2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)

Learning from Outcomes: Evidence-Based Rankings

Cynthia Dwork — Harvard University
Michael P. Kim, Omer Reingold — Stanford University
Guy N. Rothblum, Gal Yona — Weizmann Institute of Science

***Abstract*—Many selection procedures involve ordering candidates according to their qualifications. For example, a university might order applicants according to a perceived probability of graduation within four years, and then select the top 1000 applicants. In this work, we address the problem of *ranking* members of a population according to their "probability" of success, based on a training set of historical binary outcome data (e.g., graduated in four years or not). We show how to obtain rankings that satisfy a number of desirable accuracy and fairness criteria, despite the coarseness of the training data. As the task of ranking is global (the rank of every individual depends not only on their own qualifications, but also on every other individuals' qualifications), ranking is more subtle and vulnerable to manipulation than standard prediction tasks.**

Towards mitigating unfair discrimination caused by inaccuracies in rankings, we develop two parallel definitions of *evidence-based* rankings. The first definition relies on a semantic notion of *domination-compatibility*: if the training data suggest that members of a set S are more qualified (on average) than the members of T, then a ranking that favors T over S (where T *dominates* S) is blatantly inconsistent with the evidence, and likely to be discriminatory. The definition asks for domination-compatibility, not just for a pair of sets, but rather for every pair of sets from a rich collection $\mathcal{C}$ of subpopulations. The second definition aims at precluding even more general forms of discrimination; this notion of *evidence-consistency* requires that the ranking must be justified on the basis of consistency with the expectations for every set in the collection $\mathcal{C}$. Somewhat surprisingly, while evidence-consistency is a strictly stronger notion than domination-compatibility when the collection $\mathcal{C}$ is predefined, the two notions are equivalent when the collection $\mathcal{C}$ may depend on the ranking in question.

***Index Terms*—ranking; prediction; algorithmic fairness;**

I. INTRODUCTION

Since its inception as a field of study roughly one decade ago [1]–[4], research in algorithmic fairness has exploded, especially in the machine learning community [5]–[13]. Much of this work focuses on so-called "group fairness" notions, which address the relative treatment of different demographic groups. More theoretical work has advocated for "individual fairness" which, speaking intuitively, requires that people who are similar, with respect to a given classification task, should be treated similarly by classifiers for that task. Both approaches face significant challenges: group notions provide notoriously weak protections to individuals and are provably incompatible with one another; individual fairness requires task-specific similarity information for every pair of individuals, which may be unavailable. The past two years have seen exciting developments on several fronts in theoretical computer science that strive to bridge the gap between group and individual notions of fairness for the tasks of scoring, classifying, and auditing [14]–[18].

In this work, we turn our attention to fairness when *ranking* individuals based on the perceived probability of an outcome. Rankings are of interest for several reasons. First, ranking is at the heart of triage, say, in disaster relief. Second, ranking is often the underlying impetus for scoring, for example in university admissions. Third, some approaches to affirmative action involve stratifying the population according to some criterion, *e.g.*, high school (as done in California and Texas) or education level of mother [19]. Students within each stratum are ranked by grades (in California and Texas), or hours spent on homework [19], and the top-ranked students from each stratum are admitted. Fourth, studying ranking informs our understanding of what we should demand of a scoring function.

Note that in the above examples, grades and hours spent on homework are proxies for qualities that are difficult to articulate and even more difficult to measure. They may capture intuition about "probability" or "chance" of (say) graduation within 4 years, but the meaning of an individual probability has long been debated (see [20][1] and the references therein as well as the discussion in Section I-B).

In this work, we will not assume access to individual probabilities, even in the training data. Rather, we follow the approach taken for scoring functions initiated in [14] and rely for training only on 0/1 outcome data (*e.g.*, did, or did not, graduate within 4 years). In other words, even if we posit the existence of a scoring function p^* mapping each individual x to its "true probability" $p^*(x)$ of a positive outcome, these probabilities can be accessed only indirectly, e.g. by computing outcome statistics based on observational data. Note that, even ensuring evidence-consistent treatment for a relatively large and homogeneous set of individuals, all sharing a known value $p^*(x) = v$, may be impossible without knowledge about the rest of the population, as their rank may vary dramatically based on the p^* values outside the set. Despite this challenging setup, we develop definitions and methods for powerful protection against unfair discrimination.

A. Contributions and Results

Occam's Razor for Rankings. In general, no two scoring functions p^* and $\tilde{p}$ that are statistically close can be distinguished based on a small sample of outcomes, and it

---

[1] Written in response to the use of machine learning to estimate recidivism risk.

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00016

is easy to think of examples where obtaining an accurate ranking is beyond reach. For example, if half of the individuals receive a positive outcome with probability 1, but this set is computationally indistinguishable from its complement, where individuals receive a positive outcome with probability 0, then coming up with an accurate ranking will be computationally infeasible. Computational considerations aside, if the partition between 1's and 0's were truly random, then learning an accurate ranking from a bounded training sample would be information-theoretically impossible. It is, thus, natural to define an accurate ranking to be one where individuals are ranked by their values according to a function $\tilde{p}$ that is statistically close to p^*, as this is the best we can hope for.

Our first result is an Occam's Razor Theorem for (agnostically) learning rankings. We show that, given a class R of rankings, there is an algorithm that, given a sample of size growing as $\log(|R|)$, returns an approximately optimal ranking. (The running time depends polynomially on the size of R.) The proof of this theorem is qualitatively very different from the standard Occam's Razor Theorem for PAC learning: standard proofs of Occam's Razor-style results evaluate the "quality" of each hypothesis separately based on the data, and argue that a hypothesis with maximum quality is an approximate optimizer. We argue that any such approach, which considers the quality of each ranking on its own, will fail in our setting. Consider the example from above, in which half of the population has $p^*(x) = 1$ and the other half has $p^*(x) = 0$. In this case, a random ranking could disguise itself as being accurate: an observer who only sees the binary outcome data cannot distinguish the situation in which everyone is either a 0 or 1 from one consistent with $p(x) = 1/2$. Under such a p, a random ranking will appear as accurate as any other ranking, despite its inconsistencies with p^*. Instead, our proof relies on the accurate ranking revealing the inaccuracy of other rankings. Quantitatively, the theorem is different from, and a bit weaker than, the analogue for PAC learning (and we prove that this is unavoidable).

Protecting Groups. Learning rankings that are highly accurate for most individuals may be computationally or information-theoretically infeasible, and mis-ranked individuals may experience harmful outcomes. Thus, individual fairness is impossible in the setting considered by this paper, and we focus on protecting a large collection of intersecting groups (sets) of individuals. As noted above, even for a large and homogeneous set, we cannot reason about the fairness of its members' rankings in isolation. For example, suppose all members of a set S have p^* value $1/3$. Outside S, two homogeneous sets T_1 and T_0 have members with p^* values of 1 and 0, respectively. A scoring rule $\tilde{p}$ that is perfect on S, but assigns the value $1/2$ to all the members of $T_1 \cup T_0$, would still induce a ranking that dramatically downplays the fitness of S. This potential harm to the members of S cannot be detected or reasoned about without considering the outcomes of individuals outside the set S.

Despite these considerable challenges, we develop definitions and methods that provide powerful protection against unfair discrimination. Our starting point is to focus on the relative treatment of members of pairs of sets in the collection. For a simple example, consider a pair S, T of disjoint sets, and suppose that, empirically, a larger fraction of the members of S have positive outcomes (graduate within 4 years) than than do the members of T. Then a ranking that puts all elements of T (the less successful group) ahead of all those in S would be considered "unfair." In the above example, even ranking all the members of S below all the members of T_0 is unfair. Our goal is much more ambitious than defeating this simple example: we require our rankings to be simultaneously fair (defined formally below) for all pairs of groups defined by a rich collection $\mathcal{C}$ of possibly-intersecting subsets of the population.

The choice of sets in $\mathcal{C}$ is an important one, as the fairness conditions will not be guaranteed to apply to sets not in $\mathcal{C}$. But how can we ensure *awareness* of which groups should be included? Even well-intentioned algorithm designers can be ignorant of some types of discrimination, the number of potentially relevant categories may be daunting[2], and members of an historically-oppressed group may have internalized the negative stereotypes and not see their treatment for the oppression it is [22]. For these and other reasons, such as lack of resources and power, it is inappropriate to expect members of an oppressed group S to insist that S be included in $\mathcal{C}$. Our approach follows in the footsteps of [14] and follow-up works in defining sets from a complexity-theoretic perspective. As the examples above indicate, we may fail to protect sets that we cannot efficiently identify (e.g. the set of ones that are randomly mixed with the set of zeros). A natural goal that we adopt here is to protect every set that we *can* identify with some given computational resources (e.g., sets that can be defined with a small decision tree or by circuits of a given size) in accordance with two fairness notions. The key technical requirement is that the sets be fixed in advance and membership in every set can be computed from an individual's data.

Domination-Compatibility. We construct two parallel notions of increasingly strong fairness requirements. *Domination-Compatibility* aims to preclude rankings in which a qualified group is consistently undervalued in the ranking when compared to another, less qualified group. We formalize the situation where a ranking favors one group over another via the notion of *domination*. For a given ranking and equal-size sets S and T, we say that S *dominates* T if there exists a matching between S and T in which every member of S is matched to a person in T whose rank is worse. In fact, we will work with a more general notion that allows for approximate domination between sets of different sizes (Definition IV.2).

A ranking that does not exhibit this type of unfair behavior for *any* pair of sets in a class $\mathcal{C}$ is said to be $(\mathcal{C}, \alpha)$-domination-compatible.

---

[2] Social psychologist Claude Steele writes, "There exists no group on earth that is not negatively stereotyped in some way – the old, the young, northerners, southerners, WASPs, computer whiz kids, Californians, and so forth." [21].

Definition (Domination-Compatibility, informal)**.** *We say that a ranking is* $(\mathcal{C}, \alpha)$*-domination compatible if for every two subsets* $S, T \in \mathcal{C}$*:*

If S *dominates* T*, then* $\mathbf{E}_{x \sim \mathcal{D}_S}[p^*(x)] + \alpha \geq \mathbf{E}_{x \sim \mathcal{D}_T}[p^*(x)]$

where $x \sim \mathcal{D}_s$ *denotes a random unlabeled sample from the distribution on universe elements, conditioned on* $x \in S$*, and analogously for* $\mathcal{D}_T$*.*

The formal definition also accounts for approximate domination (see Definition IV.4). An important advantage of this definition is that it can both be obtained and audited from labeled data, as it only considers expectations of p^* on sets in $\mathcal{C}$.

Evidence-Consistency. *Evidence-Consistency* is specified in terms of a ranking's consistency with a scoring function that satisfies increasingly demanding *accuracy* conditions. Here, accuracy is specified with respect to expectations of the 0/1 outcomes data in a training set, which we think of as the "evidence". Thus, we require that our rankings be consistent with the evidence.

Definition (Evidence-Consistency, informal)**.** *A ranking is* $(\mathcal{C}, \alpha)$*-evidence-consistent if there exists a scoring function* $\tilde{p}$ *that is consistent with the ranking, and for which the following holds:*

$$\forall S \in \mathcal{C}, \quad \left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[p^*(x)] - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\tilde{p}(x)] \right| \leq \alpha.$$

Importantly, we show that the global consistency guarantee provided by evidence-consistency is a more powerful guarantee that implies the pairwise protections provided by domination-compatibility.

Theorem 1 (Evidence-Consistency implies Domination Compatibility, informal)**.** *If a ranking is* $(\mathcal{C}, \alpha)$*-evidence-consistent, then it is also* $(\mathcal{C}, 2\alpha)$*-domination-compatible.*

In fact, evidence-consistency is strictly stronger than domination-compatibility. In Section IV-D, we demonstrate that there are choices of $\mathcal{C}$ and α, such that there exist rankings that are $(\mathcal{C}, 0)$-domination-compatible but are not $(\mathcal{C}, \alpha)$-evidence-consistent. Further, we prove that domination-compatibility is equivalent to a significantly weaker notion of pairwise-consistency. On an intuitive level, domination-compatibility can be justified by a separate explanation for every pair of sets; evidence-consistency, however, requires a single explanation that simultaneously justifies the rankings of all sets.

Strengthening the protections through self-reference. Somewhat surprisingly, we show that even for a rich collection $\mathcal{C}$, evidence-consistency (and thus also domination-compatibility) can leave the door open to harms that directly affect sets included in the collection $\mathcal{C}$, including harms that can be audited from labeled data. As an example, let S be a set of individuals whose p^* values are all 0.8, whereas outside of S all individuals have p^* value 0.5. If we rank according to p^*, then the individuals in S should be ranked highest. We argue that for *any choice of* $\mathcal{C}$ (comprised of sufficiently large sets), there exists a blatantly unfair ranking that is $\mathcal{C}$-evidence-consistent and that systematically degrades the ranks in S.

To see this, consider a scoring function $\tilde{p}$ constructed by assigning the correct value of 0.8 to every individual in S, and for every individual $x \notin S$ sampling an outcome uniformly at random from $\{0, 1\}$. With high probability, $\tilde{p}$ has accurate expectations for all sets in $\mathcal{C}$. Thus, the ranking $r^{\tilde{p}}$ induced by $\tilde{p}$ is evidence-consistent; however, this ranking harms the members of S, who receive median ranks rather than being ranked at the top. Moreover, this harm is demonstrable from the data: if we consider the set T of individuals ranked above the members of S, we see that this set dominates S in the ranking, even though the expectation of its labels is lower! The issue is that the set T is only defined *a posteriori*, after we are given the ranking.

Motivated by this example, we strengthen both domination-compatibility and evidence-consistency by asking that they hold for a richer family of sets *defined by the ranking under consideration*. Intuitively, once a ranking is proposed, the sets that are implied by these rankings – sets of individuals that are identically ranked, which we will refer to as quantiles – become relevant. Furthermore, the quantiles within every set in $\mathcal{C}$ are also relevant.

For a collection $\mathcal{C}$ and a ranking r, we consider an augmented collection of subsets $\mathcal{C}_r$; loosely and informally, $\mathcal{C}_r$ includes the quantiles induced by the ranking r, and the intersections of each of these quantiles with every set in $S \in \mathcal{C}$. Definition V.4 provides a formal treatment. Returning to the example above, once the ranking $r^{\tilde{p}}$ is suggested, sets related to T will appear in $\mathcal{C}_r$. Thus, asking for domination-compatibility or evidence-consistency with respect to $\mathcal{C}_r$ (rather than $\mathcal{C}$), yields stronger notions of *reflexive* domination-compatibility and *reflexive* evidence-consistency (respectively). While the weaker (non-reflexive) notions were *not* equivalent to one another, the reflexive notions are equivalent!

Theorem 2 (Equivalence of Reflexive Notions)**.** *If a ranking is* $(\mathcal{C}, \alpha)$*-reflexive-evidence-consistent then it is* $(\mathcal{C}, 2\alpha)$*-reflexive-domination-compatible. If a ranking is* $(\mathcal{C}, \alpha)$*-reflexive-domination-compatible then it is* $(\mathcal{C}, \alpha)$*-reflexive-evidence-consistent.*

Learning Evidence-Consistent Rankings. Generally speaking, we can learn an evidence-consistent ranking by directly learning the scoring function required in the definition, and using the ranking that it induces. For (non-reflexive) evidence-consistency, this entails learning a function $\tilde{p} : \mathcal{X} \to [0, 1]$ that (approximately) respects all of the expectations of subsets $S \in \mathcal{C}$. The task of learning such a function has been recently studied in the context of fair prediction [14], [18], these works show how to learn such a $\tilde{p}$ from a small number of binary samples.

Reflexive evidence-consistency, however, requires the existence of a scoring function that respects all of the expectations of subsets in $\mathcal{C}_r$. This collection is defined *adaptively*: the sets are only defined after a particular ranking r is specified. The aforementioned algorithm only works for a family of sets that are fixed in advance, and thus it cannot be directly applied towards the stronger definition.

Instead, we turn our attention to the stronger notion of *multi-calibration* studied by [14]. Loosely, a function $\tilde{p}$ is multi-calibrated for a collection $\mathcal{C}$ if it is calibrated on every set in $\mathcal{C}$. Calibration, which has been well studied in the statistics literature, where weather forecasting is often a driving example, says that the fraction of positive outcomes among those elements assigned a score of $v \in [0,1]$ is equal to v, for all v simultaneously. We show that multi-calibration and reflexive evidence-consistency are closely related.

