A COMPLEXITY-THEORETIC PERSPECTIVE ON FAIRNESS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael Pum-Shin Kim

August 2020

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Omer Reingold, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Gregory Valiant

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

James Zou

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

Algorithms make predictions about people constantly. The spread of such prediction systems—from precision medicine to targeted advertising to predictive policing—has raised concerns that algorithms may perpetrate unfair discrimination, especially against individuals from minority groups. While it's easy to speculate on the risks of unfair prediction, devising an effective definition of algorithmic fairness is challenging. Most existing definitions tend toward one of two extremes—individual fairness notions provide theoretically-appealing protections but present practical challenges at scale, whereas group fairness notions are tractable but offer marginal protections. In this thesis, we propose and study a new notion—*multi-calibration*—that strengthens the guarantees of group fairness while avoiding the obstacles associated with individual fairness.

Multi-calibration requires that predictions be well-calibrated, not simply on the population as a whole but simultaneously over a rich collection of subpopulations $\mathcal{C}$. We specify this collection—which parameterizes the strength of the multi-calibration guarantee—in terms of a class of computationally-bounded functions. Multi-calibration protects every subpopulation that can be identified within the chosen computational bound. Despite such a demanding requirement, we show a generic reduction from learning a multi-calibrated predictor to (agnostic) learning over the chosen class $\mathcal{C}$. This reduction establishes the feasibility of multi-calibration: taking $\mathcal{C}$ to be a learnable class, we can achieve multi-calibration efficiently (both statistically and computationally). To better understand the requirement of multi-calibration, we turn our attention from fair prediction to fair ranking. We establish an equivalence between a semantic notion of domination-compatibility in rankings and the technical notion of multi-calibration in predictors—while conceived from different vantage points, these concepts encode the same notion of evidence-based fairness. This alternative characterization illustrates how multi-calibration affords qualitatively different protections than standard group notions.

Acknowledgments

Writing this thesis would not have been possible without the contributions and support of countless individuals. In my time as a Ph.D. student, I have had the opportunity to learn from many amazing colleagues, both at Stanford and across the world. First and foremost, I would like to acknowledge my advisor, Omer Reingold, for his constant mentorship and friendship. Omer has dedicated a nontrivial chunk of his life during the past four years to chatting with me on Hangouts, discussing everything from the technical subtleties in the proof of Theorem 7.28, to the day-to-day joys and challenges of doing a Ph.D. and more generally of being a person. Learning from Omer how to ask the right questions—about research and about life—has been an absolute privilege.

One of Omer's most important contributions to my research trajectory has been introducing me to other distinguished researchers, who quickly turned into trusted mentors. Guy Rothblum has been hugely influential in my graduate career, and in the development of my research. This thesis would not exist in its current form, but for Guy's visit to Stanford in the Summer of 2017 that kickstarted our collaboration. Since the original multi-calibration project, Guy has spent innumerable hours with me—sanity checking my proofs, telling me (frustratingly-accurately) where my writing needs work, and patiently listening to me think out loud—for which I am very grateful.

At Stanford, James Zou has taught me much of what I know about machine learning research. James's breadth and depth of knowledge are unmatched, and I'm thankful that he has spent considerable time to share a small slice of that knowledge with me. Lastly, Cynthia Dwork has been a tireless collaborator and a great advocate for our research efforts. Since its inception, Cynthia has been one of the champions of the field of algorithmic fairness. I admire Cynthia's clarity in vision—to discern the problems worth attacking a decade before the rest of the world catches up—and am truly appreciative of her continued mentorship and support.

I am also indebted to many other colleagues, who have served as excellent partners in crime, in research and shenanigans. With regards to the research, this thesis draws directly from joint works with fellow students Sumegha Garg, Amirata Ghorbani, Úrsula Hébert-Johnson, and Gal Yona. In no small way, this thesis should be viewed as the culmination of our collective efforts over the past years. With regards to the shenanigans, thanks to everyone in the Theory Group at Stanford for the never-ending coffee breaks, thrilling games of MUPI, and laughter along the way.

Finally, I would like to thank my friends and family. Without you, I would not be where or who I am today. For everything, thank you.

Contents

Part I

A Complexity-Theoretic Perspective on Fairness

Chapter 1

Overture

1.1 Introduction

Today, algorithmic predictions inform decisions across all aspects of life. Algorithms are making medical diagnoses, driving our cars, and running sophisticated content recommendation and advertising platforms. There is a growing recognition—within academic circles and the public consciousness—that algorithmic systems have a profound impact on our daily lives. The nominal promise of such algorithmic prediction systems, is that everyone stands to benefit from their use: individuals will receive predictions tailored to their needs, while decision-makers will use the data-driven predictions to improve their efficiency.

Despite the appeal of this narrative, a growing body of evidence suggests that algorithms—particularly machine-learned prediction systems—may perpetuate human biases, in ways that adversely affect underrepresented groups. From an advertising platform that decides whether to display housing ads on the basis of race [ASB$^+$19], to an "artificial intelligence" that penalizes résumés on the basis of gender [Vin18], to a facial recognition system that achieves near-perfect performance on white men but performance commensurate with random guessing on black women [BG18]: countless examples demonstrate how machine-learned models' predictions can be *unfair* to significant groups of individuals and underscore the need to address such systematic failures.

A notable concrete example comes from the "Gender Shades" study [BG18]. Motivated by anecdotal evidence of bias in facial recognition systems, the Gender Shades project demonstrated rigorously that a number of commercial face detectors exhibited significant performance gaps across different subpopulations. While face detectors achieved roughly

90% accuracy on a popular benchmark, a closer investigation revealed that the systems were significantly less accurate on women compared to men and on Black individuals compared to White—worse yet, this performance discrepancy compounded considerably when comparing White men ($\sim$ 100%) to Black women ($\sim$ 65%). The Gender Shades study substantiates the intuition that machine-learned classifiers may optimize predictions to perform well on individuals from the majority population, inadvertently hurting performance on historically-disenfranchised groups in significant ways.

Addressing unfairness. While the phenomenon of algorithmic unfairness seems to be widespread, in this thesis we focus our attention on algorithms that make predictions about people. Such *predictors* take as input data about an individual person and produce an inference or a judgement about that person. Often, these predictors are trained using supervised machine learning: given many individual-label (x, y) pairs, the machine-learned predictor p outputs $p(x')$, an estimate of the label y' for a previously-unseen individual x'. For instance, a university running college admissions might develop a predictor—based on historical application and graduation records—that given the grades, test scores, and demographic information of an applicant, outputs an estimate of their probability of graduation within four years. Based on such a predictor, the university might also develop a *classifier*, which outputs a binary decision of whether to accept or reject the individual applicant.

A common strategy for addressing undesirable discrimination in machine learning begins by identifying some form of bias in the training pipeline of a given machine-learned model—often in the training data—and then re-training the model to correct for the observed biases. Patching flawed models to address problematic behavior is a necessary practice, but does not present a sustainable long-term solution for promoting fairness in algorithmic prediction systems. Indeed, adopting such a reactive approach to mitigating harms is inherently limited: the only forms of discrimination that can be prevented are those which have already been observed.

A more subtle—but also more essential—challenge with this strategy for addressing unfairness in prediction systems is determining what we consider to be "undesirable discrimination." Already in this introduction, we have used words like "unfair" and "biased" and "discriminatory" imprecisely—suggestive of some implicitly-agreed-upon "problematic" behavior but without clear delineation. Each of these words carries a different set of connotations in common parlance, and words like "bias" and "discrimination" are also used within statistical machine learning to communicate specific technical concepts. Fundamental to any conversation about fairness in algorithmic prediction systems, we must first ask: *When we indict an algorithm for being "unfair," what do we actually mean?*

Framing and perspective. We frame our answer to this question within broader investigations into the foundations of "responsible computing." This nascent field—borne out of the theoretical computer science community—aims to model socio-technical problems formally using mathematical, statistical, and computational analysis to wrestle with social challenges that arise when algorithms interact with people. Importantly, work in this area emphasizes the importance of abstraction and definitions.

Establishing effective abstractions is one of the greatest success stories of theoretical computer science as a field of study. Our objective—by removing instance-specific details—is to capture the elements of the problem (both human and technical) most salient to the issue of algorithmic unfairness, so that our results will suggest effective general-purpose solutions. Paired with the right level of abstraction, we investigate and develop rigorous mathematical notions of fairness. Such definitions allow us to discuss the issues of algorithmic unfairness precisely, rather than simply intuitively. The eventual goal of this research program is to develop statistical and algorithmic techniques that come with *provable guarantees of fairness*, much in the way that the cryptographic protocols that secure our online communications come with formal guarantees of security.

Setting the Stage

Research efforts to formalize algorithmic fairness began roughly a decade ago [PRT08, KC11, KAS11, DHP+12]. Since these early works, research aiming to address issues of unfair discrimination in algorithmic systems has exploded, especially within the machine learning community. To date, almost all approaches to defining fairness fall into one of two paradigms: group fairness and individual fairness. Each paradigm has appeals and drawbacks, which we discuss next.

Group fairness. Most of the research on algorithmic fairness has focused on achieving so-called group fairness notions [ZWS+13, ES15, Cho17, KMR17, HPS16, BCZ+16, FSV16, KLRS17, MCPZ18, MPB+18]. Defining a notion of group fairness involves identifying a protected group, (e.g., defined on the basis of gender or race) as well as a statistic of interest, (e.g., the selection rate or statistical bias); then, the notion requires that the statistic of interest “looks right” in the protected group compared to the rest of the population.

The ease of defining and implementing these notions makes them especially appealing for machine learning practitioners; nevertheless, the broad-strokes statistical nature of group notions tend to provide very weak protections. The foundational work of [DHP$^+$12] was the first to identify the significant shortcomings of group notions. Exploiting the on-average nature of the constraints, [DHP$^+$12] demonstrated how predictors could satisfy the “letter” of these marginal statistical constraints, while still materially discriminating against individuals from the protected groups.

Complicating matters further, most notions of group fairness are known to be mutually-incompatible. For instance, spurred by a debate raised in the popular press over the COMPAS recidivism risk prediction system [ALMK16], there has been lots of recent interest in the incompatibility of two popular notions: calibration and balanced error rates [KMR17, Cho17, PRW$^+$17]. While both notions of fairness are simple to state and have obvious merits (as well as particular failure modes) no nontrivial predictor can simultaneously satisfy both notions. Such impossibility results dash any hope of strengthening the protections of group notions by enforcing different group notions on the same predictor.

Fairness through awareness. To address the shortcomings of group notions, the work of [DHP$^+$12] proposed an alternative paradigm for defining fairness, dubbed “fairness through awareness.” This framework takes the perspective that a fair classifier should *treat similar individuals similarly*, formalizing this abstract goal by assuming access to a task-specific similarity metric that encodes which pairs of individuals should receive similar predictions. The proposed *individual fairness* notion requires that if the distance between two individuals (according to the metric) is small, then their predictions cannot be very different.

While the approach of fairness through awareness offers a theoretically-principled way to allow for high-utility predictions while ensuring fairness, a challenging aspect of this approach is the assumption that the similarity metric is known for all pairs of individuals. Deciding on an appropriate metric is itself a delicate matter and could require input from sociologists, legal scholars, and specialists with domain expertise. For instance, in a loan repayment setting, a simple seemingly-objective metric might be a comparison of credit scores. A potential concern, however, is that these scores might themselves encode historical discrimination; in such cases, human judgment might be needed on a case-by-case basis.

Thus far, the challenges involved in obtaining an effective fairness metric have stymied the adoption of individual fairness within the machine learning community.

Beyond groups and individuals. Given the appeals and shortcomings of group and individuals fairness, a natural question arises:

Are there meaningful notions of fairness
that bridge the gap between group and individual notions?

In this thesis, we address this question, presenting a novel framework for defining and enforcing notions of fairness for prediction tasks that reside between group and individual notions. These *multi-group* fairness notions are defined by requiring a group fairness notion to hold, not just on the population as a whole, but instead on each group from a vast collection of diverse subpopulations. The multi-group perspective was originally introduced in two concurrent works [HKRR18, KNRW18] and has been studied subsequently in works of the author and others [KRR18, KGZ19, KNRW19, GKR19, DKR$^+$19, SCM20].

Enforcing group notions to hold over a richer class of groups naturally provides quantitatively stronger guarantees. This thesis argues that when we take the collection of subpopulations to be sufficiently-expressive, unexpectedly, multi-group fairness notions provide *qualitatively* stronger protections, akin to those of individual fairness. If achieving individual-level fairness is the information-theoretic ideal, then achieving multi-group fairness is the complexity-theoretic ideal.

1.2 A Complexity-Theoretic Perspective on Fairness

The remainder of this chapter is dedicated to overviewing the thesis as a whole. Our goal is to communicate the intuition behind our contributions, emphasizing their qualitative significance rather than the quantitative details. Ideally, the reader will be able to discern the moral of the story from the Overture alone and can refer to the later chapters for more detailed exposition. As such, we introduce some basic notation needed for the definitions and theorems. Rigorous preliminaries and assumptions are introduced at the start of Chapter 2.

Basic preliminaries. Throughout, we will denote individuals' features using $x \in \mathcal{X}$ and their binary outcome (or label) by $y \in \mathcal{Y} = \{0, 1\}$. We imagine that individuals are distributed according to a distribution $x \sim \mathcal{D}$. Given an individual x, we imagine their outcome to be distributed according to $y \sim \mathrm{Ber}(p^*(x))$ for some $p^*(x)$. Importantly, this means that the function

$$p^* : \mathcal{X} \to [0,1]$$

is the information-theoretic optimal predictor of y given x. We denote the distribution of individual-outcome pairs as $(x, y) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. Our goal will be to recover predictors $p : \mathcal{X} \to [0,1]$ that approximate p^*, in a fair manner, using a bounded set of samples from the underlying distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$.

Reasoning about p^* is a useful abstraction, allowing us to compare against the ideal predictions if we had unbounded statistical and computational resources. Throughout, we assume that the features representing individuals $\mathcal{X}$ are expressive enough that $p^* : \mathcal{X} \to [0,1]$ captures all meaningful variation across the population in the likelihood of an outcome given an individual. Importantly, p^* will only be used as a reference point—we never need direct access to p^* to learn multi-calibrated predictors.

Calibration. The bulk of this thesis is devoted to studying the flagship notion of multi-group fairness, called *multi-calibration*. The starting point for multi-calibration is the statistical condition known as group calibration. Intuitively, calibrated predictions "mean what they say" regardless of group membership: suppose that p is a calibrated predictor and x represents a certain individual's features; then, given the prediction $p(x)$, learning the individual's group membership status $x \in S$ (say, membership in the majority vs. minority population), should not change the posterior belief about their outcome $y \in \{0,1\}$. In many application domains, obtaining calibrated predictions is a necessary baseline for fair treatment. For instance, in medical settings risk scores are often involved in triage—the highest risk patients receive care first. When risk scores are poorly calibrated across demographic groups, significant harms may occur to historically-marginalized groups. For instance, racial bias in medical risk predictors affects access to treatment and eventual patient outcomes, due to under-estimating the risk for Black patients relative to similarly-sick White patients [OPVM19].

Formally, a predictor p is calibrated on a group $S \subseteq \mathcal{X}$ if for all possible values[1] $v \in [0,1]$,

[1] We avoid measure theoretic complications by assuming $\mathcal{X}$ is discrete and the support size of p is finite, handling this issue formally in subsequent chapters.

the predictions amongst the individuals $x \in S$ who receive value v are accurate on average.

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ p^*(x) \mid p(x) = v,\ x \in S\ ] \approx v.$$

First note that the optimal predictor p^* is perfectly calibrated over all subpopulations S. Because $\mathbf{E}_{\mathcal{D}} [\ p^*(x)\ ] = \mathbf{Pr}_{\mathcal{D}_{\mathcal{X} \times \mathcal{Y}}} [\ y = 1\ ]$, all calibrated predictions $p(x)$ can be meaningfully thought of as encoding a "probability" that for a given individual x, the outcome will be $y = 1$. While group calibration requires that predictions mean what they say across groups, unfortunately, predictions can be calibrated without saying very much. Indeed, calibrated predictions can ignore variation within groups leading to potentially-harmful "algorithmic stereotyping." This form of unfairness stems from the fact that predicting the average value for a group is always calibrated on the group.

In particular, suppose we enforce group calibration over two disjoint groups $S, T \subseteq \mathcal{X}$. Suppose in each group there is significant variation within p^*; for instance, perhaps $p^*(x) \in \{0, 1\}$ is a 50:50 mix of positive and negative outcomes in each group. In T, which we'll think of as the majority, we predict perfectly according to $p(x) = p^*(x)$. In S, the minority, we predict according to the mean of the optimal predictions $p(x) = 0.5 = \mathbf{E}_{\mathcal{D}} [\ p^*(x) \mid x \in S\ ]$. These predictions ignore all variation within S—stereotyping based solely on their group membership. But it's easy to verify that p is actually calibrated: amongst the individuals who receive prediction 0.5, the expectation of $p^*(x)$ is 0.5. Using the predictions of p for medical triage, as above, could lead to significant harms to the minority group S. All of the risky patients in T will receive attention before those in S; despite similar risk, they receive different predictions. The predictor p satisfies calibration, but completely overlooks the individuals in S who were "qualified" or deserving of attention.

Strengthening calibration. Group calibration fails to offer meaningful protections—even to the groups it designates as "protected"—because the constraints, which are defined marginally over the groups, do not account for meaningful variation within the groups. As in the example above, while calibration required on-average consistency over the group S, it required no such consistency for the group of qualified individuals within S.

But suppose we were able to enforce calibration over a set of highly-qualified individuals in S. Specifically, suppose we identify a set $S' \subseteq S$ where $\mathbf{E}_{\mathcal{D}} [\ p^*(x) \mid x \in S'\ ]$ is large (close to 1). Then, to be calibrated over this group, the predictions would also need to be reasonably accurate on the *individuals* within the group. Intuitively, even enforcing calibration over the group of unqualified individuals within S (i.e., likely to have $y = 0$) could help protect the qualified individuals; by breaking the symmetry within S, under-confident predictions on the qualified individuals would no longer appear calibrated on these groups. In this sense, identifying meaningful structure within S—subpopulations S' where membership in S' correlates with the outcome y—may help us protect S as a whole.

The challenge with this approach, of course, is that the groups of qualified and unqualified individuals are hard to anticipate. Indeed, the reason for developing a predictor in the first place is to estimate these groups. Thus, rather than attempting to single out the group of qualified individuals for protection, we could instead try to offer protections to as many groups as possible. This goal leads us to the definition of multi-calibration, which guarantees calibration over a collection of subpopulations $\mathcal{C}$.

Definition (Multi-Calibration)**.** *Suppose $\mathcal{C}$ is a collection of subpopulations. A predictor $\tilde{p}$ is $\mathcal{C}$-multi-calibrated if it is calibrated over all subpopulations in $\mathcal{C}$; that is, for all $S \in \mathcal{C}$ and all supported values $v \in [0,1]$*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}}[\ p^*(x) \mid \tilde{p}(x) = v,\ x \in S\ ] \approx v.$$

In other words, a multi-calibrated predictor means what it says across every subpopulation within the class $\mathcal{C}$. When we discuss multi-calibration more formally, we will discuss $(\mathcal{C}, \alpha)$-multi-calibration, where the equality must hold up to accuracy $\alpha \geq 0$. This notion of approximate multi-calibration is important when discussing how to learn multi-calibrated predictors from data, but is not essential to discuss its properties as a notion of fairness.

Technically, multi-calibration is a generalization of group calibration; for instance, taking our collection of groups to be the majority and minority population as before, $\mathcal{C} = \{S, T\}$, $\mathcal{C}$-multi-calibration is simply group calibration. The question becomes how to choose $\mathcal{C}$ to provide strong protections, while maintaining a feasible notion. Ideally, we would offer protection to *every* group; certainly, if we enforced multi-calibration for such a collection, then the set of qualified individuals would receive protection. The problem with such a proposition is that it is statistically impossible to protect all groups. In particular, protecting every group with calibration would require *individual-level* recovery of p^*. In most settings, learning the optimal predictor exactly is simply not possible.

1.2.1 Calibration for the Computationally-Identifiable Masses

Drawing inspiration from cryptography, we take $\mathcal{C}$ to be the collection of "efficiently-identifiable" subpopulations. In particular, we imagine the collection $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ to be specified by a class of functions with bounded complexity: for each subpopulation $S \in \mathcal{C}$, the characteristic function $\chi_S : \mathcal{X} \to \{0,1\}$ that determines set membership $x \in S$ can be evaluated within the class. We think of this complexity class of functions as a "computational bound." Concretely, the bound is specified by a collection of simple (i.e., low complexity) but expressive functions—for instance, small conjunctions, decision trees, linear functions, even bounded neural networks.

The collection of subpopulations can intersect in nontrivial ways and need not correspond to groups we traditionally think of as "protected." By selecting a computationally-bounded class $\mathcal{C}$, we strike a balance between simplicity and expressivity: individually, each subpopulation is simple enough to be described succinctly; collectively, the class is expressive enough to capture all subpopulations that can be understood within a given bound on statistical and computational resources. Importantly, by restricting our attention to the computationally-bounded subpopulations $\mathcal{C}$ (rather than all subpopulations), we allow for the possibility of learning $\mathcal{C}$-multi-calibrated predictors from a small set of data.

For instance, in an application where individuals are encoded as d-dimensional boolean feature vectors, $S \in \mathcal{C}$ could be the subpopulations whose characteristic function $\chi_S : \mathcal{X} \to \{0,1\}$ is defined by 3-way conjunctions over the features.

$$\mathcal{C} = \{\ S \subseteq \mathcal{X} : \chi_S(x) = x_i \wedge x_j \wedge x_k \text{ for } i,j,k \in [d]\ \}$$

Intuitively, choosing such a collection $\mathcal{C}$ goes beyond defining "protected" groups, and instead defines "meaningful" groups. Rather than ensuring calibration over the demographic groups of "women" and "Black individuals" marginally, multi-calibration over this collection $\mathcal{C}$ might require calibration over "Black women, who wear glasses," as one of the $\Theta(d^3)$ conjunctions in $\mathcal{C}$. Importantly, $\mathcal{C}$ is defined over all of the features—not just the "sensitive" features, and the more expressive the collection $\mathcal{C}$ is, the stronger the protections that are afforded by $\mathcal{C}$-multi-calibration. Even if wearing glasses isn't considered a sensitive attribute, incorporating the attribute may help to reveal structure within the data distribution, and in turn, will help to protect the protected populations.

To illustrate this point concretely, we show formally that multi-calibrated predictors guarantee accuracy on any set of qualified individuals $S \in \mathcal{C}$ that the collection identifies. Given any predictor p, we can consider naively rounding it to $\{0, 1\}$ by using a threshold of $1/2$; on any qualified set $S \in \mathcal{C}$, the rounded classifier must obtain small error.

Theorem 1. *Suppose* $\mathbf{E}_{\mathcal{D}}\left[\ p^*(x) \mid x \in S\ \right] \geq 1 - \varepsilon$ *for some* $S \in \mathcal{C}$*. Then, after naive rounding, any* $\mathcal{C}$*-multi-calibrated predictor obtains classification error on* S *of at most* $O(\varepsilon)$*.*

In other words, a key strength of multi-calibration lies in its ability to protect any set of qualified individuals that can be efficiently identified from the data at hand. An analogous inequality holds for any $S \in \mathcal{C}$ that correlates with $y = 0$ rather than $y = 1$. This structural result demonstrates the importance of taking $\mathcal{C}$ to be an expressive complexity class: if a predictor $\tilde{p}$ is multi-calibrated over a collection $\mathcal{C}$, then $\tilde{p}$ must incorporate all information about p^* that can be identified within the computational bound defined by $\mathcal{C}$. By taking $\mathcal{C}$ to be as expressive as the data and computational constraints allow, we increase our ability to identify and protect groups of highly-qualified individuals–even if they're part of a group that is disadvantaged on average.

