Stability and Multigroup Fairness in Ranking with Uncertain Predictions

Siddartha Devic [1] **Aleksandra Korolova** [2] **David Kempe** [1] **Vatsal Sharan** [1]

Abstract

Rankings are ubiquitous across many applications, from search engines to hiring committees. In practice, many rankings are derived from the output of predictors. However, when predictors trained for classification tasks have intrinsic uncertainty, it is not obvious how this uncertainty should be represented in the derived rankings. Our work considers *ranking functions*: maps from individual predictions for a classification task to distributions over rankings. We focus on two aspects of ranking functions: stability to perturbations in predictions and fairness towards both individuals and subgroups. Not only is stability an important requirement for its own sake, but — as we show — it composes harmoniously with individual fairness in the sense of Dwork et al. (2012). While deterministic ranking functions cannot be stable aside from trivial scenarios, we show that the recently proposed *uncertainty aware* (UA) ranking functions of Singh et al. (2021) are stable. Our main result is that UA rankings also achieve group fairness through successful composition with *multiaccurate* or *multicalibrated* predictors. Our work demonstrates that UA rankings naturally interpolate between group and individual level fairness guarantees, while simultaneously satisfying stability guarantees important whenever machine-learned predictions are used.

1. Introduction

Rankings underpin many modern systems: companies rank job applications (TurboHire, 2023; Geyik et al., 2019), ad marketplaces rank ads to serve to a user (Google, 2023), and social media platforms and feeds rank content (Meta, 2023). Rankings are also used to partially automate decision making in settings with limited resources or attention span (such as job candidate interview selections or ad delivery). Rankings are often derived from predictions generated by machine learning models designed and deployed on relevant classification tasks. For example, a job advertisement platform may use a model which predicts an individual's *relevance score* for each job they apply to; say, on a scale from 1 to 3 (corresponding to irrelevant, suitable, or extremely relevant); this score is then factored into the platform's ranking of job applicants shown to a company recruiter with a limited time budget.

In practice, machine learning models often predict *distributions* over classes instead of a single class. This is because predictions correspond to a belief about what the future may possibly hold, but not a certainty about what the future *will* look like. A plethora of recent research in model calibration (Guo et al., 2017; Minderer et al., 2021; Gupta & Ramdas, 2022), conformal prediction (Bastani et al., 2022; Jung et al., 2023), and uncertainty quantification (Angelopoulos & Bates, 2021) has tackled the issue of ensuring that the uncertainty estimates output by a model are meaningful, rather than artifacts of any particular training regime.

With uncertainty inherent in predictions, we argue that it is essential to revisit the question of how to meaningfully convert predictions (made in the form of distributions over classes) into rankings. Without any uncertainty — for example, if one had access to an oracle for the future — it would usually be clear what a ranking should look like: *meritocracy* would suggest that one always places the more suitable candidates higher in the ranking. However, when given only *predictions* about suitability/merit with intrinsic uncertainty, the approach for generating a meaningful ranking is less clear. After all, one must choose a ranking over candidates, e.g., in order to make an interview decision, *before* witnessing the exact suitability of each candidate, which is generally only observable after an individual works in the job for months or even years. Since an uncertain prediction can be considered a *prior* (and typically imperfect) belief on the qualifications or performance of any given individual, the fundamental task of designing a meaningful ranking algorithm *utilizing* these predictions must be reexamined.

For a meaningful derivation of rankings from predictors, we consider the following two properties to be essential:

[1]Department of Computer Science, University of Southern California [2]Department of Computer Science and Public Affairs, Princeton University. Correspondence to: Siddartha Devic <devic@usc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Anonymity. All individuals must be treated symmetrically a priori, i.e., if the predictions are permuted, then the ranking is permuted according to the same permutation.

Stability. If the predictor's distribution over classes changes only slightly (in Total Variation distance), then the corresponding induced ranking should only change slightly.

The reason for requiring anonymity is self-evident: it rules out discrimination on the basis of the identity of individuals. Stability is more nuanced: it articulates a desire to have rankings which are agnostic to small amounts of noise in the predictions for each individual. In deployed applications, a small amount of variation injected by a seeded training/test data set split or a randomized training procedure can introduce noise at the level of individual predictions (Ganesh et al., 2023). Furthermore, there will also always be at least some additional noise due to incomplete data entries, mistaken inputs, etc. (Nettleton et al., 2010). Rankings should be generally agnostic to these sources of noise: if minute noise in predictions can induce large changes in the derived ranking, the ranking cannot be very meaningful or fair to begin with. Stability can therefore be interpreted as a way to combat micro-*arbitrariness* of rankings induced by learned predictors (Cooper et al., 2024). For stability to be meaningful, we will need to shift our focus to *distributions* over rankings: utilizing randomness to deal with uncertainty will be key in achieving stability.

Anonymity can be construed as a fairness notion, but it is a very minimal one. Fairness in a stronger sense has been the focus of much recent work, both in the context of ranking (see, e.g., Singh & Joachims (2018; 2019)) and in the context of classifiers/predictors (see, e.g., Hardt et al. (2016); Caton & Haas (2020); Dwork et al. (2012); Awasthi et al. (2020)). As ML-based predictors are often used in order to ultimately produce rankings (Wang & Chen, 2012), it is a natural desideratum that the ranking function *preserve* fairness guarantees of the underlying predictor: this ensures that no additional unfairness is introduced in post-processing the classifier's output. As we will see, not all ranking functions satisfy such fairness *composition* properties.

1.1. Our Contributions

We focus on scenarios in which individuals, scored by a predictor, must be presented to a decision maker in a linear order or *ranking*. We assume that the predictions take the form of *distributions over classes*, modeling inherent uncertainty in the underlying ground truth or data. In Section 2, we define a *ranking function* as a map from such probabilistic predictions to a distribution over rankings. Figure 1 illustrates this setting with an example.

Our first (and very immediate — see Section 3.1) result is that stability naturally composes with individual fairness (Dwork et al., 2012): if the predictor is individually fair and the ranking function is stable, then the composition satisfies a natural generalization of individual fairness to rankings. This result further confirms that stability is a desirable property for a ranking function.

In light of the desirability of stability, we next investigate which ranking functions are stable. Deterministic ranking functions are natural and popular; unfortunately, we show (Section 3.2) that the only *stable* deterministic ranking functions are constant, i.e., trivial functions that output the same ranking regardless of the predictions. Further, deterministic ranking functions cannot be anonymous. Thus, one must choose between stability/anonymity and determinism, providing significant evidence in favor of randomization. With randomization, stability and anonymity both become achievable: we show (Section 3.3) that a natural adaptation of the Uncertainty Aware (UA) ranking functions of Singh et al. (2021); Devic et al. (2023) to the case of multiclass predictions of the classifier is indeed anonymous and stable.

We then investigate the fairness guarantees of UA ranking in more depth, and prove (Section 4) our main result: UA ranking naturally preserves multiaccuracy and multicalibration guarantees[1] (Kim et al., 2019; Hébert-Johnson et al., 2018). We show that when the predictor is multiaccurate (or multicalibrated), then the ranking distribution output by UA ranking satisfies a natural generalization of multiaccuracy (resp. multicalibration) towards the same groups. This result can be interpreted as an interpolation between individual and group fairness notions for ranking: as the set of subgroups the predictor is multicalibrated against becomes more refined, the UA ranking for predictions more accurately reflects the UA ranking induced by the unknown ground truth classes of individuals.

To investigate the tradeoff between fairness, stability, and utility, in Section 5, we introduce a standard ranking utility model, and show that the utility optimal ranking function cannot hope to achieve stability or fairness guarantees similar to UA. We also investigate a ranking function which provides a guaranteed tradeoff between stability/fairness and utility. We believe that this will be useful to practitioners interested in employing stable rankings in practice. Finally, in Section 6, we corroborate our theoretical results with experimental evidence.

While various notions of stability in rankings have been

[1] Multiaccuracy requires that the uncertainty estimates of a predictor are *unbiased* over a set of subgroups (combating discrimination between groups); multicalibration guarantees that the estimates are also *calibrated* over subgroups (combating discrimination between *and within* groups). These are arguably the most popular notions of fairness in settings with uncertain predictions where predictors output uncertainty estimates, since obtaining meaningful or accurate estimates at the level of individuals is usually computationally and statistically infeasible.

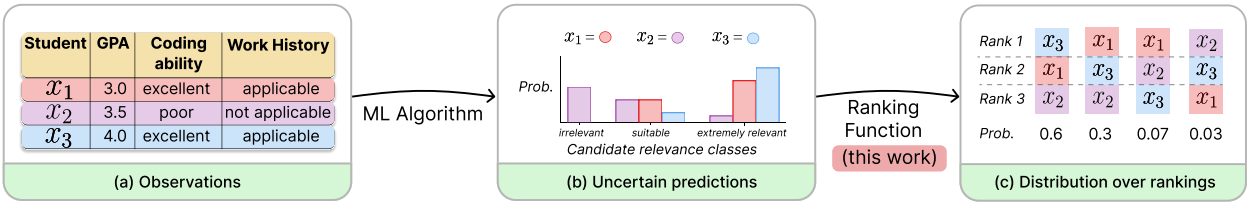

Figure 1: An overview of our setting, using students being ranked by an employer for potential interviews. Observations (a), given by the students' resumés and coding abilities, are fed into a machine learning algorithm which produces distributions (b) over the relevance classes {irrelevant, suitable, extremely relevant} for each candidate. Then, a ranking function *takes as input these uncertain predictions to produce a distribution (c) over rankings of the three candidates. Although it may appear that x_3 is the most qualified or relevant, due to inherent uncertainty in observations, a ranking function may place x_1 or x_2 at rank one with non-zero probability (c).*

proposed before (see, e.g., Asudeh et al. (2018)), our framework is unique in that it frames the rankings as induced by *predictions* of some machine learning algorithm — this ties our work more closely to modern applications. Another benefit of our definition of stability is that it makes progress towards the broader goal of rankings which compose with fair predictors.

1.2. Related Work

We defer a full discussion of related work to Appendix A.

Fairness in Ranking. By far the most relevant related work is of Dwork et al. (2019), who are also interested in fair rankings induced by predictors, but importantly restrict their focus to only deterministic rankings. Indeed, their motivating example is a setting in which small perturbations to a predictor can massively impact an induced ranking. By requiring stability of ranking functions, we approach this problem fundamentally differently: we allow (and indeed, require) non-deterministic rankings. The multiaccuracy and multicalibration guarantees of Dwork et al. (2019) for induced rankings from predictors are similar in flavor to ours; however, a fundamental difference is that we show this guarantee to be compatible with stability, and, furthermore, that our guarantees hold for each position k in the ranking.

At the intersection of group and individually fair rankings, the work of Gorantla et al. (2023) is most similar to ours. They show that one can sample from a distribution over rankings which is simultaneously individually and group fair (in a proportional representation sense) for laminar groups. In contrast, our group fairness hinges on the group-level statistical constraints of multicalibration imposed on the underlying predictor, which instead allow for potentially arbitrarily overlapping groups. For a comprehensive overview on group/individual fairness in rankings, the interested reader is referred to the survey of Zehlike et al. (2021).

Stability in Ranking. In the information retrieval literature, Asudeh et al. (2018) also study the notion of stability for rankings. They work in a setting where a *score* is calculated based on a weighted sum of the features of each item, and stability is then with respect to small changes of these weights. They furthermore state that stability is not a property of their "scoring function" (a particular weighted sum over features). In contrast, we explicitly define stability more generally as a property of our ranking function, which maps from any set of predictions (data set) to a randomized ranking. Oh et al. (2022); Bruch et al. (2020) also study the sensitivity of rankings empirically, providing experimental evidence that randomization can help stability (which they define as *robustness*) during the training of learning-to-rank models. Our theoretical and experimental results are complementary and corroborate this empirical evidence.

Uncertainty and Fairness. Rastogi & Joachims (2023); Mehrotra & Celis (2021); Mehrotra & Vishnoi (2022); Tahir et al. (2023); Guo et al. (2023) all investigate fairness in ranking and decision making under uncertainty in predictions or sensitive attributes. We work with unbiased uncertainty, and leave biased uncertainty for future investigation. Independently of the line of work on UA rankings, Shen et al. (2023) propose ranked proportionality, which shares a similar definition. Their work is in the more general setting of the assignment problem with uncertain priorities, and they focus on algorithmic approaches for achieving a variety of fairness notions simultaneously. Our work is instead focused on proving certain *properties* of rankings induced by predictors (predictors which, when stated in the language of Shen et al. (2023), may induce uncertain priorities).

Calibration and Ranking. In Section 4, we work with (multi)-calibrated predictors. The ranking community has investigated the impact of calibration of ranking models on diversity, utility, and fairness (Menon et al., 2012; Kweon et al., 2022; Yan et al., 2022; Busa-Fekete et al., 2011; DiCiccio et al., 2023). These works all attempt to infer uncertainty from the *scoring function* within a *score-and-sort* ranking model, whereas we assume that uncertainty is given in the form of machine-learned predictions.

2. Notation and Preliminaries

We write vectors in boldface. We use the standard notation $\boldsymbol{x}_{-i}$ to denote the vector $\boldsymbol{x}$ with the ith coordinate removed. For a random event $\mathcal{E}$, we write $\mathbb{1}_{\mathcal{E}}$ for the indicator function which is 1 if $\mathcal{E}$ happens, and 0 otherwise. The total variation distance of two measures μ, ν is defined as the maximum difference in probability for any event $\mathcal{E}$ under the two measures, i.e., $d_{\text{TV}}(\mu, \nu) := \max_{\mathcal{E}} |\mu(\mathcal{E}) - \nu(\mathcal{E})| = \frac{1}{2}||\mu - \nu||_1$. We will use the *entry-wise* matrix norms[2] $||M||_1 = \sum_{i,j} |M_{i,j}|$ and $||M||_\infty = \max_{i,j} |M_{i,j}|$.

$\mathcal{X}$ denotes a set of *individuals*; it contains humans, ads, service requests, or other entities towards whom fairness is desired. The elements of $\mathcal{X}$ can be labeled with labels from the finite[3] label set $[L] = \{1, 2, \ldots, L\}$. We work in the multiclass "ordinal" classification setting where the labels are sorted from most to least preferred as $L \succ L-1 \succ \cdots \succ 2 \succ 1$. This notation corresponds with the intuition that possessing a higher merit score/class is valued more by a decision maker. A common special case is $L = 2$, i.e., binary labels, where label 1 might represent irrelevant/unsuitable, while label 2 represents relevant/suitable.