Theorem 3 (Connection to multi-calibration, informal)**.** *The ranking induced by a* $(\mathcal{C},\alpha)$*-multi-calibrated function is* $(\mathcal{C},\alpha)$*-reflexive-evidence-consistent. Further, any consisten scoring function that exhibits the correct expectations defined by a* $(\mathcal{C},\alpha)$*-reflexive-evidence-consistent ranking is statistically close to being* $(\mathcal{C},\alpha)$*-multi-calibrated.*

We develop this result formally in Section V-C. Leveraging the connection to multi-calibration, we can use known algorithms for learning multi-calibrated functions to obtain reflexively evidence-consistent rankings from labeled data. For arbitrary collections $\mathcal{C}$, the learning algorithms of [14] run in polynomial time in $|\mathcal{C}|$; however, for structured, agnostically-learnable, collections $\mathcal{C}$, the running time may be improved, depending on the efficiency of the agnostic learner. This theorem gives further motivation for learning multi-calibrated scoring functions in the context of predictions.

Stronger Notions Yet? The reader may wonder if there are natural notions in the evidence-consistency and domination compatibility hierarchies. Stronger notions could always exist and exploring them is an excellent direction for further research; nevertheless, we note that we do not expect to see examples demonstrating weaknesses of reflexive evidence-consistent rankings of the sort that we demonstrated for plain evidence-consistent rankings. This is because, for a sufficiently rich family of sets, reflexive evidence-consistent rankings will be highly accurate (in the sense discussed in the context of the Occam's Razor theorem). In other words, it is natural to expect that any weakness of reflexive evidence-consistency would exploit a weakness in the family of sets $\mathcal{C}$, and would fail for a sufficiently-rich family $\mathcal{C}$.

B. Discussion

The Meaning of Probabilities. As discussed above, the notion of an individual probability $p^*(x)$ is debatable. Still, assuming some underlying scoring function p^* is a useful (and common) abstraction that aims at capturing some underlying uncertainty. We therefore follow tradition and specify our definitions based on an hypothesized p^*. We stress that all but one of our results hold if we replace p^* with the function o that assigns to individual x its outcome $o(x)$.[3]

An insight exciting to us is the perspective that the notion of Evidence-Consistency gives into the idea of p^*. Consider outcomes that are completely deterministic – half of the individuals will see a positive outcome with probability 1 and the rest will see it with probability 0. If the set of 1's is computationally indistinguishable from a uniform set, then, we argue, it is legitimate to view p^* as assigning all individuals the value $1/2$. But what if we have richer information specifying all of the expectations for a family of sets $\mathcal{C}$? By analogy to the preceding argument, any multi-calibrated scoring function $\tilde{p}$ is a legitimate candidate for the role of p^*. Thus, any evidence-consistent ranking may legitimately be considered "accurate." So even if individual probabilities are always beyond reach (when only given a sample of outcomes), we can still assign putative individual probabilities that respect a rich body of evidence.

The Choice of $\mathcal{C}$ in Light of Evidence-Consistency. Fairness, as specified in our framework, ultimately hinges on the expressive power of the sets in $\mathcal{C}$, which relies in turn on the richness of the individual data and the computational resources. To see this, consider disjoint sets of students S and T, where the students in S attend a wealthy high school and the students in T attend an impoverished school. Members of S may have access to advanced placement (AP) classes, whereas members of T may not. Thus, it may be impossible for AP-capable students in T to demonstrate their ability to excel in advanced courses. Even multi-calibration does not necessarily guarantee equal discriminative capability on S and T: there is no way for an algorithm to extract information that is not present in the data.

It is possible that we could define (if not always efficiently measure) the inadequacy of the expressive power of $\mathcal{C}$, from the perspective of this work. For example, given a ranking $\tilde{p}$ we can define, for each $S \in \mathcal{C}$, $V_S(\tilde{p})$ to be the fraction of members of S whose rank is in the top ten per cent. If the value of $V_S(\tilde{p})$ varies greatly on a pair of evidence-consistent rankings, then the evidence – as interpreted via the sets in $\mathcal{C}$ – is not reliably capturing the qualifications of the members of S: different rankings consistent with the evidence yield very different values. The variability of the set of the space of evidence-consistent rankings is closely tied to the legitimacy of viewing a multi-calibrated $\tilde{p}$ as a vector of true probabilities.

Ranking versus Predicting. In many settings, a position within a ranking is as useful as a score. For example, an experienced clinician can translate a claim that a patient is in the top 10% among the population at risk for developing a given ailment into an absolute estimate of this risk. This leads to an important observation: a ranking together with a training set of historical outcome data (the clinician's experience with previous patients) yields a scoring function. This practical insight is born out theoretically, yielding an equivalence: any scoring function immediately induces a ranking; given a ranking and sufficient training data, we can efficiently find a calibrated scoring function that induces this ranking (Section II-C).

---

[3] The only exception is the Occam's Razor result.

C. Further Related work

The most closely related work, and the technical springboard for our contributions, is the definition and construction of multi-calibrated scoring functions [14]. The approach to *fair affirmative action* proposed in [4] makes no explicit use of rankings but is "morally equivalent" to the approaches of Roemer and the universities of Texas and California mentioned above, and kindled our interest in rankings. The work of [23] follows the approach of Roemer more explicitly and aims to select individuals from different (known and non-overlapping) populations in accordance with their population-specific ranking. Unlike the present work, they assume direct access to the underlying real-valued outcomes (what we refer to as p^*).

The use of machine learning techniques to rank instances is called *learning to rank* (see also the literature on *rank aggregation* for the Web, including [24] and references therein). Within this broad literature training samples in the *pairwise approach* are ordered pairs $(x, x') \in \mathcal{X} \times \mathcal{X}$, signifying that x is of higher rank than x' under an assumed true ranking, while a training sample in the *pointwise approach* consists of a single instance $x \in \mathcal{X}$, annotated with either a numerical or ordinal score. The special case in which the scores are constrained to be binary is known as the *bipartite ranking problem* and has been in studied in [25], [26]. [27] also study the connections between prediction and ranking, proving weak regret transfer bounds (where the mapping for transforming a model from one problem to another depends on the underlying distribution) between the problems of binary classification, bipartite ranking, and class-probability estimation.

Typically, the objective in learning to rank is to minimize the probability that a randomly chosen pair (x, x') is misordered, meaning: in the true ranking x is ranked above x', but in the published ranking x' is ranked above x. Various popular ranking algorithms operate by minimizing a convex upper bound on the empirical ranking error over a class of ranking functions (see e.g. RankSVM [28] and RankBoost [26]). Recently, [29] proposed cross-AUC, a variant of the standard AUC metric that corresponds to the probability that a random positive example from one group is ranked below a random negative example from the other group. This is similar yet significantly weaker variant of our notion of domination. Finally, several recent works have considered fairness in rankings from the perspective of information retrieval, where the objective is to guarantee fair representation in search results [30]–[32].

II. RANKINGS AND PREDICTORS

In this section, we give an overview of our formal goals for learning rankings from binary outcome data. We begin with some notation and preliminaries. Then, we discuss technical issues of how we represent rankings and what it means to recover a "good" ranking from a small sample of outcome data. Finally, we show various connections between the world of ranking and that of prediction; along the way, we prove a number of lemmas and introduce concepts that will be useful throughout Sections III and IV.

a) Notation and preliminaries.: We use $\mathcal{X}$ to denote a discrete universe over individuals and $\mathcal{Y} = \{0, 1\}$ to denote the space of binary outcomes. For any function $f : \mathcal{X} \to [0, 1]$, we denote the support of f as $\mathrm{supp}(f) = \{v \in [0, 1] : \exists x \in \mathcal{X} \text{ s.t. } f(x) = v\}$.

We assume that there is a fixed, but unknown distribution $\mathcal{D}_\mathcal{X}$ over individuals; for any subset $S \subseteq \mathcal{X}$, we denote by $x \sim \mathcal{D}_S$ a random (unlabeled) sample from $\mathcal{D}_\mathcal{X}$ conditioned on $x \in S$. Given an individual, we assume there is a distribution $\mathcal{D}_{\mathcal{Y} \mid \mathcal{X}}$ over outcomes; specifically, we assume that there is some function $p^* : \mathcal{X} \to [0, 1]$ such that $y \sim \mathcal{D}_{\mathcal{Y} \mid \mathcal{X}}$ is sampled according to $\mathrm{Ber}(p^*(x))$; that is $\mathbf{Pr}[y = 1 \mid x] = p^*(x)$. Together, $\mathcal{D}_\mathcal{X}$ and $\mathcal{D}_{\mathcal{Y} \mid \mathcal{X}}$ induce a joint distribution over $\mathcal{X} \times \mathcal{Y}$. We denote by $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ a random labeled sample.

We say a *predictor* is a function $p : \mathcal{X} \to [0, 1]$ that aims to approximate p^*. Throughout, unless otherwise specified, we measure closeness to p^* in terms of ℓ_1-distances, where we let

$$\|p - p^*\|_1 = \mathop{\mathbf{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[|p(x) - p^*(x)| \right].$$

For a predictor $p : \mathcal{X} \to [0, 1]$ and a subset $S \subseteq \mathcal{X}$, we denote the (canonical) median of p over S as

$$\mathop{\mathbf{med}}_{x \sim \mathcal{D}_S} [p(x)] = \inf_{v'} \left\{ v' \in \mathop{\mathrm{argmin}}_{v \in [0,1]} \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [|v - p(x)|] \right\}.$$

We use $\|p - p'\|_\infty$ to denote $\sup_{x \in \mathcal{X}} |p(x) - p'(x)|$.

A. Rankings, predictors and recovery goal

In this work, we formalize the idea of learning rankings over $\mathcal{D}_\mathcal{X}$ from binary outcomes sampled from $\mathcal{D}_{\mathcal{Y} \mid \mathcal{X}}$. Before discussing the learning model, we discuss how we represent rankings over $\mathcal{D}_\mathcal{X}$. In the case where we have a fixed universe of individuals $\mathcal{X} = [n]$, a natural way to represent a ranking is as a permutation π, where the "best" individual $x \in \mathcal{X}$ is $\pi^{-1}(1)$ and the "worst" $\pi^{-1}(n)$. For our setting where we wish to learn a ranking over a fixed but arbitrary distribution $\mathcal{D}_\mathcal{X}$, we generalize the idea of a permutation-based ranking.

Definition II.1 (Ranking)**.** *A function $r : \mathcal{X} \to [0, 1]$ is a ranking over $\mathcal{D}_\mathcal{X}$ if for all $\tau \in \mathrm{supp}(r)$*

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_\mathcal{X}} [r(x) < \tau] = \tau.$$

We denote by $\mathcal{R} \subseteq [0, 1]^\mathcal{X}$ the set of all rankings.

Note that this definition allows rankings to specify groups of individuals at the same rank; specifically, for any threshold $\tau \in [0, 1]$, the top τ-fraction of the distribution of individuals $\mathcal{D}_\mathcal{X}$ will have $r(x) \le \tau$. This definition has the appealing property that it does not require the ranking to distinguish between every pair of individuals if there is not enough information. In particular, a ranking $r \in \mathcal{R}$ specifies equivalence classes of individuals according to their rank $r(x)$. Still, some applications may call for rankings that do not allow for ties. Formally, we say that a ranking $r \in \mathcal{R}$ is *strict* if r is injective.

Note that any ranking satisfying Definition II.1 can be turned into such a strict ranking by randomly breaking ties.

Given a set of labeled samples $(x_1, y_1), \dots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, we hope to recover a ranking that approximates the ranking according to p^*. More generally, given any predictor $p : \mathcal{X} \to [0, 1]$, we can discuss a natural ranking $r^p \in R$ that orders $\mathcal{X}$ in descending order according to their p values, defined as follows.

Definition II.2 (Induced ranking)**.** *Given a predictor* $p : \mathcal{X} \to [0, 1]$*, the induced ranking* $r^p \in \mathcal{R}$ *is defined as follows.*

$$r^p(x) = \underset{x' \sim \mathcal{D}_{\mathcal{X}}}{\mathbf{Pr}} \left[p(x') > p(x)\right]$$

Thus, if we could learn p^* exactly, then we could implement the induced ranking by comparing individuals according to predictor p^*. (Lemma II.7 below formalizes this intuitive claim.)

Approximating the true ranking. Still, from a sample of labeled data, we cannot hope to learn $p^* : \mathcal{X} \to [0, 1]$ exactly; the best approximation we could hope for is an ℓ_1-approximation. We might hope that the ℓ_1-approximate recovery of p^* would translate to approximate recovery of the induced ranking. In particular, suppose $\|p - p^*\|_1 \le \varepsilon$; what can we say about r^p as compared to r^{p^*}? We argue that we cannot make nontrivial guarantees about the closeness of r^p and r^{p^*} using standard measures of distance, like ℓ_∞ or ℓ_1. To see this, consider the following example.

Example II.3. Let $\varepsilon > 0$. For a pair of injective functions with bounded values, $\xi^*, \xi : \mathcal{X} \to [-\varepsilon/2, \varepsilon/2]$, let $p^*, p : \mathcal{X} \to [0, 1]$ be defined as follows.

$$p^*(x) = 1/2 + \xi^*(x) \qquad\qquad p(x) = 1/2 + \xi(x)$$

Note that $\|p - p^*\|_1 \le \varepsilon$, but the induced rankings could be arbitrarily different; for instance, we could take $\xi(x) = -\xi^*(x)$ for all $x \in \mathcal{X}$. In particular, the induced ranking r^p is determined entirely by the choice of ξ, which contributes at most ε to $\|p - p^*\|_1$ by construction.

In other words, very small changes in a predictor can make very large changes in the outputs of the induced ranking, and thus we cannot hope to recover a ranking r with nontrivial guarantees on $\left\|r - r^{p^*}\right\|$. Thus, to learn rankings from binary labeled data with nontrivial guarantees, we need a different notion of recovery. Note that in the example above, even though the numerical value of the induced ranking may change significantly under small changes in p (e.g. go from 0 to 1), in a sense, both rankings r^p and r^{p^*} seem reasonable because $|p^*(x) - p^*(x')|$ is very small for every pair $x, x' \in \mathcal{X} \times \mathcal{X}$. Intuitively, if p^* and p are statistically-indistinguishable – and thus, are equally valid in a standard prediction setting – then our measure of quality of a ranking should not distinguish between the induced rankings r^p and r^{p^*} that arise from these predictors. This example further highlights the motivation for allowing for non-strict rankings that allow for indistinguishable individuals to receive that same rank.

Consistent predictors. To formalize this intuition, we need to take a dual perspective: rather than evaluating the quality of a ranking r in terms of its closeness to the ranking r^{p^*} induced by p^*, we evaluate closeness by comparing p^* to a predictor that is *consistent* with r. In particular, a ranking induces a collection of predictors that respect the ordering of the ranking. Formally, we define consistency as follows.

Definition II.4 (Consistency with a ranking)**.** *For a ranking* $r : \mathcal{X} \to [0, 1]$*, a predictor* $p : \mathcal{X} \to [0, 1]$ *is consistent with* r *if for all* $x, x' \in \mathcal{X} \times \mathcal{X}$*:*

- *if* $r(x) < r(x')$*, then* $p(x) \ge p(x')$*, and*
- *if* $r(x) = r(x')$*, then* $p(x) = p(x')$*.*

We denote by $\mathcal{P}(r) \subseteq [0, 1]^{\mathcal{X}}$ *the set of all the predictors that are consistent with a ranking* r*.*

Our recovery goal focuses on consistency: a ranking r is close to optimal if there exists a predictor p_r that is consistent with r and close to p^*.