The notion. Multi-calibration is an "evidence-based" notion of fairness. This framework takes the perspective that, first and foremost, predictors should investigate the evidence—given through training data—to understand the distribution of individuals and outcomes to the extent that the statistical and computational resources allow. For a given computational bound $\mathcal{C}$, multi-calibration defines a strong set of consistency constraints that can be tested within the bound from a small set of data. Practically, the collection of meaningful groups $\mathcal{C}$ can be specified by a regulator or virtuous learner, based on the richness of available training data and computational power. By varying the complexity of $\mathcal{C}$, the guarantees of $\mathcal{C}$-multi-calibration interpolate between those of group fairness and individual fairness, strengthening as $\mathcal{C}$ becomes more expressive.

Our emphasis in much of the thesis is on the importance of multi-calibration as a notion of fairness for prediction tasks. No matter how you get your hands on a predictor—through empirical risk minimization or specialized algorithms—the definition of multi-calibration sets a rigorous standard of what we should expect from our machine-learned predictors. Multi-calibration articulates that the most meaningful populations to reason about are those that can be identified efficiently from data. And while multi-calibration only requires you to protect groups within a computationally-bounded class $\mathcal{C}$, it requires you to protect *every* such group. The contrapositive perspective sheds some light on the effectiveness of multi-calibration: if a predictor p does not satisfy multi-calibration over a computationally-bounded class $\mathcal{C}$, then there is some simple explanation (i.e., some $S \in \mathcal{C}$) for why the predictions are flawed and how to improve them.

Developing Multi-Group Fairness

Concurrent with the development of multi-calibration in [HKRR18], another group of researchers [KNRW18] also recognized the need to provide guarantees of fairness between the level of groups and individuals. Their work promoted analogous notions of *multi-group parity*[2] based on group fairness through parity of selection rates (demographic parity) and of false negative rates (equalized opportunity). Specifically, their notions require that for all subpopulations within a given class $\mathcal{C}$, the rate is similar to that of the overall population; for instance, for multi-group demographic parity the selection rate of a classifier $f : \mathcal{X} \to \{0, 1\}$ must be close to some global selection rate β for all subpopulations $S \in \mathcal{C}$.

$$\left| \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ f(x) = 1\ ] - \beta \right| \leq \alpha.$$

Yet another notion, which we introduced in [KRR18], defines a multi-group analogue of metric fairness [DHP$^+$12], which—depending on the selected fairness metric—may provide an intermediary multi-group notion. All of these multi-group fairness notions start with many of the same motivations as multi-calibration, and there are a number of technical similarities between the works of [HKRR18, KNRW18, KRR18], as well as their empirical follow-ups [KGZ19, KNRW19]. These notions also differ in important technical and conceptual ways, which we explore in Part III of this thesis.

1.3 Overview of Results

Next, we give a high-level overview of the results of this thesis. Part II of the thesis studies the feasibility of multi-calibration as a notion of fairness in theory and in experiments. Part III investigates other multi-group notions of fairness and their relation to multi-calibration. First, we characterize when achieving multi-calibration is feasible. We demonstrate that the complexity of $\mathcal{C}$-multi-calibration is tightly connected to the complexity of the class $\mathcal{C}$ in a number of senses.

[2] Note that [KNRW18] refer to their notions as "rich subgroup fairness."

- *Representing multi-calibrated predictors*—To begin, we show that the complexity of representing multi-calibrated predictors scales modestly with the complexity of representing the subpopulations $S \in \mathcal{C}$. Provided that for each $S \in \mathcal{C}$, the characteristic function $\chi_S : \mathcal{X} \to \{0, 1\}$ has a succinct description (i.e., bounded complexity), then there is a $\mathcal{C}$-multi-calibrated predictor $\tilde{p}$ with a (slightly less) succinct description. Importantly, the complexity of $\tilde{p}$ is independent of the complexity of p^*, depending only on the complexity of $\mathcal{C}$ (and the desired accuracy guarantee).

- *Learning multi-calibrated predictors*—Next, we show that the learnability of $\mathcal{C}$-multi-calibrated predictors is tightly connected to the task of learning the class $\mathcal{C}$. Statistically, we demonstrate that $\mathcal{C}$-multi-calibrated predictors can be learned from a small number of samples scaling with $\log(|\mathcal{C}|)$. Computationally, we demonstrate that learning a $\mathcal{C}$-multi-calibrated predictor is equivalent to the task of agnostic learning the class $\mathcal{C}$. Theoretically, this equivalence should be viewed as a hardness result: agnostic learning is a notorious hard problem from computational complexity theory. Practically, however, the tight equivalence yields an effective reduction from "multi-calibrated" learning to standard machine learning. In three case studies, we demonstrate the empirical effectiveness of using off-the-shelf regression tools to implement the multi-calibration framework to mitigate disparities in prediction quality.

Collectively, these results demonstrate that when we take $\mathcal{C}$ to be a class of efficiently-identifiable functions, we can learn $\mathcal{C}$-multi-calibrated predictors from data efficiently, regardless of the complexity of p^*, in a way that improves the prediction quality across all identifiable subpopulations. In the next part, we consider multi-group fairness beyond multi-calibration.

- *Fairness through "computationally-bounded" awareness*—Fairness through awareness [DHP+12] advocated treating similar individuals similarly, based on a task-specific fairness metric. When the task calls for a metric where two individuals x, x' are similar if $|p^*(x) - p^*(x')|$ is small, then multi-calibration gives a strong, feasible notion of protections—even when p^* cannot be recovered. To handle more general similarity metrics, we introduce another notion, *multi-metric fairness.* We show algorithmic feasibility of this notion from a small number of samples from the underlying metric. Multi-metric fairness gives a strong multi-group metric fairness guarantee that is achievable in settings where individual-level metric fairness is infeasible.

- *Parity, information, and multi-calibration*—We show that the guarantees of multi-calibration can be understood through the information-theoretic concept of *refinements*. This perspective allows us to better compare the guarantees between multi-calibration and other notions of multi-group fairness based on parity [KNRW18]. In all, we argue that starting with a multi-calibrated predictor, then adjusting the predictions to reduce disparity may promote representation of underserved groups while avoiding the negative delayed impacts that may arise from blindly requiring parity (as described by [LDR$^+$18]).

Finally, we turn our attention to understanding notions of fairness in rankings. Studying ranking—rather than prediction—is motivated for a number of reasons. Often, the impetus for developing a predictor is not actually to understand individuals' absolute risk, but rather to understand their rank within the population. Further, recovering a ranking requires understanding the population globally, not just within the majority. As such, formalizing the problem of fairness in rankings also helps to clarify what we should expect from predictors.

- *Evidence-based rankings*—We initiate the study of fairness in rankings. We introduce a semantic multi-group notion called *domination-compatibility*: if a ranking favors a group T over another S (i.e., if T *dominates* S), then in reality, the expected outcome of T should exceed that of S (i.e., the *evidence* about $\mathbf{E}_{\mathcal{D}_T}[\ p^*(x)\ ] \gg \mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ]$ should be consistent with the ranking). Surprisingly, we show that if we enforce this notion across a rich enough collection of subpopulations (informed by the ranking itself), then the class of $\mathcal{C}$-domination-compatible rankings is exactly the set of rankings induced by $\mathcal{C}$-multi-calibrated predictors.

The equivalence between these notions—an intuitive notion about the relative ranking of subgroups and a technical notion about absolute consistency of subgroups' predictions with the underlying risk—bolsters the perspective that obtaining multi-calibrated predictors requires globally-consistent learning across all identifiable subpopulations.

Statement of results. In what follows, we give a more detailed statement of the results from each section. For the sake of presentation we state the theorems informally, assuming that the subpopulations $S \in \mathcal{C}$ all have $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] = \Omega(1)$, and take all failure probabilities to be $\beta = \delta = \Omega(1)$. The informal theorems are substantiated through formal propositions and proofs for arbitrary settings of the relevant parameters in Chapters 2-7.

1.3.1 Learning Multi-Calibrated Predictors

As is clear from the definition, if we can obtain a highly-accurate approximation to the information-theoretic optimal p^*, then multi-calibration is feasible (for any computational class $\mathcal{C}$). A concern, however, is that the definition might be restrictive enough that learning a multi-calibrated predictor would require learning p^*. Typically, information-theoretic recovery of the optimal predictor is intractable . Thus, to establish the effectiveness of multi-calibration as a notion of fairness, we must establish that for meaningful collections $\mathcal{C}$, learning p^* is not necessary. The first set of results of the thesis demonstrate the feasibility of multi-calibration: for any true underlying optimal predictions p^*, there exists a $\mathcal{C}$-multi-calibrated predictor whose complexity depends only on $\mathcal{C}$ and the desired accuracy α—independent of p^*.

Theorem 2. *For any distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, any collection of subpopulations $\mathcal{C}$ and accuracy parameter α, there exists a $(\mathcal{C}, \alpha)$-multi-calibrated predictor whose complexity scales as $O\left(\text{complexity}(\mathcal{C}) \cdot \text{poly}(1/\alpha)\right)$.*

In other words, even if the distribution of y conditioned on x (i.e., p^*) is arbitrarily complex, the complexity of multi-calibrated predictors only scales with the complexity of the subpopulations we aim to protect and the desired calibration accuracy. Here, we state the theorem informally using an abstract measure of complexity($\mathcal{C}$). Concretely, circuit complexity [Sha49] serves as a natural measure for which the theorem applies.

Sample complexity of multi-calibration. With knowledge that multi-calibrated predictors can be represented succinctly, the next question to ask is whether we can learn them efficiently, both in terms of statistical and time complexity. In fact, Theorem 2 is a corollary of a generic algorithm we give for learning multi-calibrated predictors. The algorithm is an iterative boosting-style algorithm, and Theorem 2 follows directly from an upper bound on the number of required iterations.

Simply stated, in each iteration, the algorithm searches for a subpopulation $S \in \mathcal{C}$ where the current predictions are mis-calibrated. If such a subpopulation exists, the algorithm calibrates the predictions locally and continues; if no such subpopulation exists, the algorithm terminates returning a multi-calibrated predictor. Importantly, we show that this procedure must terminate in a bounded number of iterations, and that the re-calibration steps can be implemented with statistical validity from a small number of labeled samples.

Theorem 3. *There exists a learning algorithm that, for any collection of subpopulations $\mathcal{C}$ and accuracy parameter α, given $m = O(\mathrm{poly}(\log(|\mathcal{C}|)/\alpha))$ labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, learns a $(\mathcal{C}, \alpha)$-multi-calibrated predictor in $O(\mathrm{poly}(1/\alpha))$ iterations.*

That is, from a set of labeled samples whose cardinality grows modestly with the complexity of $\mathcal{C}$ and α, it is statistically possible to learn a $(\mathcal{C}, \alpha)$-multi-calibrated predictor. Note that the sample complexity dependence on $\mathcal{C}$ and α—growing with $\log(|\mathcal{C}|)$ and $\mathrm{poly}(1/\alpha)$—is typical of learning problems; however, the way that we achieve this upper bound is non-standard. Multi-calibration (and even group calibration) constrains predictors in a self-referential manner. Indeed, for any calibrated predictor $\tilde{p}$, the true expectation of the outcome is v, *amongst the individuals x who received score $\tilde{p}(x) = v$*. For these reasons, the standard approach for establishing sample complexity through uniform convergence arguments present significant challenges for learning multi-calibrated predictors.[3]

Instead, our algorithm works in the statistical query framework. We develop an algorithm whose termination and correctness depend on the correctness of access to a statistical query oracle. Still, when learning a multi-calibrated predictor from data, our iterative algorithm needs to reason about constraints that depend on the current predictions—and thus, on the prior access to the data set. Without care, the multi-calibration constraints lead to a classic statistical pitfall: asking future queries based on the answers to past queries. The classic tools used to guarantee statistical validity assume that all queries of the data are fixed before the analysis begins; in the worst case, allowing the data analyst to queries adaptively leads to rapid overfitting of the data set [DFH$^+$15c].

To counteract overfitting to the sample, we leverage a recently-discovered powerful connection between generalization in adaptive data analysis and differential privacy [DFH$^+$15c, DFH$^+$15b, DFH$^+$15a, BNS$^+$16, JLN$^+$20]. This line of work demonstrates how data analysis under differential privacy is robust to adaptivity in analysis: many adaptively selected queries can be answered before overfitting the sample, provided these queries are answered in a differentially private manner. By implementing the statistical mechanisms in a differentially private manner, we obtain the improved sample complexity. As a consequence of our analysis strategy, we show that multi-calibration is compatible with differential privacy, a strong notion of privacy protections for individuals in our training set.

[3] Very recent work of [SCM20] addresses the question of uniform convergence for the multi-calibration loss (i.e., the difference between sample and distributional violation to the multi-calibration constraints). In fact, their bounds are the first to establish uniform convergence bounds even for the well-studied notion of calibration.

Corollary. *For $\varepsilon = \Theta(\alpha)$ and $\delta > 0$, there exists a (ε, δ)-differentially-private algorithm achieving the guarantees of Theorem 3 for learning a $(\mathcal{C}, \alpha)$-multi-calibrated predictor.*

Computational complexity of multi-calibration. While the algorithm from Theorem 3 has bounded iteration complexity, each iteration runs through every subpopulation in the collection $S \in \mathcal{C}$. As we want to think of $\mathcal{C}$ as a large, expressive class of functions, in many application domains, running time that grows as $\Omega\left(|\mathcal{C}|\right)$ may be prohibitive. Naturally, the question to ask is when we can improve over the naive search strategy. We answer this question by showing a tight equivalence between the problem of learning $\mathcal{C}$-multi-calibrated predictors and agnostic learning the class $\mathcal{C}$.

Theorem 4. *Learning a $(\mathcal{C}, \alpha)$-multi-calibrated predictor is $O(\mathrm{poly}(1/\alpha))$-time equivalent to the problem of agnostic learning the class $\mathcal{C}$.*

The bidirectional nature of this equivalence presents both good and bad news. The bad news is that agnostic learning is a notoriously hard problem in computational complexity theory. Thus, under natural complexity and cryptographic assumptions [Val84, GGM84], learning $\mathcal{C}$-multi-calibrated predictors is intractable for sufficiently-expressive choices of $\mathcal{C}$. The good news is that the equivalence suggests an effective reduction from the seemingly-harder task of $\mathcal{C}$-multi-calibrated learning to the standard task of machine learning with the class $\mathcal{C}$. Indeed, all of practical machine learning is agnostic in nature. And despite the theoretical hardness, machine learning has proved exceptionally effective at extracting structure from noisy, complex data.

Thus, Theorem 4 suggests a practical strategy for leveraging the power of "vanilla" machine learning techniques (e.g., logistic regression or decision tree learning) to obtain multi-calibrated predictors. Unlike many novel theoretical results, this strategy is not hypothetical: we demonstrate the effectiveness of the multi-calibration using off-the-shelf machine learning tools across an array of case studies, from facial recognition to income prediction to medical diagnosis. While preliminary in nature, the experiments demonstrate concretely that the multi-group paradigm maintains the practical efficiency of standard group fairness notions, while providing significant improvements to the performance of the learned models, across a wide array of application domains. These empirical results suggest a promising future for deploying multi-calibration at scale.

1.3.2 Fairness through Computationally-Bounded Awareness

Multi-calibration applies the multi-group fairness paradigm by starting with a group notion and extending it to hold over subpopulations to strengthen the protections. A different approach to the paradigm starts with a desirable individual fairness notion and relaxes it to hold over subpopulations to improve the algorithmic complexity. We apply this approach to the metric fairness notion of [DHP$^+$12], which requires a predictor to be Lipschitz with respect to the task-specific fairness metric. Concretely, the notion requires that for all pairs of individuals $x, x' \in \mathcal{X}$

$$\left|p(x) - p(x')\right| \ \leq \ d(x, x').$$

We extend this individual notion to a multi-group notion that we call *multi-metric fairness.* Rather than enforcing the metric constraint to hold over every pair of individuals, we can require an on-average Lipschitzness condition to hold over all subpopulations from a collection $\mathcal{C}$. That is, we take an expectation over pairs of subpopulations, and for all $S, T \in \mathcal{C}$ require that

$$\mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_S \\ x' \sim \mathcal{D}_T}} \left[\ \left|p(x) - p(x')\right| \ \right] \ \leq \ \mathop{\mathbf{E}}_{\substack{x \sim \mathcal{D}_S \\ x' \sim \mathcal{D}_T}} \left[\ d(x, x') \ \right] + \tau$$

for some small constant τ. Intuitively, we think of the metric d as defining the ideal information-theoretic set of similarity constraints; thus, multi-metric fairness provides a computational relaxation of this goal. The advantage of this relaxation is that it only requires very limited access to the metric. Whereas individual metric fairness requires a metric value for all pairs of individuals, multi-metric fairness only requires the expectations over large groups, which can be estimated from a small set of samples.

Theorem 5. *Given a weak agnostic learning oracle for the class $\mathcal{C}$ and $O(\log(|\mathcal{C}|) \cdot \mathrm{poly}(1/\tau))$ samples from the metric d, there exists an oracle-efficient algorithm for learning a τ-optimal $(\mathcal{C}, d)$-multi-metric fair linear hypotheses.*

As with multi-calibration, the algorithmic result works in general, and is computationally efficient when (weak agnostic) learning over the collection $\mathcal{C}$ is efficient. Note that we can loosely view the guarantees of multi-calibration as a form of multi-metric fairness, where we take the metric to be the statistical distance $d(x, x') = |p^*(x) - p^*(x')|$. The feasibility of multi-metric fairness (over the restricted class of linear functions) shows that we can enforce multi-group notions where consistency with p^* is not necessarily the ideal.

Comparing multi-calibration and multi-group parity. Beyond the notions of multi-calibration and multi-metric fairness, a popular approach to multi-group fairness is based on group parity. Notably, [KNRW18] developed parallel notions of multi-group fairness for the notions of demographic parity and equalized opportunity. This work (concurrent with [HKRR18]) also demonstrated a tight connection between the computational complexity of learning classifiers satisfying the multi-parity notions and weak agnostic learning the class $\mathcal{C}$. Another contemporary work of [HSNL18] also takes a multi-group fairness perspective, requiring parity in accuracy across groups (i.e., no group will experience loss significantly greater than any other). Their approach—based on distributionally robust optimization—ensures this relative performance guarantee across all sufficiently-large groups. These approaches towards multi-group fairness based on parity bear a number of technical and conceptual similarities to multi-calibration.

But there are also key differences that distinguish multi-calibration as a solution concept. The first major conceptual distinction between these notions is where they "bottom out." The easiest way to see that multi-calibrated predictors exist is noting that perfect predictions (i.e., according to p^*) satisfy the multi-calibration constraints. For multi-group parity (or any notion of parity), utility and fairness are portrayed as at odds with one another, and feasibility is established by arguing that useless predictions (e.g., those that treat everyone the same) are feasible.

Concretely, this means that the role of the collection of subpopulations $\mathcal{C}$ has different interpretations for each notion. For $\mathcal{C}$-multi-calibration taking $\mathcal{C}$ to be as expressive as possible—including complex subpopulations defined over a rich set of features—improves the resulting fairness guarantee. At the extreme, if we took $\mathcal{C}$ to be defined by all efficient computations, then $\mathcal{C}$-multi-calibrated predictors are computationally indistinguishable from the information-theoretic optimal predictions p^*. For $\mathcal{C}$-multi-parity, the richer we take $\mathcal{C}$, the more likely it is that the only feasible solution is a trivial predictor that makes no distinctions between any individuals. Thus, [KNRW18] advocate defining $\mathcal{C}$ only over the set of "sensitive attributes" and auditing within intersectional protected groups.

Do no harm. The typical motivation for enforcing notions of parity is to correct for historical discrimination, possibly reflected through biased training data. The problem with correcting for these biases while learning, however, is that there may be significant qualitative differences in the solutions satisfying parity that the optimization algorithm does not account for. The effective differences between multi-group parity and multi-calibration comes into focus, when considering a recent work that introduced the idea of "delayed impact" of fairness notions [LDR+18]. This work identifies ways in which enforcing parity-based notions may actually lead to long-term negative impacts on underrepresented groups that these notions are meant to protect, due to over-selection within the groups. The model is stylized but natural, and raises a serious concerns about blindly applying parity-based notions to predictors learned through constrained optimization. Because multi-group parity intentionally requires parity to hold across a diverse set of subpopulations, it may be particularly prone to causing such negative impacts due to over-selection.

To understand the relationship between multi-calibration and multi-parity notions, as well as their potential for negative impact, we revisit the setting of [LDR+18]. We analyze the setting through the information-theoretic notion of *refinements* of predictors. We show a new interpretation of multi-calibration as simultaneous refinement over a rich class of subpopulations: multi-calibrated predictors incorporate all of the information accessible to the class. Leveraging this characterization, we demonstrate that for a natural predict-then-select strategy for devising a classifier, enforcing multi-calibration through refinement improves many quantities of interest for fair classification.

Theorem 6. *Suppose $\tilde{p}$ is a $\mathcal{C}$-multi-calibrated refinement of a predictor p. Then, for all groups $S \in \mathcal{C}$ and all selection rates within the group, the true positive rate, false positive rate, and positive predictive value over S can only improve in $\tilde{p}$.*

$$\mathrm{TPR}_S^{\tilde{p}} \geq \mathrm{TPR}_S^{p}, \qquad \mathrm{FPR}_S^{\tilde{p}} \leq \mathrm{FPR}_S^{p}, \qquad \mathrm{PPV}_S^{\tilde{p}} \geq \mathrm{PPV}_S^{p}.$$

In this sense, rather than advocating parity in predictions between subpopulations—possibly to the detriment of the absolute predictions within some groups—multi-calibration promotes improved predictions across every identifiable subpopulation. We present this result as an example of a more general "do-no-harm" phenomenon that multi-calibrated predictors exhibit. Fed trustworthy data as evidence, multi-calibrated predictors cannot deviate in ways that unexpectedly harm performance in identifiable subgroups.

1.3.3 Evidence-Based Rankings

Finally, we turn our attention to fairness when ranking individuals based on the perceived probability of an outcome. Expanding our study from prediction to ranking is of interest for several reasons. Ranking is at the heart of triage—allocating resources in disaster relief or providing timely care in emergency medicine–and is often the underlying impetus for prediction. For instance, recall our example where a college admissions committee develops a model to predict the probability that a given applicant will succeed (e.g., graduate within 4 years). Typically, the university will accept the same number of applicants every year; in this case, the predicted probability is a proxy used to elicit the most qualified applicants to accept. In actuality the admissions committee cares more about the the *relative* ordering of individuals compared to one another than the individuals' *absolute* qualifications.

This example highlights how the goals of ranking are qualitatively different than those of prediction. While prediction cares about absolute recovery of the risk, ranking cares about the relative ordering of individuals. The differences in the ranking and prediction objectives mean that small absolute errors in a predictor might result in large relative errors in the corresponding (induced) ranking. In this sense, ranking is a more global task than prediction: while notions of accuracy and fairness in predictors seem to be robust to small absolute errors on a portion of the population, intuitive notions of accuracy and fairness in rankings seem brittle to such mistakes.

Domination-Compatibility and Multi-Calibration. Our investigation of rankings begins with a simple group notion of fairness that considers the relative ranking of groups: for a given pair of sets S, T, suppose that on-average S is more qualified than T, but the ranking in question favors T above S. Such a transposition—from the true qualifications of S and T to their ordering in the ranking—represents a form of obvious systematic bias that could be audited from data. We formalize this concern through a notion, which we call *domination-compatibility.* This multi-group notion of fairness in rankings requires that for all pairs of subpopulations $S, T \in \mathcal{C}$, if the ranking favors T above S (i.e., if T *dominates* S in a sense we make formal), then the quality of T must be at least that of S, $\mathbf{E}_{\mathcal{D}_T}[\ p^*(x)\ ] \geq \mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ]$. Thus, a domination-compatible ranking must respect the "evidence" provided by the statistical tests defined by the subpopulations in $\mathcal{C}$.

Domination-compatibility is an intuitive notion of fairness that protects subpopulations relative to one another. We establish a lemma that makes a compelling argument for domination-compatibility: if a ranking violates $\mathcal{C}$-domination-compatibility, then no predictor exists that is consistent with the ranking and the statistical evidence from the class $\mathcal{C}$. By contrapositive, this observation suggests that if we start with a globally-consistent predictor—such as a multi-calibrated predictor—then the induced ranking must satisfy domination-compatibility. Surprisingly, for an appropriately-strengthened variant of domination-compatibility, the reverse implication also holds. Taking this stronger notion of *reflexive*-domination-compatible rankings, we derive a tight equivalence with multi-calibrated predictors.

