underline{\ell}(q^{(z)})$. For $\alpha < 1/4$, using $\underline{\ell}(\bar{q}(\alpha)) = \ell(\bar{q}(\alpha), z)$, we have:

  $$\ell(\bar{q}(\alpha), \mathrm{pred}(f(\alpha))) - \underline{\ell}(\bar{q}(\alpha)) \ge (1 - 2\alpha)\big(\ell(q^{(z)}, \mathrm{pred}(f(\alpha))) - \ell(q^{(z)}, z)\big) \ge \frac{1}{2}\epsilon.$$

  By uniform calibration (Theorem 30), there is δ'_z such that $\forall\alpha \in [0, \frac{1}{4})$, $\phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha)) \ge \delta'_z$ (or we would have $\mathrm{pred}(f(\alpha)) = \{z\} = \mathrm{argmin}\,\ell(\bar{q}(\alpha), .)$). We then have

  $$\int_{\alpha\in[0,1]} \big(\phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))\big)d\alpha \ge \int_{\alpha\in[0,\frac{1}{4}]} \big(\phi(\bar{q}(\alpha), f(\alpha)) - \underline{\phi}(\bar{q}(\alpha))\big)d\alpha \ge \frac{1}{4}\delta'_z$$

  Similarly if $z' \notin \mathrm{pred}(f([0,1]))$.

Taking $c' = \min(\frac{1}{4}\delta'_z, \frac{1}{4}\delta'_{z'}, \frac{\delta}{2})$ and $c = \frac{\delta^2}{2}$ gives the result. □

H.4 Local valleys of calibrated surrogate losses

We now prove that when the task loss does not have connected argmins, the set of distributions in $\mathcal{Q}$ for which $\phi(q, .)$ has bad local valleys has non-zero measure.

Proposition 51. *Under Assumption (A5), the set* $\{q \in \mathcal{Q} : \phi(q, .)$ *has bad local valleys*$\}$ *has non-zero Lebesgue measure.*

Proof. Let q_0 be a distribution such that $\mathrm{argmin}\,\ell(q_0, .)$ is disconnected, and let $z \in \mathrm{argmin}\,\ell(q_0, .)$. Let $\bar{q}$, δ_0 and η_0 be defined as in point *ii)* of Lemma 49, i.e., $\bar{q}$ is a segment between q_0 and $q^{(z)}$.

First let $\delta_1 \in (0, \delta_0)$, such that $\mathrm{pred}(\mathrm{lev}_{\delta_1}\phi(q_0, .)) = \mathrm{argmin}\,\ell(q_0, .)$. Such a δ_1 exists using the equality of argmins for calibrated surrogate losses (Theorem 28).

Let $s' \in \mathrm{lev}_{\delta_1}\phi(q_0, .)$ such that $\mathrm{pred}(s')$ is not in the same connected component of $\mathrm{argmin}\,\ell(q_0, .)$ as z. Using the Lipschitz property of $q \mapsto \phi(q, s) - \underline{\phi}(q)$ (Lemma 48), we have:

$$\forall\alpha \in [0,1], \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta), \quad \phi(q, s') - \underline{\phi}(q) < \delta_1 + (2\eta + 4\alpha)B_\Phi. \tag{9}$$

Moreover, for every $\alpha > 0$, using point *iii)* of Lemma 49 with $q := \bar{q}(\alpha)$, we can find $\delta_\alpha \in (0, \delta_1)$ and $\eta_\alpha > 0$ such that:

$$\forall s \in \mathrm{lev}_{\delta_\alpha}\phi(\bar{q}(\alpha), .), \;\; \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta_\alpha), \mathrm{pred}(s) = \{z\}.$$

Thus, for any $s' \in \mathrm{lev}_{\delta_1}\phi(q_0, .)$ such that $\mathrm{pred}(s')$ is not in the same connected component as z is not in $\mathrm{lev}_{\delta_\alpha}\ell(\bar{q}(\alpha), .)$, we have

$$\forall \alpha \in [0,1], \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta_\alpha), \quad \phi(q, s') - \underline{\phi}(q) \geq \delta_\alpha + (2\eta_\alpha + 4\alpha)B_\Phi. \tag{10}$$

Finally, since z and z' are not connected in $\operatorname{argmin} \ell(q_0, .)$, for every continuous function $f : [0,1] \to \mathbb{R}^n$ with $z \in \mathrm{pred}(f(0))$ and $z' \in \mathrm{pred}(f(1))$, we must have $\mathrm{pred}(f([0,1])) \not\subseteq \operatorname{argmin} \ell(q_0, .)$ (since continuous functions preserve connectedness), and thus, by our choice of δ_0 from Lemma 49, we have:

$$\forall \alpha \in [0,1], \forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta_0) \max_{t \in [0,1]} \big(\phi(q, f(t)) - \underline{\phi}(q)\big) \geq \delta_0. \tag{11}$$

The proof finishes by taking $\alpha = \frac{\delta_0 - \delta_1}{12B_\Phi}$, $\eta = \min(\eta_\alpha, \eta_0, \alpha)$. With these values, we have that $\forall q \in \mathcal{B}_1(\bar{q}(\alpha), \eta)$:

1. Using (9), there exists $s' \in \mathrm{lev}_{\delta_1 + (2\eta + 4\alpha)B_\Phi}\phi(q, .)$ such that $\mathrm{pred}(s')$ is in $\operatorname{argmin} \ell(q_0, .)$ but not connected to z in $\operatorname{argmin} \ell(q_0, .)$.
2. Using $\delta_1 + (2\eta + 4\alpha)B_\Phi < \delta_0$ and (11), s' above and its connected component C in $\mathrm{lev}_{\delta_1 + (2\eta + 4\alpha)B_\Phi}\phi(q, .)$ are not connected to any s such that $\mathrm{pred}(s) = \{z\}$. Thus, the connected component C is a local valley of $\ell(q, .)$,
3. Using (10), the infimum over C is suboptimal. Thus C is a bad local valley.

Thus, the measure of bad local valleys is at least the measure of $\mathcal{B}_1(\bar{q}(\alpha), \eta)$, and is thus > 0.

□