Definition II.5 (Adjacency)**.** *A ranking* r *is* ε*-adjacent to* p^* *if there exists a consistent predictor* $p_r \in \mathcal{P}(r)$ *such that* $\|p_r - p^*\|_1 \le \varepsilon$*.*

To illustrate the guarantees of adjacency as a way to evaluate the quality of a rankings, we begin by revisiting the construction given in Example II.3. While we argued that the induced ranking r^p could be almost arbitrarily far from r^{p^*} in terms of $\left\|r^p - r^{p^*}\right\|_1$, note that r^p is ε-adjacent to p^*. In fact, because $\|p - p^*\|_1 \le \varepsilon$, p acts as a "certificate" of the ε-adjacency of r^p. Thus, as desired, from the perspective of ε-adjacency, r^p and r^{p^*} are equivalent rankings. In fact, it is not hard to verify that for this example, *every ranking* $r \in \mathcal{R}$ is ε-adjacent to p^* because $|p^*(x) - p^*(x')| \le \varepsilon$ for all $x, x' \in \mathcal{X} \times \mathcal{X}$.

Thus, measuring adjacency to p^* is a more flexible notion of closeness of a ranking. To see that this notion of comparison still provides a meaningful guarantee of recovery, consider the following example.

Example II.6. Let $\varepsilon > 0$. Suppose $\mathcal{X}$ is partitioned into two equally-sized sets S, T. Let $p^* : \mathcal{X} \to [0, 1]$ be defined as follows.

$$p^*(x) = \begin{cases} 3/4 & \text{if } x \in S \\ 1/4 & \text{if } x \in T \end{cases}$$

As in Example II.3, any ranking $r \in \mathcal{R}$ that permutes individuals within S and within T, but respects the order of S before T may accrue significant differences in $\left\|r - r^{p^*}\right\|_1$, but will still be 0-adjacent to $p^* \in \mathcal{P}(r)$.

Consider, however, a ranking that does not place all of S before all of T. Intuitively, this "interleaving" is clearly undesirable: there are members $x \in S$, with significantly higher $p^*(x)$ than all of T, being ranked below members $x' \in T$. Note that adjacency formalizes this intuition: as more and more interleaving occurs in some $r \in \mathcal{R}$, the optimal $\|p - p^*\|_1$ for $p \in \mathcal{P}(r)$ increases significantly.

B. Efficiently approximating the induced ranking of a predictor

Above, we argued that given a predictor $p : \mathcal{X} \to [0,1]$, the induced ranking $r^p \in \mathcal{R}$ is a well-defined function. Still, if we want to evaluate the ranking $r^p(x)$ exactly on an individual $x \in \mathcal{X}$, then in principle, we might have to evaluate $p(x')$ for all other $x' \in \mathcal{X}$. Here, we show that given oracle access to a predictor p and a small number of unlabeled samples from $\mathcal{D}_\mathcal{X}$, we can produce an approximation $\tilde{r}^p$ of the induced ranking of p. Specifically, we produce a ranking $\tilde{r}^p$ with which $p \in \mathcal{P}(\tilde{r}^p)$ is consistent (i.e. $\tilde{r}^p(x) < \tilde{r}^p(x')$ only if $p(x) > p(x')$); further, $\tilde{r}^p$ will be a pointwise approximation to the exact induced ranking r^p (i.e. $\|r^p - \tilde{r}^p\|_\infty \leq \beta$).

Proposition II.7. *Let $\beta, \delta > 0$. For a predictor $p : \mathcal{X} \to [0,1]$, let $r^p \in \mathcal{R}$ denote the induced ranking of p. There exists an efficient algorithm that given oracle access to p and $m \geq \frac{2\log(2/\beta\delta)}{\beta^2}$ unlabeled samples $x_1, \ldots, x_m \sim \mathcal{D}_\mathcal{X}$ produces a ranking $\tilde{r}^p : \mathcal{X} \to [0,1]$ such that*

- *$p \in \mathcal{P}(\tilde{r}^p)$; specifically, $\forall x, x' \in \mathcal{X} \times \mathcal{X}$: $\tilde{r}^p(x) < \tilde{r}^p(x') \implies r^p(x) < r^p(x')$.*
- *$\|r^p - \tilde{r}^p\|_\infty \leq \beta$*

with probability at least $1-\delta$.

Proof. For a threshold $\tau \in [0,1]$, consider a Bernoulli random variable X_τ distributed according the the indicator of $\mathbf{1}[r^p(x) < \tau]$ for $x \sim \mathcal{D}_\mathcal{X}$. Note that by the definition of a ranking, the expectation $\mathbf{E}[X_\tau] = \tau$. Consider the empirical estimate over m independent samples $x_i \sim \mathcal{D}_\mathcal{X}$.

$$\bar{X}_\tau = \frac{1}{m}\sum_{i=1}^m \mathbf{1}[r^p(x_i) < \tau]$$

Let $T = \{\beta/2, \beta, \ldots, 1-\beta/2, 1\}$ be a set of $2/\beta$ equally spaced thresholds. We can use Hoeffding's inequality and a union bound to bound the probability that the empirical estimates $\bar{X}_\tau$ will be more than $\beta/2$ away from the true expectation.

$$\mathbf{Pr}\left[\left|\bar{X}_\tau - \tau\right| > \beta/2\right] \leq \exp\left(\frac{-m\beta^2}{2}\right)$$

Thus, if $m \geq \frac{2\log(2/\beta\delta)}{\beta^2}$ with probability at least $1-\delta$, the empirical estimates of $\bar{X}_\tau$ for all $2/\beta$ thresholds $\tau \in T$ will be accurate up to $\beta/2$.

Given the predictor p, we can implement a comparison oracle that given a pair of inputs $x, x' \in \mathcal{X} \times \mathcal{X}$, returns the indicator of $\mathbf{1}[p(x) > p(x')]$. Thus, given some $x \in \mathcal{X}$ and the unlabeled sample, we can estimate $r^p(x)$ as follows.

$$\begin{aligned}\tilde{r}^p(x) &= \frac{1}{m}\sum_{i=1}^m \mathbf{1}[p(x_i) > p(x)] \\ &= \frac{1}{m}\sum_{i=1}^m \mathbf{1}[r^p(x_i) < r^p(x)]\end{aligned}$$

We can bound this estimate from below and above as follows. Suppose $r^p(x) \in [\tau_-, \tau_+]$ for consecutive $\tau_- \leq \tau_+ \in T$.

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}[r^p(x_i) < \tau_-] \leq \tilde{r}^p(x) \leq \frac{1}{m}\sum_{i=1}^m \mathbf{1}[r^p(x_i) < \tau_+] \quad (1)$$

$$\begin{aligned}\mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < \tau_-] - \beta/2 &\leq \tilde{r}^p(x) \\ &\leq \mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < \tau_+] + \beta/2 \quad (2)\end{aligned}$$

$$\begin{aligned}\mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < r^p(x)] - \beta &\leq \tilde{r}^p(x) \\ &\leq \mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < r^p(x)] + \beta \quad (3)\end{aligned}$$

where (1) follows by the assumption that $r^p(x) \in [\tau_-, \tau_+]$; (2) follows by the accuracy of the empirical estimates in T; and (3) follows by the fact that $|\tau_+ - \tau_-| = \beta/2$ for all consecutive $\tau_- < \tau_+ \in T$. Thus, the empirical estimate $\tilde{r}^p(x)$ will be within β of the true $r^p(x)$. ☐

In particular, note that given the sample of unlabeled data, we can build a data structure that given oracle access to p can efficiently approximate the rank $r^p(x)$ for any $x \in \mathcal{X}$. Further, note that all of the arguments used to prove Proposition II.7 work equally well if we restrict our attention to some subset $S \subseteq \mathcal{X}$. Thus, if we have access to samples from $\mathcal{D}_S$, we can similarly evaluate the ranking of individuals within the subpopulation $\mathcal{D}_S$. Such a procedure may be useful for identifying individuals in the most qualified individuals across different subsets.

Corollary II.8. *Suppose $\beta, \delta, \tau > 0$. Given access to a predictor $p : \mathcal{X} \to [0,1]$, a subset $S \subseteq \mathcal{X}$, and $\tilde{O}\left(\log(1/\delta)/\beta^2\right)$ unlabeled samples from $\mathcal{D}_S$, there is an efficient procedure that identifies the top τ'-fraction of individuals over $\mathcal{D}_S$ for some $\tau' \in [\tau - \beta, \tau + \beta]$ with probability at least $1-\delta$.*

C. Transforming a ranking into a predictor through calibration

Next, we turn our attention to obtaining a predictor given a ranking. As discussed, given a ranking $r \in \mathcal{R}$, there may be many consistent predictors that form the collection $\mathcal{P}(r)$. Our goal will be to recover a predictor $p : \mathcal{X} \to [0,1]$ that approximates the "best" consistent predictor $p_r \in \mathcal{P}(r)$. Formally, if r is ε-adjacent to p^*, we want to compute a predictor p such that $\|p - p^*\|_1$ is close to ε. Without any further information about p^*, this goal is impossible; however, we show that a small set of labeled samples $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ provides enough information about p^* to pin down a predictor p that achieves essentially optimal $\|p - p^*\|_1$.[4] The structure of the proof will introduce a number of concepts that will be useful for identifying the best ranking in a given class (see Section III), and will motivate our notions of fairness presented in Section IV.

---

[4] A conceptually similar result is shown in [27].

Our approach to transforming a ranking $r \in \mathcal{R}$ into a predictor follows the intuition that the partition of $\mathcal{X}$ induced by a ranking, which we call the *quantiles*, identify useful structure in p^* when r is ε-adjacent to p^* (for small $\varepsilon > 0$).

Definition II.9 (Quantiles according to a ranking)**.** *For a ranking $r \in \mathcal{R}$, the quantiles of r, denoted by $\mathcal{Q}_r$, partition $\mathcal{X}$ as*

$$\mathcal{Q}_r = \{Q_{r,\tau} : \tau \in \mathrm{supp}(r)\}$$
$$\textit{where } Q_{r,\tau} = \{x \in \mathcal{X} : r(x) = \tau\}.$$

Intuitively, the quantiles of a ranking r capture the "knowledge" contained in r and the number of quantiles (i.e. the support size of r) indicates the "confidence". For example, at one extreme, the constant ranking $r_0 \in \mathcal{R}$, where $r_0(x) = 0$ for all $x \in \mathcal{X}$, has a single quantile and makes no distinctions between individuals; at the other extreme, a strict ranking has quantiles at an individual-level resolution. While the quantiles are well-defined for any ranking, operationally, we will often need to work with quantiles that are sufficiently coarse.

Definition II.10 (γ-coarse ranking)**.** *A ranking $r \in \mathcal{R}$ is γ-coarse if for all $\tau \in \mathrm{supp}(r')$,*

$$\mathbf{Pr}[x \in Q_{r,\tau}] \ge \gamma/2.$$

Note that a γ-coarse ranking is supported on at most $2/\gamma$ quantiles (where the factor of 2 is an arbitrary constant factor chosen for convenience). Importantly, given any ranking $r \in \mathcal{R}$, we can turn it into a γ-coarse ranking that approximates r. In general, such a coarse approximation will not be unique; we establish the existence of a canonical γ-coarse ranking that preserves certain structure of r.

Lemma II.11 (Canonical γ-coarse ranking)**.** *For any ranking $r \in \mathcal{R}$ and $\gamma > 0$, there exists a canonical γ-coarse ranking, denoted $r^\gamma \in \mathcal{R}$, that satisfies the following consistency properties:*

- *r' maintains consistency with r: $\mathcal{P}(r') \subseteq \mathcal{P}(r)$;*
- *for all predictors $p_r \in \mathcal{P}(r)$ consistent with r, there exists a predictor $p_r^{\gamma\text{-med}} \in \mathcal{P}(r^\gamma)$ such that $\left\|p_r - p_r^{\gamma\text{-med}}\right\|_1 \le \gamma$.*

Intuitively, we form the canonical γ-coarse ranking by merging the quantiles of r into quantiles of probability density of about γ. To maintain consistency, we need to ensure that for all x, x' such that $r(x) = r(x')$, then $r^\gamma(x) = r^\gamma(x')$, which may require quantiles of larger size, which results in some technical subtlety.

Proof. We define $r^\gamma \in \mathcal{R}$ by greedily building quantiles of probability density at least $2\gamma/3$. Starting with an index $i = 1$ and threshold $\tau_1 = 0$, let

$$Q_i \leftarrow \left\{ x : \mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r(x') \le r(x)] \ge \tau_i + 2\gamma/3 \right\} \setminus \bigcup_{j<i} Q_j;$$
$$\forall x \in Q_i : \quad r^\gamma(x) \leftarrow \inf_{x' \in Q_i} \{r(x')\};$$
$$\tau_{i+1} \leftarrow \sup_{x \in Q_i} \{r(x)\};$$
$$i \leftarrow i + 1;$$

until $\tau_i > 1 - 2\gamma/3$. Suppose at termination, $i = t$. Add the remaining $x \in \mathcal{X} \setminus \bigcup_{i \le t} Q_i$, to Q_{t-1} and set $r^\gamma(x) = \tau_t$. By construction, $r^\gamma \in \mathcal{R}$ is a γ-coarse ranking: each quantile will have probability density at least $2\gamma/3$; and any x, x' where $r(x) = r(x')$ will be included in the same quantile of r^γ.

To see the property in the proposition statement, consider some predictor $p_r \in \mathcal{P}(r)$ and the predictor $p_r^{\gamma\text{-med}} \in \mathcal{P}(r^\gamma)$ defined to give the median value of p_r over each quantile. Specifically, for each $i \in [t]$ and each $x \in Q_i$, let

$$p_r^{\gamma\text{-med}}(x) = \mathop{\mathbf{med}}_{x' \in \mathcal{D}_{Q_i}} [p_r(x')].$$

Consider that statistical distance between p_r and $p_r^{\gamma\text{-med}}$.

$$\begin{aligned}
&\left\|p_r^{\gamma\text{-med}} - p_r\right\|_1 \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[\left|p_r^{\gamma\text{-med}}(x) - p_r(x)\right|\right] \\
&= \sum_{i=1}^{t} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_\mathcal{X}} [x \in Q_i] \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_i}} \left[\left|p_r^{\gamma\text{-med}}(x) - p_r(x)\right|\right] \\
&= \sum_{i=1}^{t} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_\mathcal{X}} [x \in Q_i] \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_i}} \left[\left|\mathop{\mathbf{med}}_{x' \sim \mathcal{D}_{Q_i}} [p_r(x')] - p_r(x)\right|\right]
\end{aligned} \tag{4}$$

With this expansion of $\left\|p_r^{\gamma\text{-med}} - p_r\right\|_1$, we split the analysis of individual terms based on the the probability density of the quantiles. Note that

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_i}} \left[\left|\mathop{\mathbf{med}}_{x' \sim \mathcal{D}_{Q_i}} [p_r(x')] - p_r(x)\right|\right] \tag{5}$$
$$\le \sup_{x \in Q_i} \left|\mathop{\mathbf{med}}_{x' \sim \mathcal{D}_{Q_i}} [p_r(x')] - p_r(x)\right| \tag{6}$$
$$\le \frac{1}{2} \cdot \left(\sup_{x \in Q_i} p_r(x) - \inf_{x \in Q_i} p_r(x)\right) \tag{7}$$

where (6) follows because an expectation is always upper bounded the maximum supported value; and (7) follows by the definition of the median. Thus, for some $i \in [t]$, if $\mathbf{Pr}[x \in Q_i] \le 2\gamma$, then the contribution of the ith quantile to the sum in (4) is bounded by $\gamma \cdot \left(\sup_{x \in Q_i} p_r(x) - \inf_{x \in Q_i} p_r(x)\right)$.

On the other hand, if $\mathbf{Pr}[x \in Q_i] > 2\gamma$, we claim term can be bounded by

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_\mathcal{X}} [x \in Q_i] \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_i}} \left[\left|\mathop{\mathbf{med}}_{x' \sim \mathcal{D}_{Q_i}} [p_r(x')] - p_r(x)\right|\right]$$
$$\le \frac{2\gamma}{3} \cdot \left(\sup_{x \in Q_i} p_r(x) - \inf_{x \in Q_i} p_r(x)\right).$$

To see this, note that by the construction of r^γ, such a large quantile can only arise in r^γ if it merged a large quantile from r; in turn, this implies that $p_r^{\gamma\text{-med}}(x) = \mathbf{med}_{x'\sim\mathcal{D}_{Q_i}}[p_r(x')]$ for a large fraction of Q_i. Specifically, there can be at most $2\gamma/3$ probability mass before merging with a large quantile of r (and $2\gamma/3$ after the large quantile, in the case of Q_t by the termination condition). All other $x \in Q_i$ satisfy $p_r^{\gamma\text{-med}}(x) = \mathbf{med}_{x'\sim\mathcal{D}_{Q_i}}[p_r(x')]$, and thus, contribute 0 to $\left\|p_r^{\gamma\text{-med}} - p_r\right\|_1$. Again, by properties of the median, this means that the total contribution cannot exceed $\frac{1}{2}\cdot\left(\sup_{x\in Q_i} p_r(x) - \inf_{x\in Q_i} p_r(x)\right)$ per each of the $\leq 4\gamma/3$ probability mass.