Theorem 7. *A ranking r is $\mathcal{C}$-reflexive-domination-compatible if and only if it is the induced ranking of a $\mathcal{C}$-multi-calibrated predictor.*

In other words, any ranking that respects the ordering of the relevant subgroups derived from $\mathcal{C}$ suggests an analogous globally-consistent multi-calibrated predictor; conversely, every $\mathcal{C}$-multi-calibrated predictor $\tilde{p}$ must order subpopulations consistently relative to one another—not just overall, but also when restricting to the level sets of $\tilde{p}$.

This equivalence suggests multiple interpretations. First and most pertinent to the question of fair ranking, the result shows that if we want to satisfy a strong semantic notion of domination-compatibility fairness in rankings, it suffices to learn a multi-calibrated predictor and then turn it into a ranking. Second, this equivalence gives a new perspective on the guarantees of multi-calibration. While multi-calibration was conceived as a notion of fairness in prediction—emphasizing absolute accuracy on identifiable populations—the result shows that learning a multi-calibrated predictor requires learning a globally consistent ranking. This interpretation strongly supports the idea that multi-calibration requires a qualitatively different type of learning than earlier notions. Learning a multi-calibrated predictor implies a global understanding of all of the computationally-identifiable subpopulations and how they relate to one another.

1.4 Contents of the Thesis

This thesis draws much of its content from the following published manuscripts.

- *Calibration for the (Computationally-Identifiable) Masses.* Joint work with Úrsula Hébert-Johnson, Omer Reingold, and Guy N. Rothblum, appearing at *ICML* in 2018, [HKRR18].
- *Fairness through Computationally-Bounded Awareness.* Joint work with Omer Reingold and Guy N. Rothblum, appearing at *NeurIPS* in 2018, [KRR18].

- *Multiaccuracy: Black-Box Post-Processing for Fairness in Classification.* Joint work with Amirata Ghorbani and James Zou, appearing at *AAAI AI, Ethics, and Society* in 2019, [KGZ19].
- *Tracking and Improving Information in the Service of Fairness.* Joint work with Sumegha Garg and Omer Reingold, appearing at *EC* in 2019, [GKR19].
- *Learning from Outcomes: Evidence-Based Rankings.* Joint work with Cynthia Dwork, Omer Reingold, Guy N. Rothblum, and Gal Yona, appearing at *FOCS* in 2019, [DKR$^+$19].

The remaining chapters of the thesis are organized into two parts as follows.

Calibration for the Computationally-Identifiable Masses. Part II centers around multi-calibration, exploring the mathematical and algorithmic aspects of the notion.

- In Chapter 2, we describe the setting and formal notation that will be used throughout the thesis. We begin with a more in-depth discussion of prior notions of fairness. Then, we introduce the formal definition of multi-calibration and explore its basic properties.
- With the definition of multi-calibration in place, Chapter 3 focuses on learning multi-calibrated predictors as studied in [HKRR18]. We lay out the basic boosting-style learning framework and then turns to implementing the algorithm from a small set of labeled samples. Much of the technical content of this chapter is dedicated to answering the adaptively selected statistical queries required to achieve multi-calibration in a way that generalizes. Ultimately, this leads us to a differentially-private implementation of the learning algorithm that ensures efficient statistical validity.
- Chapter 4 explores the connections between learning multi-calibrated predictors and standard machine learning. First, we establish the theoretical equivalence between the multi-calibration auditing problem and weak agnostic learning, discovered in [HKRR18]. Then, we shift to a more practical perspective explored in [KGZ19], auditing and post-processing for multi-accuracy using off-the-shelf regression techniques. We present theoretical and empirical results from [KGZ19], demonstrating the effectiveness of the multi-calibration framework for improving predictions across underrepresented subpopulations without without hurting the overall performance.

Fairness through Computationally-Bounded Awareness. Part III explores other notions of multi-group fairness. While framed from different perspectives, we draw connections between these notions and multi-calibration.

- In Chapter 5, we present work on a different multi-group fairness notion from [KRR18]. This notion—*multi-metric fairness*—extends the individual metric fairness notion of [DHP$^+$12] to the multi-group setting. We explore the properties of this notion and present an algorithm for learning high-accuracy predictors satisfying it.

- In Chapter 6, we return to studying multi-calibrated predictors in terms of the information-theoretic concept of *refinements*. We show that the multi-calibration guarantee can be understood as a strong form of simultaneous refinement, across a rich class of calibrated predictors. Leveraging this understanding, the chapter concludes exploring a natural setting in which refined, multi-calibrated predictors may be an effective tool for reducing disparity according to other notions of fairness (e.g., selection or false negative rates).

- Finally, Chapter 7 explores evidence-based rankings and their connection to multi-calibrated predictors, as studied in [DKR$^+$19]. We introduce our learning to rank setting formally, then introduce the multi-group notions of fairness for rankings, *domination-compatibility* and *evidence-consistency*. Developing these notions to their limits, we show a strong equivalence with multi-calibration.

Part II

Calibration for the Computationally-Identifiable Masses

Chapter 2

Between Groups and Individuals

In this chapter, we introduce the main technical content of the thesis. In Section 2.1, we establish notation, nomenclature, and formal assumptions that will be used throughout. Then in Section 2.2, we give an overview of prior notions of fairness. Drawing from the foundational work of [DHP^{+}12], we highlight two main paradigms for establishing algorithmic fairness: group fairness and individual fairness. Each paradigm has its appeals as well as its drawbacks, which we discuss. Given the state of prior fairness notions, we ask the question: *What lies between group and individual notions of fairness?* Answering this question leads us to the definition of multi-calibration, which we introduce formally in Section 2.3 and study in the remainder of the thesis.

2.1 Notation and Formal Assumptions

Throughout the thesis, we will use the following notation. Let $\mathcal{X}$ to denote the population of individuals, which we assume to be a discrete domain. For each individual, we use $x \in \mathcal{X}$ denote an individual or the features/covariates of the individual interchangeably. Each individual also has an associated outcome drawn from a boolean outcome space, denote by $\mathcal{Y} = \{0, 1\}$.

We assume that $\mathcal{X}$ and $\mathcal{Y}$ are jointly distributed according to a fixed, but unknown distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. Let $\mathcal{D}$ denote the marginal distribution over individuals $\mathcal{X}$ and let $\mathcal{D}_{\mathcal{Y}|\mathcal{X}}$ denote the marginal distribution over the outcome given an individual. For a subset $S \subseteq \mathcal{X}$, we use the notational shorthand $\mathcal{D}_{S\times\mathcal{Y}}$ and $\mathcal{D}_S$ to denote the conditional distributions (joint and marginally over individuals, respectively) amongst individuals $x \in S$. For all distributions, we use the notation $x \sim \mathcal{D}$ to denote that x is a random sample distributed according to $\mathcal{D}$. To denote probabilities and expectations over these samples, we use boldface notation, as in

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}[\ x \in S\ ] \qquad\qquad \mathop{\mathbf{E}}_{x,y \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}[\ p(x) - y\ ].$$

We use $\mathbf{1}[\ P(z)\ ]$ to denote the indicator function of a predicate $P(z)$.

Our focus will be on learning predictors over the distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. A predictor $p : \mathcal{X} \to [0,1]$ is a function mapping individuals to a real-valued prediction. For a predictor p and a subset of individuals $S \subseteq \mathcal{X}$ we denote its support on S as

$$\mathrm{supp}_S(p) = \left\{ v \in [0,1] : \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S}[\ p(x) = v\ ] > 0 \right\}.$$

When clear from context we use $\mathrm{supp}(p)$ to denote the support of p over the entire domain. To focus on the definitions and issues of fairness in prediction, we avoid measure-theoretic complications by assuming that $\mathcal{X}$ is discrete and that predictors are supported on a discrete and finite set of values. As an example, this assumption means that conditional expectations of the form

$$\mathop{\mathbf{E}}_{x,y \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}[\ y \mid p(x) = v\ ]$$

are well-defined for all $v \in \mathrm{supp}(p)$. When we discuss learning multi-calibrated predictors, we work with an explicit discretization, but throughout the rest of the thesis, we will often make this assumption implicitly.

We use $p^* : \mathcal{X} \to [0,1]$ to denote the Bayes optimal predictor, defined as

$$p^*(x) = \mathop{\mathbf{Pr}}_{y \sim \mathcal{D}_{\mathcal{Y}|\mathcal{X}}}[\ y = 1 \mid x\ ].$$

$p^*(x)$ represents the "true probability" that the outcome of an individual x will be 1. In other words, our distributional assumption of $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ can be equivalently defined by the marginal distribution on individuals $\mathcal{D}$ paired with a conditional outcome distribution, where $y|x \sim \mathrm{Ber}(p^*(x))$. We use the convention that $\tilde{p} : \mathcal{X} \to [0,1]$ typically refers to a learned estimate of p^*, often a multi-calibrated predictor. When a predictor always outputs a boolean response $f : \mathcal{X} \to \{0,1\}$ we call it a binary classifier or a decision rule. Predictors can generically be turned into classifiers by thresholding or by randomly rounding.

2.2 Prior Notions of Fairness

The study of fairness in algorithms and prediction tasks emerged in the past decade, stemming from the influential work of [DHP$^+$12]. Much of this study focuses on notions of fairness that lie at one of two extremes. At one extreme, group notions tend to be intuitive to work with and compatible with existing machine learning techniques, but offer only marginal protections. At the other extreme, individual fairness offers principled protections to individuals, but presents challenges at scale. Next, we overview the most common notions of group and individual fairness and discuss their strengths and weaknesses.

2.2.1 Group Fairness

The general recipe for a notion of group fairness consists of three main components.

(1) identify a sensitive attribute, which defines "protected groups" (e.g., gender, race, sexual orientation, etc.);

(2) identify a statistic of interest;

(3) ensure that the statistic of interest "looks right" across the groups defined by the sensitive attribute.

Such a straightforward framework for defining fairness is appealing in practice, because ensuring group fairness only requires reasoning about marginal statistics of a handful of groups defined by the sensitive attribute. Often, these constraints are convex in the predictions and can be incorporated easily into existing machine learning and optimization procedures.

Parity-based fairness notions. To describe the group fairness notions, we will assume that the chosen sensitive attribute partitions the population into two groups $S, T \subseteq \mathcal{X}$. The group notions will reason about marginal statistics over the distributions $\mathcal{D}_S$ and $\mathcal{D}_T$.

One of the earliest notions of group fairness is called *demographic parity*[1] and aims to equalize the selection rates of individuals across groups.

Definition 2.1 (Demographic Parity, [DHP$^+$12]). *A binary classifier $f : \mathcal{X} \to \{0,1\}$ satisfies demographic parity if*

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S}[\ f(x) = 1\ ] = \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_T}[\ f(x) = 1\ ].$$

[1] Many works refer to this notion as *statistical parity.*

That is, the rate at which the classifier accepts individuals from S and from T (i.e. $f(x) = 1$) is equal. For instance, a company screening candidates for a job may enforce demographic parity to ensure they interview roughly the same number of men as women. Note that, in full generality, we can allow demographic parity to be achieved by selecting a randomized classification rule (distribution over deterministic classifiers) to ensure parity in selection rate (over the randomness in the classification as well). For instance, a decision rule that randomly selects individuals, so that $f(x)$ is statistically independent of x, would satisfy demographic parity.

Parity-based notions of fairness can be taken a step further, to require equality of other statistics. A popular alternative to demographic parity is *equalized opportunity*, which aims to equalize the false negative rate of a classifier.

Definition 2.2 (Equalized Opportunity, [HPS16])**.** *A binary classifier $f : \mathcal{X} \to \{0,1\}$ satisfies equalized opportunity if*

$$\mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}_{S\times\mathcal{Y}}}\left[\ f(x) = 1 \mid y = 1\ \right] = \mathop{\mathbf{Pr}}_{x,y\sim\mathcal{D}_{S\times\mathcal{Y}}}\left[\ f(x) = 1 \mid y = 1\ \right].$$

In other words, equalized opportunity aims to ensure that the selection rate of the "qualified individuals" is preserved across groups. Note that according to equalized opportunity, the set of qualified individuals is defined in an *a posteriori* sense: the qualified group ($y = 1$) is only defined after the outcome is revealed. Thus, enforcing equalized opportunity and its variants[2] makes the most sense when the group of individuals $\mathcal{X}_1 = \{x : y = 1\}$ can be estimated fairly accurately from data. Again, one trivial way to satisfy equalize opportunity is to ensure $f(x)$ is independent from x; further, note that a perfect binary classifier where $f(x) = y$ for all $x \in \mathcal{X}$ also satisfies the notion.

In general, when we define a notion of fairness that is based on the parity of statistical quantities across groups, the notion is feasible because useless predictions satisfy the notion. For instance, if our decision rule f treats all individuals identically (i.e., if $f(x)$ is statistically independent of x), then parity is satisfied but there is no reason to make individual-level predictions. Thus, with parity-based notions of fairness, the goal is typically to minimize some expected loss function ℓ over a hypothesis class $\mathcal{H}$ subject to satisfying the parity

[2] Equalized opportunity asks for parity of false negative rates. A stronger notion requires parity of both false negative and false positive rates and goes by a few names including *equalized odds* [HPS16] and *balanced error rates* [KMR17].

constraints. For instance, for demographic parity:

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & \mathop{\mathbf{E}}_{x,y \sim \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}} [\ \ell(h(x), y)\ ] \\ \text{s.t.} \quad & \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ h(x) = 1\ ] = \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_T} [\ h(x) = 1\ ]. \end{aligned}$$

Formulating the goal of fair prediction as a constrained optimization problem is natural in this context, but should be noted as a design decision. In particular, this framing suggests that fairness (encoded in parity constraints) and utility (encoded in the loss function) are inherently at odds with one another. For the program to make sense, we must also select a fixed hypothesis class $\mathcal{H}$ to optimize over. As such, the choice of hypothesis class implicitly affects the guarantees of such parity-based notions of fairness. In such a constrained mathematical program, we can always trade off fairness for improved utility along the Pareto curve by selecting different hypothesis $h \in \mathcal{H}$. Implicit in the formulation is the assumption that all hypotheses at a given level of disparity (difference in selection rates) are equally "fair."

Calibration. An alternative approach to group fairness requires absolute predictive performance across groups, rather than relative guarantees between groups. Group calibration is the primary example that follows this approach. Calibration is a widely-studied concept from the literature on forecasting [DF81, Daw82, FV97, FV98]; the formulation in the context of algorithmic fairness is due to [KMR17]. Colloquially, for a well-calibrated predictor p of skin cancer, 70% of the lesions that receive prediction $p(x) = 0.7$ will be malignant.

Definition 2.3 (Calibration). *A predictor $p : \mathcal{X} \to [0,1]$ is calibrated over a group S if for all $v \in \mathrm{supp}_S(p)$,*

$$\mathop{\mathbf{Pr}}_{x,y \sim \mathcal{D}_{S \times \mathcal{Y}}} [\ y = 1 \mid p(x) = v\ ] = v.$$

That is, calibrated predictors "mean what they say," and the prediction $p(x)$ can be meaningfully interpreted as a conditional probability that the individual's outcome will be $y = 1$ given x. Importantly, if we require calibration over the protected groups and the majority group, then we know that risk scores in each group mean the same thing. Ensuring that a predictor is calibrated helps to mitigate mistreatment that may arise due to overt statistical bias in the underlying predictions.

The simplest way to satisfy calibration is simply to give out the expected value to every individual in the group. That said, the Bayes optimal predictor p^* is also calibrated. As such, it's clear that there is a wide range of calibrated predictors for all tasks. We discuss these issues and calibration in more detail starting in Section 2.3

Incompatibility of group notions. While many of the statistical properties defined by group fairness notions feel natural and desirable, unfortunately, most notions are known to be mutually-incompatible. For instance, spurred by a debate raised in the popular press over the COMPAS recidivism risk prediction system [ALMK16], there has been lots of recent interest in the incompatibility of calibration and balanced error rates [KMR17, Cho17,PRW$^+$17]. As such, when implementing group notions of fairness, the decision-maker or regulator needs to decide which notion is most appropriate for the given context.

2.2.2 Fairness Through Awareness

In a highly-influential work, [DHP$^+$12] identified serious flaws with the group fairness approach, highlighting how a malicious decision-maker could satisfy the "letter" of group fairness notions, while still discriminating in a way that harmed individuals from the protected groups. They argued that this phenomenon was widespread, compiling a "catalog of evils," listing a number of scenarios where unfair decisions satisfy demographic parity. As research into fairness in prediction has grown, critiques of group fairness notions have continued to emerge [CG18, LDR$^+$18].

The arguments against group notions leverage the average-case nature of the constraints. Even though we require the treatment of the majority group and the protected group to "look similar," this similarity is measure *on average* over the entire group. In other words, constraints on predictions defined marginally over groups provide little to no guarantees to the individuals who receive predictions. This observation led [DHP$^+$12] to a fundamentally different approach towards defining fairness in prediction.

Individual fairness. To address the shortcomings of group notions, [DHP$^+$12] proposed an alternative paradigm for defining fairness, which they call "fairness through awareness." This framework takes the perspective that a fair classifier should *treat similar individuals similarly.* [DHP$^+$12] formalizes this abstract goal by assuming access to a task-specific similarity metric d over pairs of individuals that encodes which pairs must receive similar predictions. The proposed notion of *individual fairness* requires that if the metric distance between two individuals is small, then the predictions of a fair classifier cannot be very different. To distinguish from subsequent notions of fairness that define constraints on a per individual basis, we refer to the notion of [DHP+12] as *metric fairness.*

Definition 2.4 (Metric Fairness, [DHP+12])**.** *Let* $D : [0,1] \times [0,1] \to [0,1]$ *be a metric over predictions and let* $d : \mathcal{X} \times \mathcal{X} \to [0,1]$ *be a task-specific fairness metric over individuals. A predictor* $p : \mathcal{X} \to [0,1]$ *is metric fair (a.k.a., individually fair) if*

$$\forall x, x' \in \mathcal{X} \times \mathcal{X} : \quad D(p(x), p(x')) \leq d(x, x').$$

In other words, this Lipschitz condition—parameterized by the task-specific metric—must hold for all pairs of individuals from the population $\mathcal{X}$. Basing fairness on a similarity metric offers a flexible approach for formalizing a variety of guarantees and protections from discrimination. Importantly, the notion avoids the weaknesses of group notions, by allowing a regulator to specify the constraints on an individual-by-individual basis. Then, subject to the individual-level fairness constraints, the classifier may be chosen to maximize utility.

While the approach of fairness through awareness offers a theoretically-principled way to allow for high-utility predictions while ensuring fairness, a challenging aspect of this approach is the assumption that the similarity metric is known for all pairs of individuals. Indeed, [DHP+12] identifies this assumption as “one of the most challenging aspects” of the framework. Deciding on an appropriate metric is itself a delicate matter and could require input from sociologists, legal scholars, and specialists with domain expertise. For instance, in a loan repayment setting, a simple seemingly-objective metric might be a comparison of credit scores. A potential concern, however, is that these scores might themselves encode historical discrimination. In such a case, a more nuanced metric that incorporates human judgment might be called for.

A number of recent works have turned their attention to relaxing the assumption that the fairness metric is known and specified in full [RY18,KRR18,JKN+19,Ilv20]. These works make significant technical progress in reducing the number of similarity queries necessary to obtain PAC-style guarantees for metric fairness. Still, there are normative questions that remain unresolved in terms of who decides on the choice of the metric and on what basis. Thus, while individual fairness provides appealing protections from a theoretical computer science perspective, a number of socio-technical challenges continue to impede its adoption in practical settings.

2.3 Multi-Calibration: Beyond Protected Groups

Given the weaknesses of group fairness and the challenges of effectively implementing individual fairness, we turn our attention to what lies in between. Specifically, we ask whether we can strengthen the guarantees of group fairness notions while still maintaining their practical appeal to machine learning practitioners. We begin with a more in-depth look into calibration—its motivation as well as its failure mode as a notion of fairness. Then, we present the notion of multi-calibration, which aims to address the issues with group calibration by affording protections not only to traditionally-protected groups, but instead to every identifiable group.

Understanding calibration. A first step to addressing these issues is to obtain calibrated predictions—to ensure that predictions mean the same thing in the protected groups as in the majority population. Optimistically, we may hope that if predictions mean the same thing across groups, then the predictions would be equally meaningful across groups. This hopeful logic breaks down, however, because predictions may satisfy calibration without actually saying that much.

Consider the following toy example based on a common classification scenario. The decision-maker obtains a calibrated predictor, then turns the predictions into decisions by selecting individuals with scores $p(x) > \tau$ above some fixed group-independent threshold. For instance, suppose that a lender is willing to accept an applicant if they have at least a 0.8 chance of repaying the loan. Further, suppose there are two disjoint populations S and T. Suppose the true risk in S and T are identically distributed as a $50 : 50$ mix of $p^*(x) \in \{0.1, 0.9\}$, such that $\mathbf{E}_{\mathcal{D}_S}[\, p^*(x)\,] = \mathbf{E}_{\mathcal{D}_T}[\, p^*(x)\,] = 0.5$. While the optimal predictions p^* is the same in each group, consider the following predictor p.

$$p(x) = \begin{cases} p^*(x) & \text{if } x \in S \\ 0.5 & \text{if } x \in T. \end{cases}$$

Even though the populations are identical, the predictions are very different in S and in T. But the predictor p is actually calibrated over both S and T! To verify, note that the conditional probabilities defining the calibration constraints are accurate.

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[\ y=1 \mid p(x)=0.1\ ] = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x) \mid p^*(x)=0.1\ ] = 0.1$$

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[\ y=1 \mid p(x)=0.9\ ] = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x) \mid p^*(x)=0.9\ ] = 0.9$$

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_T}[\ y=1 \mid p(x)=0.5\ ] = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[\ p^*(x)\ ] = 0.5$$

In this case, using a fixed threshold of $\tau = 0.8$ will select every qualified individual in S and none of the individuals from T, even though, by the fact that p is calibrated, half of them were qualified.

This example highlights how calibration allows for "algorithmic stereotyping"—large groups of individuals receive similar predictions, despite variation in outcomes across the group. This type of discrimination may be the result of adversarial manipulation of predictions or may arise naturally as the result of standard machine learning training algorithms. Because standard training procedures optimize for on-average performance, machine-learned predictions tend to be confident within the majority group and tend to be under-confident in minority groups, simply due to the relative population sizes. Similar to other group notions of fairness, calibration provides marginal guarantees that may not even protect the groups designated as "protected."

2.3.1 Protecting the Computationally-Identifiable Masses

The failure of calibration to protect the predictions—even within the protected groups—stems from the fact that calibration allows for under-confident predictions. Requiring predictions to be calibrated over a disjoint set of subpopulations always allows for the predictor to give the average value within each partition. As in the example above, even if there are qualified individuals within a population, averaging over the population may lead to overlooking the qualified subpopulation. Minority populations may be especially susceptible to such marginalization, both due to historical discrimination as well as the simple fact that there are fewer minority examples to learn from.

Ideally, if we could identify the set of qualified individuals, then we could aim to protect these individuals. Anticipating the set of qualified individuals is typically challenging, if not impossible. For instance, a natural way to define the "qualified" individuals would be as those whose outcome was positive $y = 1$—this is the perspective that equalized opportunity takes. The downside with this notion is that this set is only defined *a posteriori*: whenever there is significant uncertainty in $y \sim \text{Ber}(p^*(x))$, then the set of individuals with positive outcome is essentially a random set.

Ensuring protections for a randomly drawn set is essentially equivalent to ensuring protections over *all* sufficiently-large subpopulations. The problem with hoping to ensure that we accurately represent the qualifications of all of these subpopulations is that it is information-theoretically impossible from a small sample. In particular, such a set of constraints would require recovering p^* to very high accuracy. Without strong assumptions that p^* comes from some bounded class of functions, there will always be a subpopulation on which we are inaccurate (e.g., the set of individuals on which the model errs), until we observe essentially the entire domain of individuals and outcomes. Thus, in any statistical learning setting, we need a different approach to defining fairness.