2.1. Predictors

We focus on predictors in the multiclass setting which output distributions over L labels. Let Δ_L denote the set of all distributions on $[L]$. A *(probabilistic) predictor* $f : \mathcal{X} \to \Delta_L$ is a function mapping data points to distributions over labels. We let $\boldsymbol{p} = f(x)$ denote the vector of probabilities that the predictor f assigns to individual $x \in \mathcal{X}$. For any class $\ell \in [L]$, p_ℓ denotes the probability of that class. As an example, for a probabilistic binary predictor f in the context of determining whether a candidate is qualified for a job, p_2 would capture the probability that the individual x is qualified, while $p_1 = 1 - p_2$ is the probability that x is unqualified.

Rankings involve multiple individuals, and hence multiple predictions. A *prediction for n individuals* P is an $n \times L$ matrix where each row corresponds to the distribution over labels for a particular individual. We define $\mathcal{P}_{n,L}$ to be the set of all predictions for n individuals, i.e., the set of all $n \times L$ matrices where each row is a distribution. We will frequently consider the case in which a predictor f for single individuals is applied to each of n individuals separately. For a vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ of n individuals, we write $f(\boldsymbol{x}) = (f(x_1), \ldots, f(x_n))$ for the $n \times L$ matrix of predictions for all of the n individuals.

We use the random variable λ_x to denote the (random) label of individual x; when we specifically consider an individual x_i in a vector of individuals, we abbreviate $\lambda_i := \lambda_{x_i}$. We write $N^\ell = \sum_i \mathbb{1}_{\lambda_i = \ell}$ for the random variable that is the number of individuals with label ℓ; when we use this notation, the domain of i will be clear from the context. We extend this notation to write $N^{\geq \ell} = \sum_{\ell' = \ell}^{L} N^{\ell'}$ for the number of individuals with label ℓ or better, and similarly for $N^{>\ell}$. We will sometimes restrict the count to individuals in a particular set S, and then write $N_S^\ell = \sum_{i \in S} \mathbb{1}_{\lambda_i = \ell}$, and similarly for the derived notation. In particular, we use the notation $N_{-i}^\ell = N_{[n] \setminus \{i\}}^\ell$ for the number of individuals other than i with a particular label ℓ.

2.2. Rankings and Ranking Functions

A principal would like to use predictions provided by a predictor to output a (distribution over) rankings. As examples, consider a site or service such as LinkedIn providing an employer with a ranked list of applicants to interview (Geyik et al., 2019), or an online platform deciding on the order in which to display ads or vendors to a visitor. In these settings, because *attention* is a limited resource, a common approach would have the principal *rank* the items in question based on some function of the predictions.

A *ranking* is a total order on n individuals. A *randomized* ranking is a distribution over rankings. Let $\mathcal{M}_{\text{DS}}^{n \times n}$ denote the set of all $n \times n$ doubly stochastic matrices. Each matrix $M \in \mathcal{M}_{\text{DS}}^{n \times n}$ represents a *randomized ranking*[4] over n individuals, where $M_{i,k}$ is the probability with which individual i is ranked in position k. When reasoning about random rankings, we use $\mathcal{R}_{i,k}$ to denote the random event that individual i receives position k in the ranking.

We refer to mappings from predictions to (randomized) rankings as ranking functions:

Definition 1. *A* ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n \times n}$ *maps predictions over L labels on a data set of n individuals to a randomized ranking of those individuals.*

By focusing on ranking functions, we implicitly state that the principal interacts with the data set *only* through the

[2] rather than induced norms, which are typically described by the same notation.

[3] In Appendix C, we extend our definitions and results to continuous ranking functions. Continuous labels allow us to capture, for example, the well-known Plackett-Luce (Luce, 1959; Plackett, 1975), Bradley-Terry (Bradley & Terry, 1952), and Thurstonian ranking models (Thurstone, 1994), and establish stability results for them. These models have been used with the goal of achieving stability in practice (Bruch et al., 2020), and they also have applications within preference learning more broadly (Zhu et al., 2023; Wilde et al., 2021).

[4] More precisely, it represents the *marginal probabilities* of the distribution, which can typically be implemented by many different distributions over rankings. In particular, a distribution supported on at most n^2 rankings can be found in polynomial time using the well-known Birkhoff-von Neumann result (Birkhoff, 1946). We assume that individuals care only about the probabilities with which they are ranked in each position, in which case marginal distributions sufficiently capture fairness.

predictions over labels. That is, we do not consider *listwise* learning-to-rank schemes — such as Cao et al. (2007); Xu & Li (2007) — in which the principal directly learns a function mapping data sets of individuals' features to rankings.

2.3. Desiderata of Ranking Functions

While ranking functions could be very general, there are natural requirements that make them "reasonable" to be used. In particular, we focus on the following basic properties.

Definition 2 (Anonymity)**.** *A ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is* anonymous *if every permutation $\sigma : [n] \to [n]$ of the predictions for individuals results in an identical permutation of the individuals' ranks.*

Anonymity states that the outcome for an individual depends only on their (and everyone else's) prediction, but not on the index at which the individual appeared in the data set, i.e., on their identity. As such, it is an essential fairness requirement in virtually all settings.

A second essential property of ranking functions is *stability*: that small changes in the predictions only lead to small changes in the rankings.

Definition 3 (Stability)**.** *Fix n and L. A ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is γ-*stable *if $||r(P) - r(P')||_\infty \leq \gamma \cdot ||P - P'||_1$ for all predictions $P, P' \in \mathcal{P}_{n,L}$.*

Stability is particularly important when the predictions are the result of ML-based training methods, which will always contain non-trivial amounts of noise. Indeed, the lack of stability is a well-documented and problematic aspect of many ML-systems, and has been shown not only within the fairness literature (Cooper et al., 2024), but has long been a concern for image classification models (Goodfellow et al., 2015) and more recently also LLMs (Zou et al., 2023).

Remark 4. *Our choices of the ∞-norm and 1-norm in Definition 3 are motivated in part by the result from Theorem 11 that UA ranking is 1-stable. Using standard inequalities between norms, the 1-stability of UA with respect to our choices of norms implies weaker results for the case when both sides use the $||\cdot||_1$ norm, or both use the $||\cdot||_\infty$ norm. The resulting stability guarantees are unfortunately the best that can be obtained: it is simple to modify the example given below in Proposition 12 to show that matching lower bounds hold, so no stronger guarantees can be obtained.*

3. Predictions and Rankings

We first show useful fairness consequences of stability: combining a stable ranking function with an individually fair predictor results in fair ranking outcomes. We then show that stability and anonymity are fundamentally at odds with determinism: only constant deterministic ranking functions are stable, and no deterministic ranking function is anonymous. This establishes that randomization is inherently necessary for a ranking function to be meaningful, anonymous, and stable. We then present our adaptation of the UA ranking function of Singh et al. (2021), and show that it is anonymous and stable. Deferred proofs appear in Appendix B.

3.1. Consequences of Stability

Stability implies that small changes in predictions do not change the distribution over rankings much. This has two immediate but noteworthy consequences: (1) if the predictions are made by an *individually fair* predictor, then similar individuals will be ranked similarly, and (2) as the predictions approach ground truth, the ranking distribution produced by the ranking function approaches the rankings under the ground truth.

To formalize the first claim, we recall the seminal definition of an *individually fair* predictor (Dwork et al., 2012). This notion assumes a metric d defined on $\mathcal{X}$ capturing a relevant measure of *similarity* between individuals. For $\beta > 0$, a probabilistic predictor f is (β, d)-*individually fair* if $||f(x) - f(x')||_1 \leq \beta \cdot d(x, x')$ for all $x, x' \in \mathcal{X}$.

Proposition 5. *Let $f : \mathcal{X} \to \Delta_L$ be a (β, d)-individually fair predictor, and $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ an anonymous and γ-stable ranking function. Given a data set of individuals $(x_i)_{i\in[n]}$ and their associated predictions $P = (f(x_i))_{i\in[n]} \in \mathcal{P}_{n,L}$, let $\boldsymbol{q}, \boldsymbol{q}'$ be the ith and j^{th} rows of $r(P)$, respectively. Then, $\|\boldsymbol{q} - \boldsymbol{q}'\|_\infty \leq (2\beta\gamma) \cdot d(x_i, x_j)$.*

Proof. The proof is a straightforward application of γ-stability with respect to the given prediction matrix P and a matrix P', where P' is exactly P but with rows i and j swapped (requiring the anonymity condition). This, combined with the definition of (β, d)-individual fairness for i, j, completes the proof. □

The result can be interpreted as a composition guarantee for anonymous and stable rankings with individually fair predictors: if x, x' are simultaneously in a data set, the difference in their distributions over rankings can be at most proportional to their dissimilarity under the metric d. Another interpretation is the following: Stability and individual fairness are both Lipschitz conditions, and composition of Lipschitzness implies that an individually fair predictor combined with a stable ranking will induce an individually fair ranking.

Another very straightforward but desirable consequence of stability is obtained by considering one prediction to be ground truth and the other obtained from a learned classifier.

Corollary 6. *Let $f^* : \mathcal{X} \to \Delta_L$ be the ground truth label distribution for individual x, and $f : \mathcal{X} \to \Delta_L$ the*

learned predictor. Assume that f is ϵ-accurate, satisfying that $\|f(x) - f^*(x)\|_1 \leq \epsilon$ for all $x \in \mathcal{X}$. Then, any γ-stable ranking function r guarantees that $\|r(f(\boldsymbol{x})) - r(f^*(\boldsymbol{x}))\|_\infty \leq \gamma \cdot n\epsilon$ for all $\boldsymbol{x} \in \mathcal{X}^n$.

Put differently, for any stable ranking function, accurate individual level uncertainty estimates (relative to a ground truth f^*) will induce accurate individual level rankings. Although somewhat obvious, we highlight this property of stability since the "ground truth" approach is often a central assumption in the study of machine learned predictors (Shalev-Shwartz & Ben-David, 2014).

3.2. Stability and Determinism are Incompatible

A third property which most rankings used in practice possess, and which is often considered desirable by practitioners, is determinism: that for given inputs, only one ranking (rather than a distribution over rankings) can result.

Definition 7 (Determinism)**.** *A ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is* deterministic *iff for all $P \in \mathcal{P}_{n,L}$, the resulting distribution over rankings $r(P)$ has only entries in $\{0, 1\}$.*

Perhaps the most well-known deterministic ranking function is given by the Probability Ranking Principle (PRP) of Robertson (1977). In the setting with binary predictions, this ranking function sorts individuals by decreasing probability of belonging to class 2, i.e., being qualified.

Naturally, one may ask whether a deterministic ranking function like the PRP can be stable or anonymous. Unfortunately, neither is possible, as captured by the following.

Proposition 8. *No deterministic ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is anonymous. Furthermore, for any $\gamma > 0$, any deterministic and γ-stable ranking function must be* constant*, in the sense that $|\operatorname{Im}(r)| = 1$, i.e., the ranking function outputs the same ranking for all input predictions.*

Proposition 8 formalizes the intuition that randomness is required to achieve stability. Indeed, the main results of our work also show that randomization and the resulting stability are crucial for achieving desirable *fairness* guarantees.

3.3. Uncertainty Aware Rankings

Meaningful deterministic ranking functions cannot be stable; in fact, it is not immediate that there exist (non-constant) stable ranking functions. We now show that *Uncertainty Aware (UA) Rankings*, introduced by Singh et al. (2021), are anonymous and stable.

UA rankings were originally introduced via an axiomatization of when a probabilistic ranking should be considered "fair" for given merit distributions. Devic et al. (2023) further refined this axiomatization by combining notions of meritocracy and *lifting* deterministic decision making to decision making under uncertainty.

The definition of Singh et al. (2021) assumed that merit distributions were continuous and ties occurred with probability 0. Motivated by predictors which output distributions over discrete label sets such as 1–5 or {irrelevant, suitable, extremely relevant} with a corresponding total order, we adapt the definition of UA rankings (though, as mentioned previously, in Appendix C, we also generalize all of our notions to continuous labels):

Definition 9 (Uncertainty Awareness (Singh et al., 2021))**.** *A randomized ranking $M \in \mathcal{M}_{DS}^{n\times n}$ is* uncertainty aware *for a prediction $P \in \mathcal{P}_{n,L}$ if for each individual i and position k, the entry $M_{i,k}$ is the probability that i has the kth highest label if all labels $\lambda_i \sim \boldsymbol{p}_i$ are sampled independently from the respective distributions $\boldsymbol{p}_i \in \Delta_L$, and ties are broken uniformly. Formally, conditioned on the drawn labels λ_i of all individuals i, which entail the counts N^ℓ for all labels, the probability for individual i to obtain rank k is*

$$\mathbb{P}[\mathcal{R}_{i,k} \mid \lambda_i = \ell, N^1, \ldots, N^L] = \frac{1}{N^\ell} \mathbb{1}_{N^{>\ell} < k \leq N^{\geq \ell}}. \quad (1)$$

A ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is uncertainty aware if $r(P)$ is uncertainty aware for all $P \in \mathcal{P}_{n,L}$.

Because the definition of uncertainty awareness fully prescribes the ranking distribution M for a given prediction P (as shown in Lemma 4.2 of Singh et al. (2021)), there is a unique uncertainty aware ranking function r for any given n, L; we henceforth denote it by r_{UA}.

Intuitively, the fairness of UA can be interpreted through a *possible futures* viewpoint. Given two individuals A, B, if A has more merit than B in 60% of futures (when the merits/labels of both A and B are sampled from their respective distributions), then UA implements the requirement that the allocation in the present should respect this uncertainty and give A the better rank at least 60% of the time (and B at least 40% of the time); this entails the need for randomization. We refer the reader to Devic et al. (2023); Singh et al. (2021) for a formal argument on the fairness of UA ranking.

We also remark that the definition of Singh et al. (2021) did not require the labels/merits of different individuals to be sampled *independently* from their respective distributions P_i. We add this independence requirement to facilitate connections with learning algorithms for predictors. An added benefit of the independence assumption is that it makes it possible to explicitly compute $r_{\text{UA}}(P)$ in polynomial time, as captured by the following proposition. Note that this is in contrast to the case of possibly correlated labels/merits from previous work (Singh et al., 2021; Devic et al., 2023); indeed, a main technical contribution of these works was

analyzing the loss in fairness/utility incurred due to imperfectly approximating $r_{\text{UA}}(P)$ via sampling.

Proposition 10. *There exists an algorithm which, given $P \in \mathcal{P}_{n,L}$, exactly computes $r_{UA}(P)$ in time $O(n^4 + n^3 L)$.*

While the notion and use of uncertainty aware rankings may appear to be of primarily theoretical interest, it is in fact used in practice. For example, the NBA draft lottery can be understood through the lens of uncertainty aware rankings. The *merit* of a team is its need for better choice picks, which can be (imperfectly) inferred from the team's performance in the previous season. The draft order is then obtained by a weighted lottery based on these uncertain merits.

We now present the central result of this section: that the uncertainty aware ranking function is anonymous and stable.