Picking up at (4), we continue to bound the distance by showing the sum telescopes.

$$
\begin{aligned}
(4) &\leq \sum_{i=1}^{t} \gamma \cdot \left(\sup_{x\in Q_i} p_r(x) - \inf_{x\in Q_i} p_r(x)\right)\\
&\leq \gamma \cdot \left(\sup_{x\in \mathcal{X}} p_r(x) - \inf_{x\in \mathcal{X}} p_r(x)\right) &(8)\\
&\leq \gamma
\end{aligned}
$$

where (8) follows by the fact that p_r is consistent with r so $\sup_{x\in Q_i} p_r(x) \leq \inf_{x\in Q_{i+1}} p_r(x)$ and the sum telescopes; the final inequality follows by the fact that $p_r : \mathcal{X} \to [0,1]$. ☐

Given the quantiles of a ranking, there is a natural predictor that gives the expected value of p^* on each quantile, which we call its *calibration*.

Definition II.12 (Calibration of a ranking)**.** *For a ranking $r \in \mathcal{R}$, the calibration of r is the predictor $p_r^{\mathrm{cal}} : \mathcal{X} \to [0,1]$ where for each $\tau \in \operatorname{supp}(r)$, for all $x \in Q_{r,\tau}$,*

$$
p_r^{\mathrm{cal}}(x) = \mathop{\mathbf{E}}_{x'\sim\mathcal{D}_{Q_{r,\tau}}}[p^*(x')] = \mathop{\mathbf{E}}_{x'\sim\mathcal{D}_{\mathcal{X}}}[p^*(x') \mid r(x') = \tau]\,.
$$

The γ-calibration of r is the predictor $p_r^{\gamma\text{-cal}}$, obtained by calibrating the γ-coarse ranking r^γ.

The next proposition shows that that the γ-calibration approximates the optimal consistent predictor $p_r = \operatorname{argmin}_{p\in\mathcal{P}(r)} \|p - p^*\|_1$.

Proposition II.13. *For any $r \in \mathcal{R}$ and $\gamma > 0$, let $p_r^{\gamma\text{-cal}}$ be the γ-calibration of r. If r is ε-adjacent to p^* for some $\varepsilon \geq 0$, then*

$$
\left\|p^* - p_r^{\gamma\text{-cal}}\right\|_1 \leq 2\varepsilon + 2\gamma.
$$

Proposition II.13 shows closeness between p^* and the exact γ-calibration, $p_r^{\gamma\text{-cal}}$. While in general, we can't hope to compute the calibration of a ranking exactly, given sufficiently many labeled samples from $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$, we can estimate $\mathbf{E}[p^*(x) \mid r^\gamma(x) = \tau]$ for each $\tau \in \operatorname{supp}(r^\gamma)$. Specifically, we can use the empirical expectations over the quantiles over a small set of $m \geq \tilde{\Omega}\left(\frac{\log(1/\delta)}{\gamma^4}\right)$ labeled samples; this argument is similar to formal arguments presented in the subsequent proof of Theorem III.1.

To demonstrate Proposition II.13, we first prove the following lemma, which will also be useful for establishing subsequent results.

Lemma II.14. *Suppose for $t \in \mathbb{N}$, $\mathcal{S} = \{S_i\}_{i\in[t]}$ is a partition of $\mathcal{X}$. Let $p^{\mathcal{S}} : \mathcal{X} \to [0,1]$ give the expected value of p^* on each partition; that is, for each $i \in [t]$, for $x \in S_i$, $p^{\mathcal{S}}(x) = \mathbf{E}_{x'\sim\mathcal{D}_{S_i}}[p^*(x)]$. Let $p_0^{\mathcal{S}} : \mathcal{X} \to [0,1]$ be any piecewise constant predictor over the partition $\mathcal{S}$; that is, for each $i \in [t]$, for $x \in S^i$, $p_0^{\mathcal{S}}(x) = v_i$ for some constant $v_i \in [0,1]$. Then,*

$$
\begin{aligned}
\left\|p^{\mathcal{S}} - p_0^{\mathcal{S}}\right\|_1 &\leq \left\|p^* - p_0^{\mathcal{S}}\right\|_1\,,\\
\left\|p^{\mathcal{S}} - p^*\right\|_1 &\leq 2\cdot\left\|p_0^{\mathcal{S}} - p^*\right\|_1\,.
\end{aligned}
$$

Proof. Consider $\left\|p^{\mathcal{S}} - p^*\right\|_1$. First, we apply the triangle inequality as follows.

$$
\left\|p^{\mathcal{S}} - p^*\right\|_1 \leq \left\|p^{\mathcal{S}} - p_0^{\mathcal{S}}\right\|_1 + \left\|p_0^{\mathcal{S}} - p^*\right\|_1
$$

Next, we show that $\left\|p^{\mathcal{S}} - p_0^{\mathcal{S}}\right\|_1 \leq \left\|p_0^{\mathcal{S}} - p^*\right\|_1$.

$$
\begin{aligned}
\left\|p^{\mathcal{S}} - p_0^{\mathcal{S}}\right\|_1 &= \sum_{i\in[t]} \mathop{\mathbf{Pr}}_{x\sim\mathcal{X}}[x \in S_i]\cdot\left|\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S^i}}[p^*(x)] - v_i\right|\\
&\leq \sum_{i\in[t]} \mathop{\mathbf{Pr}}_{x\sim\mathcal{X}}[x \in S_i]\cdot\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_i}}[|p^*(x) - v_i|] &(9)\\
&= \left\|p^* - p_0^{\mathcal{S}}\right\|_1
\end{aligned}
$$

where (9) follows by Jensen's inequality. ☐

With this lemma in place, we are ready to prove Proposition II.13.

Proof of Proposition II.13. For a ranking $r \in R$ that is ε-adjacent to p^*, for $\gamma > 0$, let $\mathcal{Q}_{r,\gamma}$ be the quantiles of r^γ and let $p_r^{\gamma\text{-cal}} : \mathcal{X} \to [0,1]$ be the γ-calibration of r. Let $p_r = \operatorname{argmin}_{p\in\mathcal{P}(r)} \|p - p^*\|_1$, and let $p_r^{\gamma\text{-cal}} \in \mathcal{P}(r)$ be the predictor that gives the median prediction of p_r on each γ-quantile, as in Lemma II.11. Then, we can derive the following inequalities.

$$
\begin{aligned}
\left\|p_r^{\gamma\text{-cal}} - p^*\right\|_1 &\leq 2\cdot\left\|p_r^{\gamma\text{-med}} - p^*\right\|_1 &(10)\\
&\leq 2\cdot\left(\left\|p_r^{\gamma\text{-med}} - p_r\right\|_1 + \|p_r - p^*\|_1\right)\\
&\leq 2\gamma + 2\varepsilon &(11)
\end{aligned}
$$

where (10) follows by Lemma II.14 because $p_r^{\gamma\text{-med}}$ is piecewise constant over the γ-quantiles and (11) follows by the assumption that r is ε-adjacent to p^* and Lemma II.11. ☐

Note that in Proposition II.13, when we convert an ε-adjacent ranking to a predictor, we can guarantee a predictor that is $(2\varepsilon + \gamma)$-adjacent for any constant $\gamma > 0$; further, by concentration arguments deferred to Section III, this same guarantee can be achieved using an a small random sample to estimate the γ-calibration. We argue that in our learning model, with access to binary samples $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$, the factor of 2 loss between the adjacency and the ℓ_1-distance of the recovered predictor is optimal.

Observation II.15 (Informal)**.** *For any $c < 2$, there is an $\varepsilon > 0$ and a distribution $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$, such that no algorithm that is given access to a ranking $r \in \mathcal{R}$ that is ε-adjacent to p^**

and a bounded number of labeled samples $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ *can produce a predictor* p_r *such that*

$$\|p_r - p^*\|_1 \leq c \cdot \varepsilon.$$

Proof Sketch. Let $\mathcal{X} = [N]$ be a finite universe and $\mathcal{D}_{\mathcal{X}}$ be the uniform distribution over $\mathcal{X}$. Suppose $r \in \mathcal{R}$ is the constant ranking; that is, $r(x) = 0$ for all $x \in \mathcal{X}$. We construct a hard distribution over the choice of $p^* : \mathcal{X} \to [0, 1]$, where we can bound the adjacency of r to p^*, but it is impossible to recover a predictor that always achieves the optimal ℓ_1 error.

For some $\varepsilon > 0$, let $p_\varepsilon : \mathcal{X} \to [0, 1]$ be defined as $p_\varepsilon(x) = \varepsilon$ for all $x \in \mathcal{X}$. For some subset $S \subseteq \mathcal{X}$, let $p_S : \mathcal{X} \to 0, 1$ be defined as $p_S(x) = 1$ if $x \in S$ and $p_S(x) = 0$ for $x \notin S$. Let $S_\varepsilon \subseteq \mathcal{X}$ be a random subset sampled by independently sampling $x \in S_\varepsilon$ with probability ε for each $x \in \mathcal{X}$. Then, consider the following distribution over the choice of p^*:

$$p^* = \begin{cases} p_\varepsilon & \text{w.p. } 1/2 \\ p_{S_\varepsilon} & \text{w.p. } 1/2 \end{cases}$$

for a randomly drawn S_ε. Note for a bounded set of samples (say, $o(\sqrt{N})$ samples), with probability $1 - o(1)$, there will be no $x \in \mathcal{X}$ sampled more than once; conditioned on this event, the labeled samples $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ for either choice of p^* are identically distributed.

Despite the identical distribution of labeled samples, the feasible minimizer of $\|p_r - p^*\|_1$ is not the same. In particular, because r is the constant ranking, to be consistent $p_r \in \mathcal{P}(r)$ must be constant over $\mathcal{X}$. When $p^* = p_\varepsilon$, then p_ε is the minimizer, and r is 0-adjacent to p^*. In other words, if we output any predictor p_r other than p_ε, then with probability $1/2$, then $\|p_r - p^*\|_1 > c \cdot \varepsilon$ for every constant c. Thus, to get any multiplicative approximation to the best ℓ_1 error, every algorithm must output p_ε.

But consider when $p^* = p_{S_\varepsilon}$; in this case, the constant predictor $p_0(x) = 0$ for all $x \in \mathcal{X}$ will minimize the ℓ_1 error to p^*, with $\|p_0 - p^*\|_1 \leq \varepsilon + o(1)$. Using p_ε as the estimate of p^*, we can bound the expected ℓ_1 error as follows.

$$\begin{aligned}\mathbf{E}\left[\|p_\varepsilon - p^*\|_1\right] &= \mathbf{Pr}[p^*(x) = 1] \cdot (1 - \varepsilon) + \mathbf{Pr}[p^*(x) = 0] \cdot \varepsilon \\ &= \varepsilon \cdot (1 - \varepsilon) + (1 - \varepsilon) \cdot \varepsilon \\ &= 2\varepsilon - 2\varepsilon^2\end{aligned}$$

Taking $\varepsilon > 0$ to be an arbitrarily small constant, we can see that the recovery guarantee approaches 2ε, which approaches a factor 2 worse than optimal. □

III. IDENTIFYING THE BEST RANKING

Proposition II.13 shows that given a ranking $r \in \mathcal{R}$ and a small sample of labeled data, we can recover an approximately optimal predictor that is consistent with r. Still, because we only see the realization of $y \sim \mathrm{Ber}(p^*(x))$, it is not immediately obvious how to evaluate the ℓ_1-distance between the derived predictor and p^*. Thus, from the analysis in Section II alone, given a collection of rankings $R \subseteq \mathcal{R}$, it's not clear whether we can find the best $r \in R$.

In this section, we prove an agnostic "Occam's Razor"-style theorem for rankings. That is, we show that given a class of rankings $R \subseteq \mathcal{R}$, it is information-theoretically possible to identify the (approximately) best ranking $r \in R$ from a small set of m labeled samples $(x_1, y_1) \ldots (x_m, y_m) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$. Slightly more formally, for any $\varepsilon > 0$, if there is an ε-adjacent ranking $r \in R$, we give an algorithm that runs in polynomial-time in $|R|$ and m, and returns an $O(\varepsilon)$-adjacent ranking $r' \in R$.

Because we only have access to samples of binary outcomes, our access to p^* is very limited. As such, the proof differs significantly from classic proofs of identifiability for boolean functions, as in [33], or for rankings given comparison data of the form $\mathbf{1}[p^*(x) > p^*(x')]$. Indeed, given an individual sample $(x, y) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$, we cannot reliably determine any conclusive information about $p^*(x)$. With a small sample complexity, it is exceedingly unlikely to see any $x \sim \mathcal{D}_{\mathcal{X}}$ twice, let alone enough times to accurately estimate the bias. Further, as discussed earlier, even if we could learn p^* *exactly* on significant portions of $\mathcal{D}_{\mathcal{X}}$, if there are non-trivial portions where we are still uncertain, it is impossible to extract a globally consistent ranking.

While the result is self-contained and does not directly impact the subsequent discussion of learning evidence-consistent rankings, it introduces some key insights about how to extract information about the "true" ranking induced by p^* from binary outcomes. In particular, the proof hinges on the fact that the empirical expectations of *outcomes* on (sufficiently-large) subsets of $\mathcal{X}$ will concentrate around their expectation. Further, the proof clarifies the intuition that the rankings in the class R can help to identify structure in the true ranking r^{p^*}, even if r^{p^*} is not in the class. This intuition is paramount to developing our strongest notion of reflexive evidence-consistency in Section IV.

Theorem III.1. *Suppose* R *is a class of rankings such that there exists an* ε*-adjacent* $r \in R$*. For any* $\gamma, \delta > 0$*, there is an algorithm that given* $m \geq \tilde{\Omega}\left(\frac{\log(|R|/\delta)}{\gamma^5}\right)$ *labeled samples* $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ *with probability at least* $1 - \delta$ *produces some* $r' \in R$ *that is* $(3\varepsilon + \gamma)$*-adjacent. The algorithm runs in* $\mathrm{poly}(|R|, m)$ *time.*

Proof. For $\gamma > 0$, we will show how to recover a $(3\varepsilon + c \cdot \gamma)$-adjacent ranking for some constant c; the theorem follows by choosing $\gamma' = \gamma/c$, losing only a constant factor in the sample complexity. For each $r \in R$, let r^γ denote its canonical γ-coarse ranking. For every two rankings $r \in R$ and $q \in R$, consider the predictor $p_{rq}^{\gamma\text{-cal}} : \mathcal{X} \to [0, 1]$ defined to give the expected value of p^* over each of the intersections of quantiles according to r^γ and q^γ, where for all $x \in Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}$,

$$p_{rq}^{\gamma\text{-cal}}(x) = \mathop{\mathbf{E}}_{x' \sim \mathcal{D}}\left[p^*(x') \mid r^\gamma(x') = \tau,\ q^\gamma(x') = \sigma\right].$$

For each $q \in R$, we define the following loss function.

$$L_R(q) = \min_{p \in P(q)} \max_{r \in R} \left\|p_{qr}^{\gamma\text{-cal}} - p\right\|_1$$

The theorem follows by showing that the $r \in R$ that minimizes (the empirical estimate of) L_R is an approximately optimal ranking over R. In particular, suppose for every $q, r \in R \times R$, we can find a empirical estimate of $p^{\gamma\text{-cal}}_{qr}$, which we denote $\hat{p}_{qr} : \mathcal{X} \to [0,1]$, that satisfies $\left\|\hat{p}_{qr} - p^{\gamma\text{-cal}}_{qr}\right\|_1 \le O(\gamma)$. Let the empirical loss function $\hat{L}_R(q)$ be defined as

$$\hat{L}_R(q) = \min_{p \in P(q)} \max_{r \in R} \left\|\hat{p}_{qr} - p\right\|_1 .$$

We bound the distance of the ranking that minimizes the empirical loss to p^*.