Multi-calibration takes a different perspective on how to define qualified individuals. Instead of requiring protections on a group defined after the outcomes are revealed, multi-calibration requires *a priori* protections for the set of populations that could reasonably be identified from the given data. Rather than trying to ensure calibration over all subpopulations, we relax the goal to ensure calibration over all "meaningful" subpopulations. Specifically, we consider a subpopulation to be worthy of protection if it can be *identified efficiently* from the data.

Slightly more formally, consider a collection of subpopulations $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$. Multi-calibration with respect to a class $\mathcal{C}$ will require that a predictor is well-calibrated simultaneously over each $S \in \mathcal{C}$. While we can define multi-calibration over any collection $\mathcal{C}$, we will think of $\mathcal{C}$ as being defined by a simple, but expressive *computational class.* Specifically, for any subpopulation $S \subseteq \mathcal{X}$, we can consider its characteristic function

$$\chi_S(x) = \begin{cases} 1 & x \in S \\ 0 & x \notin S. \end{cases}$$

Conversely, any class of boolean functions $\mathcal{C}$ induces a collection of subpopulations by imagining that each $c \in \mathcal{C}$ defines the characteristic function of some subpopulations $S_c \subseteq \mathcal{X}$. We can define such a collection with respect to any computational class $\mathcal{C}$, e.g., small conjunctions, decision trees, or linear functions. As we take $\mathcal{C}$ to be more and more expressive, the protections of multi-calibration become stronger, as we are forced to reason about subpopulations that may be relevant to the task at hand. In particular, the collection $\mathcal{C}$ shouldn't be thought of as defining "protected" groups so much as "meaningful" groups that will help to identify structure within p^*.

Formal definitions. With the intuition for multi-calibration in place, we are ready to define the notion formally. First, we formally define a useful collection of subpopulations which we refer to as the level sets of p.

Definition 2.5 (Level sets)**.** *Given a predictor $p : \mathcal{X} \to [0,1]$ and a subpopulation $S \subseteq \mathcal{S}$, for each $v \in \mathrm{supp}(p)$, let $S_v = \{x \in S : p(x) = v\}$. The sub-level sets of the predictor p on the subpopulation S are*

$$\{S_v : v \in \mathrm{supp}(p)\}\,.$$

The level sets of p are the sub-level sets over $\mathcal{X}$.

Intuitively, calibration requires on-average consistency over the level sets with the underlying Bayes optimal predictor p^*. Formally, we define approximate calibration as follows.

Definition 2.6 (Approximate Calibration)**.** *Suppose a predictor $p : \mathcal{X} \to [0,1]$ has finite support $s = |\mathrm{supp}(p)|$. For $\alpha \geq 0$ and a subset $S \subseteq \mathcal{X}$, p is α-calibrated over S if for all $v \in \mathrm{supp}(p)$ such that $\mathbf{Pr}_{x \sim \mathcal{D}_S}\left[\ p(x) = v\ \right] > \alpha/s$,*

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

Note that in this notion, we only consider the level sets that are sufficiently large. This condition is largely a technical requirement, due to the fact that it is statistically impossible to reason about groups if they are too small. Note, however, that the choice to parameterize the approximation factor by the support size is consistent with the intuition that predictions should "mean what they say." Specifically, a predictor that gives out lots of distinct supported values must be confident about smaller level sets than a predictor that gives out only a few values. Overall, excluding these small level sets of density at most α/s can introduce α additional error in the predictions over $\mathcal{D}_S$. We will think of α as a small constant throughout, so in most applications this will be an acceptable and unavoidable degree of error.

With this more technical definition in place, we define $(\mathcal{C}, \alpha)$-multi-calibration as approximate calibration over every subpopulation defined by $\mathcal{C}$.

Definition 2.7 (Multi-Calibration)**.** *Let $\alpha \geq 0$. For a collection of subsets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \alpha)$-multi-calibrated if for all $S \in \mathcal{C}$, $\tilde{p}$ is α-calibrated over S.*

By taking an expressive class of subpopulations, multi-calibration strengthens the requirements of calibration significantly. Intuitively, when we take $\mathcal{C}$ to consist of many overlapping subpopulations, defined by a function class capable of identifying interesting patterns within an individual's data, then $\mathcal{C}$-multi-calibration requires a predictor to make sense globally—not just marginally over a partition of the input space.

While the multi-calibration constraints are stringent, especially as $\mathcal{C}$ becomes progressively more complex, the notion is always feasible. This fact follows from the observation that p^* itself is $(\mathcal{C}, 0)$-multi-calibrated for any choice of $\mathcal{C}$. Indeed, p^* is perfectly calibrated over any subpopulation $S \subseteq \mathcal{X}$:

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

In this sense, a multi-calibrated predictor $\tilde{p}$ can be viewed as a *computational relaxation* of the information-theoretic optimal p^*. Specifically, we can frame multi-calibration in the language of pseudorandomness. Imagining $\mathcal{C}$ as a computational class, each subpopulation $S \in \mathcal{C}$—equivalently a boolean circuit $\chi_S : \mathcal{X} \to \{0,1\}$—defines a collection of statistical tests on the level sets of a predictor p: for each $v \in \mathrm{supp}(p)$

$$\left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}} \left[\ \chi_S(x) \cdot (p^*(x) - v) \mid p(x) = v\ \right] \right| \leq \alpha \cdot \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}} \left[\ \chi_S(x) = 1\ \right].$$

None of these statistical tests can significantly distinguish a $(\mathcal{C}, \alpha)$-multi-calibrated predictor from p^*—each tests passes up to the α approximation. In this sense, $\tilde{p}$ can be viewed as computationally-indistinguishable from p^* by the class of calibration tests defined within the class $\mathcal{C}$.

2.3.2 Multi-Accuracy

Often, it will be sufficient (and significantly easier) to work with a simpler notion of fairness, which we refer to as *multi-accuracy.* Multi-accuracy relaxes the guarantee of multi-calibration, and requires that a predictor be accurate in expectation (unbiased) over each subpopulation defined by $\mathcal{C}$.

Definition 2.8 (Multi-Accuracy). *Let $\alpha \geq 0$. For a collection of subsets $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \alpha)$-multi-accurate if for all $S \in \mathcal{C}$,*

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

By a straightforward argument, multi-calibration implies multi-accuracy. While immediately true in the exact case, when we work with approximate calibration, we lose an additional α when translating the quantitative guarantee.

Proposition 2.9. *If a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{C}, \alpha)$-multi-calibrated, then $\tilde{p}$ is $(\mathcal{C}, 2\alpha)$-multi-accurate.*

To gain familiarity with both notions, we include a proof of the proposition.

Proof. Suppose that $\tilde{p}$ is a $(\mathcal{C}, \alpha)$-multi-calibrated predictor with $s = |\mathrm{supp}(p)|$. For any subset $S \in \mathcal{C}$, consider the true expected outcome over $\mathcal{D}_S$. Denote by $L_S(\tilde{p})$ the large supported values,

$$L_S(\tilde{p}) = \left\{ v \in \mathrm{supp}(\tilde{p}) : \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) = v\ ] > \alpha/s \right\},$$

and let $R_S(\tilde{p}) = \mathrm{supp}(\tilde{p}) \setminus L_S(\tilde{p})$.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x)\ ] \\
&\quad = \sum_{v \in \mathrm{supp}_S(\tilde{p})} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) = v\ ] \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x) \mid \tilde{p}(x) = v\ ] && (2.1) \\
&\quad = \sum_{v \in L_S(\tilde{p})} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) = v\ ] \cdot \left(\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x) \mid \tilde{p}(x) = v\ ] \right) \\
&\qquad\qquad + \sum_{v \in R_S(\tilde{p})} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) = v\ ] \cdot \left(\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x) \mid \tilde{p}(x) = v\ ] \right) && (2.2) \\
&\quad \leq \sum_{v \in L_S(\tilde{p})} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) = v\ ] \cdot \left(\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) \mid \tilde{p}(x) = v\ ] + \alpha \right) + |R_S(\tilde{p})| \cdot \frac{\alpha}{s} && (2.3) \\
&\quad \leq \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x)\ ] + 2\alpha && (2.4)
\end{aligned}$$

where (2.1) follows by iterated expectations; (2.2) follows by expanding the summation across the partition of supported values; (2.3) follows by the assumption that $\tilde{p}$ is multi-calibrated and the definition of α-calibration; and (2.4) follows again by iterated expectations and the fact that $R_S(\tilde{p}) \subseteq \mathrm{supp}(\tilde{p})$, so $|R_S(\tilde{p})| \leq s$. Thus, we have shown that for each $S \in \mathcal{C}$

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x)\ ] - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x)\ ] \leq 2\alpha.$$

The reverse inequality follows similarly, so $\tilde{p}$ is $(\mathcal{C}, 2\alpha)$-multi-accurate. □

Most of this thesis focuses on multi-calibration for its strengths. In principle, multi-accuracy provides weaker protections than multi-calibration and can be abused by making arbitrary distinctions within subpopulations while still satisfying the correct overall expectations. Still, despite the fact that multi-accuracy is a weaker notion, we can still prove strong protections when the class $\mathcal{C}$ captures structure within p^*.

Multi-accuracy guarantees that the predictions of a classifier appear unbiased over a rich class of subpopulations. From a practical machine learning perspective, we often prefer guarantees in terms of the classification accuracy, not just the bias. Recall, the classification error or zero-one loss of a classifier f over a subpopulation S is defined as follow.

$$\mathrm{er}_S(f) = \mathop{\mathbf{Pr}}_{(x,y) \sim \mathcal{D}_{S \times \mathcal{Y}}} [\ f(x) \neq y\ ]$$

For a predictor $p : \mathcal{X} \to [0,1]$, we define a rounded classifier f^p as follows based on a threshold of 1/2.

$$f^p(x) = \mathbf{1}[\ p(x) > 1/2\ ]$$

Intuitively, as we take the collection $\mathcal{C}$ to include a more diverse set of subpopulations, the guarantees of simultaneous unbiasedness should become stronger, and guarantee that a useful classifier can be extracted. This intuition is formalized in the following proposition.

Proposition 2.10. *Let $\alpha, \varepsilon > 0$. For any $S \in \mathcal{C}$, suppose $\mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ] \geq 1 - \varepsilon$. For a $(\mathcal{C}, \alpha)$-multi-accurate predictor $\tilde{p}$, the classification error on S of the classifier $f^{\tilde{p}}$ is upper bounded as*

$$\mathrm{er}_S(f^{\tilde{p}}) \leq 3\varepsilon + 2\alpha.$$

Proof. Consider some $S \in \mathcal{C}$ such that $\mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ] \geq 1 - \varepsilon$. By $(\mathcal{C}, \alpha)$-multi-accuracy, this means that $\mathbf{E}_{\mathcal{D}_S}[\ \tilde{p}(x)\ ] \geq 1 - \varepsilon - \alpha$. With this upper bound on the expectation, we can analyze the probability that $f^{\tilde{p}}(x) = 0$ over $\mathcal{D}_S$; this will upper bound the probability of false negatives. By Bayes' rule,

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x)\ \right] = \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) < 1/2\ \right] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) \mid \tilde{p}(x) < 1/2\ \right]& \\ + \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) \geq 1/2\ \right] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) \mid \tilde{p}(x) \geq 1/2\ \right].&\end{aligned}$$

For notational convenience, let

$$p_0 = \mathop{\mathbf{Pr}}_{\mathcal{D}_S}\left[\ \tilde{p}(x) < 1/2\ \right] \quad v_0 = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) \mid \tilde{p}(x) < 1/2\ \right] \quad v_1 = \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x) \mid \tilde{p}(x) \geq 1/2\ \right].$$

We can rearrange to obtain the following bound.

$$\begin{aligned} p_0 &= \frac{v_1 - \mathbf{E}_{x\sim\mathcal{D}_S}\left[\ \tilde{p}(x)\ \right]}{v_1 - v_0} \\ &\leq \frac{1 - (1 - \varepsilon - \alpha)}{1 - v_0} \\ &\leq 2 \cdot (\varepsilon + \alpha) \end{aligned}$$

where the inequalities follow by the fact that $v_0 < 1/2$, $v_1 \leq 1$, and the assumed bound on the expectation from multi-accuracy.

Finally note from the original assumption, we know that $\mathbf{Pr}_{\mathcal{D}_S}\left[\ y = 0\ \right] \leq \varepsilon$; thus, the probability of false positives cannot exceed ε. In total, we bound the error as the probability of false positive or false negative as $3\varepsilon + 2\alpha$. □

This proposition highlights the intuition that the strength of multi-calibration lies in its ability to detect the subpopulation of the *a priori* "qualified individuals;" that is, a set $S \in \mathcal{C}$ where $\mathbf{E}_{\mathcal{D}_S}\left[\ p^*(x)\ \right]$ is close to 1. Of course, an analogous statement could be proved about identifying "unqualified individuals" as well. The point is that when $\mathcal{C}$ identifies meaningful structure in p^*—when there are sets $S \in \mathcal{C}$ that correlate nontrivially with the expected outcome of interest—then any predictor that is multi-accurate over $\mathcal{C}$ must differentiate nontrivially between the set where $y = 0$ and $y = 1$, even before these groups are revealed. In this sense, the more expressive we take the class $\mathcal{C}$—and thus, the stronger the *a priori* guarantee we give—the more likely $\mathcal{C}$ is to contain a subpopulation that correlates strongly with the set of *a posteriori* qualified individuals.

Chapter Notes

Work on fairness in algorithms and machine learning has exploded since the publication of [DHP$^+$12]. A comprehensive survey of this emerging field is beyond the scope of this thesis. For an overview of the growing list of notions and approaches to fairness, we recommend [MPB$^+$18] as well as [BHN19].

Multi-calibration and multi-accuracy[3] were first defined and studied in a joint work with Úrsula Hébert-Johnson, Omer Reingold, and Guy N. Rothblum [HKRR18]. While intuitively part of the motivation for multi-calibration and multi-accuracy, a variant of Proposition 2.10 first appeared in a follow-up work to [HKRR18], joint work with Amirata Ghorbani and James Zou [KGZ19].

[3] [HKRR18] refers to multi-accuracy as "multi-accuracy-in-expectation", or "multi-AE" for short.

Chapter 3

Learning Multi-Calibrated Predictors

Having introduced multi-calibration, we turn our attention to learning multi-calibrated predictors. If feasible, multi-calibration provides strong guarantees of population-wide learning—no subpopulation identified within $\mathcal{C}$ can be overlooked. Still, the question remains: *When is multi-calibration feasible?*

To start, if the true risk function $p^* : \mathcal{X} \to [0, 1]$ is simple enough to learn to high precision, then multi-calibration will be attainable. As we've argued, p^* is multi-calibrated with respect to every collection of subpopulations $\mathcal{C}$. Often, however, p^* will not be directly learnable. In such cases, we may worry that multi-calibration will be infeasible: perhaps the multi-calibration constraints are restrictive enough that learning p^* is *necessary* to obtain a multi-calibrated predictor. In this chapter, we describe a learning algorithm that shows this fear is misplaced. We demonstrate that learning a $(\mathcal{C}, \alpha)$-multi-calibrated predictor is possible with complexity (in terms of necessary data and running time) scaling as a function of $\mathcal{C}$ and α—independent of p^*.

We describe and analyze the algorithm for learning multi-calibrated predictors across the next sections. In Section 3.1, we describe our iterative algorithm in the statistical query framework. At this level of abstraction, we can analyze the iteration complexity, which affects the running time as well as the eventual model complexity. Then in Section 3.2, we show how to implement the statistical query oracle from a small set of samples. Due to the self-referential calibration constraints, standard uniform convergence arguments do not suffice to guarantee validity. Instead in Section 3.3, we leverage a line of work showing how

Algorithm 1 Multi-Calibrate Pseudocode

```
Repeat:
  if there exists $S \in \mathcal{C}$ and $v$ s.t. $p_t$ is significantly mis-calibrated on $S$ then
    $p_{t+1} \leftarrow$ re-calibrate on $\{x \in S : p_t(x) = v\}$
    continue                                    // next iteration upon update
  end if
  return $\tilde{p} = p_t$                          // return when calibrated
```

differentially private algorithms can be used to maintain validity in adaptive data analysis to obtain efficient sample complexity for learning a multi-calibrated predictor.

Representing predictors. When describing the learning algorithm, we aim to be as generic as possible, without explicit dependence on the way that the features encoding individuals $x \in \mathcal{X}$ or the functions $c \in \mathcal{C}$ are represented. Still, to give a formal treatment, we must discuss some technical elements of the way that we represent real-valued predictors. Specifically, we will learn predictors $p : \mathcal{X} \to [0,1]$ of finite discrete support. Recall, the definition of approximate calibration is parameterized by the support size $s = |\text{supp}(p)|$. As we'll see, discretizing the interval $[0,1]$ into $s = \Theta(1/\alpha)$ values suffices to ensure α-calibration simultaneously across all $S \in \mathcal{C}$ with minimal overhead.

3.1 Learning a Multi-Calibrated Predictor

At its core, the algorithmic approach to learning a multi-calibrated predictor is one of the simplest imaginable, and can be viewed as a variant of the boosting algorithm given in [TTV09]. The algorithm starts with a trivial predictor $p_0 : \mathcal{X} \to [0,1]$. Then, we begin an iterative procedure that in the tth iteration checks whether there is a set $S \in \mathcal{C}$ on which the current predictor p_t is mis-calibrated. If there is, then we update the predictions to better reflect $\mathbf{E}[p^*(x)]$; else, we terminate guaranteed that the final predictor is well-calibrated on all $S \in \mathcal{C}$. We describe the generic template for such an iterative algorithm in Algorithm 1.

While simple enough to state, the technicalities in implementing this algorithm efficiently from a small set of samples take up the remainder of the chapter. Intuitively, if the algorithm terminates and each component is implemented correctly, the correctness is immediate; the returned predictor passes all of the possible multi-calibration tests. As such, our analysis focuses on an implementation of the components that guarantees correctness and termination in a small number of iterations.

3.1.1 Statistical Query Algorithm

To begin, we describe an implementation of the framework sketched above in the statistical query model. Assuming access to certain distributional queries, we show that we can implement the framework sketched in Algorithm 1 in a finite number of iterations. After presenting the statistical query algorithm, we will show how to implement the necessary oracles from a small set of samples.

Abstracting the interaction with labeled samples away, we allow the algorithm to ask for approximate expectations of $p^*(x)$ on subpopulations over the distribution $\mathcal{D}$. More generally, we define these statistical queries for bounded functions over the domain.

Definition 3.1 (Statistical query)**.** *Suppose for $b > 0$, $q : \mathcal{X} \to [-b, b]$ is a bounded function. For a subset $S \subseteq \mathcal{X}$, a statistical query on q over the distribution $\mathcal{D}$ estimates the true expected outcome*

$$q(S) = \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ q(x)\ ] .$$

Given a sequence of statistical queries $(q^1, S^1), \ldots, (q^T, S^T)$ and $\alpha, \beta > 0$, a (possibly-randomized) function $\mathrm{SQ}_{\mathcal{D},\alpha} : \{\mathcal{X} \to [-b, b]\} \times \{0,1\}^{\mathcal{X}} \to [0,1]$ satisfies the statistical query guarantee with (α, β)-distributional accuracy if with probability at least $1 - \beta$

$$\left|\ \mathrm{SQ}_{\mathcal{D},\alpha}(q^t, S^t) - q^t(S^t)\ \right| \leq \alpha$$

for each $t \in [T]$.

We drop explicit reference to β as we will always set β to be negligible; the correctness of the algorithm is predicated on the queries being correct. Note that this definition of statistical queries is tailored to the definition we use for approximate calibration. Specifically, α-calibration over a subset S requires *relative* error of at most α (i.e., scaled by the density of S in $\mathcal{D}$). Thus, we define statistical queries to be α-accurate on expectation queries over restrictions of the distribution, $\mathcal{D}_S$. Traditionally, statistical queries are defined to guarantee *absolute* error (i.e., always measured with respect to an expectation over $\mathcal{D}$).

Recall that the definition of multi-calibration only requires that we be well-calibrated over sub-level sets $\{x \in \mathcal{S} : \tilde{p}(x) = v\}$ provided $\mathbf{Pr}_{\mathcal{D}_S} [\ \tilde{p}(x) = v\ ] > \alpha/s$ for a support-s

predictor $\tilde{p}$. This technicality, while sensible as a notion of approximation is reasonable, is used primarily to ensure that our algorithm terminates in a finite number of iterations and samples. Thus, we will also assume access to a certain density query oracle $\mathrm{DQ}_{\mathcal{D},\tau} : \{0,1\}^{\mathcal{X}} \to \{✗, ✓\}$ satisfying the property

- if $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \le \tau/2$, then $\mathrm{DQ}_{\mathcal{D},\tau}(S) =$ ✗
- if $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] > \tau$, then $\mathrm{DQ}_{\mathcal{D},\tau}(S) =$ ✓

with high probability. With these oracles in place, we describe the statistical query algorithm in Algorithm 2.

As in the sketch, the algorithm searches over each $S \in \mathcal{C}$ and $v \in \Lambda$. If a given sub-level set is a large enough fraction of $\mathcal{D}_S$ to be considered, then the algorithm proceeds to make a statistical query of p^*. When we query the expectation over the set of individuals $x \in S$ where $p_t(x) = v$, then we see the statistical query tells us the degree to which the sub-level set is miscalibrated. Specifically, p_t violates $(\mathcal{C}, \alpha)$-multi-calibration if there is some set where

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

Querying the expected value with sufficient accuracy, we update the predictor from p_t to p_{t+1} if the miscalibration exceeds α.

Remark. *As a technicality, we assume that $\mathrm{SQ}_{\mathcal{D},\alpha}$ returns a multiple of α. This ensures that the support of the predictor remains Λ and does not grow in cardinality with updates. We take care to ensure this property holds when implementing the oracle from samples.*

Analysis of algorithm. We analyze Algorithm 2 showing that—assuming the statistical and query guarantees—the algorithm terminates in a bounded number of iterations returning a multi-calibrated predictor $\tilde{p}$. First, we observe that if Algorithm 2 terminates, then $\tilde{p}$ is a $(\mathcal{C}, \alpha)$-multi-calibrated predictor.