Theorem 11. *Let $r_{UA} : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ be the UA ranking function for n individuals and L labels. r_{UA} is anonymous and 1-stable.*

Proof Sketch. That UA ranking is anonymous follows from the definition since it treats all individuals based only on their predictions, and not based on their position in the prediction matrix. The difficult part is proving that UA ranking is stable: this is because the ranking of any given individual depends not only on their own prediction, but the predictions for the other $n-1$ individuals. Lemma 19 is crucial in showing stability: it states that any individual — conditioned on that individual having a given label — will have a similar ranking under two different prediction matrices P and Q, where the difference in the rankings is bounded by the total variation distance between the distributions for all other individuals under P and Q. With Lemma 19 in hand, simply applying the law of total conditional expectation will let us prove Theorem 11. The difficulty lies in proving Lemma 19.

To prove Lemma 19, we use an insight into the conditional decomposition of the UA ranking probability for a given individual i at rank k (Proposition 18). This proposition allows us to consider the rank of individual i by considering the rank of i conditioned on i having a particular label. The proof of Proposition 18 is simply a counting argument.

With this conditional decomposition in hand, we are able to apply a coupling argument for the event that the individual i in question obtains a particular rank when other individuals' classes are sampled with respect to either P or Q. In particular, we can say that under P and Q, the labels of every individual except i will behave similarly if the total variation distance is small. □

Our stability analysis is essentially tight up to a factor of 2, ruling out the possibility of, for example, $\frac{1}{n}$-stability:

Proposition 12. *For any $L \geq 3$ and $n \geq 2$, r_{UA} is not γ-stable for any $\gamma < \frac{1}{2}$.*

Finally, as the number of labels increases, the predictor can provide the ranking function with more fine-grained information, which should allow the ranking function to produce a wider class of distributions over rankings. The next proposition, stated informally here and formally as Proposition 21, shows that this is indeed the case:

Proposition 13. *The expressivity of UA ranking functions is strictly increasing in L.*

4. Multigroup Fairness Guarantees

We now present our main result: UA rankings naturally compose with multiaccurate and multicalibrated predictors.

We have shown that UA rankings are stable, and, furthermore, that stable rankings compose harmoniously with individually fair predictors. Corollary 6 demonstrates that an accurate predictor, at the individual level, can combine with a stable ranking function (such as UA) to induce a ranking which is close to that of the underlying ground truth. In practice, however, obtaining accurate uncertainty estimates at the individual level is too strong of an assumption for arbitrary data sets of individuals. This is because such a requirement is equivalent to learning the *Bayes optimal predictor* (the true distribution over the labels conditioned on the features of an individual) (Shalev-Shwartz & Ben-David, 2014, Section 3.2.1), which generally requires the number of samples or running time to be exponential in the dimensionality of the features used for prediction, which can be statistically or computationally infeasible.

Instead, we focus on obtaining a coarser guarantee for UA rankings at the level of *subgroups* of the domain for data sets *sampled* i.i.d. from a distribution $\mathcal{D}$ over individuals (instead of arbitrary data sets). The i.i.d. assumption is a standard setting for machine learning, and has proven useful in many practical settings. Our contributions are tightly connected to the statistical group-fairness conditions of *multiaccuracy* and *multicalibration* (Kim et al., 2019; Hébert-Johnson et al., 2018). Our guarantees will be meaningful since they directly imply that, relative to an underlying ground truth, unbiased predictors will induce unbiased rankings.

4.1. Group-wise Accuracy Guarantees

We first recall the definitions of multiaccuracy and multicalibration from the fair machine learning literature. Then, we state our result on the average accuracy of rankings induced by multiaccurate and multicalibrated predictors when compared to rankings induced from nature.

Definition 14 (Multiaccuracy/Multicalibration (Kim et al., 2019; Hébert-Johnson et al., 2018))**.** *Let $\mathcal{D}$ be a distribution*

over individuals $\mathcal{X}$. Let f^* be the ground truth distribution of labels, i.e., $f^*(x)$ is the true distribution of labels for individual x, while f is the predictor, so $f(x)$ is the predicted label distribution.

Let $\mathcal{C}$ be a collection of sets for which the predictor is to be multiaccurate/multicalibrated. Let $\alpha \geq 0$ be a parameter for how far from fully accurate/calibrated the predictor is allowed to be. When writing $\mathbb{E}[\boldsymbol{v}]$ for a vector-valued quantity $\boldsymbol{v}$, we mean the coordinate-wise expectations. Let $\Delta_{f,g}(x) = f(x) - g(x)$ denote the difference in predictions on datapoint x between functions f, g.

*(1) f is $(\mathcal{C}, \alpha)$-*multiaccurate *if for every set $S \in \mathcal{C}$,*

$$\|\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{x\in S} \cdot \Delta_{f^*,f}(x)\right]\|_\infty \leq \alpha.$$

That is, for each of the given groups $S \in \mathcal{C}$ and each label, the expected probability mass on that label is approximately the same for the predictor as for the ground truth.

*(2) Let some interval width $\delta \in (0, 1]$ be given, such that $1/\delta$ is an integer. f is $(\mathcal{C}, \alpha, \delta)$-*multicalibrated *if for every set $S \in \mathcal{C}$ and vector $(j_1, j_2, \ldots, j_L) \in \{0, 1, \ldots, 1/\delta - 1\}^L$,*

$$\|\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{x\in S} \cdot \mathbb{1}_{\forall \ell,\ f(x)_\ell \in [j_\ell\cdot\delta,(j_\ell+1)\cdot\delta)} \cdot \Delta_{f^*,f}(x)\right]\|_\infty \leq \alpha.$$

That is, in addition to fixing a group, even if we also fix a (rough) interval within which the predicted probability mass must lie, the predictor still has to be close to the ground truth, for each possible label.

Each set $S \in \mathcal{C}$ represents a protected group of import in the underlying population $\mathcal{X}$. The sets in $\mathcal{C}$ can be complex, overlapping, nested, laminar, etc., since both multiaccuracy and multicalibration with respect to $\mathcal{C}$ are — in contrast to other notions of group fairness in supervised learning such as equalized odds (Awasthi et al., 2020) or equality of opportunity (Hardt et al., 2016) — statistically sound in that they are *consistent* with the underlying ground truth f^*. There are a variety of learning and post-processing algorithms which achieve multiaccuracy/ multicalibration (Hébert-Johnson et al., 2018; Kim et al., 2019; Gopalan et al., 2022; Haghtalab et al., 2023).

We now present the central contribution of this section: multicalibration and multiaccuracy for a predictor f guarantee that on a per group basis, UA rankings derived from f will be close to UA rankings derived from the ground truth f^*.

Theorem 15. *Let $\mathcal{D}$ be a distribution over individuals $\mathcal{X}$. Let f^* be the ground truth distribution of labels, and f a predictor. For any n, let $\mathcal{D}^n$ be the distribution obtained from drawing a vector of n i.i.d. samples from $\mathcal{D}$.*

Let $\mathcal{C}$ be a collection of sets with $\mathcal{X} \in \mathcal{C}$, the sets of individuals for which the predictor f will be assumed to be multiaccurate/multcalibrated. Let $\alpha \geq 0$ be a parameter for how far from fully accurate/calibrated the predictor is allowed to be. Define $\Delta^{i,k}_{f,g}(\boldsymbol{x}) = \mathbb{P}_{r_{UA}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{UA}(g(\boldsymbol{x}))}[\mathcal{R}_{i,k}]$ as the deviation of UA ranking on the probabilities of event $\mathcal{R}_{i,k}$ under the predictions given by predictors f, g on dataset $\boldsymbol{x}$. Then, we have that:

(1) If f is $(\mathcal{C}, \alpha)$-multiaccurate, then the following holds for all sets $S \in \mathcal{C}$ and $k \in [n]$:

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim Unif([n])}\left[\mathbb{1}_{x_i\in S} \cdot \Delta^{i,k}_{f^*,f}(\boldsymbol{x})\right]\right| \leq Ln\alpha.$$

(2) If f is $(\mathcal{C}, \alpha, \delta)$-multicalibrated and $(\{\mathcal{X}\}, \alpha)$-multiaccurate, then for every set $S \in \mathcal{C}$, vector $(j_1, j_2, \ldots, j_L) \in \{0, 1, \ldots, 1/\delta - 1\}^L$, and $k \in [n]$:

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim Unif([n])}\left[\mathbb{1}_{x_i\in S} \cdot \mathbb{1}_{f(x_i)_\ell \in [j_\ell\cdot\delta,(j_\ell+1)\cdot\delta)\ for\ all\ \ell} \cdot \Delta^{i,k}_{f^*,f}(\boldsymbol{x})\right]\right| \leq Ln\alpha.$$

Proof Sketch. The proofs for both the multiaccuracy and multicalibration parts of the theorem are essentially identical. The key insight is that when all individuals are sampled from an underlying distribution $\mathcal{D}$, the distribution of competing individuals that i faces is the same as the aggregate distribution in the population under $\mathcal{D}$. By the multiaccuracy / multicalibration assumptions over the entire population, this distribution is sufficiently well estimated by the predictor f. This analysis applies conditioned on the particular label of individual i. It is then combined with the conditional decomposition of Proposition 18 and the total variation bound of Lemma 19. □

Theorem 15 can intuitively be thought of as the following: a predictor which is unbiased on average over a collection of subgroups $\mathcal{C}$ will induce an uncertainty aware ranking which, for those subgroups, has a similar outcome to the (usually inaccessible) uncertainty aware ranking induced by the ground truth label distribution. The multicalibration guarantee refines this to hold for not only subgroups, but intervals of predictions of the predictor f within that subgroup. Part (2) of Theorem 15 requires f be simultaneously $(\mathcal{C}, \alpha, \delta)$-multicalibrated and $(\{\mathcal{X}\}, \alpha)$-multiaccurate; that is, f is unbiased on average across individuals sampled from $\mathcal{D}$. This combination of properties can be achieved by, e.g., the algorithm of Gopalan et al. (2022). In Appendix D, we detail further computational/statistical considerations, and also demonstrate how Theorem 15 can be seen as an interpolation result between individual and group fairness.

5. Ranking Functions and Utility

Most online marketplaces utilizing rankings and ranking functions are also concerned with utility or revenue. In

this section, we introduce a natural class of utility models inspired by the literature standard (Taylor et al., 2008), and prove that the optimal utility ranking function cannot achieve the stability or group fairness guarantees that UA rankings enjoy.

Definition 16 (Utility Model)**.** *Let $w_1 \geq w_2 \geq \cdots \geq w_n \geq 0$ be position weights, and $\tau : \Delta_L \to \mathbb{R}_{\geq 0}$ a function we call the* class utility map*, which determines how predictions are mapped to utilities. The utility $U(P, r)$ of a prediction matrix $P \in \mathcal{P}_{n,L}$ under a ranking function r is*

$$\mathbb{E}_{\sigma \sim r(P)} \sum_{k=1}^{n} w_k \cdot \tau(\boldsymbol{p}_{\sigma(k)}) = \sum_{i=1}^{n} \sum_{k=1}^{n} \left[r(P)_{i,k} \cdot w_k \cdot \tau(\boldsymbol{p}_i) \right].$$

A particularly natural and common type of class utility map is the expected utility $\tau(\boldsymbol{p}) = \sum_{\ell=1}^{L} v_\ell \cdot p_\ell$, where $v_L > v_{L-1} > \cdots > v_1 \geq 0$ are the utilities for labels $\ell \in [L]$. Combined with the weights $w_k = 1/\log_2(1+k)$, this class captures DCG (Järvelin & Kekäläinen, 2002), for example.

The ranking function which achieves optimal utility will depend on τ; we denote it r^{τ}_{opt}. In Appendix E, we show strong impossibility results as Propositions 27 and 28, summarized informally here:

Proposition 17. *Even for binary labels and expected utility, r^{τ}_{opt} is unstable, and cannot satisfy multiaccuracy fairness guarantees akin to Theorem 15.*

Since it is typically desirable to obtain provable guarantees quantifying the tradeoff between fairness and utility, in Appendix F we introduce a relaxation of stability/fairness akin to a definition from Singh et al. (2021). We show that a ranking function r^{ϕ}_{mix} which randomizes between r^{τ}_{opt} and r_{UA} with parameter $\phi \in [0, 1]$ achieves provable guarantees for both utility and approximate stability/fairness.

6. Experiments on Stability and Utility

We run experiments to complement our theoretical results and further investigate the fairness-utility tradeoff. Our results demonstrate that r_{UA} is far more stable than r^{τ}_{opt} in practice, and also achieves higher utility than two baseline ranking functions. We use the US Census data set `ACS` (Ding et al., 2021) and the student dropout task `Enrollment` (Martins et al., 2021). Table 1 shows the stability of r_{UA} and r^{τ}_{opt} to noise introduced by neural networks trained with SGD, averaged over multiple runs. In Table 2, we report the utility of r_{UA}, the uniform ranking r_{unif} assigning each individual to each rank with equal probability, and Plackett-Luce rankings r_{PL} (Plackett, 1975; Luce, 1959). The utility is normalized such that r^{τ}_{opt} always achieves utility 1. Further experimental details are deferred to Appendix G.

Quantity	`ACS`	`Enrollment`
$\|r_{\text{UA}}(P) - r_{\text{UA}}(P')\|_\infty$	0.011 ± 0.002	0.012 ± 0.002
$\|r^{\tau}_{\text{opt}}(P) - r^{\tau}_{\text{opt}}(P')\|_\infty$	0.947 ± 0.224	0.680 ± 0.466
$\|P - P'\|_1$	0.971 ± 0.582	0.653 ± 0.453

Table 1: Measured stability over 30 neural network training runs (15 pairs of networks) for 10 data sets of $n = 30$ individuals each. ∞-norm of UA deviation being bounded above by $\|P - P'\|_1$ confirms stability of UA (Theorem 11). Instability of the ranking r^{τ}_{opt} is also demonstrated (Proposition 17).

	$n = 20$	$n = 40$	$n = 60$
r_{UA}	$\mathbf{0.726} \pm 0.027$	$\mathbf{0.724} \pm 0.027$	$\mathbf{0.727} \pm 0.020$
r_{PL}	0.616 ± 0.038	0.621 ± 0.028	0.624 ± 0.023
r_{unif}	0.540 ± 0.043	0.548 ± 0.029	0.550 ± 0.030
r_{UA}	$\mathbf{0.852} \pm 0.030$	$\mathbf{0.860} \pm 0.023$	$\mathbf{0.857} \pm 0.018$
r_{PL}	0.755 ± 0.041	0.767 ± 0.027	0.767 ± 0.025
r_{unif}	0.552 ± 0.052	0.561 ± 0.037	0.562 ± 0.033

Table 2: Normalized utility achieved by r_{UA}, r_{unif}, and r_{PL} for $n = 20, 40,$ and 60 random individuals from the test set of `ACS` (top 3 rows) and `Enrollment` (bottom 3 rows). Mean/std taken across 30 neural network training runs. UA outperforms the uniform and PL ranking.