Suppose $q = \operatorname{argmin}_{r \in R} \hat{L}_R(r)$ is the minimizer of $\hat{L}_R$ over R and let $p_q \in \mathcal{P}(q)$ denote a consistent predictor for q that achieves the minimum value of $\max_{r \in R} \left\|\hat{p}_{qr} - p_q\right\|_1$. Further, let $r^* = \operatorname{argmin}_{r \in R} \min_{p \in \mathcal{P}(r)} \left\|p - p^*\right\|_1$ be the optimal ranking in R; specifically, we assume that r^* is ε-adjacent to p^* for some $\varepsilon > 0$. Let $p_{r^*} \in \mathcal{P}(r^*)$ denote a consistent predictor for r^* such that $\left\|p^* - p_{r^*}\right\|_1 = \varepsilon$.

Using the triangle inequality, we expand the ℓ_1-distance between p_q and p^* as

$$\begin{aligned} &\left\|p_q - p^*\right\|_1 \\ &\le \left\|p_q - \hat{p}_{qr^*}\right\|_1 + \left\|\hat{p}_{qr^*} - p^{\gamma\text{-cal}}_{qr^*}\right\|_1 + \left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1 \\ &\le \hat{L}_R(q) + O(\gamma) + \left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1 \end{aligned}$$

where $\left\|p_q - \hat{p}_{qr^*}\right\|_1 \le \hat{L}_R(q) = \max_{r \in R} \left\|p_q - \hat{p}_{qr}\right\|_1$ by the definition of p_q, and $\left\|\hat{p}_{qr} - p^{\gamma\text{-cal}}_{qr}\right\|_1 \le O(\gamma)$ by assumption. We will bound $\hat{L}_R(q)$ and $\left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1$ separately.

For $r \in R$, let $p^{\gamma\text{-med}}_r \in \mathcal{P}(r)$ be the canonical predictor associated with r^γ from Lemma II.11. Then, by the fact that q minimizes $\hat{L}_R$, we can bound $\hat{L}_R(q)$ as follows.

$$\begin{aligned} \hat{L}_R(q) &\le \hat{L}_R(r^*) \\ &= \min_{p \in \mathcal{P}(r^*)} \max_{r \in R} \left\|\hat{p}_{r^*r} - p\right\|_1 \\ &\le \max_{r \in R} \left\|\hat{p}_{r^*r} - p^{\gamma\text{-med}}_{r^*}\right\|_1 && (12) \\ &\le \max_{r \in R} \left\{ \left\|\hat{p}_{r^*r} - p^{\gamma\text{-cal}}_{r^*r}\right\|_1 + \left\|p^{\gamma\text{-cal}}_{r^*r} - p^{\gamma\text{-med}}_{r^*}\right\|_1 \right\} \\ &\le O(\gamma) + \left\|p^* - p^{\gamma\text{-med}}_{r^*}\right\|_1 && (13) \\ &\le \left\|p^* - p_{r^*}\right\|_1 + \left\|p_{r^*} - p^{\gamma\text{-med}}_{r^*}\right\|_1 + O(\gamma) \\ &\le \varepsilon + O(\gamma) && (14) \end{aligned}$$

where (12) follows by the fact that $p^{\gamma\text{-med}}_{r^*} \in \mathcal{P}(r^*)$; (13) follows by the assumption that $\left\|\hat{p}_{r^*r} - p^{\gamma\text{-cal}}_{r^*r}\right\|_1 \le O(\gamma)$ and applying Lemma II.14 to bound $\left\|p^{\gamma\text{-cal}}_{r^*r} - p^{\gamma\text{-med}}_{r^*}\right\|_1$ because for each partition defined by the quantiles of $r^{*\gamma}$ and r^γ, $p^{\gamma\text{-cal}}_{r^*r}$ gives the expectation over the partition and $p^{\gamma\text{-med}}_{r^*}$ is piecewise constant; finally, (14) follows by the assumption that $\left\|p^* - p_{r^*}\right\|_1 \le \varepsilon$ and applying Proposition II.11 to p_{r^*}.

Next, we bound $\left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1$. Let $p^{\gamma\text{-med}}_{qr^*}$ denote the predictor that gives the median value of p_{r^*} over the partition defined by $p^{\gamma\text{-cal}}_{qr^*}$. Specifically, for all $x \in \mathcal{X}$ such that $p^{\gamma\text{-cal}}_{qr^*}(x) = v$,

$$p^{\gamma\text{-med}}_{qr^*}(x) = \mathop{\mathbf{med}}_{x' \sim \mathcal{D}} \left[p_{r^*}(x') \mid p^{\gamma\text{-cal}}_{qr^*}(x') = v \right].$$

Then, we can bound $\left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1$ as follows.

$$\begin{aligned} &\left\|p^{\gamma\text{-cal}}_{qr^*} - p^*\right\|_1 && (15) \\ &\le 2 \cdot \left\|p^{\gamma\text{-med}}_{qr^*} - p^*\right\|_1 && (16) \\ &\le 2 \cdot \left(\left\|p^{\gamma\text{-med}}_{qr^*} - p_{r^*}\right\|_1 + \left\|p_{r^*} - p^*\right\|_1 \right) \\ &\le 2 \cdot \left(\left\|p^{\gamma\text{-med}}_{r^*} - p_{r^*}\right\|_1 + \left\|p_{r^*} - p^*\right\|_1 \right) && (17) \\ &\le 2\gamma + 2\varepsilon. && (18) \end{aligned}$$

where (16) follows by Lemma II.14 applied to the piecewise constant predictor $p^{\gamma\text{-med}}_{qr^*}$; (17) follows by the observation that $\left\|p^{\gamma\text{-med}}_{qr^*} - p_{r^*}\right\|_1 \le \left\|p^{\gamma\text{-med}}_{r^*} - p_{r^*}\right\|_1$, which can be seen by the convex optimization interpretation of the median as the minimizer of ℓ_1; and (18) follows again by the assumption that r^* is ε-adjacent to p^* and $\left\|p^{\gamma\text{-med}}_{r^*} - p_{r^*}\right\|_1 \le \gamma$ by Lemma II.11. Thus, we see that $\left\|p_q - p^*\right\|_1 \le 3\varepsilon + O(\gamma)$.

Thus, it remains to bound the sample complexity necessary to recover for each $r, q \in R \times R$ a $\hat{p}_{rq} : \mathcal{X} \to [0,1]$ such that

$$\left\|\hat{p}_{rq} - p^{\gamma\text{-cal}}_{rq}\right\|_1 \le O(\gamma).$$

Note that to bound this ℓ_1-error, it suffices to estimate the statistical queries

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}} \left[p^*(x) \mid r^\gamma(x) = \tau,\ q^\gamma(x) = \sigma \right]$$

up to γ additive error for each $r, q \in R \times R$ and $\tau \in \operatorname{supp}(r^\gamma)$, $\sigma \in \operatorname{supp}(q^\gamma)$.

First, suppose that for each $r, q \in R \times R$ and $\tau \in \operatorname{supp}(r^\gamma)$, $\sigma \in \operatorname{supp}(q^\gamma)$, we can obtain s labeled samples directly over each of the subsets of interest $(x_1, y_1), \ldots, (x_s, y_s) \sim \mathcal{D}_{Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}, \mathcal{Y}}$. Then, applying Hoeffding's inequality, we can bound the probability that the empirical estimate on the sample deviates significantly from the actual expectation.

$$\begin{aligned} \mathbf{Pr} \left[\left| \frac{1}{s} \sum_{i=1}^{s} y_i - \mathop{\mathbf{E}}_{x \sim \mathcal{D}} \left[p^*(x) \mid x \in Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma} \right] \right| > \gamma \right] \\ \le 2 \exp\left(-2s\gamma^2\right). \end{aligned}$$

Thus, if $s \ge \frac{\log(2/\delta_0)}{2\gamma^2}$, then the probability the estimate deviates by more than γ is at most δ_0. Because there are at most $O\left(\frac{|R|^2}{\gamma^2}\right)$ statistical queries to estimate, then given $s \ge \Omega\left(\frac{\log(|R|^2/\gamma^2\delta)}{\gamma^2}\right)$ on each subset of interest, by a union bound, all of the expectations will be accurate up to γ with probability at least $1 - \delta/2$.

Next, we argue that we can exclude intersections of quantiles $Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}$ that are smaller than γ^3 probability mass; this allows us to bound the sample complexity from $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$ necessary to guarantee that each subset of interest has at least

s samples. Note that for each $r, q \in R \times R$, the predictor $\hat{p}_{rq}$ will be supported on at most $\Omega\left(1/\gamma^2\right)$ values. Thus, in order to obtain a γ additive ℓ_1-approximation, it suffices to provide guarantees on the estimates for the sets where such that

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{X}}\left[x \in Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}\right] \geq \gamma^3.$$

In particular, ignoring all sets where $\mathbf{Pr}_{x\sim\mathcal{D}_\mathcal{X}}\left[x \in Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}\right] < \gamma^3$ incurs at most an additional γ^3 ℓ_1-error per set, so $\gamma^3 \cdot O\left(1/\gamma^2\right) f = O(\gamma)$ overall.

Thus, we may assume that for every intersection of interest $\mathbf{Pr}_{x\sim\mathcal{D}_\mathcal{X}}[x \in Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}] \geq \gamma^3$. Again, by Hoeffding's inequality, if we take $l \geq \frac{\log(2s/\delta)}{\gamma^3}$ the probability that every sample misses such a set is at most $\delta/2s$. Thus, if we take $m \geq sl \geq \tilde{\Omega}\left(\frac{\log(|R|/\delta)}{\gamma^5}\right)$ samples, then another union bound shows that with probability at least $1-\delta/2$, each $Q_{r^\gamma,\tau} \cap Q_{q^\gamma,\sigma}$ in our collection will have at least s samples. Thus, with probability at least $1-\delta$, every estimate will be accurate up to $O(\gamma)$ additive error. □

IV. EVIDENCE-BASED RANKINGS

Section III shows that, information-theoretically, given a class of rankings $R \subseteq \mathcal{R}$, we can identify an approximately optimal ranking $r \in R$. Still, when the class R isn't sufficiently-expressive to contain a ranking that is ε-adjacent to p^* for small ε, approximate recovery may not be enough to guarantee the *fairness* of the eventual ranking. Consider the following simple example that illustrates how an ε-adjacent ranking allows for a subset of fraction ε to be significantly mistreated.

Example IV.1. Let $\varepsilon > 0$ and $S_1 \subseteq \mathcal{X}$ be a subset such that $\mathbf{Pr}_{x\sim\mathcal{D}_\mathcal{X}}\left[x \in S_1\right] \leq \varepsilon$. Suppose $p^*, p : \mathcal{X} \to [0,1]$ is defined as follows.

$$p^*(x) = \begin{cases} 1 & \text{if } x \in S_1 \\ 0.01 & \text{otherwise} \end{cases} \quad p(x) = \begin{cases} 0 & \text{if } x \in S_1 \\ 0.01 & \text{otherwise} \end{cases}$$

Note that in the induced ranking of p^*, S_1 is the top ε-fraction, but in the induced ranking of p, S_1 is the bottom ε-fraction. Still, despite the fact that the rank of S_1 has moved arbitrarily far, $\|p - p^*\|_1 \leq \varepsilon$.

This example highlights the fact that ε-adjacency, while a reasonable recovery goal, is not enough to guarantee fair treatment for groups of size less than ε. Note, however, that such a blatant mistreatment can be detected from the 0/1 data we have at hand! This motivates a further study of fair rankings, that aims to protect sufficiently large subsets of $\mathcal{X}$.

A. Domination-Compatibility and Evidence-Consistency

The example above demonstrated one way in which significant groups can be blatantly mistreated. Intuitively, if a "fair" ranking gives preference to a subset S over another subset T, we would expect that S should be more qualified than T in terms of p^*, at least on average. We begin by formalizing what we mean when we say that a ranking r gives preference to S over T, which we refer to as *domination*.

Definition IV.2 (Domination)**.** *Let $S, T \subseteq \mathcal{X}$ be two subsets and $\gamma \geq 0$. For a ranking $r \in \mathcal{R}$, we say that S γ-dominates T in r if for all thresholds $\tau \in [0,1]$,*

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[r(x) < \tau] + \gamma \geq \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_T}[r(x) < \tau].$$

That is, S dominates T if for every threshold $\tau \in [0,1]$, the fraction (with respect to $\mathcal{D}$) of individuals from S that are ranked below τ is at least as large as the fraction of individuals in T, up to a slack of γ.

Intuitively, there is a natural combinatorial interpretation of the domination condition in terms of matchings. Specifically, in the special case where S and T are discrete sets of equal cardinality and the distribution of interest $\mathcal{D}_\mathcal{X}$ is the uniform distribution, then S γ-dominates T if, after discarding a γ-fraction of the individuals from each group, there exists a perfect matching $M : S \to T$ in which where every $x \in S$ is matched to some $M(x) \in T$, whose rank in r is no better than that of x; that is, $r(x) \leq r(M(x))$. We use Definition IV.2 because it allows for comparison between S and T that are arbitrarily-intersecting subsets of arbitrary probability densities.

We argue that domination formally captures the intuition that a ranking strongly prefers one subset over another. In particular, the following lemma shows that if S dominates T in a ranking r, then every consistent predictor $p \in \mathcal{P}(r)$, favors S over T on average.

Lemma IV.3. *If S γ-dominates T in r, then for every $p \in \mathcal{P}(r)$,*

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[p(x)\right] + \gamma \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}\left[p(x)\right].$$

Proof. For a ranking $r \in \mathcal{R}$, let $p \in \mathcal{P}(r)$ be consistent predictor. By consistency, for each $v \in \operatorname{supp}(p)$, there exists some $\tau_v \in \operatorname{supp}(r)$ (the minimum τ where $r(x) = \tau$ and $p(x) = v$) such that for any subset $S \subseteq \mathcal{X}$

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}\left[p(x) > v\right] = \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}\left[r(x) < \tau_v\right].$$

Suppose S γ-dominates T. Consider the difference in expectations of $p(x)$ under $\mathcal{D}_S$ and $\mathcal{D}_T$, which we expand using the identity for nonnegative random variables $\mathbf{E}[X] = \int_v \mathbf{Pr}[X > v]dv$.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}\left[p(x)\right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[p(x)\right] \\
&= \int_0^1 \left(\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_T}[p(x) > v] - \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[p(x) > v]\right) dv \\
&= \int_0^1 \left(\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_T}[r(x) < \tau_v] - \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[r(x) < \tau_v]\right) dv \\
&\leq \gamma
\end{aligned}$$

where the final inequality bounds the difference in probabilities by γ-domination. □

Lemma IV.3 suggests a natural group fairness notion for rankings. Suppose $\mathbf{E}_{x\sim\mathcal{D}_T}[p^*(x)]$ is significantly larger than

$\mathbf{E}_{x\sim\mathcal{D}_S}[p^*(x)]$ but S γ-dominates T for some small γ. Then, Lemma IV.3 show that no consistent predictor $p \in \mathcal{P}(r)$ can respect the true potential of S and T, even on average! Such a reversal under r – where the expected potential of T is higher than that of S, but S dominates T in r – represents a form of blatant discrimination against T: either the individuals of T are being significantly undervalued or the individuals in S are being overvalued by the ranking r.

With this in mind, a baseline notion of fairness for a ranking r would be that r does not exhibit any such blatant reversals for any pair of subsets from some rich collection $\mathcal{C}$; formally, we call this notion *domination-compatibility*.

Definition IV.4 (Domination-compatibility). *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets and $\alpha \geq 0$. A ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-domination-compatible if for all pairs of subsets $S, T \in \mathcal{C} \times \mathcal{C}$ and for every $\gamma \geq 0$, if S γ-dominates T in r, then*

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] + (\gamma+\alpha) \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p^*(x)]$$

Looking ahead, since the expectations $\mathbf{E}_{x\sim\mathcal{D}_S}[p^*(x)]$ and $\mathbf{E}_{x\sim\mathcal{D}_T}[p^*(x)]$ will eventually be estimated from the sample of binary labels, the definition allows for an additional additive slack of α.