Proposition 3.2. *Suppose Algorithm 2 returns a predictor $\tilde{p}$. Then, $\tilde{p}$ is $(\mathcal{C}, \alpha)$-multi-calibrated with all but negligible probability.*

Proof. Note that by construction of Λ, $\tilde{p}$ is supported on at most $s = 4/\alpha$ possible values. Recall, by the definition of $(\mathcal{C}, \alpha)$-multi-calibration and α-calibration, we need to reason

Algorithm 2 SQ-Multi-Calibrate

```
Given:
  $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$                    // collection of subpopulations
  $\alpha > 0$                                   // approximation parameter
  $\gamma > 0$                                   // density lower bound
  $\mathrm{SQ}_{\mathcal{D},\alpha}$                               // statistical query oracle
  $\mathrm{DQ}_{\mathcal{D},\tau}$                               // density query oracle

Initialize:
  $\tau \leftarrow \alpha^2\gamma/4$
  $\Lambda \leftarrow \left\{0, \frac{\alpha}{4}, \frac{\alpha}{2}, \ldots, \left\lfloor\frac{4}{\alpha}\right\rfloor \cdot \frac{\alpha}{4}\right\}$      // discretize interval [0,1]
  $\forall x \in \mathcal{X}: \quad p_0(x) \leftarrow 1/2$                // initialize uniformly

Repeat: for $t = 0, 1, \ldots$
  for each $S \in \mathcal{C}$ and each $v \in \Lambda$ do
    $S_v \leftarrow \{x \in S : p_t(x) = v\}$                 // consider level sets
    if $\mathrm{DQ}_{\mathcal{D},\tau}(S_v) =$ ✗ then
      continue to next $S_v$                        // only consider large sets
    end if
    $u \leftarrow \mathrm{SQ}_{\mathcal{D},\alpha/4}(p^*, S_v)$                  // query expectation
    if $|v - u| > 3\alpha/4$ then
      $\forall x \in S_v: \quad p_{t+1}(x) \leftarrow u$         // test and update for calibration
      $\forall x \in \mathcal{X} \setminus S_v: \quad p_{t+1}(x) \leftarrow p_t(x)$
      break and continue to $t \leftarrow t+1$         // new iteration upon update
    end if
  end for
  return $\tilde{p} = p_t$                          // return when no update occurs
```

about the sufficiently-large level sets $\{x \in S : \tilde{p}(x) = v\}$ where

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}_S}[\ \tilde{p}(x) = v\ ] > \frac{\alpha}{s} \geq \alpha^2/4.$$

If $\tilde{p}$ is returned, then the final iteration certifies that for all $S \in \mathcal{C}$ and all values $v \in \Lambda$,

$$\text{if } \mathrm{DQ}_{\mathcal{D},\tau}(S_v) = \checkmark, \text{ then } \left|v = \mathrm{SQ}_{\mathcal{D},\alpha}(p^*; S_v)\right| \leq 3\alpha/4.$$

With all but negligible probability, if $\mathbf{Pr}_{x\sim\mathcal{D}_S}[\ p(x) = v\ ] > \alpha^2/4$, then the density query $\mathrm{DQ}_{\mathcal{D},\tau}(S_v) = \checkmark$ for all S_v. This follows by taking $\tau = \alpha^2\gamma/4$, the assumed density lower bound of γ for all $S \in \mathcal{C}$, and the assumed density query guarantee. Thus, for each such S_v, the iteration continues to test the expectation of p^*. By the statistical query accuracy guarantee, we know that with all but negligible probability $u = \mathrm{SQ}_{\mathcal{D},\alpha/4}(p^*, S_v)$ satisfies

$$\left| u - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ p^*(x) \mid \tilde{p}(x) = v\ ] \right| \leq \alpha/4.$$

If $|u - v| \leq 3\alpha/4$ for each of these sets, then

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

Thus, any returned $\tilde{p}$ satisfies $(\mathcal{C}, \alpha)$-multi-calibration. □

Next, we argue that Algorithm 2 is guaranteed to terminate and return a predictor $\tilde{p}$ in a bounded number of iterations.

Lemma 3.3. *Let $\alpha, \gamma > 0$. Suppose that $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations such that $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$ for all $S \in \mathcal{C}$. Assuming access to a density query oracle $\mathrm{DQ}_{\mathcal{D}}$ and statistical query algorithm $\mathrm{SQ}_{\mathcal{D}}$, Algorithm 2 returns a $(\mathcal{C}, \alpha)$-multi-calibrated predictor $\tilde{p}$ in at most $O\left(\frac{1}{\alpha^4 \gamma}\right)$ iterations.*

Proof. We prove Lemma 3.3 using a potential argument. Specifically, we will track the expected squared error between p_t and p^*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}} \left[\ (p^*(x) - p_t(x))^2\ \right].$$

We show that in each iteration, the update from p_t to p_{t+1} causes a significant reduction in the expected squared error. Noting that the squared error is a nonnegative potential that cannot exceed 1 (by the fact that predictors are bounded in the interval $[0,1]$), progress at each iteration implies a bounded number of iterations.

First, we note that for every non-terminating iteration t, there is some $S \in \mathcal{C}$ and $v \in \Lambda$ such that $|u - v| > 3\alpha/4$ for $u = \mathrm{SQ}_{\mathcal{D},\alpha/4}(p^*, S_v)$. Thus, the only changes from p_t to p_{t+1} occur on $x \in S$ where $p_t(x) = v$ to $p_{t+1}(x) = u$. Thus, we consider the differences in squared error conditioning on $x \in S$ and $p_t(x) = v$, which we denote $\Delta_{t,S,v}$.

$$\begin{aligned}
\Delta_{t,S,v} &= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\ (p^*(x) - p_t(x))^2 \ \middle|\ p_t(x) = v\ \right] - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\ (p^*(x) - p_{t+1}(x))^2 \ \middle|\ p_t(x) = v\ \right] \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\ (p^*(x) - v)^2 \ \middle|\ p_t(x) = v\ \right] - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\ (p^*(x) - u))^2 \ \middle|\ p_t(x) = v\ \right] \\
&= v^2 - u^2 - 2 \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ] \cdot (v - u) \\
&= \left(v + u - 2 \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ]\right) \cdot (v - u) \\
&= (v - u)^2 - 2 \cdot \left(\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ] - u\right) \cdot (v - u) \qquad (3.1)
\end{aligned}$$

For notational convenience, we introduce the following variables.

$$d_{vp^*} = v - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ] \qquad d_{up^*} = u - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ]$$

By the update conditions and statistical query guarantee, we can derive the following inequalities.

$$\alpha/2 \leq |v - u| \leq d_{vp^*} + \alpha/4 \qquad (3.2)$$

$$|v - u| \leq \frac{3}{2} \cdot |d_{vp^*}| \qquad (3.3)$$

$$d_{vp^*} \cdot (v - u) > 0, \qquad (3.4)$$

where (3.3) follows from (3.2) and (3.4) indicates that updating from v to u moves in the same direction as updating from v to the true expectation of p^*.

Then, rearranging (3.1), we obtain the following bound.

$$\begin{aligned}
(3.1) &= (v - u)^2 - 2 \cdot ((v - u) - d_{vp^*}) \cdot (v - u) \\
&= 2d_{vp^*} \cdot (v - u) - (v - u)^2 \\
&\geq \frac{(v - u)^2}{3} \qquad (3.5)
\end{aligned}$$

where (3.5) follows by (3.3) and (3.4). By (3.2) $\Delta_{t,S,v} \geq \alpha^2/12$.

To track the change to the overall potential, we need to scale this conditional expectation probability that $x \in S$ and $p_t(x) = v$. We denote the overall potential change at the tth iteration as Δ_t.

$$\begin{aligned}\Delta_t &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ (p^*(x) - p_t(x))^2\ \right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ (p^*(x) - p_{t+1}(x))^2\ \right] \\ &= \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x \in S_v\ ] \cdot \Delta_{t,S,v}\end{aligned}$$

We know that if S_v is updated, then it passed the density query test. By the density query oracle guarantee, we know that if $\mathrm{DQ}_{\mathcal{D}}(S_v, \alpha^2\gamma/4) = \checkmark$, then $\mathbf{Pr}_{x\sim\mathcal{D}}[\ x \in S_v\ ] > \alpha^2\gamma/8$. Thus, in all, we know that the progress at each iteration is lower bounded by

$$\Delta_t > \frac{\alpha^4\gamma}{96}.$$

Thus, Algorithm 2 terminates and returns $\tilde{p}$ after at most $O\left(\frac{1}{\alpha^4\gamma}\right)$ iterations. □

In order to obtain to complete the description of an algorithm for learning a multi-calibrated predictor from samples, it remains to demonstrate how to implement the density query and statistical query oracles from a small set of samples. Even without such an implementation, however, bounding the iteration complexity of Algorithm 2 actually provides an upper bound on the representation complexity of multi-calibrated predictors. In particular, as a corollary of Lemma 3.3, we can obtain a bound on the circuit complexity needed to represent a $(\mathcal{C}, \alpha)$-multi-calibrated predictor $\tilde{p}$. Importantly, the bound only depends on the circuit complexity of $\mathcal{C}$ and the approximation parameters—not on the complexity of representing p^*. Thus, multi-calibrated predictors can always be represented efficiently, even when the underlying true risk is arbitrarily complex.

Theorem 3.4 (Corollary of Lemma 3.3, informal)**.** *Suppose $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations with circuit complexity c; that is, for each $S \in \mathcal{C}$, the characteristic function $\chi_S : \mathcal{X} \to \{0,1\}$ is computable by a circuit of size at most c. Then, for $\alpha, \gamma > 0$, if $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$ for all $S \in \mathcal{C}$, there exists a $(\mathcal{C}, \alpha)$-multi-calibrated predictor $\tilde{p}$, such that the circuit complexity of $\tilde{p}$ is at most $O\left(c/\alpha^4\gamma\right)$.*

We state Theorem 3.4 informally, ignoring the details of how we represent the $\Theta(\alpha)$-accurate real-values and perform functions like addition. The takeaway is that the complexity of representing multi-calibrated functions depends on the complexity of $\mathcal{C}$. This is another way in which $\mathcal{C}$ must define "efficiently-identifiable" subpopulations in order to be effective: if we cannot efficiently compute set membership for $S \in \mathcal{C}$, then it is not clear how to enforce multi-calibration over $\mathcal{C}$. Conversely, the theorem says that if we can compute set membership efficiently, then the multi-calibrated predictor resulting from Algorithm 2 will be efficient to evaluate. This fact follows immediately from the bound on the iteration complexity of Algorithm 2 and by observing that the only operations necessary to compute p_t at each iteration are addition, testing equality over values in Λ, and evaluating set membership for the updated sets.

3.2 Answering Multi-Calibration Statistical Queries

In this section, we open the black boxes of the query oracles used in Algorithm 2. We will show how to implement the oracles using the empirical expectations from samples. Due to the adaptive nature of the queries of Algorithm 2—as p_t changes, the set of possible queries changes—we need to take care in translating the guarantees for a single query into a bound for all of the queries. In this section, we establish a baseline using the naive strategy of taking a fresh sample after every update. In Section 3.3, we show how to improve the labeled sample complexity by implementing the statistical query oracle in a differentially private manner.

3.2.1 Implementing Oracles from Small Set of Samples

We begin by showing how to implement the density and statistical query oracles, for a fixed query set, from a small set of samples. After analyzing how to answer a single query, we turn to bounding the number of samples needed to answer all of the queries asked in Algorithm 2.

Implementing the density query oracle. We argue that the density query oracle can be implemented given access to unlabeled samples from the distribution $\mathcal{D}$. Specifically, the oracle is described in Algorithm 3. Taking m sufficiently large, the implementation satisfies the density query oracle guarantees with high probability.

Lemma 3.5. *Given a subset $S \subseteq \mathcal{X}$, threshold $\tau > 0$, failure probability $\beta > 0$, and m independent unlabeled samples from $\mathcal{D}$, Algorithm 3 satisfies the density query oracle guarantee with probability at least $1 - \beta$ provided $m \geq \dfrac{32 \cdot \log(1/\beta)}{\tau}$.*

Algorithm 3 Density Query Oracle

```
On query $\mathrm{DQ}_{\mathcal{D},\tau}(S)$:
  Draw $x_1, \ldots, x_m \sim \mathcal{D}$                 // independent unlabeled samples
  $\hat{\gamma}_S \leftarrow \frac{1}{m} \cdot \sum_{i=1}^{m} \mathbf{1}[\ x \in S\ ]$          // compute empirical density
  if $\hat{\gamma}_S > 3\tau/4$ then
    return ✓                                       // sufficiently dense
  end if
  return ✗
```

Proof. Let $\gamma_S = \mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ]$ denote the true density of the set S over $\mathcal{D}$. To establish the lemma, we prove the follow claims that correspond to the density query oracle guarantees. With probability at least $1 - \beta$, over the randomness in the samples:

(a) if $\gamma_S \leq \tau/2$, then $\hat{\gamma}_S \leq 3\tau/4$

(b) if $\gamma_S > \tau$, then $\hat{\gamma}_S > 3\tau/4$

Noting that $\hat{\gamma}_S$ is a empirical expectation of m independent nonnegative random variables bounded by 1 with expectation γ_S, both claims will follow by an application of Chernoff's inequalities (Theorem A.1).

(a) First, suppose that $\gamma_S \leq \tau/2$. Then, we bound the probability that $\hat{\gamma}_S$ exceeds $3\tau/4$ as follows.

$$\mathbf{Pr}[\ \hat{\gamma}_S > 3\tau/4\ ] = \mathbf{Pr}\left[\ \hat{\gamma}_S > \frac{3\tau}{4\gamma_S} \cdot \gamma_S\ \right]$$

Letting $\Delta = \left(\frac{3\tau}{4\gamma_S} - 1\right)$ and applying Chernoff's inequality, we obtain

$$\mathbf{Pr}[\ \hat{\gamma}_S > 3\tau/4\ ] \leq \exp\left(\frac{-\Delta^2}{2+\Delta} \cdot \gamma_S \cdot m\right).$$

Under the assumption that $\gamma_S \leq \tau/2$, then Δ is bounded as

$$\Delta = \left(\frac{3\tau}{4\gamma_S} - 1\right) \geq \left(\frac{3\tau}{4 \cdot (\tau/2)} - 1\right) = 1/2.$$

Thus, we suppose $\Delta \geq 1/2$. Then,

$$\begin{aligned}\exp\left(\frac{-\Delta^2}{2+\Delta}\cdot\gamma_S\cdot m\right) &\leq \exp\left(\frac{-\Delta^2\cdot\gamma_S}{5\Delta}\cdot m\right)\\ &= \exp\left(\frac{-\Delta\cdot\gamma_S}{5}\cdot m\right).\end{aligned}$$

Again, leveraging the assumption that $\gamma_S \leq \tau/2$, we can expand the numerator in the exponential as follows.

$$\begin{aligned}\Delta\cdot\gamma_S &= \left(\frac{3\tau}{4\gamma_S}-1\right)\cdot\gamma_S\\ &= \frac{3\tau}{4}-\gamma_S\\ &\geq \frac{\tau}{4}\end{aligned}$$

Thus, we can obtain the desired bound of β failure probability by taking

$$\mathbf{Pr}\left[\ \hat{\gamma}_S > 3\tau/4\ \right] \leq \exp\left(\frac{-\tau\cdot m}{20}\right) \leq \beta.$$

(b) Next, suppose that $\gamma_S > \tau$. Then, we bound the probability that $\hat{\gamma}_S$ is at most $3\tau/4$ as

$$\mathbf{Pr}\left[\ \hat{\gamma}_S \leq 3\tau/4\ \right] \leq \mathbf{Pr}\left[\ \hat{\gamma}_S \leq \frac{3}{4}\cdot\gamma_S\ \right],$$

by the assumption that $\gamma_S > \tau$ and the fact that the probability of tail events is monotonically decreasing away from the true mean. This time, we can apply Chernoff's inequality in the other direction with $\Delta = 1/4$:

$$\begin{aligned}\mathbf{Pr}\left[\ \hat{\gamma}_S \leq 3\tau/4\ \right] &\leq \exp\left(\frac{-(1/4)^2}{2}\cdot\gamma_S\cdot m\right)\\ &\leq \exp\left(\frac{-\tau\cdot m}{32}\right) \qquad (3.6)\\ &\leq \beta\end{aligned}$$

where (3.6) follows by the assumption on $\gamma_S > \tau$.

Thus, taking $m \geq \dfrac{32\cdot\log(1/\beta)}{\tau}$ suffices to provide the density query oracle guarantee for $\mathrm{DQ}_{\mathcal{D}}(S,\tau)$ with probability at least $1-\beta$ in either case. □

Algorithm 4 Statistical Query Oracle

```
On query $\mathrm{SQ}_{\mathcal{D},\alpha}(p^*, S)$:
  Draw $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$        // independent labeled samples
  $m_S \leftarrow \sum_{i=1}^{m} \mathbf{1}\{x_i \in S\}$
  $\hat{y}_S \leftarrow \frac{1}{m_S} \cdot \sum_{i=1}^{m} \mathbf{1}[\ x_i \in S\ ] \cdot y_i$        // compute empirical expectation
  $u \leftarrow$ round $\hat{y}_S$ to nearest multiple of $\alpha$
  return $u$
```

Implementing the statistical query oracle. Next, we turn our attention to implementing the statistical query oracle from labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. The implementation in Algorithm 4 tests and reports the empirical average of the statistical query.

Lemma 3.6. *For a subset $S \subseteq \mathcal{X}$, let $\gamma_S = \mathbf{Pr}_{x\sim\mathcal{D}}[\ x \in S\ ]$. Given $0 < \alpha < 1/2$, failure probability $\beta > 0$, and m independent samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, Algorithm 4 satisfies the (α, β)-distributional accuracy statistical query guarantee, provided $m \geq \dfrac{4 \cdot \log(2/\beta)}{\alpha^2 \cdot \gamma_S}$.*

Proof. First, we argue that with high probability $m_S \geq \gamma_S \cdot m/2$. By Chernoff's inequality,

$$\mathbf{Pr}[\ m_S < \gamma_S \cdot m/2\ ] \leq \exp\left(-\frac{\gamma_S \cdot m}{8}\right),$$

so for $m \geq \dfrac{8 \cdot \log(2/\beta)}{\gamma_S}$, then $m_S \geq \gamma_S \cdot m/2$ with probability at least $1 - \beta/2$. Next, we condition on the event where there are at least $m_S \geq \gamma_S \cdot m/2$ draws where $x \in S$. $\hat{y}_S$ is an unbiased estimate of $\mathbf{E}_{\mathcal{D}_{S\times\mathcal{Y}}}[\ y\ ] = \mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ]$. Further, $\hat{y}_S$ is the sum of m_S independent random variables bounded in the range $\{0, 1\}$. Thus, we can bound the probability that the estimate deviates by more than α using Hoeffding's inequality.

$$\begin{aligned}
&\mathbf{Pr}\left[\ \left|\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x)\ ] - \hat{y}_S\right| > \alpha\ \middle|\ m_S \geq t\ \right] \\
&\qquad \leq \mathbf{Pr}\left[\ \left|\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x)\ ] - \hat{y}_S\right| > \alpha\ \middle|\ m_S = t\ \right] \\
&\qquad \leq \exp\left(-\frac{\alpha^2 \cdot t}{2}\right) \qquad (3.7)
\end{aligned}$$

where (3.7) follows by Hoeffding's Inequality (Theorem A.3). For $t = \gamma_S \cdot m/2$, then for $m \geq \dfrac{4 \cdot \log(2/\beta)}{\alpha^2 \gamma_S}$ we have that with probability at least $1 - \beta/2$, $|\hat{p}_S - \hat{y}_S|$ is α-accurate.

To satisfy the statistical query requirement, we take m to guarantee that $\hat{y}_S$ is $\alpha/2$-accurate and then round to the nearest multiple of α. Thus, we can apply a union bound to conclude that provided

$$m \geq \max\left\{\frac{8 \cdot \log(2/\beta)}{\gamma_S}, \frac{4 \cdot \log(2/\beta)}{\alpha^2 \gamma_S}\right\} = \Theta\left(\frac{\log(1/\beta)}{\alpha^2 \gamma_S}\right),$$

the statistical query oracle guarantee is satisfied with probability at least $1-\beta$. The claimed bound follows by assuming $\alpha < 1/2$ is nontrivial. □

Answering the adaptive queries. With the sample complexities from Lemmas 3.5 and 3.6 in place, we can establish upper bounds on the sample complexity needed to answer a fixed set of queries. Specifically, suppose we want to guarantee validity on queries over a fixed collection of subpopulations $\mathcal{C}$; then, we can apply the lemmas taking $\beta = \beta_0 / |\mathcal{C}|$, then applying a union bound to bound the probability of failure by β_0.

The problem with applying this argument to all of the queries made by Algorithm 2 is that the set of possible queries is not fixed. Multi-calibration requires us to reason about the accuracy of the level sets of the predictor in question. At the tth iteration, we make queries on sets defined as $S_v = S \cap \{x : p_t(x) = v\}$. Importantly, the queries over the level at the tth iteration depends on the $t-1$ prior updates we've made. In other words, the queries we ask are chosen adaptively based on the earlier statistical queries. Classically, to maintain validity over adaptive statistical queries, we sample a fresh set of data every time we update the set of queries we may ask. In the context of Algorithm 2, the set of queries changes each time we update the predictions. Thus, to guarantee validity, we can take a fresh set of data per iteration, resulting in the following bound.

Proposition 3.7 (Naive sample complexity). *Let $\alpha, \beta, \gamma > 0$. Suppose that $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations such that $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$ for all $S \in \mathcal{C}$. Then, given access to m_u independent unlabeled samples from $\mathcal{D}$ and m_ℓ independent labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, there is an implementation of Algorithm 2 that returns a $(\mathcal{C}, \alpha)$-multi-calibrated predictor with probability at least $1-\beta$, for $m_u = O\left(\frac{\log(|\mathcal{C}| / \alpha\beta\gamma)}{\alpha^6\gamma^2}\right)$ and $m_\ell = O\left(\frac{\log(|\mathcal{C}| / \alpha\beta\gamma)}{\alpha^8\gamma^2}\right)$.*

Proof. The claim follows by running Algorithm 2 with $\mathrm{DQ}_{\mathcal{D}}$ implemented by Algorithm 3 and $\mathrm{SQ}_{\mathcal{D}}$ implemented by Algorithm 4. The general approach will be to feed Algorithm 3 and 4 fresh samples at the beginning of each iteration with sufficiently many samples to guarantee validity within the iteration. The overall complexity follows from the bound of $T = O(1/\alpha^4\gamma)$ on the iteration complexity.

Note that for each $t = 0, \dots, T$, at the start of iteration t, there is a fixed set of density and statistical queries that the algorithm can make. Specifically, these queries are determined by the subsets $\mathcal{C}^t$ defined as

$$\mathcal{C}^t = \left\{ S_v^t = S \cap \{x : p_t(x) = v\} : \forall S \in \mathcal{C}, \forall v \in \Lambda \right\}. \tag{3.8}$$

By the discretization defined in Λ, $|\mathcal{C}_t| \leq |\mathcal{C}| / \alpha$.

(Density queries) In each iteration, Algorithm 2 tests if each $S_v^t \in \mathcal{C}^t$ has density at least $\tau = \alpha^2\gamma$. Thus, by Lemma 3.5, given $m_0 = O\left(\frac{\log(1/\beta_0)}{\alpha^2\gamma}\right)$ unlabeled samples per iteration, each query to $\mathrm{DQ}_\mathcal{D}$ will be correct with probability at least $1 - \beta_0$. By union bounding against each potential query $S_v^t \in \mathcal{C}^t$ per iteration, and then against each iteration $t \in [T]$, if we take $\beta_0 = O\left(\frac{\beta}{|\mathcal{C}^t| \cdot T}\right)$, then with probability at least $1 - \beta$ every query across the algorithm will be correct. Thus, taking $\beta_0 = O\left(\frac{\alpha^5\beta\gamma}{|\mathcal{C}|}\right)$ the stated bound on m_u follows.

(Statistical queries) Once we've certified that a set $S_v^t \in \mathcal{C}^t$ has density at least $\gamma_{S_v^t} \geq \alpha^2\gamma$, then we may ask a statistical query. By Lemma 3.6, given $m_0 = O\left(\frac{\log(1/\beta_0)}{\alpha^2 \cdot \alpha^2\gamma}\right)$ labeled samples per iteration, each query to $\mathrm{SQ}_\mathcal{D}$ will be correct with probability at least $1 - \beta_0$. Selecting β_0 as above, the stated bound on m_ℓ follows. □

3.3 Improved Sample Complexity via Differential Privacy

As we've discussed, the multi-calibration statistical queries asked in Algorithm 2 are chosen adaptively. In other words, the decisions about which subpopulations to query changes as a function of the answers to prior queries. From a classical statistical perspective, this form of adaptive data analysis is a recipe for statistical invalidity due to overfitting. A recent line of work, however, has investigated techniques for guaranteeing the validity of conclusions of an adaptive statistical analyst. Specifically, the research program investigates how differential privacy [DMNS06, DR14]—a concept of algorithmic stability used to guarantee the privacy of individuals within a database—can be used to prevent overfitting. Starting with [DFH$^+$15a, DFH$^+$15c, DFH$^+$15b], many works from the past few years [RZ16, BNS$^+$16, JLN$^+$20] have pinned down quite elegantly the way that differential privacy can be used to improve statistical validity in adaptive data analysis.

At a high-level, differential privacy guarantees that upon viewing the output of a mechanism, an adversary cannot effectively distinguish whether an individual's datum was or was not included as an element of the mechanism's data. Intuitively, if an adaptive data analyst could overfit to the training data significantly, then running the analysis could be used as an effective statistical test to break differential privacy. Much technical work has gone into turning this intuition into formal theorem statements, culminating in the work of [JLN$^+$20], who give an intuitive proof of the "transfer" theorem showing that a differentially private, sample-accurate mechanism must also be distributionally-accurate. Most pertinent to our application is the following theorem, implicit in [JLN$^+$20].