7. Conclusion

Stability of ranking functions is a natural desideratum to prevent large deviations arising in rankings from noise in learned classifiers; combined with individually fair predictions, it results in fair rankings. Stability is achieved by the natural Uncertainty Aware Ranking Functions, which also preserve group fairness guarantees of their underlying classifiers. An interesting direction for future work is to find a general sufficient condition for ranking functions which implies this preservation of group fairness.

Another important extension is to consider correlations between the labels of different individuals, and whether analogous individual/group fairness guarantees can still be provided in this case. This would apply to both the actual prediction/ranking time and to the implications on the learned predictors when they are learned from correlated data.

Impact Statement

The work advances understanding of fairness when using ranking functions. The societal consequences of the work are shared with other work on improving algorithmic fairness; none need to be specifically highlighted here.

Acknowledgements

SD was supported by the Department of Defense (DOD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. This work was also funded in part by NSF awards #1916153, #2333448, #1956435, #1943584, #2344925, and #2239265, and an Amazon Research Award 2023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Amazon.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Preprint, 2021. 1

Asudeh, A., Jagadish, H. V., Miklau, G., and Stoyanovich, J. On obtaining stable rankings. In *Proc. 44th Intl. Conf. on Very Large Data Bases*, volume 12, pp. 237–250, 2018. 3, 15

Awasthi, P., Kleindessner, M., and Morgenstern, J. Equalized odds postprocessing under imperfect group information. In *Proc. 23rd Intl. Conf. on Artificial Intelligence and Statistics*, pp. 1770–1780. PMLR, 2020. 2, 8

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Practical adversarial multivalid conformal prediction. In *Proc. 36th Advances in Neural Information Processing Systems*, 2022. 1

Birkhoff, G. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946. 4

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4

Bruch, S., Han, S., Bendersky, M., and Najork, M. A stochastic treatment of learning to rank scoring functions. In *Proc. 13th ACM Intl. Conf. on Web Search and Data Mining*, pp. 61–69, 2020. 3, 4, 15, 21, 26

Busa-Fekete, R., Kégl, B., Éltetö, T., and Szarvas, G. Ranking by calibrated adaboost. In *Proceedings of the Learning to Rank Challenge*, pp. 37–48. PMLR, 2011. 3, 15

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proc. 24th Intl. Conf. on Machine Learning*, pp. 129–136, 2007. 5

Caton, S. and Haas, C. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020. 2

Cohen, D., Mitra, B., Lesota, O., Rekabsaz, N., and Eickhoff, C. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *Proc. 44th Intl. Conf. on Research and Development in Information Retrieval (SIGIR)*, pp. 654–664, 2021. 14

Cooper, A. F., Lee, K., Barocas, S., De Sa, C., Sen, S., and Zhang, B. Is my prediction arbitrary? Measuring self-consistency in fair classification. In *Proc. 38th AAAI Conf. on Artificial Intelligence*, 2024. 2, 5

Devic, S., Kempe, D., Sharan, V., and Korolova, A. Fairness in matching under uncertainty. In *Proc. 40th Intl. Conf. on Machine Learning*, volume 202, pp. 7775–7794, 2023. 2, 6, 14, 25

DiCiccio, C., Hsu, B., Yu, Y., Nandy, P., and Basu, K. Detection and mitigation of algorithmic bias via predictive parity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1801–1816, 2023. 3, 15

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In *Proc. 35th Advances in Neural Information Processing Systems*, 2021. 9, 25

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proc. 3rd Innovations in Theoretical Computer Science*, pp. 214–226, 2012. 1, 2, 5

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Learning from outcomes: Evidence-based rankings. In *Proc. 60th IEEE Symp. on Foundations of Computer Science*, pp. 106–125. IEEE, 2019. 3, 14

Ganesh, P., Chang, H., Strobel, M., and Shokri, R. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1789–1800, 2023. 2

García-Soriano, D. and Bonchi, F. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 436–446, 2021. 14

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proc. 25th Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 2221–2231, 2019. 1, 4

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, 2015. URL http://arxiv.org/abs/1412.6572. 5

Google. Ads help: About ad rank, 2023. URL https://support.google.com/google-ads/answer/1722122?hl=en. Accessed: 2023-09-28. 1

Gopalan, P., Kim, M. P., Singhal, M. A., and Zhao, S. Low-degree multicalibration. In *Proc. 35th Conference on Learning Theory*, pp. 3193–3234. PMLR, 2022. 8, 21, 22

Gorantla, S., Mehrotra, A., Deshpande, A., and Louis, A. Sampling individually-fair rankings that are always group fair. In Rossi, F., Das, S., Davis, J., Firth-Butterfield, K., and John, A. (eds.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023*, pp. 205–216. ACM, 2023. 3, 14

Guiver, J. and Snelson, E. Learning to rank with softrank and gaussian processes. In *Proc. 31st Intl. Conf. on Research and Development in Information Retrieval (SIGIR)*, pp. 259–266, 2008. 14

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proc. 34th Intl. Conf. on Machine Learning*, pp. 1321–1330. PMLR, 2017. 1

Guo, R., Ton, J.-F., and Liu, Y. Fair learning to rank with distribution-free risk control. Preprint, 2023. 3, 14

Gupta, C. and Ramdas, A. Top-label calibration and multiclass-to-binary reductions. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=WqoBaaPHS-. 1

Haghtalab, N., Jordan, M. I., and Zhao, E. A unifying perspective on multi-calibration: Unleashing game dynamics for multi-objective learning. *Proc. 37th Advances in Neural Information Processing Systems*, 36, 2023. 8

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Proc. 30th Advances in Neural Information Processing Systems*, 29, 2016. 2, 8

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 2015. 25

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proc. 35th Intl. Conf. on Machine Learning*, pp. 1939–1948. PMLR, 2018. 2, 7, 8, 21, 22

Heuss, M., Cohen, D., Mansoury, M., Rijke, M. d., and Eickhoff, C. Predictive uncertainty-based bias mitigation in ranking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 762–772, 2023. 14

Huang, L. and Vishnoi, N. Stable and fair classification. In *Proc. 36th Intl. Conf. on Machine Learning*, 2019. 15

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 9

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Batch multivalid conformal prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 1

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019. 2, 7, 8, 14, 22

Korevaar, H., McConnell, C., Tong, E., Brinkman, E., Shine, A., Abbas, M., Metevier, B., Corbett-Davies, S., and El-Arini, K. Matched pair calibration for ranking fairness. *arXiv preprint arXiv:2306.03775*, 2023. 15

Kweon, W., Kang, S., and Yu, H. Obtaining calibrated probabilities with personalized ranking models. In *Proc. 36th AAAI Conf. on Artificial Intelligence*, volume 36, pp. 4083–4091, 2022. 3, 14

Luce, R. D. *Individual choice behavior*. Courier Corporation, 1959. 4, 9, 21, 26

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., and Realinho, V. Early prediction of student's performance in higher education: A case study. In *Trends and Applications in Information Systems and Technologies: Volume 1 9*, pp. 166–175. Springer, 2021. 9, 25

Mehrotra, A. and Celis, L. E. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 237–248, 2021. 3, 14

Mehrotra, A. and Vishnoi, N. Fair ranking with noisy protected attributes. *Proc. 36th Advances in Neural Information Processing Systems*, 35:31711–31725, 2022. 3, 14

Menon, A. K., Jiang, X. J., Vembu, S., Elkan, C., and Ohno-Machado, L. Predicting accurate probabilities with a ranking loss. In *Proc. 29th Intl. Conf. on Machine Learning*, volume 2012, pp. 703, 2012. 3, 14

Meta. Our approach to facebook feed ranking, 2023. URL https://transparency.fb.com/features/ranking-and-content/. Accessed: 2023-09-28. 1

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Proc. 35th Advances in Neural Information Processing Systems*, 34: 15682–15694, 2021. 1

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. Pairwise fairness for ranking and regression. In *Proc. 34th AAAI Conf. on Artificial Intelligence*, volume 34, pp. 5248–5255, 2020. 15

Nettleton, D. F., Orriols-Puig, A., and Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33:275–306, 2010. 2

Oh, S., Ustun, B., McAuley, J., and Kumar, S. Rank list sensitivity of recommender systems to interaction perturbations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1584–1594, 2022. 3, 15

Oh, S., Ustun, B., McAuley, J., and Kumar, S. Finest: Stabilizing recommendations by rank-preserving fine-tuning. *arXiv preprint arXiv:2402.03481*, 2024. 15

Penha, G. and Hauff, C. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 160–170, 2021. 14

Pitoura, E., Stefanidis, K., and Koutrika, G. Fairness in rankings and recommendations: an overview. In *Proc. 48th Intl. Conf. on Very Large Data Bases*, 2022. 23

Plackett, R. L. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24 (2):193–202, 1975. 4, 9, 21, 26

Rastogi, R. and Joachims, T. Fair ranking under disparate uncertainty. Preprint, 2023. 3, 14

Robertson, S. E. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977. 6

Shabat, E., Cohen, L., and Mansour, Y. Sample complexity of uniform convergence for multicalibration. *Proc. 34th Advances in Neural Information Processing Systems*, 33: 13331–13340, 2020. 22

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 6, 7, 22

Shen, Z., Wang, Z., Zhu, X., Fain, B., and Munagala, K. Fairness in the assignment problem with uncertain priorities. In *Proc. 22nd Intl. Conf. on Autonomous Agents and Multiagent Systems*, pp. 188–196, 2023. 3, 14

Singh, A. and Joachims, T. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2219–2228, 2018. 2

Singh, A. and Joachims, T. Policy learning for fairness in ranking. *Proc. 33rd Advances in Neural Information Processing Systems*, 32, 2019. 2, 23

Singh, A., Kempe, D., and Joachims, T. Fairness in ranking under uncertainty. In *Proc. 35th Advances in Neural Information Processing Systems*, pp. 11896–11908, 2021. 1, 2, 5, 6, 9, 14, 24, 25, 26

Soliman, M. A. and Ilyas, I. F. Ranking with uncertain scores. In *2009 IEEE 25th international conference on data engineering*, pp. 317–328. IEEE, 2009. 14

Tahir, A., Cheng, L., and Liu, H. Fairness through aleatoric uncertainty. Preprint, 2023. 3, 14

Tang, B., Koçyiğit, Ç., Rice, E., and Vayanos, P. Learning optimal and fair policies for online allocation of scarce societal resources from data collected in deployment. *arXiv preprint arXiv:2311.13765*, 2023. 14

Taylor, M., Guiver, J., Robertson, S., and Minka, T. Softrank: optimizing non-smooth rank metrics. In *Proc. 1st ACM Intl. Conf. on Web Search and Data Mining*, pp. 77–86, 2008. 9

Thurstone, L. L. A law of comparative judgment. *Psychological review*, 101(2):266, 1994. 4, 21

TurboHire. Unleashing the power of ai & automation to effortlessly discover the best talent, 2023. URL https://turbohire.co/features/talent-screening/#candidate-scoring. Accessed: 2023-10-09. 1

Wang, C.-J. and Chen, H.-H. Learning to predict the cost-per-click for your ad words. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2291–2294, 2012. 2

Wilde, N., Bıyık, E., Sadigh, D., and Smith, S. L. Learning reward functions from scale feedback. *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021. 4

Xu, J. and Li, H. Adarank: a boosting algorithm for information retrieval. In *Proc. 30th Intl. Conf. on Research and Development in Information Retrieval (SIGIR)*, pp. 391–398, 2007. 5

Yan, L., Qin, Z., Wang, X., Bendersky, M., and Najork, M. Scale calibration of deep ranking models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4300–4309, 2022. 3, 14

Yang, T., Luo, C., Lu, H., Gupta, P., Yin, B., and Ai, Q. Can clicks be both labels and features? unbiased behavior feature collection and uncertainty-aware learning to rank. In *Proc. 45th Intl. Conf. on Research and Development in Information Retrieval (SIGIR)*, pp. 6–17, 2022. 14

Yang, T., Xu, Z., Wang, Z., Tran, A., and Ai, Q. Marginal-certainty-aware fair ranking algorithm. In *Proc. 16th ACM Intl. Conf. on Web Search and Data Mining*, pp. 24–32, 2023. 14

Zehlike, M., Yang, K., and Stoyanovich, J. Fairness in ranking: A survey. Preprint, 2021. URL https://arxiv.org/abs/2103.14000. 3, 14

Zhu, B., Jordan, M., and Jiao, J. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proc. 40th Intl. Conf. on Machine Learning*, pp. 43037–43067. PMLR, 2023. 4

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. Preprint, 2023. 5

A. Full Discussion of Related Work

Fairness in Ranking. By far the most relevant related work is of Dwork et al. (2019), who are also interested in fair rankings induced by predictors, but importantly restrict their focus to only deterministic rankings (where a better prediction means that an individual will always receive a higher rank) induced by probabilistic binary predictors. Indeed, their motivating example is a setting in which small perturbations to a predictor can massively impact an induced ranking. By requiring stability of ranking functions, we approach this problem fundamentally differently: we allow (and indeed, require) non-deterministic rankings. The multiaccuracy and multicalibration guarantees of Dwork et al. (2019) for induced rankings from predictors are similar in flavor to ours; however, a fundamental difference is that we show this guarantee to be compatible with stability, and, furthermore, that our guarantees hold for each position k in the ranking.

At the intersection of group and individually fair rankings, the work of Gorantla et al. (2023) is most similar to ours. They show that one can sample from a distribution over rankings which is simultaneously individually and group fair (in a proportional representation sense) for laminar groups. In contrast, our group fairness hinges on the group-level statistical constraints of multicalibration imposed on the underlying predictor, which instead allow for potentially arbitrarily overlapping groups. García-Soriano & Bonchi (2021) also work at the intersection of group and individual fairness in rankings, although their group fairness constraints require that certain groups get representation amongst the top-k positions in the ranking, for all $k \in [n]$. Both of these works and ours more broadly explore the interplay between group and individual fairness constraints.

There is far too rich a literature on group and individually fair rankings to cover here, so we restrict attention to only works related to uncertainty and fairness; for a more comprehensive overview, the interested reader is referred to the survey of Zehlike et al. (2021).

Uncertainty in Rankings. Rastogi & Joachims (2023) investigate fairness in uncertainty aware rankings when the uncertainty estimates themselves may be biased for different subgroups. We work in the simpler setting in which we assume that uncertainty estimates are themselves unbiased. Mehrotra & Celis (2021) and Mehrotra & Vishnoi (2022) investigate uncertain protected attributes in the settings of subset selection and ranking, respectively. We do not assume that anything is known about individuals' protected attributes; instead, we only require utilizing the output of a group-fair (multiaccurate) predictor in Section 4. Training such a predictor, however, will require certain knowledge of protected attributes (see, e.g., Kim et al. (2019)).