A $(\mathcal{C},\alpha)$-domination-compatible ranking r guarantees that if S dominates T in r, then the true expectation of p^* over S is not significantly lower than that over T. Intuitively, the fact that S receives preferential treatment compared to T in r is "justified" by $\mathcal{D}_{\mathcal{Y}\,|\,\mathcal{X}}$.

As discussed, one reason that violating the domination-compatibility criteria seems so objectionable is that there does not exist *any* consistent predictor $p \in \mathcal{P}(r)$ that exhibits the true expectations on the identified sets $S \in \mathcal{C}$. This observation motivates a notion of fair rankings from the perspective of consistent predictors, which we call *evidence-consistency*. Evidence-Consistency goes a step further than domination-compatibility and requires that a consistent predictor exists that exhibits the correct expectations of p^* for every subset in the collection.

Definition IV.5 (Evidence-Consistency). *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$. A ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-evidence-consistent if there exists a consistent predictor $\tilde{p} \in \mathcal{P}(r)$ where for every $S \in \mathcal{C}$,*

$$\left|\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\tilde{p}(x)]\right| \leq \alpha.$$

In other words, a ranking r is evidence-consistent with respect to a class $\mathcal{C}$ if there is a consistent predictor $p \in \mathcal{P}(r)$ that cannot be refuted using the statistical tests defined by the class $\mathcal{C}$. If $\mathcal{C}$ represents the collection of tests that can be feasibly carried out (from a computational or statistical perspective), then from this perspective, an evidence-consistent ranking is a plausible candidate for the ranking induced by p^*.

As a definition, a ranking that is evidence-consistent over a class $\mathcal{C}$ provides a guarantee of consistency to p^* that is parameterized by the expressiveness of $\mathcal{C}$; for a fixed value of α, the richer the class $\mathcal{C}$, the stronger the guarantee provided by consistency with the actual expectations. Viewing $\mathcal{C}$ as a complexity class of "efficiently-identifiable" subsets, evidence-consistency guarantees that no inconsistencies in the ranking can be identified within the computational bound specified by $\mathcal{C}$.

B. Evidence-Consistency implies Domination-Compatibility

By requiring a globally-consistent predictor that respects the expectations defined by subsets $S \in \mathcal{C}$, evidence-consistency guarantees that the ranking does not misrepresent the (average) potential of any $S \in \mathcal{C}$ compared to another $T \in \mathcal{C}$. In particular, if a ranking satisfies evidence-consistency with respect to a class $\mathcal{C}$ then it also satisfies domination-compatibility with respect to the class.

Theorem IV.6 (Formal restatement of Theorem 1). *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. If a ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-evidence-consistent, then r is $(\mathcal{C},2\alpha)$-domination-compatible.*

Proof. Suppose for $\alpha \geq 0$ a ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-evidence-consistent. Let $S, T \in \mathcal{C}$ be two sets where S γ-dominates T, for some $\gamma \geq 0$.

By the definition of evidence-consistency, we know that there exists a predictor $p_r \in P(r)$ such that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p_r(x)] - \alpha$$
$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p^*(x)] \leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p_r(x)] + \alpha$$

Further, by Lemma IV.3, because S γ-dominates T, we know that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p_r(x)] \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p_r(x)] + \gamma.$$

Combining the three inequalities, we can derive the following inequality.

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p^*(x)] + \gamma - 2\alpha$$

Thus, for every pair $S, T \subseteq \mathcal{C} \times \mathcal{C}$ where S γ-dominates T, the expectation of p^* over S and T satisfy the domination-compatibility requirement with additive slack 2α. $\square$

C. Learning an evidence-consistent ranking

With the above implication in place, one way to learn a ranking that satisfies $(\mathcal{C},\alpha)$-domination-compatibility is to first learn a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ that respects all of the expectations of subsets $S \in \mathcal{C}$ (up to $\alpha/2$ tolerance), and then convert $\tilde{p}$ into its induced ranking $\tilde{r}$. To see this, recall that the predictor $\tilde{p} \in \mathcal{P}(\tilde{r})$ is consistent with its induced ranking. Further, because such a $\tilde{p}$ exhibits the correct expectations over the collection $\mathcal{C}$, it can witness the predictor required in the definition evidence-consistency; that is, $\tilde{p}$ certifies that its induced ranking $\tilde{r}$ is $(\mathcal{C},\alpha/2)$-evidence-consistent.

The task of learning a predictor that respects the expectations over a collection of sets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ has been studied recently in the context of fair prediction [14], [18].[5] These works, which refer to such a condition as $(\mathcal{C}, \alpha)$-*multi-accuracy*, show how to learn such a $\tilde{p}$ from a small number of binary samples.

Proposition IV.7 ([14]). *Let $\alpha, \gamma, \delta > 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a fixed collection of subsets. There is an algorithm that given $m \geq \tilde{\Omega}\left(\frac{\log(|\mathcal{C}|/\delta)}{\gamma\alpha^2}\right)$ labeled samples $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ learns a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ such that with probability $1-\delta$, for every $S \in \mathcal{C}$ where $\mathbf{Pr}_{x\sim\mathcal{D}_{\mathcal{X}}}[x \in S] \geq \gamma$,*

$$\left| \underset{x\sim\mathcal{D}_S}{\mathbf{E}}[p^*(x)] - \underset{x\sim\mathcal{D}_S}{\mathbf{E}}[\tilde{p}(x)] \right| \leq \alpha.$$

The algorithm runs in $\mathrm{poly}(|\mathcal{C}|, m)$ time.

[14] also show that for structured classes $\mathcal{C}$, the running time of the algorithm can be improved, by reducing the task of learning a $(\mathcal{C}, \alpha)$-multi-accurate predictor to the task of agnostic learning the class $\mathcal{C}$ in the sense of [35], [36].

D. A separation between domination-compatibility and evidence-consistency

We conclude this section by showing that when the sets we aim to protect are predefined, domination-compatibility is a strictly weaker notion than evidence-consistency. Specifically, while evidence-consistency implies domination compatibility, the reverse implication does not hold. The following examples demonstrates a ranking that is $\mathcal{C}$-domination-compatible but is not $\mathcal{C}$-evidence-consistent.

Example IV.8. Let $\mathcal{X}$ be a universe of 100 individuals, split into two disjoint sets A and B, each of size 50. Let $\mathcal{D}_{\mathcal{X}}$ be the uniform distribution on $\mathcal{X}$. Further assume that there are two subsets $A' \subset A$, $B' \subset B$, each of size 10. Define p^* as follows:

$$p^\star(x) = \begin{cases} 1.0 & x \in B \\ 0.0 & x \in A' \\ 0.5 & x \in A - A' \end{cases}$$

Let $\mathcal{C} = \{A, B, C\}$, where $C = A' \cup B'$. Then the true expectations are, $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_A}}[p^*(x)] = 0.4$, $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_B}}[p^*(x)] = 1.0$, and $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_C}}[p^*(x)] = 0.5$. Now, consider the ranking $r : B - B' \succ C \succ A - A'$.

Note that r is $\mathcal{C}$-domination compatible, because the domination criterion holds for every two sets in $\mathcal{C}$. Indeed: for $\{A, B\}$, B 0-dominates A in r and the true expectation of B is greater than the true expectation of A; for $\{A, C\}$, C 0-dominates A in r and the true expectation of C is greater than the true expectation of A; finally, for $\{B, C\}$, B 0-dominates C in r and the true expectation of B is greater than the true expectation of C.

On the other hand, we claim that r isn't $(\mathcal{C}, \alpha)$-evidence-consistent, for every $\alpha < 0.1$. Fix $\alpha < 0.1$ and assume for contradiction that it is, and let $p \in \mathcal{P}(r)$ be a predictor that is simultaneously $\alpha-$consistent with the expectations of $\{A, B, C\}$. To maintain consistency with $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_B}}[p^*(x)] = 1.0$, $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_{B'}}}[p(x)] \geq 1-\alpha$. This implies that $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_{B'}}}[p(x)] \geq 1 - 5\alpha$ (because $|B'| = 0.2 \cdot |B|$). To maintain consistency with $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_C}}[p^*(x)] = 0.5$, this implies that $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_{A'}}}[p(x)] \leq 3\alpha$. Finally, to maintain consistency with $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_A}}[p^*(x)] = 0.4$, $\mathbf{E}_{x\sim\mathcal{D}_{\mathcal{X}_{A-A'}}}[p(x)] \geq 0.5 - 2\alpha$. But note that since the members of A' are ranked before the members of $A - A'$ by r, the fact that $p \in \mathcal{P}(r)$ means that the scores of A' should be greater-equal than the scores of $A - A'$, or $3\alpha \geq 0.5 - 2\alpha$, or that $\alpha \geq 0.1$, which is a contradiction.

Pairwise-consistency. In fact, $\mathcal{C}$-domination-compatibility is equivalent to a significantly weaker notion, which we refer to as *pairwise-consistency*.

Definition IV.9 (Pairwise-consistency)**.** *Let $\mathcal{C}$ be a family of sets over $\mathcal{X}$. A ranking $r \in \mathcal{R}$ satisfies $(\mathcal{C}, \alpha)$-pairwise-consistency if for every two sets $S, T \in \mathcal{C} \times \mathcal{C}$, there exists a predictor $p \in \mathcal{P}(r)$ such that*

$$\left| \left(\underset{x\sim D_S}{\mathbf{E}}[p^*(x)] - \underset{x\sim D_S}{\mathbf{E}}[p(x)] \right) - \left(\underset{x\sim D_T}{\mathbf{E}}[p^*(x)] - \underset{x\sim D_T}{\mathbf{E}}[p(x)] \right) \right| \leq \alpha \quad (19)$$

Note that definition relaxes evidence-consistency in two aspects. First, while evidence-consistency requires that there is a single, global predictor p to be simultaneously accurate in expectation for all sets, pairwise-consistency instead requires that for every pair of sets, there exists a predictor consistent with expectations; in particular, there may be a different consistent predictor for each pair of sets separately. This switch of quantifiers represents a significantly weaker requirement, similar in spirit to the domination criterion that only compares two sets at a time. Second, while in Definition (IV.5) the consistent predictor had to have approximately accurate expectations on both S and T, here it is only required to not distort the *relative* distance between their expectations.

With this definition in place, we can show that domination-compatibility is equivalent to this weaker notion of pairwise-consistency.

Theorem IV.10. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. A ranking $r \in \mathcal{R}$ is $(\mathcal{C}, \alpha)$-pairwise-consistency if and only if r is $(\mathcal{C}, \alpha)$-domination-compatible.*

Proof. First, we show that pairwise-consistency implies domination-compatibility. Let $\alpha \geq 0$ and assume that a ranking r satisfies $(\mathcal{C}, \alpha)$-pairwise-consistency. Let $S, T \in \mathcal{C}$ be two sets where S γ-dominates T, for some $\gamma \geq 0$. From pairwise-consistency, we know that there exists a predictor $p \in P(r)$ for which the condition in Equation (19) holds. In

---

[5] Earlier work in pseudorandomness studied the question of existence and circuit complexity of such predictors [34].

particular, this implies that

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim D_S}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim D_T}[p^*(x)] \\ \geq \left(\mathop{\mathbf{E}}_{x\sim D_S}[p(x)] - \mathop{\mathbf{E}}_{x\sim D_T}[p(x)]\right) - \alpha\end{aligned}$$

Lemma (IV.3), on the other hand, guarantees that $\mathbf{E}_{x\sim\mathcal{D}_S}[p(x)] - \mathbf{E}_{x\sim\mathcal{D}_T}[p(x)] \geq \gamma$. Together, these two facts imply that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] + (\gamma+\alpha) \geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p^*(x)]$$

which concludes the proof of the first direction.

Next, we show that domination-compatibility implies pairwise-consistency. Let $\alpha \geq 0$ and assume that r satisfies $(\mathcal{C},\alpha)$-domination-compatibility. Let $S, T \in \mathcal{C}$ be any two sets such that $\mathbf{E}_{x\sim D_T}[p^*(x)] \geq \mathbf{E}_{x\sim D_S}[p^*(x)]$. We split into cases as follows.

First, consider the case that $\mathbf{E}_{x\sim D_T}[p^*(x)] - \mathbf{E}_{x\sim D_S}[p^*(x)] \leq \alpha$. Consider the predictor $p(x) \triangleq \mathbf{E}_{x\sim D_T}[p^*(x)] - \mathbf{E}_{x\sim D_S}[p^*(x)]$. Since it treats everyone identically, it is consistent with any ranking, and in particular with r. Also, by definition, $\mathbf{E}_{x\sim D_T}[p(x)] - \mathbf{E}_{x\sim D_S}[p(x)] = 0$, so

$$\begin{aligned}&\left|\left(\mathop{\mathbf{E}}_{x\sim D_S}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim D_S}[p(x)]\right)\right. \\ &\qquad\left. - \left(\mathop{\mathbf{E}}_{x\sim D_T}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim D_T}[p(x)]\right)\right| \\ &= \mathop{\mathbf{E}}_{x\sim D_T}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim D_S}[p^*(x)] \leq \alpha\end{aligned}$$

r satisfies pairwise-consistency constraint in this case.

Next, consider the case that $\mathbf{E}_{x\sim D_T}[p^*(x)] > \mathbf{E}_{x\sim D_S}[p^*(x)] + \alpha$. Denote $\mathbf{E}_{x\sim D_T}[p^*(x)] = \mathbf{E}_{x\sim D_S}[p^*(x)] + (\alpha+\gamma)$ where $\gamma > 0$.

Observe that the fact that r satisfies $(\mathcal{C},\alpha)$-domination-compatibility implies that in this case, S does not γ-dominate T. This implies that there exists some threshold $\tau \in [0,1]$ for which

$$\Delta^\tau \triangleq \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_T}[r(x) < \tau] - \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[r(x) < \tau] > \gamma \tag{20}$$

Next, choose ε as follows: if $\Delta^\tau \leq \gamma + 2\alpha$, let $\varepsilon = 1$. Otherwise, choose ε such that $\frac{\gamma}{\Delta^\tau} \leq \varepsilon \leq \frac{\gamma+2\alpha}{\Delta^\tau}$. Note that $0 < \varepsilon \leq 1$. Define the following predictor p. If $r(x) \geq \tau$, $p(x) = 0$. If $r(x) < \tau$, then $p(x) = \varepsilon$. Now, by definition:

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p(x)] &= \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[r(x) < \tau] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p(x) \mid r(x) < \tau] \\ &= \varepsilon \cdot \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[r(x) < \tau]\end{aligned}$$

Similarly, $\mathbf{E}_{x\sim\mathcal{D}_T}[p(x)] = \varepsilon \cdot \mathbf{Pr}_{x\sim\mathcal{D}_T}[r(x) < \tau]$. Thus,

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p(x)] = \varepsilon \cdot \Delta^\tau$$

To conclude the proof, observe that $\gamma < \varepsilon \cdot \Delta^\tau \leq \gamma + 2\alpha$. This is from the definition of ε, as well as Equation (20). We therefore have:

$$\begin{aligned}\gamma \leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p(x)] \leq \gamma + 2\alpha \\ \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x)] = \gamma + \alpha\end{aligned}$$

Thus, the absolute value of the difference between the two is smaller than α, and so by definition, r satisfies $(\mathcal{C},\alpha)$-pairwise-consistency. □

V. PROTECTING QUANTILES YIELDS STRONGER EVIDENCE-BASED NOTIONS

The results of Section IV-A establish that the strength of evidence-consistency hinges on the expressiveness of $\mathcal{C}$; the richer $\mathcal{C}$ is, the stronger the "semantic" protections provided by consistency with the actual expectations in sets in $\mathcal{C}$. Somewhat surprisingly, we argue that these protections may be too weak, even for a rich class $\mathcal{C}$. Indeed, any approach that only explicitly protects a *predefined* collection of subpopulations can leave the door open to abuses, including ones that we show can be audited from labeled data. In this section, we build up to increasingly stronger notions of both domination compatibility and evidence-consistency, which we refer to as reflexive domination compatibility and reflexive evidence-consistency.