Theorem 3.8 (Implicit in [JLN$^+$20])**.** *Suppose $\alpha, \beta, \varepsilon, \delta, \Delta > 0$. For a distribution $\mathcal{D}$ supported on a discrete universe $\mathcal{Z}$, consider a sequence of bounded statistical queries over functions $q_1, \ldots, q_T : \mathcal{Z} \to [0, \Delta]$ such that for all $t \in [T]$,*

$$\mathop{\mathbf{E}}_{z \sim \mathcal{D}}[\ q_t(z)\ ] \in [0, 1].$$

Suppose that a mechanism M, given m samples $z_1, \ldots, z_m \sim \mathcal{D}$, answers the sequence of queries while satisfying the following properties:

- *M is (ε, δ)-differentially private;*
- *M is (α, β)-sample accurate; that is, with probability at least $1 - \beta$, for all queries q_t,*

$$\left| \frac{1}{m} \sum_{i=1}^{m} q_t(z_i) - M(q_t) \right| \leq \alpha;$$

- *and $M(q_t) \leq 1$ for all $t \in [T]$.*

Then for every $c, d > 0$, the mechanism M is (α', β')-distributionally accurate for $\alpha' = \alpha + c + (e^{\varepsilon} - 1) + 2d$ and $\beta' = \beta/c + \delta/d$; that is, with probability at least $1 - \beta'$, for all queries q_t,

$$\left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}}[\ q_t(x)\ ] - M(q_t) \right| \leq \alpha'.$$

Note that by setting the parameters in Theorem 3.8 as $\alpha = \alpha_0, \varepsilon = \alpha_0/2, c = \alpha, d = \alpha/2$, and $\beta = \alpha\beta_0, \delta = \alpha\beta_0$, then we obtain a $(4\alpha_0, 3\beta_0)$-distributionally accurate mechanism for answering the sequence of queries.

Algorithm 5 DP-Multi-Calibrate

```
Repeat: for $t = 0, 1, \ldots$
  for each $S \in \mathcal{C}$ and each $v \in \Lambda$ do
    $S_v \leftarrow \{x \in S : p_t(x) = v\}$                    // consider level sets
    $\hat{\gamma}_{S,v} \leftarrow \mathrm{DQ}_{\mathcal{D},\tau,\sigma}(S_v)$
    if $\hat{\gamma}_{S,v} < \tau$ then
      continue to next $S_v$                                     // only consider large sets
    end if
    $\mu \leftarrow 1/\hat{\gamma}_{S,v}$
    $\Delta \leftarrow \text{DP-SQ}_{\mathcal{D},\alpha}(\delta_{v,S,\mu})$     // DP query expectation
    if $\Delta > 0$ then
      $\forall x \in S_v : \quad p_{t+1}(x) \leftarrow v - \Delta$            // test and update for calibration
      $\forall x \in \mathcal{X} \setminus S_v : \quad p_{t+1}(x) \leftarrow p_t(x)$
      break and continue to $t \leftarrow t + 1$                  // new iteration upon update
    end if
  end for
  return $\tilde{p} = p_t$                                       // return when no update occurs
```

Improved Sample Complexity. In this section, we will re-implement the statistical query oracle in a manner that guarantees differential privacy, even given the adaptivity in the chosen queries. Our mechanism will be based on the private multiplicative weights framework of [HR10], which is designed to answer many statistical queries accurately, provided the number of rounds of adaptivity is bounded. In all, we will prove the following statement about the revised Algorithm 5, completing our analysis of the sample complexity of learning a $(\mathcal{C}, \alpha)$-multi-calibrated predictor.

Theorem 3.9. *Let $\alpha, \beta, \gamma > 0$. Suppose that $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations such that for all $S \in \mathcal{C}$, $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$. Then, given access to m_u independent unlabeled samples from $\mathcal{D}$ and m_ℓ independent labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, there is an implementation of Algorithm 5 that returns a $(\mathcal{C}, \alpha)$-multi-calibrated predictor with probability at least $1 - \beta$ provided*

$$m_u \geq \Omega\left(\frac{\log(|\mathcal{C}|/\beta)}{\alpha^8\gamma^2}\right) \qquad\qquad m_\ell \geq \Omega\left(\frac{\log(|\mathcal{C}|/\alpha\beta\gamma)^{3/2}}{\alpha^6\gamma^{3/2}}\right).$$

While the sample complexity in Theorem 3.9 does not improve over Proposition 3.7 in all parameter settings, in many reasonable settings, the approach applying differential privacy will be improved. In particular, for the multi-calibration guarantee to give strong protections over the collection of subpopulations $\mathcal{C}$, we would typically take α to be a small constant and protect populations that may only be in the distribution with a small γ probability. Thus, saving a $1/\alpha^2\gamma^{1/2}$-factor is a considerable savings regardless of the chosen $\mathcal{C}$, and typically will outweigh the additional $\log(|\mathcal{C}|/\beta)^{1/2}$-factor. Further, in many settings the distribution $\mathcal{D}$ is well-understood; in this case, the additional unlabeled samples may be significantly cheaper than obtaining labeled samples.

Overview of approach. Due to the nature of the multi-calibration constraints and the way that Algorithm 2 asks queries, our application of the main DP-transfer theorem will be slightly indirect. Recall that in principle, we would like to query the value of $\mathbf{E}_{x\sim\mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ]$; if it is far from v, then we want to issue an update. To apply the transfer theorem we need to adjust the algorithm as follows.

First, when asking density queries of the sub-level-sets $\mathrm{DQ}_{\mathcal{D},\tau,\sigma}(S_v)$, we will require a much tighter estimate of the probability $\gamma_{S,v} = \mathbf{Pr}_{\mathcal{D}}[\ x \in S_v\ ]$. In particular, we require an $e^{\pm\sigma}$ multiplicative approximation for $\sigma = O(\alpha)$ rather than a constant, and have the oracle return the estimate $\hat{\gamma}_{S,v}$. We require more unlabeled samples in Algorithm 5 compared to Algorithm 2 to perform this more stringent density estimation.

Then, with the estimate $\hat{\gamma}_{S,v}$ in hand, we can issue a statistical query that simulates the desired query. Specifically, we will query the following difference function. For a value $v \in [0,1]$ and a sensitivity parameter μ, we let the function $\delta_{v,\mu} : \mathcal{X} \to [0,1]$ be defined as

$$\delta_{v,S,\mu}(x) = \mu \cdot (v - p^*(x)) \cdot \mathbf{1}[\ x \in S\ ].$$

If we take a statistical query over the function $\delta_{v,S,\mu}$—based on the choice of $\mu = \hat{\gamma}_{S,v}$—is to approximate the mis-calibration on S_v.

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ \mu \cdot (v - p^*(x)) \cdot \mathbf{1}[\ x \in S\ ]\ ] &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ \frac{(v - p^*(x)) \cdot \mathbf{1}[\ x \in S_v\ ]}{\hat{\gamma}_{S,v}}\ \right] \\ &\approx v - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ].\end{aligned}$$

Some care must be taken to prove the claimed approximations formally, but the main technical hurdle is to implement the statistical query algorithm in a way that we can apply the transfer theorem for statistical generalization.

Specifically, with the estimate $\hat{\gamma}_{S,v}$ in hand, we will implement the statistical query by answering the query to guarantee (α, β)-sample accuracy; that is, accuracy to the empirical estimate

$$u = \frac{1}{m} \sum_{i=1}^{m} \frac{y_i \cdot \mathbf{1}[\ x_i \in S_v\ ]}{\hat{\gamma}_{S,v}}.$$

Importantly, estimating the density $\hat{\gamma}_{S,v}$ non-privately and separately from the statistical query allows our interaction with the labeled sample to have fixed, bounded sensitivity on a query whose answer never needs to exceed 1. Thus, if we can implement the mechanism to satisfy (ε, δ)-differential privacy, we can then apply the transfer theorem to guarantee that the mechanism's responses will be distributionally accurate.

The details. As we will see, implementing the statistical query oracle to answering the multi-calibration queries while satisfying differential privacy is a bit delicate. In particular, to guarantee differential privacy, our technique will require a bound on the number of rounds of adaptively chosen queries; conversely, to guarantee a bounded number of rounds, we need to guarantee that the algorithm only suffers "privacy loss" upon an update to the predictive model. Thus, we use a separate unlabeled sample to estimate the density queries; this way, we will only guarantee privacy over the labeled sample (sufficient to guarantee generalization) but will never use any of our privacy budget on the density queries.

First, we show that given a sufficiently accurate estimate of the density $\mathbf{Pr}_{\mathcal{D}}[\ x \in S_v\ ]$, we can answer the multi-calibration queries as a bounded-sensitivity "absolute" statistical query (i.e., as an expectation over $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ without conditioning).

Lemma 3.10. *Suppose* $\gamma_{S,v} = \mathbf{Pr}_{\mathcal{D}}[\ x \in S_v\ ]$ *and let* $\left|\log\left(\frac{\hat{\gamma}_{S,v}}{\gamma_{S,v}}\right)\right| \le \sigma < 1/2$*. Then,*

$$\left| \mathop{\mathbf{E}}_{x,y\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \frac{y \cdot \mathbf{1}[\ x \in S_v\ ]}{\hat{\gamma}_{S,v}}\ \right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x) \mid p_t(x) = v\ ] \right| \le 2\sigma.$$

Proof. Leveraging the bounded ratio between $\hat{\gamma}_{S,v}$ and $\gamma_{S,v}$, we bound the expectation as follows.

$$\begin{aligned}
\mathop{\mathbf{E}}_{x,y\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\frac{y\cdot\mathbf{1}[\,x\in S_v\,]}{\hat{\gamma}_{S,v}}\right] &\le e^{\sigma}\cdot\mathop{\mathbf{E}}_{x,y\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\frac{y\cdot\mathbf{1}[\,x\in S_v\,]}{\gamma_{S,v}}\right]\\
&= e^{\sigma}\cdot\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\,p^*(x)\mid p_t(x)=v\,]\\
&\le (1+2\sigma)\cdot\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\,p^*(x)\mid p_t(x)=v\,]\\
&\le \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\,p^*(x)\mid p_t(x)=v\,]+2\sigma.
\end{aligned}$$

The other direction of the inequality follows by a similar argument. □

Thus, taking $\sigma = \alpha/4$, to estimate the statistical query $\mathbf{E}_{\mathcal{D}_S}[\,p^*(x)\mid p_t(x)=v\,]$, we obtain the unlabeled sample complexity again by a Chernoff bound, similar to Lemma 3.5. With such a density query oracle in place, it suffices to implement statistical queries for $\delta_{v,S,\mu}$.

Differentially Private Statistical Queries. The mechanism we build essentially follows the private multiplicative weights mechanism of [HR10]. We adapt the proof of privacy and accuracy to our setting; specifically, our proof follows the analysis from [Vad17] quite closely. Our goal in this section is not to give a comprehensive overview of DP, but rather provide a high-level proof of the result given a working knowledge of DP. For a comprehensive background on differential privacy, the author recommends [DR14, Vad17].

Differential privacy is a strong notion of stability for data analysis algorithms. A DP mechanism for interacting with a database guarantees that small changes to the database cannot have profound impacts on the distribution of outcomes.

Definition 3.11 (Differential Privacy). *Fix $\varepsilon,\delta>0$. For a discrete domain $\mathcal{Z}$ and a family of queries $\mathcal{Q}\subseteq\{q:\mathcal{Z}\to[0,1]\}$, a mechanism $M:\mathcal{Z}^m\times\mathcal{Q}\to[0,1]$ is (ε,δ)-differentially private if for all neighboring databases $Z,Z'\subseteq\mathcal{Z}$ where $\|Z-Z'\|_0=1$ and every query $q\in\mathcal{Q}$, for all events T*

$$\mathbf{Pr}[\,M(Z,q)\in T\,]\le e^{\varepsilon}\cdot\mathbf{Pr}\left[\,M(Z',q)\in T\,\right]+\delta$$

where the probability is taken over the randomness in the mechanism. A mechanism is ε-differentially private if it is $(\varepsilon,0)$-differentially private.

A key appeal of differential private mechanisms is that they allow for modular design. In particular, given the outputs of multiple DP mechanisms over the same data, the total privacy loss degrades gracefully. Further, DP is closed under post-processing—that is, for any function f, if the release of X is (ε, δ)-DP, then the release of $f(X)$ is equally privacy-preserving (ε, δ)-DP. Formally, our analysis relies upon the following advanced composition theorem.

Theorem 3.12 (Advanced composition, [DRV10]). *Suppose $M_1, \ldots, M_T$ are T (possibly adpatively chosen) ε_0-differentially private mechanisms for $T < 1/\varepsilon_0^2$. Then, $M = (M_1, \ldots, M_T)$ is (ε, δ)-differentially private for any $\delta > 0$ and $\varepsilon = O(\varepsilon_0 \cdot \sqrt{T \cdot \log(1/\delta)})$.*

With advanced composition and post-processing in place, the last preliminary we need to describe our mechanism is the Laplace mechanism, which guarantees ε-DP for bounded sensitivity queries. We say a function Q has input sensitivity of Δ if for all neighboring databases $\|Z - Z'\|_0 = 1$,

$$\left|Q(Z) - Q(Z')\right| \leq \Delta.$$

Theorem 3.13 (Laplace Mechanism, [DMNS06]). *Suppose $Q : \mathcal{Z}^* \to [0, 1]$ is a function of a database with input sensitivity Δ. Then, given a database $Z \in \mathcal{Z}^*$, a mechanism that releases*

$$Q(Z) + \mathrm{Lap}\left(\frac{\Delta}{\varepsilon}\right)$$

is ε-differentially private.

Here, $L \sim \mathrm{Lap}(B)$ denotes a random variable drawn from the Laplace distribution with mean 0 and scale parameter B. The main fact that we need about the Laplace distribution is the following tail bound.

Fact (Laplace tails). *Let $L \sim \mathrm{Lap}\,(B)$ be a random variable distributed according to the Laplace distribution with mean 0 and scale B. The magnitude of the random variable $|L|$ is distributed as follows.*

$$\mathbf{Pr}\left[\ |L| > \tau\ \right] = \exp\left(\frac{-\tau}{B}\right)$$

The mechanism. With these preliminaries in place, we're now able to describe the mechanism for interacting with the labeled data. The mechanism can be viewed as a composition of multiple Laplace mechanisms, designed to allow for the release of many queries. Importantly, when the query value is sufficiently small (i.e., the predictor p_t is already quite

Algorithm 6 Private SQ Oracle

```
Initialize:
  Given $\alpha, \gamma, \varepsilon_0 > 0$ and $m \in \mathbb{N}$
  Draw $Z \leftarrow (x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$      // independent labeled samples
  $\mu_0 \leftarrow 2/\alpha^2\gamma$      // worst-case sensitivity
  $\tau \leftarrow \alpha/2 + \mathrm{Lap}(\mu_0/\varepsilon_0 m)$      // randomize error tolerance

On query DP-SQ$_{\mathcal{D},\alpha}(\delta_{v,S,\mu}, \mathcal{X})$:
  $u \leftarrow \frac{1}{m} \cdot \sum_{i=1}^{m} \mu \cdot y_i \cdot \mathbf{1}[\, x_i \in S \,]$
  $\hat{\delta}_{v,S,\mu} \leftarrow |v - u|$
  if $\hat{\delta}_{v,S,\mu} + \mathrm{Lap}(\mu_0/\varepsilon_0 m) \leq \tau$ then
    return 0      // v is sufficiently accurate
  end if
  $\tau \leftarrow \alpha/2 + \mathrm{Lap}(\mu_0/\varepsilon_0 m)$      // re-randomize after release
  return $\hat{\delta}_{v,S,\mu} + \mathrm{Lap}(\mu/\varepsilon_0 m)$   // round to nearest multiple of alpha in [0,1]
```

accurate on the sub-level set S_v) we return 0 with high probability. In such cases, very little information about the database is actually released. The main privacy loss occurs when we return a non-zero $\hat{\delta}_{v,S,\mu}$. But recall, these are also the iterations upon which we update p_t. Intuitively, as long as we make sufficient progress towards multi-calibration in each of these iterations, Algorithm 5 will terminate with a multi-calibrated predictor before we use up our privacy budget.

We give a description of the mechanism in Algorithm 6. To begin the analysis, note that for a given a run of Algorithm 5, we can split the sequence of queries into T rounds. At the start of each round $t \in [T]$, we fix the set of possible queries that we might ask to be $\mathcal{C}^t$ as defined in (3.8). Within each round, the algorithm continues to ask statistical queries from $\mathcal{C}^t$ until the response is non-zero. Then, the model p_t is updated based on the statistical query and the round ends. Importantly, the response to every query in the round is 0, except for the final response in Λ.

We argue that Algorithm 6 is a DP mechanism to handle such sequences of queries. Specifically, we will analyze Algorithm 6 as separate ε_0-DP sub-mechanisms for each round that will guarantee accuracy within the round with high probability. Then to analyze the privacy of the entire algorithm, we apply advanced composition to the T rounds and obtain an (ε, δ)-DP mechanism for $\varepsilon = \varepsilon_0 \cdot \sqrt{T \cdot \log(1/\delta)}$.

Further, for the right choice of ε_0 and m, we can guarantee that the responses are sufficiently sample accurate. Applying the transfer theorem, we will show that the entire sequence of queries are answered sufficiently accurately for Algorithm 5 from a bounded m. First, we argue that if we set ε_0 and m appropriately, we can guarantee sample accuracy.

Proposition 3.14 (Sample Accuracy)**.** *Suppose $m \geq \frac{2 \cdot \log(1/\beta_0)}{\alpha^3 \gamma \varepsilon_0}$. Then, every non-zero response is $\alpha/16$-sample accurate.*

Proof. Suppose DP-SQ$_{\mathcal{D},\alpha}(\delta_{v,S,\mu})$ responds with a non-zero response, i.e., with $\hat{\delta}_{v,S,\mu}$ + Lap$\,(\mu/\varepsilon_0 m)$. This returned value is the empirical estimate of $\delta_{v,S,\mu}$ evaluated on the sample plus Laplace noise distributed as $L \sim$ Lap$\,(\mu/\varepsilon_0 m)$. Thus, an upper bound on the error comes from the magnitude of this noise using the Laplace tails bound.

$$\begin{aligned}\mathbf{Pr}\left[\ |L| > \alpha/16\ \right] &= \exp\left(-\alpha \cdot \varepsilon_0 m/\mu\right) \\ &\geq \exp\left(-\alpha^3 \gamma \varepsilon_0 m/2\right)\end{aligned}$$

Thus, to upper bound this probability by some β_0, it suffices to take m as

$$m \geq \frac{2 \cdot \log(1/\beta_0)}{\alpha^3 \gamma \varepsilon_0}.$$

□

With sample accuracy in place, we establish that the mechanism satisfies differential privacy.

Proposition 3.15 (Privacy)**.** *Suppose $\varepsilon_0 < O\left(\varepsilon\alpha^2\gamma^{1/2}/\sqrt{\log(1/\delta)}\right)$. Then, Algorithm 6 can answer all $T = O(1/\alpha^4\gamma)$ queries while satisfying (ε, δ)-differential privacy for $\varepsilon = \Theta(\alpha)$.*

Proof. The proof follows closely the private multiplicative weights privacy proof. We break the analysis into rounds, showing that each iteration satisfies $O(\varepsilon_0)$-DP. The bound follows by applying advanced composition for $T = O(1/\alpha^4\gamma)$ rounds with the goal of $\varepsilon = \Theta(\alpha)$. We can verify that when we solve for the necessary ε_0 to obtain (ε, δ)-DP overall

$$\varepsilon = O(\varepsilon_0 \cdot \sqrt{T \cdot \log(1/\delta)}),$$

then the number of rounds is indeed $T < 1/\varepsilon_0^2$.

We note that each iteration of the mechanism can be viewed as the composition of a constant number of $O(\varepsilon_0)$-DP instances of the Laplace mechanism. To see this, note that if Algorithm 5 asks a query on S_v, then $\hat{\gamma}_{S,v} > \alpha^2\gamma/2$, so the sensitivity μ/m of each empirical query on m samples is at most $2/\alpha^2\gamma m = \mu_0/m$.

The main observation is as follows: fixing an ordering over the queries at the start of each round $(v,S,\mu)_1,\ldots,(v,S,\mu)_k$, the probability of observing the sequence of responses $\mathrm{DQ}_{\mathcal{D},\alpha}(\delta_{(v,S,\mu)_1}) = 0,\ldots,\mathrm{DQ}_{\mathcal{D},\alpha}(\delta_{(v,S,\mu)_{k-1}}) = 0, \mathrm{DQ}_{\mathcal{D},\alpha}(\delta_{(v,S,\mu)_1}) \neq 0$ is essentially the same under a small change to the underlying data set. Specifically, fix a sequence of outputs based on an underlying data set Z, and suppose we change the set by a single element. Then, we show a small change to the randomly-selected threshold τ will recover the same sequence of 0-responses. Consider fixing the random draws of Laplace random variables where $L_i \sim \mathrm{Lap}(\mu_0/\varepsilon_0 m)$ Consider the maximum over $i < k$ of the noisy difference computation.

$$\delta_0 = \max_{i<k}\left\{\hat{\delta}_{(v,S,\mu)_i} + L_i - \tau\right\} = \max_{i<k}\left\{\hat{\delta}_{(v,S,\mu)_i} + L_i - \alpha/2\right\} + \mathrm{Lap}(\mu_0/\varepsilon_0 m)$$

Specifically, note that after fixing the sequence of L_i's this maximum is a query with at most μ_0 sensitivity. Thus, answering it with additional Laplace noise $\mathrm{Lap}(\mu_0/\varepsilon_0 m)$ will be ε_0-DP. Conditioning on the event where $\delta_0 < 0$, we proceed to the test on the kth query. This query passes the test and returns a non-zero value only if the difference is sufficiently large.

$$\begin{aligned} 0 &< \delta_{(v,S,\mu)_k} - \tau + \mathrm{Lap}(\mu_0/\varepsilon_0 m) \\ &= \delta_{(v,S,\mu)_k} - \max_{i<k}\left\{\hat{\delta}_{(v,S,\mu)_i} + L_i\right\} + \delta_0 + \mathrm{Lap}(\mu_0/\varepsilon_0 m) \end{aligned}$$

With the $\{L_i\}$ and δ_0 values fixed, this threshold test is a $2\mu/m$-sensitive query over the data set; thus, the Laplace noise ensures $2\varepsilon_0$-DP. Finally, the release of the query itself is ε_0-DP by using fresh Laplace noise. □

Note that with these propositions in place, we can apply the transfer theorem. While we established sample accuracy for the non-zero queries, this is equivalent to establishing sample accuracy on queries of the form

$$\max\left\{0, \delta_{v,S,\mu}(Z) - \alpha/2\right\}.$$

Thus, applying the transfer theorem, we obtain the following lemma.

Lemma 3.16. *Suppose the parameters of Algorithm 6 satisfy the conditions in Propostions 3.14 and 3.15. Then,* DP-SQ$_{\mathcal{D},\alpha}$ *satisfies the following properties.*

- *if* $v - \mathbf{E}_{x\sim\mathcal{D}}[\, p^*(x) \mid p_t(x) = v \,] \le \alpha/4$, *then* DP-SQ$_{\mathcal{D},\alpha}(\delta_{v,S,\mu}) = 0$
- *if* $v - \mathbf{E}_{x\sim\mathcal{D}}[\, p^*(x) \mid p_t(x) = v \,] > \alpha$, *then* DP-SQ$_{\mathcal{D},\alpha}(\delta_{v,S,\mu}) > 3\alpha/4$.

Theorem 3.9 follows as a corollary of this lemma and the analysis of the statistical query algorithm from the prior section.

Chapter Notes

The results of this chapter were originally proved in [HKRR18]. Importantly, we correct an erratum from [HKRR18] in the stated sample complexity of Theorem 3.9, and present the corrected proof in detail. The author is grateful to Lee Cohen and Yishay Mansour for identifying the original error in preparing their work [SCM20].