Independently of the line of work on UA rankings (Devic et al., 2023; Singh et al., 2021), Shen et al. (2023) propose ranked proportionality, which shares a similar definition. Their work is in the more general setting of the assignment problem with uncertain priorities, and they focus on algorithmic approaches for achieving a variety of fairness notions simultaneously. Tang et al. (2023) also consider the (fair) assignment problem and its connections with calibration. Our work is instead focused on proving certain *properties* of rankings induced by predictors (predictors which, when stated in the language of Shen et al. (2023), may induce uncertain priorities). More generally in fairness in uncertain decision making, Tahir et al. (2023) consider how different sources of uncertainty can impact fairness. Guo et al. (2023) utilize conformal prediction techniques to (feasibly) train fair learn-to-rank models, and are also partially interested in a similar notion of stability as ours.

Guiver & Snelson (2008); Soliman & Ilyas (2009); Yang et al. (2022); Cohen et al. (2021); Penha & Hauff (2021) all also work in the area of ranking with uncertain scores or preferences. In contrast to these works, we simultaneously consider uncertainty, fairness, and stability of rankings. Heuss et al. (2023) also model uncertainty with a Bayesian framework that allows them to apply their method post-hoc to arbitrary retrieval models in hopes of reducing bias. Perhaps most relevant is the work of Yang et al. (2023), who examine rankings, utility, fairness, and uncertainty simultaneously. They find that modeling uncertainty can actually *improve* utility in some cases, relative to other fair ranking metrics.

Calibration and Ranking. In Section 4, we work with (multi)-calibrated predictors. Within the ranking community, there has been some investigation into the impact of calibration of ranking models. Menon et al. (2012) initiated this study, attempting to obtain predicted probabilities based on the score output of a ranking model. Kweon et al. (2022) work in a similar setting, but refine the method of obtaining predicted probabilities. Yan et al. (2022) work in the *score-and-sort* model where a scoring function is learned to score each individual, and a ranking function is derived by sorting the individuals according to their scores. Yan et al. (2022) aim to ensure that the scoring model is calibrated with respect to some external property. These works all attempt to infer uncertainty from the scoring function, whereas we assume that uncertainty is

given in the form of machine-learned predictions. Busa-Fekete et al. (2011) show that calibration for ranking functions can help increase diversity of rankings. More recently, DiCiccio et al. (2023) show that conditional predictive parity (a notion which appears to be related to multicalibration) can help decrease bias in rankings. These works all highlight the benefits of using calibrated predictive models for ranking, outside of the guarantees that we provide. Korevaar et al. (2023) relate calibration and exposure in rankings by comparing the rankings attained by subgroups with similar score distributions.

Stability in Rankings. In the information retrieval literature, Asudeh et al. (2018) also study the notion of stability for rankings. They work in a setting where a *score* is calculated based on a weighted sum of features of each item, and stability is then with respect to small changes of these weights. However, their notion of stability is based on geometric intuition for their scoring function and its dual, and only holds for any fixed data set. They furthermore state that stability is not a property of their "scoring function" (particular weighted sum over features). In contrast, we are explicitly defining stability as a property of our ranking function, which maps from any set of predictions (data set) to a randomized ranking. Oh et al. (2022) also study the sensitivity of rankings; however, their context is slightly different: they examine stability with respect to *user interactions* with, e.g., a recommendation system. In a very recent followup work (Oh et al., 2024), the same authors also provide an algorithm to empirically achieve stability in that setting. Bruch et al. (2020) provide experimental evidence showing that randomization can help stability (which they define as *robustness*) during the training of learning-to-rank models. Our theoretical results are complementary and corroborate the empirical evidence of Bruch et al. (2020) that randomized rankings are more robust to noise than deterministic ones.

Finally, in terms of the interplay between prediction systems and rankings, the work of Narasimhan et al. (2020) is perhaps most relevant; they show that the ranking problem can be considered as a pairwise binary classification problem between items to determine which item should be placed at a higher rank.

We also point out the work of Huang & Vishnoi (2019) who suggest studying rankings from the point of view of stability, albeit their motivation stems from stability in the learning theory community.

B. Omitted Proofs

B.1. Proof of Proposition 8

Proof. To prove the impossibility of anonymity, consider any prediction matrix $P = (\boldsymbol{p})_{i\in[n]}$ with identical predictions $\boldsymbol{p}$ for each individual (e.g., $\boldsymbol{p} = (1, 0, \ldots, 0) \in \Delta_L$). Then, any deterministic ranking function r must order the individuals based only on their indices in P, since they all have identical predictions. For any permutation $\sigma : [n] \to [n]$, let P_σ represent applying σ to the rows of P. Since the ranking function can only depend on the input matrix, and r is deterministic, we have that $r(P) = r(P_\sigma)$. However, by the definition of anonymity (Definition 2), the permutation σ on the rows of P should produce the permutation $r(P)_\sigma$ on the individuals in the resulting ranking, which is a contradiction. Therefore, r is not anonymous.

To prove the instability result, we prove the contrapositive. Let r be deterministic and non-constant, and $\gamma > 0$. We will show that r is not γ-stable. Because r is non-constant, there exist $P, P' \in \mathcal{P}_{n,L}$ with $r(P) \neq r(P')$. Consider the straight line $Q(\beta) = \beta P + (1-\beta)P'$, for $\beta \in [0, 1]$. Because $\mathcal{P}_{n,L}$ is convex, $Q(\beta) \in \mathcal{P}_{n,L}$ for all $\beta \in [0, 1]$. Let $\beta^* = \inf\{\beta \mid Q(\beta) = P'\}$; β^* is well-defined because $Q(1) = P'$. By definition, $Q(\beta) \neq P'$ for all $\beta < \beta^*$, and if $\beta^* = 0$, then $Q(\beta^*) = P \neq P'$. On the other hand, by the definition of the infimum, for every $\delta > 0$, there is a $\delta' < \delta$ with $Q(\beta^* + \delta') = P'$. Thus, we obtain arbitrarily close pairs $\beta' \leq \beta^* < \beta''$ with $Q(\beta') \neq Q(\beta'')$. Because r is deterministic, all entries of $r(Q(\beta'))$ and $r(Q(\beta''))$ are in $\{0, 1\}$, implying that $||r(Q(\beta')) - r(Q(\beta''))||_\infty \geq 1$. On the other hand, $||Q(\beta') - Q(\beta'')||_1 \leq ||2\delta(P' - P)||_1 \to 0$ as $\delta \to 0$. By choosing δ small enough (as a function of γ), this implies that r is not γ-stable, completing the proof. □

While we prove instability (i.e., the second part of Proposition 8) by considering a straight line between two different prediction matrices P, P', this is not essential. By considering any path (curve in $\mathbb{R}^{n\times L}$) connecting P, P' and its parametrization by $\beta \in [0, 1]$, the exact same proof still works. This shows that even if we consider only a subset of possible predictions, so long as the subset is path-connected[5], a deterministic stable ranking function must be constant. This extends the proposition to settings where prediction strategies may output only certain (path-connected) subsets of predictions, due

[5]Recall that a set A is path-connected if for every pair of elements $x, y \in A$ there exists a continuous path between x and y which is entirely contained within A.

to, for example, intrinsic preferences or implicit bias of a particular learning algorithm.

B.2. Proof of Proposition 10

The proof of Proposition 10 and Theorem 11 will use the following Proposition, which shows that by taking into account the randomness of the draws of labels and the tie breaking, the rank distribution produced by UA ranking can be summarized as follows:

Proposition 18. *Let $P \in \mathcal{P}_{n,L}$ be a prediction, and $r_{UA}(P)$ the ranking distribution produced by r_{UA} for P. Then, the probability of individual $i \in [n]$ being ranked in position $k \in [n]$ is:*

$$\mathbb{P}_{r_{UA}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] = \sum_{j=0}^{n-1} \frac{1}{j+1} \cdot \mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k], \tag{2}$$

$$\mathbb{P}_{r_{UA}(P)}[\mathcal{R}_{i,k}] = \sum_{\ell} p_{i,\ell} \cdot \mathbb{P}_{r_{UA}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]. \tag{3}$$

Proof. For the first part, we observe that the probability of $\mathcal{R}_{i,k}$ in (1) depends only on N^{ℓ} and $N^{>\ell}$. By considering all the possible values of $N^{\geq \ell}$ for which (1) gives a non-zero probability, we obtain that

$$\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] = \sum_{j=1}^{n} \frac{1}{j} \cdot \mathbb{P}_P[N^{\ell} = j \text{ and } N^{>\ell} < k \leq N^{\geq \ell} \mid \lambda_i = \ell].$$

The result is then obtained by noticing that conditioned on $\lambda_i = \ell$, we have $N_{-i}^{\ell} = N^{\ell} - 1$ and $N_{-i}^{\ell'} = N^{\ell'}$ for all $\ell' \neq \ell$. The second part of the proposition simply states the law of total probability. □

We are now ready to prove Proposition 10.

Proof of Proposition 10. The two parts of Proposition 18 combined imply that in order to compute row i of $r_{\text{UA}}(P)$, it is sufficient to compute $\mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k]$ for all pairs $(j,k) \in [m]$. This is accomplished by a dynamic program similar to a standard undergraduate exercise, which is to compute a Poisson Binomial distribution explicitly.

For notational convenience, assume that $i = n$; this is solely to avoid a special case in the recurrence, and also without loss of generality by anonymity of the UA rule. For any $t, j, j' \in \{0, 1, \ldots, m-1\}$, let

$$A(t, j, j') = \mathbb{P}_P[N_{\{1,\ldots,t\}}^{\ell} = j \text{ and } N_{\{1,\ldots,t\}}^{>\ell} = j']$$

be the probability that among the first t individuals, exactly j have label ℓ, and exactly j' have a label strictly better than ℓ. From these values, we can then construct the necessary quantities as

$$\mathbb{P}_P[N_{-n}^{\ell} = j \text{ and } k-(j+1) \leq N_{-n}^{>\ell} < k] = \sum_{j'=k-(j+1)}^{k-1} A(n-1, j, j').$$

We give the recurrence relationship for the $A(t, j, j')$. The base cases are that

$$A(0,0,0) = 1$$
$$A(0,j,j') = 0 \text{ if } j + j' \neq 0,$$

because with no individuals, the only possible numbers of individuals with given labels is 0.

Now consider $A(t, j, j')$ for $t \geq 1$. With probability $p_{t,\ell}$, individual t has label ℓ, in which case the desired event happens when $j-1$ individuals among the first $t-1$ have label ℓ, and j' have labels strictly better than ℓ. With probability $\sum_{\ell' > \ell} p_{t,\ell'}$, individual t has a label strictly better than ℓ, in which case the desired event happens when j individuals among the first $t-1$ have label ℓ, and $j'-1$ have labels strictly better than ℓ. Finally, with probability $\sum_{\ell' < \ell} p_{t,\ell'}$, individual t has a label

strictly worse than ℓ, in which case the desired event happens when j individuals among the first $t-1$ have label ℓ, and j' have labels strictly better than ℓ. These three cases disjointly cover all possibilities for the label of t, so we have derived the following recurrence:

$$A(t,j,j') = p_{t,\ell} \cdot A(t-1,j-1,j') + \sum_{\ell'>\ell} p_{t,\ell'} \cdot A(t-1,j,j'-1) + \sum_{\ell'<\ell} p_{t,\ell'} \cdot A(t-1,j,j').$$

Here, to avoid case distinctions for whether j and/or j' are 0, we treat $A(t,j,j') = 0$ whenever j or j' are negative.

Notice that for any fixed t, all values $\sum_{\ell'>\ell} p_{t,\ell'}$, being prefix sums, can be pre-computed in time $O(L)$. Thus, for all ℓ, t, the precomputation can be performed in time $O(nL)$.

Then, any one entry $A(t,j,j')$ can be computed in constant time from previously computed values. Because the table has size $O(n^3)$, the total computation takes time $O(n^3)$. Summing over all possible values of i, the total time to compute the entire ranking distribution $r_{\text{UA}}(P)$ is $O(n^4 + n^2 L)$. Finally, the post-processing of computing the $\mathbb{P}_P[N_{-i}^{\ell} = j$ and $k-(j+1) \leq N_{-i}^{>\ell} < k]$ for all k, for fixed i, ℓ, j, can be implemented in time $O(n)$ by using differences of prefix sums; thus, all values can be computed in time $O(n^3 L)$. This gives a total time of $O(n^4 + n^3 L)$. □

B.3. Proof of Theorem 11

The following lemma is a key part of the proof of stability in Theorem 11; it bounds how different the probabilities for individual i obtaining rank k can be under two different prediction matrices, as a function of how similar these matrices are:

Lemma 19. *Let $P, Q \in \mathcal{P}_{n,L}$ be two different prediction matrices. For any individual i, let $\boldsymbol{p}_i, \boldsymbol{q}_i$ be the ith rows of P, Q, respectively, i.e., the label distributions of individual i under the two predictions. Let i be an individual, $k \in [n]$ a position, and $\ell \in [L]$ a label. Then,*

$$\left| \mathbb{P}_{r_{UA}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{UA}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] \right| \leq 2 \cdot \sum_{i' \neq i} d_{TV}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'}).$$

The proof of Lemma 19 uses the following lemma, bounding the total variation distance of sums of random variables in terms of the total variation distances of the individual variables.

Lemma 20. *Let $X_i \sim \boldsymbol{p}_i, Y_i \sim \boldsymbol{q}_i$ for $i = 1, \ldots, n$ be independent categorical random variables, and $X = \sum_{i=1}^n X_i, Y = \sum_{i=1}^n Y_i$. Let $\boldsymbol{p}, \boldsymbol{q}$ be the respective distributions of X, Y. Then, $d_{TV}(\boldsymbol{p}, \boldsymbol{q}) \leq \sum_{i=1}^n d_{TV}(\boldsymbol{p}_i, \boldsymbol{q}_i)$.*

Proof. Consider a maximal coupling between each X_i and the corresponding Y_i. By the Coupling Lemma, we then have that $\mathbb{P}[X_i \neq Y_i] = d_{\text{TV}}(\boldsymbol{p}_i, \boldsymbol{q}_i)$, and $d_{\text{TV}}(\boldsymbol{p}, \boldsymbol{q}) \leq \mathbb{P}[X \neq Y]$. Now, by a union bound over all i, we obtain that

$$d_{\text{TV}}(\boldsymbol{p}, \boldsymbol{q}) \leq \mathbb{P}[X \neq Y] \leq \sum_i \mathbb{P}[X_i \neq Y_i] = \sum_i d_{\text{TV}}(\boldsymbol{p}_i, \boldsymbol{q}_i),$$

completing the proof. □

Proof of Lemma 19. First, by Equation (2) in the first part of Proposition 18,

$$\mathbb{P}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] = \sum_{j=0}^{n-1} \frac{1}{j+1} \cdot \mathbb{P}[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k].$$

Let

$$B = \left\{ j \mid \mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k] \geq \mathbb{P}_Q[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k] \right\};$$

note that B is not a random variable, but simply determined by the distributions P, Q.