To highlight the potential weakness of evidence-consistency, consider the following example, reiterated from the introduction. Suppose $\mathcal{C}$ has two sets S and S' where the learner knows that $\mathbf{E}_{x\sim\mathcal{D}_S}[p^*(x)] = 0.8$ and $\mathbf{E}_{x\sim\mathcal{D}_{S'}}[p^*(x)] = 0.5$. In order to promote the individuals of S' (potentially unfairly) above the those of S, an adversary could rank (an arbitrary) half of S' first, followed by S, then the remainder of S', while maintaining $(\mathcal{C},0)$-evidence-consistency. In particular, the predictor $\tilde{p}$ that gives $\tilde{p}(x) = 1.0$ to the first half of $x \in S'$, $\tilde{p}(x) = 0.8$ to all of $x \in S$, and $\tilde{p}(x) = 0.0$ to the remaining $x \in S'$ is consistent with the ranking and satisfies all of the expectations defined by $\mathcal{C}$. Such adversarial manipulation of evidence-consistency is possible regardless of the structure of p^* within S and S'.

In fact, this failure to satisfy domination-compatibility for subpopulations defined by the ranking is not only a problem of adversarial manipulation of this notion. We argue that such violations can arise unintentionally, even from rankings learned from data in seemingly-objective ways. Continuing the example with two sets, suppose S and S' have nontrivial intersection, where $T = S \cap S'$; again, let $\mathbf{E}_{x\sim\mathcal{D}_S}[p^*(x)] = 0.8$ and $\mathbf{E}_{x\sim\mathcal{D}_{S'}}[p^*(x)] = 0.5$. Given these expectations, a natural ranking might put $S \setminus T$ first, followed by T, then $S' \setminus T$.[6] Nevertheless, it could be the case that T contains the strongest members of the population, with $\mathbf{E}_{x\sim\mathcal{D}_T}[p^*(x)] > 0.8$. In this case, the proposed ranking would violate domination-compatibility between S and T. Indeed, this example exploits

---

[6] For example, such an ordering is the induced ranking of the maximum entropy predictor.

the fact that the relevant subset T is not included in $\mathcal{C}$. Still, with the ranking in hand, T is identifiable; T is one of the quantiles! This example shows that without explicitly considering the quantiles of the ranking themselves, violations of domination-compatibility between the sets defined by the ranking may arise in insidious ways.

A. Ordering the quantiles via domination-compatibility

These examples demonstrate that while a $(\mathcal{C}, \alpha)$-evidence-consistent ranking r provides strong overall protections for the sets in $\mathcal{C}$, it provides limited guarantees to sets defined by r itself, specifically the quantiles, which may intersect nontrivially with the sets in $\mathcal{C}$. This observation motivates enforcing some notion of consistency, not just with respect to $\mathcal{C}$, but also to ensure the quantiles of r are ordered according to their expected p^* value. We argue that a ranking that satisfies domination-compatibility over its quantiles satisfies a certain approximate ordering property.

Recall, for a ranking $r \in \mathcal{R}$ we denote the quantiles of r as $\mathcal{Q}_r = \{Q_{r,\tau} : \tau \in \operatorname{supp}(r)\}$, where $Q_{r,\tau} = \{x : r(x) = \tau\}$ for each $\tau \in \operatorname{supp}(r)$. We observe that because the quantiles partition $\mathcal{X}$ according to the order implied by the ranking, then ordering the quantiles $Q_{r,\tau}$ and $Q_{r,\tau'}$ by domination corresponds to the total order induced by comparing the corresponding ranks τ and τ'.

Lemma V.1. *Let $r \in \mathcal{R}$ be ranking. Suppose for $\tau, \tau' \in \operatorname{supp}(r)$, $S_\tau \subseteq Q_{r,\tau}$ and $T_{\tau'} \subseteq Q_{r,\tau'}$ are each subsets of a quantile. If $\tau \le \tau'$, then S_τ 0-dominates $T_{\tau'}$.*

Proof. The proof of the lemma follows immediately from the definition of quantiles and γ-domination. For a ranking $r \in \mathcal{R}$, let $\tau \le \tau' \in [0,1]$ and let $S_\tau \subseteq Q_{r,\tau}$ and $T_{\tau'} \subseteq Q_{r,\tau'}$. We argue that for all thresholds $\sigma \in [0,1]$

$$\Pr_{x \sim \mathcal{D}_{S_\tau}}[r(x) \le \sigma] \ge \Pr_{x \sim \mathcal{D}_{T_{\tau'}}}[r(x) \le \sigma].$$

In particular, by the fact that the ranking r is constant on each quantile, the statement is equivalent to

$$\mathbf{1}[\tau \le \sigma] \ge \mathbf{1}[\tau' \le \sigma],$$

which holds for all σ by the assumption that $\tau \le \tau'$. □

As such, requiring a ranking to satisfy domination-compatibility over its quantiles implies the quantiles are (approximately) correctly ordered according to their expectations.

Corollary V.2. *Suppose a ranking is $(\mathcal{Q}_r, \alpha)$-domination-compatible. Then, for all $\tau < \tau' \in \operatorname{supp}(r)$,*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_{r,\tau}}}[p^*(x)] \ge \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_{r,\tau'}}}[p^*(x)] - \alpha.$$

Note that the motivating examples from above fail to satisfy $(\mathcal{Q}_r, \alpha)$-domination-compatibility for sufficiently small $\alpha > 0$. In particular, in each example, there is a pair of quantiles where $\tau < \tau'$, but $\mathbf{E}_{x \sim \mathcal{D}_{Q_{r,\tau}}}[p^*(x)]$ is significantly smaller than $\mathbf{E}_{x \sim \mathcal{D}_{Q_{r,\tau'}}}[p^*(x)]$, violating the ordering condition of Corollary V.2.

With this mind, one way to augment the notions of domination-compatibility and evidence-consistency from Section IV would be to add the quantiles to the set $\mathcal{C}$ to protect. Specifically, we could require a new evidence-based notion $(\mathcal{C} \cup \mathcal{Q}_r, \alpha)$-evidence-consistency that would imply $(\mathcal{C} \cup \mathcal{Q}_r, 2\alpha)$-domination-compatibility by Theorem IV.6. Such a notion is strong enough to mitigate the concerns raised in the examples so far; still, we argue that simply adding the quantiles to the collection of sets to protect may not be enough. In particular, some of the attacks demonstrating the weaknesses of evidence-consistency can be "scaled down" to work, not at the level of $\mathcal{X}$, but within the subpopulations $S \in \mathcal{C}$.

Example V.3. Suppose $\mathcal{X}$ is partitioned into two (disjoint) sets: T, consisting of 80% of $\mathcal{X}$, and S. Suppose further that S is partitioned into two equally-sized sets, S_0 and S_1, and similarly T is partitioned into two equally sized sets T_0, T_1. Define p^* as follows:

$$p^*(x) = \begin{cases} 1.0 & x \in S_1 \cup T_1 \\ 0.0 & x \in S_0 \cup T_0 \end{cases}$$

And suppose that $\mathcal{C} = \{S, T\}$. Now, consider the following ranking

$$r(x) = \begin{cases} 0 & x \in T_1 \cup S_0 \\ 0.5 & x \in T_0 \cup S_1 \end{cases}$$

That is, r correctly identifies the top individuals in the majority T, but "flips" the ranking within S. We claim that this ranking satisfies domination-compatibility even if we include the quantiles. The quantiles induced by r are $Z_{r,0} = T_1 \cup S_0$ and $Q_{r,0.5} = T_0 \cup S_1$. Thus,

$$\mathcal{C} \cup \mathcal{Q}_r = \{S, T, T_1 \cup S_0, T_0 \cup S_1\}$$

Whose true expectations under p^* are $\mathbf{E}_{x \sim \mathcal{D}_S}[p^*(x)] = 0.5$, $\mathbf{E}_{x \sim \mathcal{D}_T}[p^*(x)] = 0.5$, $\mathbf{E}_{x \sim \mathcal{D}_{Q_{r,0}}}[p^*(x)] = 0.8$, and $\mathbf{E}_{x \sim \mathcal{D}_{Q_{r,0.5}}}[p^*(x)] = 0.2$.

It's left to verify that r is $(\mathcal{C} \cup \mathcal{Q}_r)$-domination-compatibility. Note that $Q_{r,0}$ 0-dominates all the other three sets, and indeed has the highest expectation. Note that since S, T both have the same expectation, the domination criterion will always hold. Finally, T, S both 0-dominate $Q_{r,0.5}$ and indeed have a higher expectation.

B. Incorporating the quantiles into evidence-based notions

Example V.3 demonstrates that without a very expressive class $\mathcal{C}$, simply adding the quantiles to the set $\mathcal{C}$ over which we require evidence-consistency may still suffer from undesirable transpositions of subgroups *within* the sets defined in $\mathcal{C}$. In this section, we show a way to incorporate the quantiles to provide a much stronger guarantee. Intuitively, rather than protecting the union of the set system $\mathcal{C}$ with the quantiles $\mathcal{Q}_r$, we protect the intersections of sets $S \in \mathcal{C}$ and each $Q_{r,\tau} \in \mathcal{Q}_r$.

Given a collection of subsets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$, a ranking $r \in \mathcal{R}$, and an approximation parameter α, consider the following set system derived by intersecting subsets $S \in \mathcal{C}$ with those defined by the quantiles of r.

Definition V.4 (Quantile-augmented collection). *Let $\alpha \geq 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$. For a ranking $r \in \mathcal{R}$ with $|\operatorname{supp}(r)| = s \in \mathbb{N}$, the α-quantile-augmented collection of r, denoted $\mathcal{C}_r^\alpha \subseteq \{0,1\}^{\mathcal{X}}$, is a collection of subsets defined as follows.*

$$\mathcal{C}_r^\alpha = \left\{ S_{r,\tau} : \begin{array}{c} S \in \mathcal{C} \cup \{\mathcal{X}\}, \\ \tau \in \operatorname{supp}(r), \\ \Pr\limits_{x \sim S}[x \in S_{r,\tau}] \geq \alpha/s \end{array} \right\} \text{ where } S_{r,\tau} = S \cap Q_{r,\tau}.$$

We make two remarks about the collection $\mathcal{C}_r^\alpha$. First, note that because we consider $S = \mathcal{X}$ as one of our sets, the reflexive level sets are a superset of the level sets. Second, note that we exclude quantiles that are sufficiently small, anticipating the fact that we wish to learn such rankings from random samples. Indeed, if $\mathbf{Pr}_{x\sim\mathcal{X}}\left[x \in \mathcal{X}_{r,\tau}\right]$ is very small, then we might not see any individuals from $\mathcal{X}_{r,\tau}$ in our random sample. Note, however, that the size of sets that we deem too small is parameterized by the support size of r; as the support size increases, we become more stringent, requiring that smaller quantiles satisfy the evidence-based constraints. If the ranking is γ-coarse for sufficiently large γ compared to α, then we will not exclude any of the quantiles.

As before, we can consider domination-compatibility and evidence-consistency with respect to this augemented collection of sets.

Definition V.5 (Reflexive domination-compatibility). *Let $\alpha \geq 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets. A ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-domination-compatible if it is $(\mathcal{C}_r^\alpha,\alpha)$-domination-compatible.*

Definition V.6 (Reflexive evidence-consistency). *Let $\alpha \geq 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets. A ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent if r is $(\mathcal{C}_r^\alpha,\alpha)$-evidence-consistent; that is, if there exists some $\tilde{p} \in \mathcal{P}(r)$ such that for all $S_\tau \in \mathcal{C}_r^\alpha$,*

$$\left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_\tau}}\left[p^*(x)\right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_\tau}}\left[\tilde{p}(x)\right] \right| \leq \alpha.$$

Again, by the characterization of these stronger "reflexive" notions as domination-compatibility and evidence-consistency over a richer collection of sets, the fact that reflexive-evidence-consistency implies reflexive-domination-compatibility follows as a corollary of Theorem IV.6.

Corollary V.7. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. If a ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent, then r is $(\mathcal{C},2\alpha)$-reflexive-domination-compatible.*

Recall that without augmenting the class $\mathcal{C}$, domination-compatibility is a strictly weaker notion than evidence-consistency. The main result of this section shows that reflexive-domination-compatibility is *equivalent* to reflexive-evidence-consistency. That is, any ranking that satisfies the domination-compatibility conditions for a rich enough class of sets (informed by the ranking itself) implies the existence of a globally consistent predictor that has the correct expectation on the same class of sets.

Theorem V.8. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. If a ranking $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-domination-compatible, then r is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent.*

Proof. Suppose $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-domination-compatible. For a subset $S \subseteq \mathcal{X}$ and $\tau \in \operatorname{supp}(r)$, let $S_{r,\tau} = S \cap Q_{r,\tau}$ denote the intersection of S with the quantile $Q_{r,\tau}$. To demonstrate that r is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent, we construct a predictor $\tilde{p} \in \mathcal{P}(r)$ that has the (approximately) correct expectations on all $S_\tau \in \mathcal{C}_r^\alpha$. Consider the predictor $\tilde{p} : \mathcal{X} \to [0,1]$ that for all $x \in Q_{r,\tau}$ gives:

$$\tilde{p}(x) = \min_{\tau' : \tau' \leq \tau \in \operatorname{supp}(r)} \mathop{\mathbf{E}}_{x' \sim \mathcal{D}_{Q_{r,\tau'}}}\left[p^*(x')\right].$$

First, we argue that $\tilde{p}$ is consistent with r. Note that for any $x, x' \in \mathcal{X} \times \mathcal{X}$ where $r(x) = r(x')$, $\tilde{p}(x) = \tilde{p}(x')$. Second, consider two $x, x' \in \mathcal{X} \times \mathcal{X}$ where $r(x) = \tau$ and $r(x') = \tau'$ for $\tau < \tau'$. By the definition of $\tilde{p}$, which assigns the $x \in Q_{r,\tau}$ the minimum expectation of p^* over all $Q_{r,\tau'}$ for $\tau' \leq \tau \in \operatorname{supp}(r)$, we know that $\tilde{p}(x')$ cannot exceed $\tilde{p}(x)$; that is, $\tilde{p}(x) \geq \tilde{p}(x')$.

It remains to show that $\tilde{p}$ satisfies the evidence-consistency constraints on the expectations of $S_{r,\tau}$. We break the inequality in Definition V.6 into two inequalities. Specifically, we show that for all $S \in \mathcal{C} \cup \{\mathcal{X}\}$, for all $\tau \in \operatorname{supp}(r)$,

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[\tilde{p}(x)\right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[p^*(x)\right] \geq -\alpha \tag{21}$$

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[\tilde{p}(x)\right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[p^*(x)\right] \leq \alpha \tag{22}$$

First, note that $\tilde{p}$ is constant over the quantiles of r, so for all $x \in Q_{r,\tau}$:

$$\tilde{p}(x) = \mathop{\mathbf{E}}_{x'\sim\mathcal{D}_{S_{r,\tau}}}\left[\tilde{p}(x')\right] = \min_{\tau' : \tau' \leq \tau \in \operatorname{supp}(r)} \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_{r,\tau'}}}\left[p^*(x)\right].$$

We can see (21) by invoking Lemma V.1 to argue that $Q_{r,\tau'}$ 0-dominates $S_{r,\tau}$ for all $\tau' \leq \tau \in \operatorname{supp}(r)$. Thus, by $(\mathcal{C}_r^\alpha,\alpha)$-domination-compatibility, we derive

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[\tilde{p}(x)\right] &= \min_{\tau' : \tau' \leq \tau \in \operatorname{supp}(r)} \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_{r,\tau'}}}\left[p^*(x)\right] \\
&\geq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[p^*(x)\right] - \alpha.
\end{aligned}$$

Similarly, we can see (22) by invoking Lemma V.1 to argue that $S_{r,\tau}$ 0-dominates $Q_{r,\tau}$. Again, by $(\mathcal{C}_r^\alpha,\alpha)$-domination-compatibility, we derive

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[\tilde{p}(x)\right] &= \min_{\tau' : \tau' \leq \tau \in \operatorname{supp}(r)} \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{Q_{r,\tau'}}}\left[p^*(x)\right] \\
&\leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{Q_{r,\tau}}}\left[p^*(x)\right] \\
&\leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}\left[p^*(x)\right] + \alpha
\end{aligned}$$

Thus, $(\mathcal{C},\alpha)$-reflexive-domination-compatibility implies $(\mathcal{C},\alpha)$-reflexive-evidence-consistency. $\square$

As such, the following equivalence between reflexive domination-compatibility and reflexive evidence-consistency holds.