Chapter 4

Multi-Calibration Auditing and Post-Processing

In Chapter 3, we established bounded sample complexity learning algorithms for obtaining multi-calibrated predictors. While the algorithms obtained efficiency in the number of samples used, a drawback is that the algorithms run in linear time in the cardinality of the collection of subpopulations we wish to protect. Note that the complexity of each iteration is dominated by searching for whether there exists some $S \in \mathcal{C}$ and $v \in \text{supp}(p)$ where the predictions are mis-calibrated. In other words, the problem of learning a multi-calibrated predictor can be (Turing) reduced the problem of auditing for multi-calibration—deciding whether a given predictor satisfies multi-calibration.

Auditing is a natural problem on its own: if an ML service provider claims their models are "fair," clients may naturally want to verify the claims themselves. In this chapter, we study the problem of auditing for multi-calibration. We show that the auditing problem is computationally equivalent to the problem of agnostic learning the class $\mathcal{C}$. This equivalence is both good news and bad news. Pessimistically, agnostic learning is a notoriously hard problem in theory—cryptography schemes are based on its hardness for certain classes of functions. Indeed, this reduction shows that under plausible cryptographic assumptions [GGM84, BR17], the running time dependence on $|\mathcal{C}|$ of Algorithm 2 cannot be improved significantly for arbitrary classes $\mathcal{C}$. Optimistically, all of practical machine learning is agnostic and is still remarkably effective at finding patterns in the data for tasks of interest. This direction shows a reduction from the problem of learning for multi-calibration to the problem of "vanilla" machine learning.

An important aspect of our auditing setup is that it only uses black-box access to the predictor in question. In fact, we observe that the reduction from learning to auditing works perfectly well even if we don't know the internals of the predictor. This observation suggests an alternative paradigm to learning a multi-calibrated predictor directly. First, we learn a highly-sophisticated model using plentiful (but possibly biased) data, that aims to approximate p^* without concern for fairness. Then, using this pre-trained unfair model, we can audit for multi-calibration by learning where the model is miscalibrated, using a held-out unbiased data set drawn from the distribution of interest $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. Using the result of auditing, we can update the model until it passes and thus, is multi-calibrated.

We begin this chapter with the formal equivalence between auditing and learning. Then, we turn to discussing more practical aspects of auditing and post-processing. In particular, we conclude the section with an empirical case study revisiting the Gender Shades study, showing how the multi-calibration framework can help to address the observed performance disparities across demographic groups.

4.1 Auditing via Weak Agnostic Learning

In this section, we demonstrate a connection between the goal of learning multi-calibrated predictors and weak agnostic learning, introduced in the literature on agnostic boosting [BDLM01, KMV08, KK09, Fel10]. At its core, the algorithmic framework presented in Chapter 3 relies upon the ability to audit a given predictor to determine whether it is sufficiently calibrated on sub-level sets or not. Formally, we can define the auditing problem as follows.

Problem (Multi-Calibration Auditing)**.** *Fix some $\alpha, \gamma, \sigma > 0$ where $\alpha\gamma > \sigma$ and collection of subpopulations $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ where for all $S \in \mathcal{C}$, $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$. Given a predictor $p : \mathcal{X} \to [0,1]$ with support $s = |\mathrm{supp}(p)|$, if there is any $S \in \mathcal{C}$ and $v \in \Lambda$ such that*

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

for sufficiently large sub-level sets $\mathbf{Pr}_{\mathcal{D}_S}[\ p(x) = v\ ] > \alpha/s$, return some $S' \subseteq \mathcal{X}$ where

$$\left| \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}\left[\ x \in S'\ \right] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S'}}\left[\ (p^*(x) - p(x))\ \right] \right| > \sigma/s.$$

In other words, if the given model p is not multi-calibrated, then the auditor is required to return a subpopulation $S' \subseteq \mathcal{X}$ on which the model is biased. By the assumption that the underlying model is not multi-calibrated, we know that such a subpopulation must exist. Note that for the auditing to be effective, we need σ to be polynomially related to our accuracy parameters; that way, we can correct the model into a multi-calibrated one in polynomially many rounds of auditing as in Algorithm 2.

We will connect this problem to the problem of detecting correlations between the miscalibration and the subpouplations within $\mathcal{C}$. For this discussion, it will be useful to consider an equivalent representation of subpopulations $S \subseteq \mathcal{X}$ as boolean concepts over $\{-1, 1\}$. Specifically, for each $S \in \mathcal{C}$ there exists a boolean function $c_S : \mathcal{X} \to \{-1, 1\}$ such that for all $x \in \mathcal{X}$

$$c_S(x) = \begin{cases} 1 & x \in S \\ -1 & x \notin S. \end{cases}$$

We will overload notation, referring to a concept $c \in \mathcal{C}$, imagining the collection of subpopulations / concepts $\mathcal{C}$ independent of the representation. We will connect the problem of finding a set $S \in \mathcal{C}$ on which a predictor p violates calibration to the problem of learning over the concept class $\mathcal{C}$ over the distribution $\mathcal{D}$.

Weak agnostic learning is a problem that centers around detecting correlations between arbitrary labels and concepts $c \in \mathcal{C}$. In our results, we will work with the *distribution-specific* weak agnostic learners of [Fel10].[1] For notational convenience, we will use the following inner product notation to represent the correlation of two functions $f, g : \mathcal{X} \to [-1, 1]$ over $\mathcal{D}$.

$$\langle f, g \rangle_{\mathcal{D}} = \mathop{\mathbf{E}}_{x \sim \mathcal{D}}[\ f(x) \cdot g(x)\ ]$$

Formally, we define weak agnostic learning as the following promise problem.

Problem ((ρ, τ)-Weak agnostic learning)**.** *Fix some $\rho > \tau > 0$ and a concept class $\mathcal{C} \subseteq \{\mathcal{X} \to \{-1, 1\}\}$, and suppose $f : \mathcal{X} \to [-1, 1]$ is an arbitrary labeling function. Promised that there is some $c \in \mathcal{C}$ with correlation $\langle c, f \rangle_{\mathcal{D}} > \rho$, return some $h : \mathcal{X} \to [-1, 1]$ with correlation at least $\langle h, f \rangle_{\mathcal{D}} > \tau$.*

Typically, we take ρ, τ to be inverse polynomial in the parameters of interest (e.g., the

[1] Often, such learners are defined in terms of their error rates rather than correlations; the definitions are equivalent up to factors of 2 in ρ and τ. Also, we will always work with a hypothesis class $\mathcal{H} = [-1, 1]^{\mathcal{X}}$ the set of functions from $\mathcal{X}$ to $[-1, 1]$, so we fix this class in the definition.

dimension of the input data). The literature on agnostic boosting shows that given such settings of parameters, weak agnostic learners can be boosted to strong agnostic learners after polynomially many iterations. Importantly, while we denote the labels $f(x)$ as a function of x, agnostic learning does not assume that these labels come from any class of bounded complexity—the labels may be arbitrary. Intuitively, if there is a concept $c \in \mathcal{C}$ that correlates nontrivially with the observed labels, then a weak agnostic learner must return a hypothesis h (not necessarily from $\mathcal{C}$), that is also nontrivially correlated with the observed labels.

4.1.1 Multi-calibration from weak agnostic learning

First, we show how we can use a weak agnostic learner to solve the multi-calibration auditing problem. Specifically, we aim to solve the following problem. Rather than issuing separate statistical queries for each S_v, we will use the learner to perform the auditing. If the auditing fails, the learner will return a hypothesis that can be used as an update that makes significant progress towards multi-calibration. Our focus in this section is on time complexity, so our discussion of sample complexity will be informal. We assume that the weak agnostic learner is guaranteed to given access to m samples for some sufficiently large $m \in \mathbb{N}$, and take at least this many labeled samples for the multi-calibration auditing problem.

Lemma 4.1. *Suppose $\mathcal{L}$ solves the (ρ, τ)-weak agnostic learning problem for a concept class $\mathcal{C}$ from labeled samples. Let $\mathcal{C}_\gamma \subseteq \{0,1\}^{\mathcal{X}}$ be the collection of subpopulations $S \in \mathcal{C}$ such that $\mathbf{Pr}_{\mathcal{D}}[\ x \in S\ ] \geq \gamma$. Then, given access to labeled samples $(x, y) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, we can audit a predictor $p : \mathcal{X} \to [0,1]$ for $(\mathcal{C}_\gamma, \alpha)$-multi-calibration auditing using $2 \cdot |\mathrm{supp}(p)|$ calls to $\mathcal{L}$ for $\rho = \alpha\gamma/s$ and $\tau \geq \sigma/s$.*

Proof. For some finite discrete support $\Lambda \subseteq [0,1]$, suppose we are given a predictor $p : \mathcal{X} \to \Lambda$. Our goal is to determine whether p satisfies $(\mathcal{C}_\gamma, \alpha)$-multi-calibration and if so, to return a hypothesis demonstrating a subpopulation where p is biased. We will demonstrate a slightly relaxed goal, by allowing the auditor to return a hypothesis $h : \mathcal{X} \to [-1,1]$, rather than a boolean $c_S : \mathcal{X} \to \{-1,1\}$. To start, we will simply verify that p is approximately calibrated overall. For each $v \in \Lambda$, we will estimate

$$\Delta_v = \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ p^*(x) \mid p(x) = v\ ] - v.$$

This can be carried out effectively from a small number of samples. Take $\varepsilon = \alpha\gamma \geq \sigma$. If for any of these level sets, the magnitude of the difference exceeds a threshold $\mathbf{Pr}_{\mathcal{D}}\left[\ p(x) = v\ \right] \cdot |\Delta_v| > \varepsilon/s$, then we can simply return the level set $\{x \in \mathcal{X} : p(x) = v\}$ as evidence of mis-calibration.

Thus, in subsequent analysis, we will assume that the model is approximately calibrated up to this absolute error of ε/s. With knowledge that p is well-calibrated overall, then for each $v \in \Lambda$ separately, we will take a new labeled sample $\{(x_1, y_m), \ldots, (x_m, y_m)\}$ and consider relabeling each x_i according to $\ell_v(x_i)$ as follows.

$$\ell_{v,y}(x_i) = \mathbf{1}\left[\ p(x_i) = v\ \right] \cdot (v - y_i)$$

Then, we will pass the samples $\{(x_i, \ell_{v,y}(x_i))\}$ to the weak agnostic learner $\mathcal{L}$. We show that if p violates $(\mathcal{C}, \alpha)$-multi-calibration by over-estimating the true value on the sub-level set $S_v = \{x \in S : p(x) = v\}$, then the correlation between $\ell_{v,y}$ and $c_S \in \mathcal{C}$ will be sufficiently large to satisfy the weak agnostic learning promise. (The other direction will follow similarly by taking the labels to be $-\ell_{v,y}(x)$.) Let $T = \mathcal{X} \setminus S$; note that we can write $c_S(x) = \mathbf{1}\left[\ x \in S\ \right] - \mathbf{1}\left[\ x \in T\ \right] = 2 \cdot \mathbf{1}\left[\ x \in S\ \right] - 1$.

$$\begin{aligned}
\langle \ell_{v,y}, c_S \rangle &= \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \ell_{v,y}(x) \cdot c_S(x)\ \right] \\
&= \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \mathbf{1}\left[\ x \in S_v\ \right] \cdot (v - y)\ \right] - \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \mathbf{1}\left[\ x \in T_v\ \right] \cdot (v - y)\ \right] \\
&= 2 \cdot \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \mathbf{1}\left[\ x \in S_v\ \right] \cdot (v - y)\ \right] - \mathop{\mathbf{E}}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}\left[\ \mathbf{1}\left[\ p(x) = v\ \right] \cdot (v - y)\ \right] \\
&= 2 \cdot \mathop{\mathbf{E}}_{\mathcal{D}}\left[\ \mathbf{1}\left[\ x \in S_v\ \right] \cdot (v - p^*(x))\ \right] - \mathop{\mathbf{E}}_{\mathcal{D}}\left[\ \mathbf{1}\left[\ p(x) = v\ \right] \cdot (v - p^*(x))\ \right] \\
&\geq 2 \cdot \alpha\gamma/s - \varepsilon/s
\end{aligned}$$

Again, taking $\varepsilon = \alpha\gamma$ shows that the correlation is at least $\rho \geq \alpha\gamma/s$. Thus, the weak agnostic learning promise is satisfied, and we obtain a hypothesis $h : \mathcal{X} \to [-1, 1]$ such that

$$\begin{aligned}
\sigma/s > \tau &\leq \langle \ell_{v,y}, h \rangle \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}}\left[\ h(x) \cdot \mathbf{1}\left[\ p(x) = v\ \right] \cdot (v - y)\ \right] \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}}\left[\ h(x) \cdot \mathbf{1}\left[\ p(x) = v\ \right] \cdot (p(x) - p^*(x)\ \right].
\end{aligned}$$

Thus, returning the function $h_v : \mathcal{X} \to [-1, 1]$ defined as

$$h_v(x) = h(x) \cdot \mathbf{1}[\ p(x) = v\ ]$$

will satisfy the (relaxed) multi-calibration auditing guarantee. □

4.1.2 Weak agnostic learning from multi-calibration

Proof. Suppose there exists some $c \in \mathcal{C}$ such that $\langle c, f\rangle_{\mathcal{D}} > \rho$. We will show how to construct a predictor $p : \mathcal{X} \to [0, 1]$ from labeled samples $\{(x_1, f(x_1)), \ldots, (x_m, f(x_m))\}$ such that auditing for multi-calibration will find some $h : \mathcal{X} \to [-1, 1]$ with correlation at least $\langle h, f\rangle_{\mathcal{D}} > \tau$, solving the weak agnostic learning problem. First, we define a function $p^* : \mathcal{X} \to [0, 1]$ in terms of $f : \mathcal{X} \to [-1, 1]$ and let μ denote its mean over $\mathcal{D}$.

$$p^*(x) = \frac{f(x) + 1}{2} \qquad\qquad \mu = \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ p^*(x)\ ] = \frac{1}{2} \cdot (\langle 1, f\rangle_{\mathcal{D}} + 1)$$

First, suppose $\mu < 1/2 - \sigma/4$. We argue that in this case, the constant hypothesis $h(x) = -1$ satisfies the weak agnostic guarantee.

$$\langle h, f\rangle_{\mathcal{D}} = -\langle 1, f\rangle_{\mathcal{D}} = 1 - 2 \cdot \mu \geq \sigma/2.$$

Thus, we will begin by testing this condition using the samples labeled according to f; if μ is sufficiently small, then we return the constant function that correlates sufficiently with the label.

Proceeding, we assume that $\mu > 1/2 - \sigma/4$. Consider for any $c \in \mathcal{C}$, the correlation between $\langle c, f\rangle_{\mathcal{D}}$ and let $S_c \subseteq \mathcal{X}$ be the corresponding subpopulation. We show that if the correlation with c is sufficiently large, then the constant predictor $p(x) = 1/2$ violates multi-calibration on S_c.

$$\begin{aligned}
\langle c, f\rangle_{\mathcal{D}} &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ c(x) \cdot f(x)\ ] \\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ \ (2 \cdot \mathbf{1}[\ x \in S_c\ ] - 1) \cdot (2 \cdot p^*(x) - 1)\ \ ] \\
&= 4 \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}}[\ \mathbf{1}[\ x \in S_c\ ] \cdot p^*(x)\ ] - 2 \cdot \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x \in S_c\ ] - 2 \cdot \mu + 1 \\
&= 4 \cdot \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x \in S_c\ ] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x) - 1/2\ ] - 2 \cdot \mu + 1
\end{aligned}$$

Let $p(x) = 1/2$ be the trivial constant predictor, and ensure that $\rho \geq \sigma$. Thus, under the weak agnostic promise that there exists some $c \in \mathcal{C}$ where $\langle c, f\rangle_{\mathcal{D}} > \rho$, we see the following inequality

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{D}_{S_c}}[\, p^*(x) - p(x) \,] &= \frac{\langle c, f\rangle_{\mathcal{D}} + 2\mu - 1}{4 \cdot \mathbf{Pr}_{x\sim\mathcal{D}}[\, x \in S_c \,]}\\ &\geq \frac{\rho - \sigma/2}{4}\\ &\geq \rho/8\end{aligned}$$

where we note that $\mathbf{Pr}_{\mathcal{D}}[\, x \in S_c \,] \leq 1$ and apply the assumption that $2\mu - 1 \geq -\sigma/2 \geq -\rho/2$. As such, if we audit the predictor $p(x) = 1/2$ for $(\mathcal{C}, \alpha)$-multi-calibration for $\alpha = \rho/4$, then the auditing promise is satisfied, and we will receive a subpopulation $S \subseteq \mathcal{X}$ such that

$$\left| \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\, x \in S \,] \cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\, p^*(x) - 1/2 \,] \right| > \sigma.$$

By the same argument as above, the hypothesis $h_S(x) = 2 \cdot \mathbf{1}[\, x \in S \,] - 1$ correlates well with f.

$$\langle h_S, f\rangle \geq \sigma/8$$

Thus, we return a hypothesis satisfying the (ρ, τ)-weak agnostic guarantee for $\tau = \sigma/8$. □

Note that the proof did not rely upon the full power of multi-calibration auditing. In particular, because the predictor we audited only had a single supported element, we were really testing whether there was any subpopulation bias. In this sense, auditing for multi-accuracy would suffice and is even more tightly connected to the problem of weak agnostic learning. We further explore the power of multi-accuracy auditing in the next sections.

4.2 Black-Box Post-Processing

With the theoretical understanding that weak agnostic learning suffices to obtain multi-calibrated predictors, we turn to a practical question of how to use standard machine learning techniques to improve the subpopulation calibration of predictive models. We will focus on the problem of auditing, electing to focus on the simpler notion of multi-accuracy. As we'll see, multi-accuracy auditing and post-processing can still provide some strong theoretical guarantees and appears to be quite effective in experiments.

We focus on a setting that is common in practice but distinct from much of the other literature on fairness in classification. Suppose we are given black-box access to a classifier, p_0, and a relatively small "validation set" of labeled samples drawn from the representative distribution $\mathcal{D}$; our goal is to audit p_0 to determine whether the predictor satisfies multi-accuracy using off-the-shelf machine learning techniques. If auditing reveals that the predictor does not satisfy multi-accuracy, we will apply the framework established in Chapter 3 to post-process p_0 to produce a new model p that is multi-accurate. Importantly, we show that with some care, we can post-process the model without adversely affecting the predictions on subpopulations where p_0 was already accurate.

Even if the initial classifier p_0 was trained in good faith, it may still exhibit biases on significant subpopulations when evaluated on samples from $\mathcal{D}$. This setting can arise when minority populations are underrepresented in the distribution used to train p_0 compared to the desired distribution $\mathcal{D}$, as in the Gender Shades study [BG18]. In general, we make no assumptions about how f_0 was trained or implemented. In particular, f_0 may be an adversarially-chosen classifier, which explicitly aims to give erroneous predictions within some protected subpopulation while satisfying marginal statistical notions of fairness. Thus, the post-processing strategy we present can be viewed as a form of black-box transfer learning—leveraging the initial performance on a potentially biased distribution to efficiently obtain a model that performs well across subpopulations on the distribution of interest.

4.2.1 Do-No-Harm Post-Processing for Multi-Accuracy

In this section, we describe Multi-Accuracy-Boost (Algorithm 7) for post-processing a pre-trained model to achieve multi-accuracy. The algorithm follows the same framework as Algorithm 1 for learning a multi-calibrated predictor. With an eye for practical applications of the algorithmic framework, we leverage the connection to learning from Section 4.1 and describe Algorithm 7 in terms of an auditor instead of a search over a class $\mathcal{C}$. Multi-Accuracy-Boost is given black-box access to an initial hypothesis $p_0 : \mathcal{X} \to [0,1]$ and uses a learning algorithm $\mathcal{L}$ to audit the model for violations to the multi-accuracy condition. When the auditor finds violations, we can use the hypothesis returned to update the predictions, post-processing until there are no significant violations. Note that we elect to use the multiplicative weights framework, which tracks more closely to cross-entropy loss minimization used in practice for binary classification. Indeed, it is possible to analyze the convergence properties of the algorithm in terms of the expected cross-entropy loss (or KL-divergence from p^*) rather than in terms of the regression loss (squared error) as in Chapter 3. The original work introducing Multi-Accuracy-Boost [KGZ19] includes a complete convergence analysis.

Residual functions. Algorithm 7 also takes a "residual function" as input. If we apply the reduction from Section 4.1 with an agnostic learner for $\mathcal{C}$, then it makes sense to use the residual labeling function from that reduction.

$$\ell(p, x, y) = p(x) - y$$

Intuitively, if we are able to learn a close approximation h to this absolute residual function, then taking steps along the residual will bring us closer to y.

$$\exp(-h(x)) \cdot p(x) \approx p(x) - h(x) \approx y + p(x) - p(x) = y.$$

In practice, instead of labeling with the absolute residual directly, we could also label our examples with a different residual function. For instance, for the classification setting, a natural choice of residual would be the partial derivative function of the cross-entropy loss with respect to the predictions

$$\begin{aligned} h(p, x, y) &= y \cdot \log(p(x)) + (1 - y) \cdot \log(1 - p(x)) \\ \ell(p, x, y) &= \frac{\partial h(p, x, y)}{\partial p(x)} = \frac{1}{1 - p(x) - y(x)}. \end{aligned}$$

which grows rapidly in magnitude as the absolute residual $|p(x) - y(x)|$ grows towards 1. Running Algorithm 7 with this gradient-based residual function is similar in spirit to gradient boosting techniques [MBBF00, Fri01], which interpret boosting algorithms as running gradient descent on an appropriate cost-functional.

Do-No-Harm. Importantly, we also adapt the framework so that Algorithm 7 exhibits what we call the "do-no-harm" guarantee; informally, if p_0 has low classification error on some subpopulation $S \subseteq \mathcal{X}$ identifiable by $\mathcal{L}$, then the resulting classification error on S cannot increase significantly. To achieve this guarantee, Algorithm 7 starts by partitioning the input space $\mathcal{X}$ based on the initial classifier p_0 into $\mathcal{X}^{(0)} = \{x \in \mathcal{X} : p_0(x) < 1/2\}$ and

Algorithm 7 Multi-Accuracy-Boost

```
Given:
  $p_0 : \mathcal{X} \to \{0,1\}$                                          // initial classifier
  $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^m \to \{h : \mathcal{X} \to [-1,1]\}$          // learning algorithm as auditor
  $\ell : \{p : \mathcal{X} \to [0,1]\} \times \mathcal{X} \times \mathcal{Y} \to [-1,1]$              // residual function
  $\tau = \alpha\gamma > 0$                                        // absolute approximation parameter
  $\eta > 0$                                                             // step size

Initialize:
  $\mathcal{X}^{(0)} \leftarrow \{x \in \mathcal{X} : p_0(x) < 1/2\}$           // Partition by initial classification
  $\mathcal{X}^{(1)} \leftarrow \{x \in \mathcal{X} : p_0(x) \geq 1/2\}$

Repeat: for $t = 0, 1, \ldots$
  $Z_t \leftarrow \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$                   // refresh data
  $h \leftarrow \mathcal{L}\left(\{\ell(p_t, x_i, y_i) : i \in [m]\}\right)$                  // learn residuals
  $h^{(0)} \leftarrow \mathcal{L}\left(\left\{\mathbf{1}\left[\ x_i \in \mathcal{X}^{(0)}\ \right] \cdot \ell(p_t, x_i, y_i) : i \in [m]\right\}\right)$
  $h^{(1)} \leftarrow \mathcal{L}\left(\left\{\mathbf{1}\left[\ x_i \in \mathcal{X}^{(1)}\ \right] \cdot \ell(p_t, x_i, y_i) : i \in [m]\right\}\right)$
  $h^{\max} \leftarrow \mathrm{argmax}_{h' \in \{h, h^{(0)}, h^{(1)}\}} \{\langle h', p_t - y \rangle\}$      // max correlation with residual
  if $\langle h^{\max}, p_t - y \rangle < \tau$ then
    return $\tilde{p} = p_t$                            // return when max residual is small
  end if
  $p_{t+1}(x) \propto \exp(-\eta h^{\max}) \cdot p_t(x)$                        // update and continue
```

$\mathcal{X}^{(1)} = \{x \in \mathcal{X} : p_0(x) \geq 1/2\}$ Partitioning the search space $\mathcal{X}$ based on the predictions of p_0 helps to ensure that the $\tilde{p}$ we output maintains the initial accuracy of p_0. Intuitively, the initial hypothesis may make false positive predictions and false negative predictions for very different reasons, even if in both cases the reason is simple enough to be identified by the auditor. More technically, the partition allows us to search over just the initially-positive-labeled examples (negative, respectively) for a way to improve the classifier; these subpopulations (and their intersections with $\mathcal{C}$ may be significantly more complex that $\mathcal{C}$ itself. A similar strategy was explored theoretically in the context of multi-calibration in [HKRR18].[2]

[2] The original work on multi-calibration referred to an analogous property as "best-in-class" predictions. In the context of multi-calibration for regression, [HKRR18] showed that partitioning the input space based on the level-sets of p_0, then performing multi-calibration preserved the squared error of the original predictor. To maintain classification accuracy, it suffices to partition the space based on the initial binary label.