We substitute the characterization (2) for both P and Q, and use the triangle inequality as well as the fact that $\frac{1}{j+1} \leq 1$, to give us that

$$\begin{aligned}
&\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right| \\
&\quad \leq \sum_{j=0}^{n-1} \left|\mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k]\right| \\
&\quad = \sum_{j=0, j\in B}^{n-1} \left(\mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k]\right) \\
&\qquad + \sum_{j=0, j\notin B}^{n-1} \left(\mathbb{P}_Q[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_P[N_{-i}^{\ell} = j \text{ and } k-(j+1) \leq N_{-i}^{>\ell} < k]\right) \\
&\quad = \left(\mathbb{P}_P[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right) \\
&\qquad + \left(\mathbb{P}_Q[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_P[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right) \\
&\quad = \left|\mathbb{P}_P[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right| \\
&\qquad + \left|\mathbb{P}_P[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right|. \qquad (4)
\end{aligned}$$

Consider the (vector-valued) random variable $(N_{-i}^{>\ell}, N_{-i}^{\ell})$, and let μ, ν denote its distribution under P, Q, respectively.

Because $[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]$ and $[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]$ are events that can be expressed in terms of this random variable, the definition of total variation distance implies that

$$\begin{aligned}
&\left|\mathbb{P}_P[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} \in B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right| \leq d_{\text{TV}}(\mu, \nu), \\
&\left|\mathbb{P}_P[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k] - \mathbb{P}_Q[N_{-i}^{\ell} \notin B \text{ and } k-(N_{-i}^{\ell}+1) \leq N_{-i}^{>\ell} < k]\right| \leq d_{\text{TV}}(\mu, \nu).
\end{aligned}$$

To bound $d_{\text{TV}}(\mu, \nu)$, associate with each individual $i' \neq i$ the 2-dimensional (random) vector $\boldsymbol{v}_{i'} = (\mathbb{1}_{\lambda_{i'} > \ell}, \mathbb{1}_{\lambda_{i'} = \ell})$. Then, $(N_{-i}^{>\ell}, N_{-i}^{\ell}) = \sum_{i' \neq i} \boldsymbol{v}_{i'}$.

For a fixed $i' \neq i$, consider the distribution of $\boldsymbol{v}_{i'}$ under $\boldsymbol{p}_{i'}$ and $\boldsymbol{q}_{i'}$. The total variation distance between these distributions is at most $d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'})$, because the vectors can differ only when the labels of i' differ. By Lemma 20, we thus obtain that $d_{\text{TV}}(\mu, \nu) \leq \sum_{i' \neq i} d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'})$.

Substituting this bound back into (4), we now obtain that

$$\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right| \leq 2 \cdot \sum_{i' \neq i} d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'}),$$

completing the proof. □

With Lemma 19 in hand, we are ready to prove Theorem 11.

Proof of Theorem 11. First, the UA ranking rule is obviously anonymous, simply by its (symmetric) definition which treats all indices identically. Thus, we focus on proving stability.

Now, let any individual i and rank k be given. By Equation (3) in the second part of Proposition 18, $\mathbb{P}[\mathcal{R}_{i,k}] = \sum_{\ell} p_{i,\ell} \cdot \mathbb{P}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]$. Now consider two different predictors P, Q. We bound the difference in probabilities for i to be ranked in position k as follows:

$$
\begin{aligned}
&\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k}]\right| \\
&\quad \le \sum_{\ell} \left|p_{i,\ell} \cdot \mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - q_{i,\ell} \cdot \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right| \\
&\quad \le \sum_{\ell} |p_{i,\ell} - q_{i,\ell}| \cdot \mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] + q_{i,\ell} \cdot \left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right| \\
&\quad \le \sum_{\ell} |p_{i,\ell} - q_{i,\ell}| + q_{i,\ell} \cdot \left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right|. \qquad (5)
\end{aligned}
$$

By Lemma 19, we can bound $\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \lambda_i = \ell]\right| \le 2 \cdot \sum_{i' \neq i} d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'})$.

Substituting this bound back into (5), we now obtain that

$$
\begin{aligned}
|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k}]| &\le \sum_{\ell} |p_{i,\ell} - q_{i,\ell}| + \sum_{\ell} q_{i,\ell} \cdot 2 \cdot \sum_{i' \neq i} d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'}) \\
&= \frac{1}{2} d_{\text{TV}}(\boldsymbol{p}_i, \boldsymbol{q}_i) + 2 \cdot \sum_{i' \neq i} d_{\text{TV}}(\boldsymbol{p}_{i'}, \boldsymbol{q}_{i'}) \\
&\le ||P - Q||_1.
\end{aligned}
$$

In the final step, we absorbed the total variation distance for i into the sum for $i' \neq i$ (which has a larger coefficient), and used that the 1-norm is exactly twice the total variation distance. □

B.4. Proof of Proposition 13

We state the formal version of Proposition 13, and prove it here.

Proposition 21. *The expressivity of uncertainty aware ranking functions is strictly increasing in L. More formally, let $n \ge L$, and $r^L_{UA} : \mathcal{P}_{n,L} \to \mathcal{M}^{n\times n}_{DS}, r^{L-1}_{UA} : \mathcal{P}_{n,(L-1)} \to \mathcal{M}^{n\times n}_{DS}$ be the corresponding UA ranking functions. Then, $r^L_{UA}(\mathcal{P}_{n,L}) \supsetneq r^{L-1}_{UA}(\mathcal{P}_{n,(L-1)})$.*

Proof. First, to see monotonicity, notice that adding a column of all 0 entries, i.e., an unused label L, does not change the behavior of r_{UA}. For any $P \in \mathcal{P}_{n,(L-1)}$, writing $[P, \mathbf{0}] \in \mathcal{P}_{n,L}$ for this matrix, we have $r^L_{\text{UA}}([P, \mathbf{0}]) = r^{L-1}_{\text{UA}}(P)$, implying that $r^L_{\text{UA}}(\mathcal{P}_{n,L}) \supseteq r^{L-1}_{\text{UA}}(\mathcal{P}_{n,L-1})$.

To prove strictness of inclusion, consider a prediction $P = J_n$ over $n = L$ individuals. Here, J_n is the $n \times n$ row-reversed identity matrix with ones along the anti-diagonal, so individual i is deterministically known to have label $L - i + 1$. Then $r_{\text{UA}}(P) = I_n$ for the $n \times n$ identity matrix I_n, i.e., individual i is ranked deterministically in position i, and we have proved that $I_n \in r^L_{\text{UA}}(\mathcal{P}_{n,L})$. Note that due to the tie breaking of r_{UA}, to achieve a deterministic ranking, no two individuals must ever have the same label, i.e., the supports of the $n = L$ rows of any prediction matrix Q yielding $r_{\text{UA}}(Q) = I_n$ must be disjoint. This implies that Q must have at least L columns, i.e., $I_n \notin r^{L-1}_{\text{UA}}(\mathcal{P}_{n,L-1})$, completing the proof of strictness of inclusion. □

B.5. Proof of Proposition 12

Proof. Let $n \ge 2$ be given. We only consider $L = 3$; for any $L > 3$, it suffices by Proposition 21 to embed the following instance and ignore the extra labels. Consider the prediction matrix P with individual 1 having prediction $\boldsymbol{p}_1 = (1/2, 0, 1/2)$, and individuals 2 through n having prediction $\boldsymbol{p}_2 = (0, 1, 0)$. Similarly, the prediction matrix P' will have individual 1 with prediction $\boldsymbol{p}'_1 = (1, 0, 0)$ and individuals 2 through n with prediction $\boldsymbol{p}_2$. That is, P and P' are identical except for individual 1.

Let $M = r_{\text{UA}}(P), M' = r_{\text{UA}}(P')$ be the resulting probabilities for placing individuals in specific positions. Since all individuals except individual 1 have deterministic qualifications, under P, there is a 50% probability that individual 1 is

ranked last, so $M_{1,n} = 1/2$, whereas $M'_{1,n} = 0$. Therefore, we have the following.

$$\|r_{\text{UA}}(P) - r_{\text{UA}}(P')\|_\infty \geq |M_{1,n} - M'_{1,n}| = 1/2, \qquad \|P - P'\|_1 = 1.$$

Thus, r_{UA} cannot be γ-stable for any $\gamma < 1/2$. □

C. Extension to Continuous Ranking Functions

An extension to continuous label spaces is straightforward. In fact, some aspects of the calculations are simplified, because for independent draws from continuous distributions, the probability of ties in labels is 0. We begin by defining the modification of concepts precisely.

The set of labels is now $L = \mathbb{R}$; this is without loss of generality, as labels that cannot occur can be assigned density 0 by the predictor. A predictor f now outputs an absolutely continuous distribution $f(x) = p_x$ over $L = \mathbb{R}$ for the data point x; i.e., a distribution which has a probability density function (PDF). We denote the PDF at $\ell \in \mathbb{R}$ by $p_x(\ell)$, so that the label λ_x of individual x is drawn from p_x, i.e., $\lambda_x \sim p_x$. We now use $\Delta_\mathbb{R}$ to denote the set of all absolutely continuous distributions over $\mathbb{R}$, so that we can still regard f as a mapping $f : \mathcal{X} \to \Delta_L$.

As before, we use $\mathcal{P}_{n,\mathbb{R}}$ to denote the set of all possible combinations of predictions based on the n data points, i.e., all n-dimensional vectors $P = (p_1, p_2, \ldots, p_n)$ such that each $p_i \in \Delta_\mathbb{R}$ is an absolutely continuous probability distribution. Then, the definition of a ranking function stays virtually unchanged:

Definition 22 (Continuous Ranking Function)**.** *A ranking function $r : \mathcal{P}_{n,\mathbb{R}} \to \mathcal{M}_{DS}^{n\times n}$ maps an n-dimensional vector P of absolutely continuous probability distributions to a distribution M over rankings.*

The definition of uncertainty awareness now extends in the natural way:

Definition 23 (Uncertainty Awareness for Continuous Ranking Functions)**.** *A randomized ranking $M \in \mathcal{M}_{DS}^{n\times n}$ is* uncertainty aware *for a vector $P \in \mathcal{P}_{n,\mathbb{R}}$ if for each individual i and position k, the entry $M_{i,k}$ is the probability that i has the kth highest label if all labels $\lambda_i \sim p_i$ are sampled* independently *from the respective distributions p_i. The ranking function r is uncertainty aware if $r(P)$ is uncertainty aware for all $P \in \mathcal{P}_{n,\mathbb{R}}$.*

Notice that this definition is identical to Definition 9, but since the labels λ_i are drawn independently from continuous distributions, the probability of ties is 0.

Stability is now defined as before:

Definition 24 (Stability)**.** *Fix n. A ranking function $r : \mathcal{P}_{n,\mathbb{R}} \to \mathcal{M}_{DS}^{n\times n}$ is γ-*stable *if $||r(P) - r(Q)||_\infty \leq \gamma \cdot \sum_{i=1}^n ||p_i - q_i||_1$ for all $P, Q \in \mathcal{P}_{n,\mathbb{R}}$, where we use the standard definition $||p||_1 = \int_\mathbb{R} |p(x)| \,\mathrm{d}x$.*

Theorem 25. *The continuous variant of UA ranking is 1-stable.*

Proof. First, fix one vector $P \in \mathcal{P}_{n,\mathbb{R}}$ of label distributions. Let $\boldsymbol{\lambda} \sim P$ be a vector of labels, with each entry λ_i drawn independently from its associated density p_i. We use $P(\boldsymbol{\lambda})$ to denote the joint (product) distribution over $\boldsymbol{\lambda}$. Because the event of any two λ_i, λ_j being equal has measure 0, we will focus our analysis on the case when all λ_i are distinct. Formally, we define $\mathcal{T} := \{(\ell_1, \ldots, \ell_n) \in \mathbb{R}^n \mid \ell_i \neq \ell_j \text{ for all } i \neq j\}$ to be the set of all label vectors all of whose entries are different. Then, the preceding observation about measure 0 can be stated precisely as $\mathbb{P}_{\boldsymbol{\lambda}\sim P}[\mathcal{T}] = 1$.

For any individual $i \in [n]$, the labels λ_i induce the counts $N_{\boldsymbol{\lambda}}^{>i}$ of individuals more qualified than i. Under UA ranking, if $\boldsymbol{\lambda} \in \mathcal{T}$, then individual i obtains rank $N_{\boldsymbol{\lambda}}^{>i} + 1$, giving conditional probabilities

$$\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] = \begin{cases} 1 & \text{if } N_{\boldsymbol{\lambda}}^{>i} = k - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that this is much simpler than Equation (2) in Proposition 18, which required a careful treatment of ties. Furthermore, observe that the probability is determined completely by $\boldsymbol{\lambda}$, and thus independent of the choice of P.

By the law of total probability, and because $\mathbb{P}_{\boldsymbol{\lambda}\sim P}[\boldsymbol{\lambda} \notin \mathcal{T}] = 0$, the probability of individual i being ranked in position k is

$$\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k}] = \int_\mathcal{T} \mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot P(\boldsymbol{\lambda}) \,\mathrm{d}\boldsymbol{\lambda}.$$

Next, let P, Q be two vectors of label densities, and $P(\boldsymbol{\lambda}), Q(\boldsymbol{\lambda})$ the associated joint product distributions. Using that $\overline{\mathcal{T}}$ has measure 0 under both P, Q, we can then write the difference between the outcomes of UA ranking under P and Q as follows:

$$\begin{aligned}
&\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k}]\right| \\
&= \left|\int_{\mathcal{T}} \left(\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot P(\boldsymbol{\lambda}) - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot Q(\boldsymbol{\lambda})\right) \mathrm{d}\boldsymbol{\lambda}\right| \\
&\leq \int_{\mathcal{T}} \left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot P(\boldsymbol{\lambda}) - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot Q(\boldsymbol{\lambda})\right| \mathrm{d}\boldsymbol{\lambda} \\
&\leq \int_{\mathcal{T}} \left(\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] \cdot |P(\boldsymbol{\lambda}) - Q(\boldsymbol{\lambda})| + \left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}]\right| \cdot Q(\boldsymbol{\lambda})\right) \mathrm{d}\boldsymbol{\lambda} \\
&\leq \int_{\mathcal{T}} |P(\boldsymbol{\lambda}) - Q(\boldsymbol{\lambda})| \, \mathrm{d}\boldsymbol{\lambda} + \int_{\mathcal{T}} \left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}]\right| \cdot Q(\boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\lambda}.
\end{aligned}$$

We now bound the two integrals separately. For the first integral, we use the fact that the measures P, Q are product measures, and that the total variation distance between two product measures — and hence the $||\cdot||_1$ norm — is upper-bounded by the sum of the total variation distances (respectively, $||\cdot||_1$ norms) for each variable in the product measure. This implies that

$$\int_{\mathcal{T}} |P(\boldsymbol{\lambda}) - Q(\boldsymbol{\lambda})| \, \mathrm{d}\boldsymbol{\lambda} \leq \sum_{j=1}^{n} ||p_j - q_j||_1.$$

Lastly, we show that the second integral is 0. This follows simply from the prior observation that $\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k} \mid \boldsymbol{\lambda}]$ is fully determined by $\boldsymbol{\lambda}$, and independent of P, implying that the term under the integral is always 0.