Theorem V.9 (Restatement of Theorem 2)**.** *If a ranking is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent, then it is $(\mathcal{C},2\alpha)$-reflexive-domination-compatible. If a ranking is $(\mathcal{C},\alpha)$-reflexive-domination-compatible, then it is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent.*

C. Reflexive evidence-based rankings and multi-calibrated predictors

Next, we turn our attention to learning a reflexive-evidence-consistent ranking. Note that while $(\mathcal{C},\alpha)$-reflexive-evidence-consistency is defined similarly to $(\mathcal{C},\alpha)$-evidence-consistency, in terms of a predictor exhibiting the correct expectations on a collection of subsets, we cannot apply the algorithm from Proposition IV.7 directly. The problem with this approach is that Proposition IV.7 assumes that $\mathcal{C} \subseteq 0,1^{\mathcal{X}}$ is a fixed, nonadaptive collection of subsets. Note that the augmented collection $\mathcal{C}_r^{\alpha}$ is defined *adpatively*; that is, these sets are defined in terms of the ranking r and are not well-specified until r is specified. Thus, we need a different approach for learning such a predictor.

To this end, we turn to the concept of *calibration* studied in scoring and prediction, and recently in the context of fair prediction [14], [37]. Intuitively, a predictor is well-calibrated if the probability of $y = 1$ over the individuals who receive score $v \in [0,1]$ is actually close to v. At two extremes, the optimal predictor p^* and the "average" predictor $p(x) = \mathbf{E}_{x' \sim \mathcal{D}_{\mathcal{X}}}[p^*(x')]$ for all $x \in \mathcal{X}$ are both calibrated. Formally, we work with the following technical definition of approximate calibration.

Definition V.10 (Calibrated predictor)**.** *Suppose $p : \mathcal{X} \to [0,1]$ is a predictor and $|\mathrm{supp}(p)| = s$. For $\alpha > 0$ and a subset $S \subseteq \mathcal{X}$, $p : \mathcal{X} \to [0,1]$ is α-calibrated over S if for all $v \in \mathrm{supp}(p)$, if $\mathbf{Pr}_{x\sim\mathcal{D}_S}[p(x) = v] \geq \alpha/s$, then*

$$\left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x) \mid p(x) = v] - v \right| \leq \alpha.$$

We say that a predictor $p : \mathcal{X} \to [0,1]$ is α-calibrated if it is α-calibrated over $\mathcal{X}$. Note that a calibrated predictor provides guarantees about the expectations on the level sets defined by the predictor. Still, reflexive evidence-consistency requires a ranking to reason about the intersections of the quantiles with every $S \in \mathcal{C}$. An analogous strengthening of calibration, referred to as *multicalibration*, was introduced and studied in [14].

Definition V.11 (Multi-calibration)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$ and let $\alpha \geq 0$. A predictor $p : \mathcal{X} \to [0,1]$ is $(\mathcal{C},\alpha)$-multi-calibrated if p is α-calibrated on every $S \in \mathcal{C} \cup \{\mathcal{X}\}$.*

Analogous to reflexive evidence-consistency, a predictor that is multi-calibrated over a class $\mathcal{C}$ provides strong consistency guarantees on the expectations defined by the intersection of sets $S \in \mathcal{C}$ and the level sets defined by the predictor. We show that this analogy can be made formal and that a $(\mathcal{C},\alpha)$-multi-calibrated predictor induces a $(\mathcal{C},\alpha)$-reflexive-evidence-consistent ranking.

Proposition V.12. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. If a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C},\alpha)$-multi-calibrated, then its induced ranking $r^{\tilde{p}}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent.*

Proof. For any predictor $p : \mathcal{X} \to [0,1]$, let $r^p \in \mathcal{R}$ denote its induced ranking, and for convenience, for every $v \in \mathrm{supp}(p)$, let $p^{-1}(v)$ denote some canonical member $x \in \{x' \in \mathcal{X} : p(x') = v\}$. Note that by the definition of an induced ranking for every $x, x' \in \mathcal{X}$, $p(x) = p(x')$ if and only if $r^p(x) = r^p(x')$; thus $|\mathrm{supp}(p)| = |\mathrm{supp}(r^p)| = s$. This means that for any subset $S \subseteq \mathcal{X}$ and for any $v \in \mathrm{supp}(p)$, there exists a unique $\tau_v \in \mathrm{supp}(r^p)$ such that

$$S_{p,v} \triangleq \{x \in S : p(x) = v\} = \{x \in S : r^p(x) = \tau_v\} = S_{r^p,\tau_v}$$

where $\tau_v = r^p(p^{-1}(v))$. Thus, suppose $\tilde{p} : \mathcal{X} \to [0,1]$ is a $(\mathcal{C},\alpha)$-multi-calibrated predictor. Using the bijection above and the definition of multi-calibration, this means that for all $S \in \mathcal{C}$ and $\tau_v \in \mathrm{supp}(r^{\tilde{p}})$ where $\mathbf{Pr}_{x\sim\mathcal{D}_S}\left[r^{\tilde{p}}(x) = \tau_v\right] \geq \alpha/s$,

$$\begin{aligned}
&\left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r^{\tilde{p}},\tau_v}}}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r^{\tilde{p}},\tau_v}}}[\tilde{p}(x)] \right| \\
&= \left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{\tilde{p},v}}}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{\tilde{p},v}}}[\tilde{p}(x)] \right| \\
&\leq \alpha.
\end{aligned}$$

Thus, $r^{\tilde{p}}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent. $\square$

As such, an algorithm that learns a $(\mathcal{C},\alpha)$-multi-calibrated predictor can be used to learn a $(\mathcal{C},\alpha)$-reflexive-evidence-consistent ranking. [14] provide such an algorithm.

Proposition V.13 ([14])**.** *Let $\alpha,\gamma,\delta > 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a fixed collection of subsets. There is an algorithm that given $m \geq \tilde{\Omega}\left(\frac{\log(|\mathcal{C}|/\delta)}{\gamma^{3/2}\alpha^{11/2}}\right)$ labeled samples $(x_1,y_1),\ldots,(x_m,y_m) \sim \mathcal{D}_{\mathcal{X},\mathcal{Y}}$ learns a $(\mathcal{C},\alpha)$-multi-calibrated predictor $\tilde{p} : \mathcal{X} \to [0,1]$ with probability $1-\delta$. The algorithm runs in $\mathrm{poly}(|\mathcal{C}|, m)$ time.*

Again, as with $(\mathcal{C},\alpha)$-multi-accuracy, the running time of the algorithm for learning a $(\mathcal{C},\alpha)$-multi-calibrated predictor can be improved for agnostically learnable $\mathcal{C}$.

Thus, the algorithm of [14] demonstrates that we can learn a $(\mathcal{C},\alpha)$-reflexive-unassailable predictor from labeled outcome data. Note, however, that the definition of reflexive evidence-consistency does not explicitly require that the predictor $\tilde{p} \in \mathcal{P}(r)$ be multi-calibrated, so it is not immediately obvious whether learning a multi-calibrated predictor and converting it to a ranking is the best way to learn a reflexive evidence-consistent ranking. Next, we show that any algorithm that learns a $(\mathcal{C},\alpha)$-reflexive-evidence-consistent ranking, paired with a small set of labeled samples, implies an algorithm for learning a multi-calibrated predictor. In particular, we show that the predictor $\tilde{p}$ that witness the reflexive evidence-consistency of r (essentially) must be multi-calibrated.

Proposition V.14. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subsets over $\mathcal{X}$ and let $\alpha \geq 0$. Suppose $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent. For any consistent predictor $\tilde{p} \in \mathcal{P}(r)$ where for all $S_{r,\tau} \in \mathcal{C}_r^{\alpha}$*

$$\left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}[p^*(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}[\tilde{p}(x)] \right| \leq \alpha,$$

and for any constant $\varepsilon > 0$, there exists a predictor $\tilde{q} \in \mathcal{P}(r)$ such that $\|\tilde{p}-\tilde{q}\|_\infty \leq \varepsilon$ and $\tilde{q}$ is $(\mathcal{C},\alpha+\varepsilon)$-multi-calibrated.

Proof. Suppose $\tilde{p} \in \mathcal{P}(r)$; then by consistency with r, for all $x, x' \in \mathcal{X}$ if $r(x) = r(x')$, then $\tilde{p}(x) = \tilde{p}(x')$. Suppose that for $r(x) = \tau$, $\tilde{p}(x) = v_\tau$. Consider defining $\tilde{q} \in \mathcal{P}(r)$ by mapping each x where $r(x) = \tau$ to a unique u_τ where $|u_\tau - v_\tau| \leq \varepsilon$ for some arbitrarily small constant $\varepsilon > 0$; thus, $|\mathrm{supp}(r)| = |\mathrm{supp}(\tilde{q})| = s$ and $\|\tilde{p}-\tilde{q}\|_\infty \leq \varepsilon$. Then, there is a bijection between the augmented sets $S_{r,\tau} \in \mathcal{C}_r^{\alpha}$ and the sets induced by the level sets of $\tilde{q}$.

$$S_{r,\tau} = \{x \in S : r(x) = \tau\} = \{x \in S : \tilde{q}(x) = u_\tau\} \triangleq S_{\tilde{q},u_\tau}$$

Thus, suppose $r \in \mathcal{R}$ is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent. Using the bijection above and the definition of evidence-consistency, this means that for all $S \in \mathcal{C}$ and $u_\tau \in \mathrm{supp}(\tilde{q})$ where $\mathbf{Pr}_{x\sim\mathcal{D}_S}[\tilde{q}(x) = u_\tau] \geq \alpha/s$,

$$\begin{aligned}
&\left| u_\tau - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x) \mid \tilde{q}(x) = u_\tau] \right| \\
&= \left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\tilde{q}(x) \mid \tilde{q}(x) = u_\tau] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x) \mid \tilde{q}(x) = u_\tau] \right| \\
&\leq \left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\tilde{p}(x) \mid \tilde{q}(x) = u_\tau] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[p^*(x) \mid \tilde{q}(x) = u_\tau] \right| + \varepsilon \\
&= \left| \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}[\tilde{p}(x)] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_{r,\tau}}}[p^*(x)] \right| + \varepsilon \\
&\leq \alpha + \varepsilon.
\end{aligned}$$

As such, $\tilde{q}$ is $(\mathcal{C},\alpha+\varepsilon)$-multi-calibrated. ☐

We remark that the proposition shows that for an evidence-consistent ranking r, every $\tilde{p} \in \mathcal{P}(r)$ is statistically close to a multi-calibrated predictor. This is largely a technicality; maintaining the bijection between the quantiles of r and the level sets of $\tilde{q}$ ensures that any sets $S_{r,\tau}$ where $\mathbf{Pr}_{x\sim\mathcal{D}_S}[r(x) = \tau] < \alpha/s$ (for which we have no guarantees) remain small enough in $\mathbf{Pr}_{x\sim\mathcal{D}_S}[\tilde{q}(x) = u_\tau] < \alpha/s$ that we need not provide a guarantee on their expectation. We also remark that a similar proof shows that the calibration of an $(\mathcal{C},\alpha)$-reflexive-evidence-consistent ranking (even if it is not consistent) is $(\mathcal{C}, 2\alpha+\varepsilon)$-multi-calibrated.

Thus, we can conclude the following tight connection between the notions of reflexive evidence-consistency and multi-calibration.

Theorem V.15 (Restatement of Theorem 3)**.** *The ranking induced by a $(\mathcal{C},\alpha)$-multi-calibrated function is $(\mathcal{C},\alpha)$-reflexive-evidence-consistent. Further, any consisten scoring function that exhibits the correct expectations defined by a $(\mathcal{C},\alpha)$-reflexive-evidence-consistent ranking is statistically close to being $(\mathcal{C},\alpha)$-multi-calibrated.*

This theorem establishes the fact that that reflexive domination-compatibility, reflexive evidence-consistency, and multi-calibration are all tightly connected concepts of fairness. We can interpret the theorem from the perspective of ranking or from the perspective of prediction. First and most pertinent to the present work, the theorem shows that in order to learn a ranking that satisfies our strongest notion of fairness, it is (essentially) necessary and sufficient to learn a multi-calibrated predictor. On the other hand, when the goal is to learn a fair and accurate predictor, this result shows that multi-calibrated predictors satisfy strong, desirable non-transposition properties. As we've discussed, ranking is an inherently global task; thus, the result supports the intuitive idea that in order to satisfy multi-calibration, learning a predictor that performs well on the majority population is not sufficient, but instead global learning is required.

ACKNOWLEDGMENT

CD was supported in part by NSF CCF-1763665, the Sloan Foundation, the Radcliffe Institute for Advanced Study, the Simons Institute, and Microsoft Research. MPK was supported in part by the CISPA Center for Information Security and the Stanford Data Science Initiative; part of this work was completed while visiting the Weizmann Institute of Science. OR was supported in part by VMwares University Research Fund and NSF CCF-1763311. GNR received funding for this project from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 819702) and from the ISF (grant number 5219/17). GY was supported by the ISF (grant number 5219/17).

REFERENCES

[1] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 2008, pp. 560–568.

[2] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2011.

[3] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, 2011, pp. 643–650.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," in *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, 2012, pp. 214–226.

[5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[6] H. Edwards and A. Storkey, "Censoring representations with an adversary," *arXiv preprint arXiv:1511.05897*, 2015.

[7] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *International Conference on Machine Learning*, 2018, pp. 3381–3390.

[8] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[10] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[11] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.

[12] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[13] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[14] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *International Conference on Machine Learning*, 2018.

[15] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 2569–2577.

[16] G. Rothblum and G. Yona, "Probably approximately metric-fair learning," in *International Conference on Machine Learning*, 2018.

[17] M. P. Kim, O. Reingold, and G. N. Rothblum, "Fairness through computationally-bounded awareness," in *Advances in Neural Information Processing Systems*, 2018, pp. 4847–4857.

[18] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," *AAAI AI, Ethics, and Society*, 2019.

[19] J. E. Roemer *et al.*, *Equality of opportunity*. Harvard University Press Cambridge, MA, 1998, no. 331.2/R62e.

[20] P. Dawid, "On individual risk," *Synthese*, vol. 194, no. 9, pp. 3445–3474, 2017, also available on ArXiv, 2014.

[21] C. M. Steele, *Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time)*. WW Norton & Company, 2011.

[22] J. T. Jost and M. R. Banaji, "The role of stereotyping in system-justification and the production of false consciousness," *British journal of social psychology*, vol. 33, no. 1, pp. 1–27, 1994.

[23] M. Kearns, A. Roth, and Z. S. Wu, "Meritocratic fairness for cross-population selection," in *ICML*, 2017.

[24] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview."

[25] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the roc curve," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 393–425, 2005.

[26] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research*, vol. 4, no. Nov, pp. 933–969, 2003.

[27] H. Narasimhan and S. Agarwal, "On the relationship between binary classification, bipartite ranking, and binary class probability estimation," in *Advances in Neural Information Processing Systems*, 2013, pp. 2913–2921.

[28] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[29] N. Kallus and A. Zhou, "The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric," *arXiv preprint arXiv:1902.05826*, 2019.

[30] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 2017, p. 22.

[31] L. E. Celis, D. Straszak, and N. K. Vishnoi, "Ranking with fairness constraints," *arXiv preprint arXiv:1704.06840*, 2017.

[32] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 2219–2228.

[33] M. J. Kearns and U. Vazirani, *An introduction to computational learning theory*. MIT press, 1994.

[34] L. Trevisan, M. Tulsiani, and S. Vadhan, "Regularity, boosting, and efficiently simulating every high-entropy distribution," in *2009 24th Annual IEEE Conference on Computational Complexity*. IEEE, 2009, pp. 126–136.

[35] A. T. Kalai, Y. Mansour, and E. Verbin, "On agnostic boosting and parity learning," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM, 2008, pp. 629–638.

[36] V. Feldman, "Distribution-specific agnostic boosting," *arXiv preprint arXiv:0909.2927*, 2009.

[37] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 2017.