We have shown that for any i, k, the difference $\left|\mathbb{P}_{r_{\text{UA}}(P)}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{\text{UA}}(Q)}[\mathcal{R}_{i,k}]\right| \leq \sum_{j=1}^{n} ||p_j - q_j||_1$. It follows that the same holds for the maximum over all i, k given by $\|r(P) - r(Q)\|_\infty$. Hence, we have completed the 1-stability proof. □

Remark 26. *Theorem 25 also implies that the well-known Plackett-Luce Plackett (1975); Luce (1959) ranking function is stable. This is because of the "Gumbel trick" Bruch et al. (2020), which reframes the Plackett-Luce ranking procedure as sorting by relevance scores with added Gumbel noise (discussed in more detail in Appendix G.2). This in fact creates a continuous density, to which we can apply Theorem 25. In other words, the Plackett-Luce model is a special case of UA ranking in which the merit distribution is implicitly assumed to be of the following form: each individual i has an observed merit estimate, and the true merit is assumed to be this estimate plus i.i.d. Gumbel noise.*

Theorem 25 can also be applied to show stability of the Thurstonian ranking model (Thurstone, 1994), which assumes that relevance scores are drawn from particular normal distributions; as with the Plackett-Luce model, the Thurstonian ranking model is in fact a special case of UA ranking with particular choices of merit distributions.

D. Multicalibration as Interpolating between Individual and Group Fairness

In contrast to learning accurate individual-level estimates (the Bayes optimal predictor), multiaccuracy/multicalibration can be achieved in time and samples polynomial in the number of sets in $\mathcal{C}$, or, more generally, polynomial in measures of complexity of $\mathcal{C}$ such as its VC-dimension (Hébert-Johnson et al., 2018; Gopalan et al., 2022).

Notice that if we define $\mathcal{C}_{\text{Bayes}} = \{\{x\} \mid x \in \mathcal{X}\}$ to be the set of all singleton groups, then $(\mathcal{C}_{\text{Bayes}}, \alpha)$-multiaccuracy guarantees increasingly accurate predictions for all individuals as $\alpha \to 0$, and $\alpha = 0$ recovers the Bayes optimal classifier, i.e., the ground truth f^*. By varying the level of granularity of the collection $\mathcal{C}$, the learned $(\mathcal{C}, \alpha)$-multiaccurate or multicalibrated predictor represents a finer or coarser approximation of the Bayes optimal classifier; thus, Theorem 15 guarantees that the induced ranking effectively *interpolates* between individual and group-level fair rankings at the granularity defined by $\mathcal{C}$.

Nonetheless, as previously noted, it is usually unreasonable to expect multiaccuracy (or multicalibration) at the level of $(\mathcal{C}_{\text{Bayes}}, \alpha)$ as $\alpha \to 0$. This is due to information and computational constraints: multicalibration algorithms must use

$\Omega(\text{poly}(|\mathcal{C}|, 1/\delta, 1/\alpha))$ samples to learn a multiaccurate/multicalibrated predictor (Shabat et al., 2020).[6] In addition to the sample complexity requirements, the class $\mathcal{C}$ must be *agnostic PAC learnable* (Shalev-Shwartz & Ben-David, 2014, Section 3.2). This is a stringent requirement which rarely holds for complex collections such as $\mathcal{C}_{\text{Bayes}}$. In practice, we envision Theorem 15 to be used with sufficiently simple classes $\mathcal{C}$, such as conjunctions of categorical features and intervals of numeric features (e.g., "women in the age range 45–65"). Working at this level of granularity not only permits efficient algorithms for obtaining multiaccurate/multicalibrated predictors, but also guarantees that the derived rankings will be unbiased for meaningful protected groups of individuals.

E. A Utility-Optimal Ranking Function Cannot be Stable or Fair

The ranking function which achieves optimal utility will clearly depend on τ; we denote it by r^{τ}_{opt}. It can be simply described as the ranking function which deterministically orders the individuals by decreasing values $\tau(\boldsymbol{p}_i)$; recall that $\boldsymbol{p}_i$ is the ith row of $P \in \mathcal{P}_{n,L}$. We now show that in general, r^{τ}_{opt} is not stable, which demonstrates a necessity to trade off notions of utility and stability.

Proposition 27. *Even for binary labels ($L = 2$) and expected utility map τ, the utility-maximizing map r^{τ}_{opt} is unstable.*

Proof. The example is standard in the literature. Assume that $v_2 > v_1$. For any $\epsilon \in (0, \frac{1}{2})$, define

$$P_\epsilon = \begin{pmatrix} \frac{1}{2}+\epsilon & \frac{1}{2}-\epsilon \\ \frac{1}{2}-\epsilon & \frac{1}{2}+\epsilon \end{pmatrix} \qquad P'_\epsilon = \begin{pmatrix} \frac{1}{2}-\epsilon & \frac{1}{2}+\epsilon \\ \frac{1}{2}+\epsilon & \frac{1}{2}-\epsilon \end{pmatrix}.$$

Then, $r^{\tau}_{\text{opt}}(P_\epsilon)$ deterministically ranks 2 ahead of 1, while $r^{\tau}_{\text{opt}}(P'_\epsilon)$ deterministically ranks 1 ahead of 2. As a result, $\|r^{\tau}_{\text{opt}}(P_\epsilon) - r^{\tau}_{\text{opt}}(P'_\epsilon)\|_\infty = 1$, while $\|P_\epsilon - P'_\epsilon\|_1 = 8\epsilon \to 0$ as $\epsilon \to 0$. This proves instability of r^{τ}_{opt}. □

Next, we show that for an extremely simple class of instances, namely, when there are two types of individuals with uniform distribution, binary labels, identical uniform ground truth distribution over the two labels for both types, and groups which are just the two singleton types, the utility-maximizing ranking function r^{τ}_{opt} cannot approach optimal multigroup fairness, never mind how close to perfectly multiaccurate the predictor gets.

Proposition 28. *Let $\mathcal{X} = \{1, 2\}$ be a domain of two "types" of individuals. Let $\mathcal{D}$ be the uniform distribution over those two types. Let $L = 2$, i.e., we consider binary labels, and the ground truth label distribution is $f^*(x) = (\frac{1}{2}, \frac{1}{2})$ for both $x \in \{1, 2\}$, i.e., under the ground truth, both types are equally likely to be good and bad. Let $\mathcal{C} = \{\{1\}, \{2\}\}$ be the collection of singleton subgroups.*

For any $\alpha \in (0, \frac{1}{2})$, let f_α be the predictor with predictions $f_\alpha(1) = (\frac{1}{2} - \alpha, \frac{1}{2} + \alpha)$ and $f_\alpha(2) = (\frac{1}{2} + \alpha, \frac{1}{2} - \alpha)$. (That is, f_α slightly overestimates the quality of type 1, and slightly underestimates the quality of type 2.) Let τ be any utility map strictly preferring higher labels, i.e., any utility map with $\tau((\frac{1}{2} - \alpha, \frac{1}{2} + \alpha)) > \tau((\frac{1}{2} + \alpha, \frac{1}{2} - \alpha))$. Let r^{τ}_{opt} be any utility-maximizing ranking function for the utility map τ.

Then, f_α is $(\mathcal{C}, \alpha/2)$-multiaccurate, yet for every number n of individuals, the group fairness under r^{τ}_{opt} towards the group $S = \{1\}$ for assignment to the top (most valuable) position in the ranking is the following:

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim \mathit{Unif}([n])}\left[\mathbb{1}_{x_i\in\{1\}} \cdot \left(\mathbb{P}_{r^{\tau}_{opt}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,1}] - \mathbb{P}_{r^{\tau}_{opt}(f_\alpha(\boldsymbol{x}))}[\mathcal{R}_{i,1}]\right)\right]\right| = \frac{1}{n} \cdot \left(\frac{1}{2} - 2^{-n}\right). \tag{6}$$

In particular, for any fixed n, the quantity stays bounded away from 0, even as $\alpha \to 0$.

The intuition for Proposition 28 is similar to that for Proposition 8: for the utility-maximizing ranking function, an arbitrarily small (but non-zero) predictive mistake can induce large variations in the resulting ranking distribution, preventing it from preserving group fairness of its predictor.

Proof. We first verify that f_α is $(\mathcal{C}, \alpha/2)$-multiaccurate. For $S = \{1\}$ or $S = \{2\}$, we have that

$$\|\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{x\in S} \cdot (f_\alpha(x) - f^*(x))\right]\|_\infty \leq \frac{1}{2} \cdot \max(|\frac{1}{2} - (\frac{1}{2} - \alpha)|, |\frac{1}{2} - (\frac{1}{2} + \alpha)|) = \alpha/2.$$

[6] We refer the reader to the original papers Kim et al. (2019); Hébert-Johnson et al. (2018) for algorithms achieving multiaccuracy and multicalibration, respectively, or Gopalan et al. (2022) for a unified approach.

The rest of the proof focuses on the group fairness analysis, i.e., proving Equation (6). We first consider the term $\mathbb{P}_{r^{\tau}_{\text{opt}}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,1}]$. First observe that under the ground truth classifier f^*, we have that $f^*(\boldsymbol{x}) = \frac{1}{2} \cdot \mathbb{1}_{n\times 2}$ for all $\boldsymbol{x}$; here $\mathbb{1}_{n\times 2}$ denotes the $n \times 2$ all-ones matrix. Under this input matrix, r^{τ}_{opt} must have some distribution $\boldsymbol{q} = (q_1, \ldots, q_n)$ over which individual is assigned the top rank; crucially for our analysis, because the ranking function *only* observes this matrix $\frac{1}{2} \cdot \mathbb{1}_{n\times 2}$, it must use the same distribution for all type vectors $\boldsymbol{x}$. We thus conclude that $\mathbb{P}_{r^{\tau}_{\text{opt}}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,1}] = q_i$ for all type vectors $\boldsymbol{x}$.

Next, we consider the term $\mathbb{P}_{r^{\tau}_{\text{opt}}(f_\alpha(\boldsymbol{x}))}[\mathcal{R}_{i,1}]$. Focus on any type vector $\boldsymbol{x} \neq (2, 2, 2, \ldots, 2)$, i.e., a vector that has at least one individual of type 1. Because $\tau(f_\alpha(1)) > \tau(f_\alpha(2))$, and r^{τ}_{opt} is utility-maximizing for the utility map τ, $r^{\tau}_{\text{opt}}(f_\alpha(\boldsymbol{x}))$ must rank all individuals of type 1 (of whom there is at least one) ahead of all individuals of type 2. From this, we obtain that $\sum_{i=1}^{n} \mathbb{1}_{x_i \in S} \cdot \mathbb{P}_{r^{\tau}_{\text{opt}}(f_\alpha(\boldsymbol{x}))}[\mathcal{R}_{i,1}] = 1$ for all $\boldsymbol{x} \neq (2, 2, 2, \ldots, 2)$.

Next, we write out the expectation from Equation (6). We use that the terms $\mathbb{1}_{x_i \in S} = 0$ for all i when $\boldsymbol{x} = (2, 2, \ldots, 2)$, which allows us to drop this term from the sum. We then use that each $\boldsymbol{x}$ under the i.i.d. uniform type distribution is drawn with probability 2^{-n}, and substitute our preceding calculations for the probabilities. This gives us the following:

$$
\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim\text{Unif}([n])}\left[\mathbb{1}_{x_i\in S}\cdot\left(\mathbb{P}_{r^{\tau}_{\text{opt}}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,1}] - \mathbb{P}_{r^{\tau}_{\text{opt}}(f(\boldsymbol{x}))}[\mathcal{R}_{i,1}]\right)\right]\right| \\
&\quad = \left|\sum_{\boldsymbol{x}}\sum_{i=1}^{n} \Pr_{\boldsymbol{x}\sim\mathcal{D}^n}[\boldsymbol{x}]\cdot\frac{1}{n}\cdot\mathbb{1}_{x_i\in S}\cdot\left(\mathbb{P}_{r^{\tau}_{\text{opt}}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,1}] - \mathbb{P}_{r^{\tau}_{\text{opt}}(f(\boldsymbol{x}))}[\mathcal{R}_{i,1}]\right)\right| \\
&\quad = \left|2^{-n}\cdot\frac{1}{n}\cdot\sum_{\boldsymbol{x}\neq(2,2,\ldots,2)}\left(\sum_{i=1}^{n}\mathbb{1}_{x_i\in S}\cdot q_i - \sum_{i=1}^{n}\mathbb{1}_{x_i\in S}\cdot\mathbb{P}_{r^{\tau}_{\text{opt}}(f(\boldsymbol{x}))}[\mathcal{R}_{i,1}]\right)\right| \\
&\quad = 2^{-n}\cdot\frac{1}{n}\cdot\left|\sum_{\boldsymbol{x}\neq(2,2,\ldots,2)}\left(\left(\sum_{i=1}^{n}\mathbb{1}_{x_i\in S}\cdot q_i\right) - 1\right)\right| \\
&\quad = 2^{-n}\cdot\frac{1}{n}\cdot\left|\left(\sum_{i=1}^{n} q_i\cdot\sum_{\boldsymbol{x}\neq(2,2,\ldots,2)}\mathbb{1}_{x_i\in S}\right) - (2^n - 1)\right| \\
&\quad \overset{(\star)}{=} 2^{-n}\cdot\frac{1}{n}\cdot\left|\left(\sum_{i=1}^{n} q_i\cdot 2^{n-1}\right) - (2^n - 1)\right| \\
&\quad = 2^{-n}\cdot\frac{1}{n}\cdot\left|2^{n-1} - (2^n - 1)\right| \\
&\quad = \frac{1}{n}\cdot\left(\frac{1}{2} - 2^{-n}\right).
\end{aligned}
$$

In the step labeled $(\star)$, we used that there are exactly 2^{n-1} vectors with $x_i = 1$; the following step used that the q_i, defining a probability distribution, sum to 1. This completes the proof. □

F. Stability-Utility and Fairness-Utility Tradeoffs

In both theory and practice, it is often necessary to trade off utility against other desiderata, such as fairness or stability (see, for example, Singh & Joachims (2019); Pitoura et al. (2022)): if achieving fairness/stability comes at a huge price in utility, it may become economically infeasible to implement fair or stable rankings. In this section, we introduce and discuss a class r^{ϕ}_{mix} of ranking functions which provide a quantifiable tradeoff between the objectives. r^{ϕ}_{mix} linearly interpolates between r_{UA} and r^{τ}_{opt} with a trade-off parameter $\phi \in [0, 1]$, chosen by the ranking/mechanism designer. We show that this interpolation naturally leads to r^{ϕ}_{mix} satisfying *approximate* stability and fairness, while providing lower-bound guarantees on the utility. While the proofs are relatively straightforward, we believe that practitioners may find this class of ranking functions useful in practical applications where the stringent requirement of 1-stability may not be necessary. The interested

reader is referred to Singh et al. (2021) for additional discussion on how to choose ϕ appropriately[7].

We first formally define the notion of *approximate* stability.

Definition 29. *Fix n and L. A ranking function $r : \mathcal{P}_{n,L} \to \mathcal{M}_{DS}^{n\times n}$ is* (γ, α)-approximately stable *if $||r(P) - r(P')||_\infty \leq \gamma \cdot ||P - P'||_1 + \alpha$ for all predictions $P, P' \in \mathcal{P}_{n,L}$.*

Notice that $(\gamma, 0)$-approximate stability recovers our stability notion from Definition 3. Approximate stability is a relaxation which allows for additive slack in the dependence on $\|P - P'\|_1$. An additive slack relaxation is natural and akin to, for example, (ϵ, δ)-differential privacy (when compared to pure differential privacy). We show that a simple mixture of UA and the optimal utility ranking satisfies the following approximate stability and utility guarantee.

Proposition 30. *Fix a utility map τ. Let r_{mix}^{ϕ} be the ranking function which randomizes between r_{UA} (with probability ϕ), and r_{opt}^{τ} (with probability $1-\phi$). Then r_{mix}^{ϕ} is $(\phi, 1-\phi)$-approximately stable. Furthermore, for any $P \in \mathcal{P}_{n,L}$, we have that $U(P, r_{mix}^{\phi}) = \phi \cdot U(P, r_{UA}) + (1-\phi) \cdot U(P, r_{opt}^{\tau})$.*

Proof. We first show approximate stability. For any $P, P' \in \mathcal{P}_{n,L}$, we have the following.

$$\begin{aligned}
\|r_{\text{mix}}^{\phi}(P) - r_{\text{mix}}^{\phi}(P')\|_\infty &= \|\phi r_{\text{UA}}(P) + (1-\phi) r_{\text{opt}}^{\tau}(P) - \phi r_{\text{UA}}(P') - (1-\phi) r_{\text{opt}}^{\tau}(P')\|_\infty \\
&\leq \phi \|r_{\text{UA}}(P) - r_{\text{UA}}(P')\|_\infty + (1-\phi)\|r_{\text{opt}}^{\tau}(P) - r_{\text{opt}}^{\tau}(P')\|_\infty \\
&\leq \phi \|P - P'\|_1 + (1-\phi).
\end{aligned}$$

The last line used the 1-stability of r_{UA} (proved in Theorem 11), as well as the fact that the $\|\cdot\|_\infty$-norm difference of doubly stochastic matrices is at most 1.

The claim about utility is simply linearity of expectations. □

It is straightforward to show that a similar approximate fairness guarantee holds for r_{mix}^{ϕ}, again due to its linearity.

Proposition 31. *Let $\mathcal{D}$ be a distribution over individuals $\mathcal{X}$. Let f^* be the ground truth distribution of labels, and f a predictor. For any n, let $\mathcal{D}^n$ be the distribution obtained from drawing a vector of n i.i.d. samples from $\mathcal{D}$.*

Let $\mathcal{C}$ be a collection of sets with $\mathcal{X} \in \mathcal{C}$, the sets of individuals for which the predictor f will be assumed to be multiaccurate/multcalibrated. Let $\alpha \geq 0$ be a parameter for how far from fully accurate/calibrated the predictor is allowed to be.

1. *If f is $(\mathcal{C}, \alpha)$-multiaccurate, then the following holds for all sets $S \in \mathcal{C}$ and $k \in [n]$:*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim Unif([n])}\left[\mathbb{1}_{x_i\in S}\cdot\left(\mathbb{P}_{r_{mix}^{\phi}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{mix}^{\phi}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}]\right)\right]\right| \leq \phi L n\alpha + 1 - \phi.$$

2. *Let some interval width $\delta \in (0,1]$ be given, such that $1/\delta$ is an integer. If f is $(\mathcal{C}, \alpha, \delta)$-multicalibrated and $(\{\mathcal{X}\}, \alpha)$-multiaccurate, then for every set $S \in \mathcal{C}$, vector $(j_1, j_2, \ldots, j_L) \in \{0, 1, \ldots, 1/\delta - 1\}^L$, and $k \in [n]$:*

$$\left|\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^n, i\sim Unif([n])}\left[\mathbb{1}_{x_i\in S}\cdot\mathbb{1}_{f(x_i)_\ell\in[j_\ell\cdot\delta,(j_\ell+1)\cdot\delta)\text{ for all }\ell}\cdot\left(\mathbb{P}_{r_{UA}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{UA}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}]\right)\right]\right| \leq \phi L n\alpha + 1 - \phi.$$

Proof. The proof follows from the following computation due to linearity of r_{mix}^{ϕ}:

$$\begin{aligned}
\mathbb{P}_{r_{\text{mix}}^{\phi}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,k}] - \mathbb{P}_{r_{\text{mix}}^{\phi}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}] &= \phi\cdot\mathbb{P}_{r_{\text{UA}}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,k}] + (1-\phi)\cdot\mathbb{P}_{r_{\text{opt}}^{\tau}(f^*(\boldsymbol{x}))}[\mathcal{R}_{i,k}] \\
&\quad - \left(\phi\cdot\mathbb{P}_{r_{\text{UA}}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}] + (1-\phi)\cdot\mathbb{P}_{r_{\text{opt}}^{\tau}(f(\boldsymbol{x}))}[\mathcal{R}_{i,k}]\right)
\end{aligned}$$

Applying this decomposition, then applying linearity of expectation, applying the triangle inequality, and then using Theorem 15 completes the proof of both parts. □

[7] We note that our approximate fairness guarantee in Proposition 31 on multiaccuracy/multicalibration with an additive slack is different from the ϕ-approximate fairness of Singh et al. (2021), which is a multiplicative notion. Indeed, both hold simultaneously for r_{mix}^{ϕ}.

G. Experimental Results and Details

We ran experiments on the US census data set `ACS` curated by Ding et al. (2021) and the student dropout task `Enrollment` introduced by Martins et al. (2021) in the UCI data set Repository.

In our experiments, we demonstrate empirically that the stability guarantees of UA rankings hold when using multiclass predictions. Furthermore, we find that in practice, the stability guarantees offered by UA ranking may be much better than 1-stability (or the 1/2-stability worst-case lower bound in Proposition 12), and show that the utility loss suffered by r_{UA} is reasonable. Although our experiments are relatively simplistic and are not the main focus of our work, they demonstrate that UA rankings have relatively good performance in terms of utility: they outperform not only a uniformly random baseline ranking, but even the Plackett-Luce distribution. At the same time, they also retain the provable fairness and stability properties.

Related Experiments. Previous work (Devic et al., 2023; Singh et al., 2021) also contain experiments demonstrating the utility and utility-fairness tradeoff of UA ranking functions. This past work assumed real-valued predictions (as opposed to the multiclass predictions in our work). Singh et al. (2021) actually deployed a paper recommendation system at a large computer science conference using UA rankings to demonstrate the viability of the method in practice. Furthermore, their experiments on the `MovieLens` data set (Harper & Konstan, 2015) demonstrate that UA ranking — corresponding to a fairness parameter of $\phi = 1$ in their work — can achieve nearly 99% of the optimal utility given by r^{τ}_{opt} in some applications. Devic et al. (2023) show the viability of UA rankings in a matching setting, running experiments on an online dating data set.

G.1. Stability Against SGD Noise in Neural Network Training

Given our focus on the combination of ranking functions with noisy predictions derived from ML-based classifiers, we first investigate experimentally the stability of UA and utility-maximizing rankings under a natural model of prediction noise. In particular, one of the most common sources of noise in predictions is the randomness in the training procedure (such as SGD). We designed a natural experiment by comparing the behavior of ranking functions under predictors learned with different randomly seeded SGD training runs. Our focus is on understanding if or to what extent the stability of UA and other ranking functions will exceed the worst-case theoretical guarantees in such quasi-realistic settings.

First, we describe the data sets. In `ACS`, the prediction target is the binary variable of whether a person is employed or not (after filtering to individuals in the age range 16–90). For computational reasons, we restrict our experiments to a subset of the data for California with parameters `survey year`='2018', `horizon`='1-Year' and `survey`='person'. These parameters are standard when using `ACS` for testing algorithmic fairness methods, due to the large amount of available data. (See, e.g., the GitHub repository of Ding et al. (2021).) We are left with 378,817 entries, and use an 80/20 train/test split. In `Enrollment`, the target is a multiclass variable for whether an individual is an enrolled, graduated, or dropout student. After cleaning the data, we are left with 4,424 entries, on which we use an 80/20 train/test split.

Since we want to compare the stability of r_{UA} against that of r^{τ}_{opt}, we next define simple and natural utilities. For `ACS`, we take class 2 to correspond to employment, and class 1 to correspond to unemployment (class 2 $\succ$ class 1). We define $\tau(\boldsymbol{p}) = p_2$, i.e., the probability of employment. For `Enrollment`, we take class 1 to be that the student has dropped out, class 2 to be enrolled, and class 3 to be graduated (class 3 $\succ$ class 2 $\succ$ class 1), and define $\tau(\boldsymbol{p}) = p_1 + 2 \cdot p_2 + 3 \cdot p_3$.

We train 30 simple three-layer MLP neural networks on the `ACS` data set, which we divide into 15 pairs of networks. Each pair of networks is initialized with the same (random) weight matrix, then trained separately with SGD. This introduces noise into the final trained neural network weights, and consequently the predictions. That is, each pair of networks has similar test accuracy, but will output different probabilities on some (or all) individuals. On `ACS`, all networks achieve between 75–80% train and test accuracy. We perform the identical procedure for the `Enrollment` data set, where the networks all achieve between 70–75% train and test accuracy (due to less data being available).

Let (f_i, g_i), for $i = 1, \ldots, 15$, be the classifiers corresponding to a given pair of networks trained from the same initialization but with different noise due to SGD. To test the stability of the ranking functions r_{UA} and r^{τ}_{opt}, for each pair (f_i, g_i), we randomly select 30 individuals from the test set as the data set of individuals $\boldsymbol{x}$, and obtain the (probabilistic) predictions $P = f_i(\boldsymbol{x})$ and $P' = g_i(\boldsymbol{x})$. We then run r^{τ}_{opt} and r_{UA} on these two prediction matrices, logging the deviations of the rankings and the resulting value of $\|P - P'\|_1$. For each pair of networks (f_i, g_i), we repeat this procedure 10 times with different randomly selected subsets of individuals. In Table 1, we report the average and standard deviation of this experiment.

The results in Table 1 demonstrate that $\|P - P'\|_1$ dominates the value of $\|r_{\text{UA}}(P) - r_{\text{UA}}(P')\|_\infty$; this behavior persisted through all of our many training runs. We conclude that UA rankings are extremely stable in the face of noise introduced during the learning of a predictor (drastically surpassing our 1-stability bound). Our results also confirm that instability of the optimal ranking function r^τ_{opt} is not only a theoretical possibility, but prevalent when working with real data. To see this, notice that the mean value of $\|r^\tau_{\text{opt}}(P) - r^\tau_{\text{opt}}(P')\|_\infty$ is two orders of magnitude larger than for r_{UA}. For the `Enrollment` data set, it even exceeds the mean value of $\|P - P'\|_1$, implying that $\gamma \leq 1$ is impossible. For the `ACS` data set, the norms are very comparable, meaning that $\gamma \ll 1$ is impossible; in fact, consistent with the large standard deviations, there are multiple instances illustrating that γ must be significantly larger than 1 for both data sets.

G.2. Utility

Next, we measure the utility attained by the different ranking functions. The utility map τ for both data sets is the same one as in Appendix G.1. In addition to the two rankings functions of primary interest, we also consider the following two baselines: (1) r_{unif}, the ranking function which places the n individuals in uniformly random order; and (2) r_{PL}, the Plackett-Luce (PL) ranking defined by Luce's axiom (Plackett, 1975; Luce, 1959).

The PL model, similar to UA, defines a distribution over rankings. At a high level, in each iteration i, the item for the ith position is sampled based on a softmax mapping of all remaining items' relevance scores. More precisely, in each iteration t, let M_t be the set of individuals not yet placed in the ranking, with $M_1 = [n]$ in the first iteration. Then, in each iteration $t = 1, \ldots, n$, individual $i \in M_t$ is placed in position t with probability

$$\mathbb{P}[\mathcal{R}_{i,t}] = \frac{\exp\left(\tau(\boldsymbol{p}_i)\right)}{\sum_{j \in M_t} \exp\left(\tau(\boldsymbol{p}_j)\right)}.$$

To efficiently compute the PL ranking, we use the (now standard) Gumbel trick from Bruch et al. (2020). That is, to sample one ranking from the PL ranking distribution $r_{\text{PL}}(P)$, we sort the individuals in decreasing order of $\tau(\boldsymbol{p}_i) + \gamma_i$, where each $\gamma_i \sim \text{Gumbel}(0, 1)$ independently. We average over 100k samples from the PL ranking distribution to compute $r_{\text{PL}}(P)$.

To measure utility, we use the DCG position weights $w_k = 1/\log_2(k+1)$. In order to make the scales of the utilities more meaningful in our comparisons, we normalize all utilities to lie in $[0, 1]$. Thereto, let r^τ_{min} be the worst-utility ranking, obtained by ordering the individuals by increasing relevance score (i.e., the individual of lowest utility is deterministically placed first). For a ranking function r, we compute the normalized utility score as follows:

$$\tilde{U}(P, r) = \frac{r(P) - r^\tau_{\text{min}}(P)}{r^\tau_{\text{opt}}(P) - r^\tau_{\text{min}}(P)}.$$

In Table 2, we report the mean and standard deviation over 30 neural network training runs of the normalized utility $\tilde{U}$ for each of the ranking functions discussed above. For each neural network and associated prediction function f, we randomly sample $n = 20, 40$, and 60 individuals from the test set. Then, we construct the prediction matrix $P = (f(x_i))_{i \in [n]}$, and report $\tilde{U}(P, r)$ for each ranking function r. We find that UA ranking outperforms the uniform and PL ranking in each experimental instance. However, UA ranking is *not* guaranteed to always outperform the uniform ranking. One can carefully construct instances in which the "safe bet" individual provides more utility than an individual who has a low probability of being a "moonshot" candidate (further discussed in Singh et al. (2021)). Such an instance crucially depends on the specific choice of τ.