Proposition 4.2 (Do-No-Harm)**.** *For a predictor $p_0 : \mathcal{X} \to [0,1]$ and a collection of subpopulations $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$, define an augmented collection $\mathcal{C}^{p_0}$ as follows.*

$$\mathcal{X}^{(0)} = \{x \in \mathcal{X} : p_0(x) < 1/2\} \qquad \mathcal{X}^{(1)} = \{x \in \mathcal{X} : p_0(x) \geq 1/2\}$$
$$\mathcal{C}^{p_0} = \mathcal{C} \cup \left\{S \cap \mathcal{X}^{(0)} : S \in \mathcal{C}\right\} \cup \left\{S \cap \mathcal{X}^{(1)} : S \in \mathcal{C}\right\}$$

Suppose $\tilde{p}$ is $(\mathcal{C}^{p_0}, \alpha)$-multi-accurate. Then, for any $S \in \mathcal{C}$,

$$\mathrm{er}_S(f^{\tilde{p}}) \leq 3 \cdot \mathrm{er}_S(f^{p_0}) + 2\alpha$$

where $f^{\tilde{p}}, f^{p_0}$ are the rounded classifiers of $\tilde{p}$ and p_0, respectively.

Proof. Proposition 4.2 is a direct corollary of Proposition 2.10. Note that we can rewrite the classification error over S as an average of the false positive and false negative rates on S. Let $\varepsilon = \mathrm{er}_S(f^{p_0})$, $\gamma^{(0)} = \mathbf{Pr}_{\mathcal{D}_S}[\ p_0(x) < 1/2\ ]$, $\gamma^{(1)} = \mathbf{Pr}_{\mathcal{D}_S}[\ p_0(x) \geq 1/2\ ]$, and $\varepsilon_{\mathrm{FN}}, \varepsilon_{\mathrm{FP}}$ represent the false error rates; then

$$\varepsilon = \gamma^{(0)} \cdot \varepsilon_{\mathrm{FN}} + \gamma^{(1)} \cdot \varepsilon_{\mathrm{FP}}.$$

Consider the subsets of S defined by the initial predictor.

$$S^{(0)} = \{x \in S : p_0(x) < 1/2\} \qquad S^{(1)} = \{x \in S : p_0(x) \geq 1/2\}$$

Both of these sets are contained in $\mathcal{C}^{p_0}$. Thus, for any $(\mathcal{C}^{p_0}, \alpha)$-multi-accurate predictor we can apply Proposition 2.10 to bound the resulting classification error on each set by

$$\begin{aligned}\mathrm{er}_S(f^{\tilde{p}}) &\leq p_0 \cdot (3\varepsilon_{\mathrm{FN}} + 2\alpha) + p_1 \cdot (3\varepsilon_{\mathrm{FP}} + 2\alpha)\\ &\leq 3\varepsilon + 2\alpha.\end{aligned}$$

Applying this proposition to the output of Algorithm 7 establishes the Do-No-Harm property. □

4.3 Empirical Study: Revisiting "Gender Shades"

We evaluate the empirical performance of Multi-Accuracy-Boost in three case studies. The first and most in-depth case study aims to emulate the conditions of the Gender Shades study [BG18], to test the effectiveness of multi-accuracy auditing and post-processing on this important real-world example. In Section 4.3.1, we show experimental results for auditing using two different validation data sets. In particular, one data set is fairly unbalanced and similar to the data used to train, while the other data set was developed in the Gender Shades study and is very balanced. For each experiment, we report for various subpopulations, the population percentage in $\mathcal{D}$, accuracies of the initial model, our black-box post-processed model, and white-box benchmarks.

4.3.1 Multi-accuracy improves gender detection

In this case study, we replicate the conditions of the Gender Shades study [BG18] to evaluate the effectiveness of the multi-accuracy framework in a realistic setting. For our initial model, we train an Inception-ResNet-v1 [SIVA17] gender classification model using the CelebA data set with more than 200,000 face images [LLWT15]. The resulting test accuracy on CelebA for binary gender classification is 98.4%.

We applied Multi-Accuracy-Boost to this p_0 using two different auditing distributions. In the first case, we audit using data from the LFW+a[3] set [WHT11, HRBLM07], which has similar demographic breakdowns as CelebA (i.e. $\mathcal{D} \approx \mathcal{D}_0$). In the second case, we audit using the PPB data set (developed in [BG18]) which has balanced representation across gender and race (i.e. $\mathcal{D} \neq \mathcal{D}_0$). These experiments allows us to track the effectiveness of Multi-Accuracy-Boost as the representation of minority subpopulations changes. In both cases, the auditor is "blind"—it is not explicitly given the race or gender of any individual—and knows nothing about the inner workings of the classifier. Specifically, we take the auditor to perform ridge regression to fit the cross-entropy gradient.[4] Instead of training the auditor on raw input pixels, we use the low dimensional representation of the input images derived by a variational autoencoder (VAE) trained on CelebA dataset using Facenet [SKP15] library.

To test the initial performance of p_0, we evaluated on a random subset of the LFW+a

[3] We fixed the original data set's label noise for gender and race.

[4] To help avoid outliers, we smooth the loss and use a quadratic approximation for points with very large residual.

	All	**F**	**M**	**B**	**N**	**BF**	**BM**	**NF**	**NM**
$\mathcal{D}$	100	21.0	79.0	4.9	95.1	2.1	2.7	18.8	76.3
p_0	5.4	23.1	0.7	10.2	5.1	20.4	2.1	23.4	0.6
MA	4.1	11.3	3.2	6.0	4.9	8.2	4.3	11.7	3.2
RT	3.8	11.2	1.9	7.5	3.7	11.6	4.3	11.1	1.8

Table 4.1: **Results for LFW+a gender classification.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; p_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multi-Accuracy-Boost; RT denotes the classification error (%) of the model after retraining on $\mathcal{D}$.

data containing 6,880 face images, each of which is labeled with both gender and race—black (**B**) and non-black (**N**). For gender classification on LFW+a, p_0 achieves 94.4% accuracy. Even though the overall accuracy is high, the error rate is much worse for females (23.1%) compared to males (0.7%) and worse for blacks (10.2%) compared to non-blacks (5.1 %); these results are qualitatively very similar to those observed by the commercial gender detection systems studied in [BG18]. We applied Multi-Accuracy-Boost, which converged in 7 iterations. The resulting classifier's classification error in minority subpopulations was substantially reduced, even though the auditing distribution was similar as the training distribution.

We compare Multi-Accuracy-Boost against a strong white-box baseline. Here, we retrain the network of p_0 using the audit set. Specifically, we retrain the last two layers of the network, which gives the best results amongst retraining methods. We emphasize that this baseline requires white-box access to p_0, which is often infeasible in practice. Multi-Accuracy-Boost accesses p_0 only as a black-box without any additional demographic information, and still achieves comparable, if not improved, error rates compared to retraining. We report the overall classification accuracy as well as accuracy on different subpopulations—e.g. **BF** indicates black female—in Table 4.1.

The second face dataset, PPB, in addition to being more balanced, is much smaller; thus, this experiment can be viewed as a stress test, evaluating the data efficiency of our post-processing technique. The test set has 415 individuals and the audit set has size 855. PPB annotates each face as dark (**D**) or light-skinned (**L**). As with LFW+a, we evaluated the test accuracy of the original p_0, the multi-accurate post-processed classifier, and retrained classifier on each subgroup. Multi-Accuracy-Boost converged in 5 iterations and again, substantially reduced error despite a small audit set and the lack of annotation about race or skin color (Table 4.2).

	All	**F**	**M**	**D**	**L**	**DF**	**DM**	**LF**	**LM**
$\mathcal{D}$	100	44.6	55.4	46.4	53.6	21.4	25.0	23.2	30.4
p_0	9.9	21.6	0.4	18.8	2.2	39.8	1.0	5.2	0.0
MA	3.9	6.5	1.8	7.3	0.9	12.5	2.9	1.0	0.8
RT	2.2	3.8	0.9	4.2	0.4	6.8	1.9	1.0	0.0

Table 4.2: **Results for the PPB gender classification data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; p_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multi-Accuracy-Boost; RT denotes the classification error (%) of the model after retraining on $\mathcal{D}$.

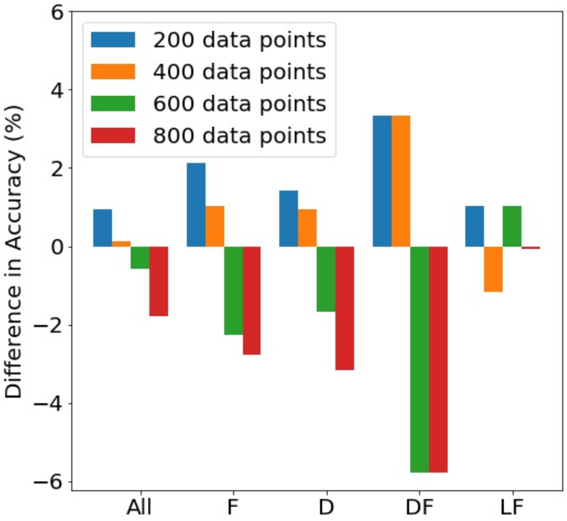

Figure 4.1: **Multi-accuracy vs. Retraining**: Difference in classification accuracy is plotted on the vertical axis; this difference represents the accuracy advantage of Multi-Accuracy-Boost compared to retraining. As the size of the audit set shrinks, Multi-Accuracy-Boost has better performance both in overall accuracy and accuracy of the subgroups with the most initial bias because it is more data efficient.

To further test the data efficiency of Multi-Accuracy-Boost, we evaluate the effect of audit set size on the resulting accuracy of each method. In Fig. 4.1, we report the performance of Multi-Accuracy-Boost versus the white-box retraining method for different sizes of audit set. The plot displays the difference in accuracy for the overall population along with the subgroups that suffered the most initial bias. It shows that the performance of Multi-Accuracy-Boost may actually be better than the white-box retraining baseline when validation data is especially scarce.

Representation matters. As discussed earlier, in the reported gender detection experiments, we audit for multi-accuracy using ridge regression over an encoding of images produced by a variational autoencoder. Using the representation of images produced by this encoding intuitively makes sense, as the autoencoder's reconstruction objective aims to preserve as much information about the image as possible while reducing the dimension. Still, we may wonder whether multi-accuracy auditing over a different representation of the images would perform better. In particular, since we are interested in improving the accuracy on the gender detection task, it seems plausible that a representation of the images based on the internal layers of the initial prediction network might preserve the information salient to gender detection more effectively.

We investigate the importance of the representation used to audit empirically. In particular, we also evaluate the performance of Multi-Accuracy-Boost using the same auditor run over two other sets of features, given by the last-layer and the second-to-last layer of the initial prediction residual network p_0. In Table 4.3, we show that using the unsupervised VAE representation yields the best results. Still, the representations from the last layers are competitive with that of the VAE, and in some subpopulations are better.

Collectively, these findings bolster the argument for "fairness through awareness," which advocates that in order to make fair predictions, sensitive information (like race or gender) should be given to the (trustworthy) classifier. While none of these representations explicitly encode sensitive group information, the VAE representation does preserve information about the original input, for instance skin color, that seems useful in understanding the group status. The prediction network is trained to have the best prediction accuracy (on an unbalanced training data set), and thus, the representations from the network reasonably may contain less information about these sensitive features. These results suggest that the effectiveness of multi-accuracy does depend on the representation of inputs used for auditing, but so long as the representation is sufficiently expressive, Multi-Accuracy-Boost may be robust to the exact encoding of the features.

Multi-accuracy auditing as diagnostic. As was shown in [BG18], models trained in good faith on unbalanced data may exhibit significant biases on the minority populations. For instance, the initial classification error on black females is significant, whereas on white males, it is near 0. Importantly, the only way we were able to report these accuracy disparities was by having access to a rich data set where gender and race were labeled.

	All	**F**	**M**	**D**	**L**	**DF**	**DM**	**LF**	**LM**
LFW+a:									
VAE	4.1	11.3	3.2	6.0	4.9	8.2	4.3	11.7	3.2
R_{1,p_0}	4.9	13.6	2.6	6.3	4.9	8.8	4.3	14.1	2.6
R_{2,p_0}	4.5	12.6	2.4	6.3	4.4	8.8	4.3	13.1	2.3
PPB:									
VAE	3.9	6.5	1.8	7.3	0.9	12.5	2.9	1.0	0.8
R_{1,p_0}	4.3	7.6	1.7	7.8	1.3	13.6	2.9	2.1	0.8
R_{2,p_0}	5.1	9.7	1.3	9.4	1.3	17.0	2.9	3.1	0.0

Table 4.3: **Effect of representation on the Multi-Accuracy-Boost performance** VAE denotes the denotes the classification error (%) using the VAE representation; R_{1,p_0} denotes the classification error (%) using the classifier's last layer representation, R_{2,p_0} denotes the classification error (%) using the classifier's second to last layer representation

Often, this demographic information will not be available; indeed, the CelebA images are not labeled with race information, and as such, we were unable to evaluate the subpopulation classification accuracy on this set. Thus, practitioners may be faced with a problem: even if they know their model is making undesirable mistakes, it may not be clear if these mistakes are concentrated on specific subpopulations. Without understanding the demographics on which the model is erring, collecting additional (biased) training data may not actually improve performance across the board.

We demonstrate that multi-accuracy auditing may serve as an effective diagnostic and interpretation tool to help developers identify systematic biases in their models. The idea is simple: the auditor returns a hypothesis h that essentially "scores" individual inputs x by how wrong the prediction $p_0(x)$ is. If we consider the magnitude of their scores $|h(x)|$, then we may understand better the biases that the encoder is discovering.

As an example, we test this idea on the PPB data set, evaluating the test images' representations with the hypotheses the auditor returns. In Figure 4.2, we display the images in the test set that get the highest and lowest effect ($|h(x)|$ large and $|h(x)| \approx 0$, respectively) according to the hypothesis returned by the auditor during the first round of auditing. The three highest-scoring images (top row) are all women, both black and white. Interestingly, all of the least active images (bottom row) are men in suits, suggesting that suits may be a highly predictive feature of being a man according to the original classifier, p_0. In this sense, multi-accuracy auditing may provide a useful exploratory tool for developers to interact with the data and their trained models to better understand where (and why) the model may be underperforming.

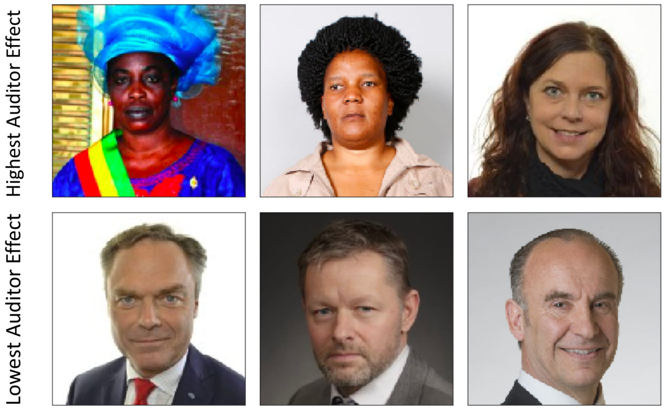

Figure 4.2: **Interpreting Auditors** Here, we depict the PPB test images with the highest and lowest activation according to the hypothesis trained in the first round of auditing. The images with the highest auditor effects corresponds to images where the auditor detects the largest bias in the classifier.

4.3.2 Additional case studies

Multi-accuracy auditing and post-processing is applicable broadly in supervised learning tasks, not just in image classification applications. We demonstrate the effectiveness of Multi-Accuracy-Boost in two other settings: the adult income prediction task and a semi-synthetic disease prediction task.

Adult Income Prediction For the first case study, we utilize the adult income prediction data set [Koh96] with 45,222 samples and 14 attributes (after removing subjects with unknown attributes) for the task of binary prediction of income more than $50k for the two major groups of Black and White. We remove the sensitive features of gender—female (**F**) and male (**M**) and race (for the two major groups)—black (**B**) and white (**W**)—from the data, to simulate settings where sensitive features are not available to the algorithm training. We trained a base algorithm, p_0, which is a neural network with two hidden layers on 27,145 randomly selected individuals. The test set consists of an independent set of 15,060 persons.

We audit using a decision tree regression model (max depth 5) $\mathcal{L}_{\mathrm{dt}}$ to fit the residual $p(x) - y(x)$. $\mathcal{L}_{\mathrm{dt}}$ receives samples of validation data drawn from the same distribution as training; that is $\mathcal{D} = \mathcal{D}_0$. In particular, we post-process with 3,017 individuals sampled from the same adult income dataset (disjoint from the training set of p_0). The auditor is given the same features as the original prediction model, and thus, is not given the gender

	All	**F**	**M**	**B**	**W**	**BF**	**BM**	**WF**	**WM**
$\mathcal{D}$	100.0	32.3	67.7	90.3	9.7	4.8	4.9	27.4	62.9
p_0	19.3	9.3	24.2	10.5	20.3	4.8	15.8	9.8	24.9
MA	14.7	7.2	18.3	9.4	15.0	4.5	13.9	7.3	18.3
SS	19.7	9.5	24.6	10.5	19.9	5.5	15.3	10.2	25.3

Table 4.4: **Results for Adult Income Data Set** $\mathcal{D}$ denotes the percentages of each population in the data distribution; p_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multi-Accuracy-Boost; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.

or race of any individual. We evaluate the post-processed classifier on the same independent test set. Multi-Accuracy-Boost converges in 50 iterations with $\eta = 1$.

As a baseline, we trained four separate neural networks with the same architecture as before (two hidden layers) for each of the four subgroups using the audit data. As shown in Table 4.4, multi-accuracy post-processing achieves better accuracy both in aggregate and for each of the subgroups. Importantly, the subgroup-specific models requires explicit access to the sensitive features of gender and race. Training a classifier for each subgroup, or explicitly adding subgroup accuracy into the training objective, assumes that the subgroup is already identified in the data. This is not feasible in the many applications where, say, race or more granular categories are not given. Even when the subgroups are identified, we often do not have enough samples to train accurate classifiers on each subgroup separately. This example illustrates that multi-accuracy can help to boost the overall accuracy of a black-box predictor in a data efficient manner.

Semi-Synthetic Disease Prediction We design a disease prediction task based on real individuals, where the phenotype to disease relation is designed to be different for different subgroups, in order to simulate a challenging setting. We used 40,000 individuals sampled from the UK Biobank [SGA$^+$15]. Each individual contains 60 phenotype features. To generate a synthetic disease outcome for each subgroup, we divided the data set into four groups based on gender—male (**M**) and female (**F**)—and age—young (**Y**) and old (**O**). For each subgroup, we create synthetic binary labels using a different polynomial function of the input features with different levels of difficulty. The polynomial function orders are 1, 4, 2, and 6 for OF, OM, YF, and YM subgroups respectively.

For p_0, we trained a neural network with two hidden layers on 32,000 individuals, without

	All	**F**	**M**	**O**	**Y**	**OF**	**OM**	**YF**	**YM**
$\mathcal{D}$	100	39.6	60.4	34.6	65.4	15.0	19.7	24.6	40.7
p_0	18.9	29.4	12.2	21.9	17.3	36.8	10.9	24.9	12.8
MA	16.0	24.1	10.7	16.4	15.7	26.5	9.0	22.7	11.6
SS	19.5	32.4	11.0	22.1	18.1	37.6	10.3	29.3	11.3

Table 4.5: **Results for UK Biobank semi-synthetic data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; p_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with Multi-Accuracy-Boost; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.

using the gender and age features. Hyperparameter search was done for the best weight-decay and drop-out parameters. The p_0 we discover performs moderately well on every subpopulation, with the exception of old females (**OF**) where the classification error is significantly higher. Note that this subpopulation had the least representation in $\mathcal{D}_0$. Again, we audit using $\mathcal{L}_{\text{dt}}$ to run decision tree regression with validation data samples drawn from $\mathcal{D} = \mathcal{D}_0$. Specifically, the auditor receives a sample of 4,000 individuals without the gender or age features. As a baseline, we trained a separate classifier for each of the subgroups using the same audit data. As Table 4.5 shows, Multi-Accuracy-Boost significantly lowers the classification error in the old female population.

Chapter Notes

This chapter is based off of results originally reported in [HKRR18, KGZ19]. The equivalence between multi-calibration auditing and weak agnostic learning was first discovered by [HKRR18]. Inspired by the theoretical foundations developed in [HKRR18], the multi-accuracy auditing and post-processing framework was developed in joint follow-up work with Amirata Ghorbani and James Zou [KGZ19]. Related concurrent works [KNRW18, KNRW19] study multi-group parity notions of fairness, also connecting the problem of auditing to weak agnostic learning (both in theory and in experiments). Our empirical investigations on muti-accuracy auditing can also be viewed as studying information-fairness tradeoffs in prediction tasks, and is particularly related to the work on "fair representations" [ZWS$^+$13, ES15, MCPZ18, CMJ$^+$19].

Subsequent to our work, post-processing for multi-accuracy has been explored and applied in clinical settings. [BDR$^+$19] studied how existing healthcare risk assessments could be post-processed to ensure calibration across minority subpopulations within the Clalit healthcare system. The multi-accuracy post-processing framework was also used in recent efforts to develop algorithmic predictions for COVID-19 risk in Israel [BRA$^+$20]. With limited data about the risk factors for COVID, the Israeli researchers post-processed a sophisticated flu predictor developed by Clalit to better match the marginal statistics available about COVID.

A different approach to subgroup fairness in classification is studied by [DIKL17]. This work investigates the question of how to learn a "decoupled" classifier, where separate classifiers are learned for each subgroup and then combined to achieve a desired notion of fairness. Decoupling the classification problem requires that important subgroups are identified in the features and that the groups we wish to protect are partitioned by these attributes. Even if this information is available, it may not always be obvious which subpopulations require special attention.

Part III

Fairness through Computationally-Bounded Awareness

Chapter 5

Multi-Group Fairness Beyond Calibration

Multi-calibration takes the approach that the starting point for fair prediction should be to represent the true underlying risk as accurately as possible within a computational bound. Sometimes, however, we may be concerned that fairness and accuracy are at odds with one another. For instance, if the data themselves contain historical biases, then learning a predictor that accurately reflects these data may further historical discriminations. In this chapter, we turn our attention to notions of multi-group fairness that are not directly tied towards accurately representing p^*. The bulk of the chapter is dedicated towards understanding a multi-group relaxation of the individual metric fairness notion of [DHP+12]. We conclude by discussing some other notions of multi-group fairness and their relation to multi-calibration. In Chapter 6, we discuss how multi-calibration can be a useful tool for understanding and addressing discrimination in prediction, even when the fairness goal is not based on accuracy with p^*.

5.1 Metric Fairness Beyond Individuals

In the influential work "Fairness through Awareness," [DHP+12] proposed a framework to resolve the apparent conflict between utility and fairness. This framework takes the perspective that a fair classifier should treat similar individuals similarly and formalizes this abstract goal by assuming access to a task-specific similarity metric d on pairs of individuals. The proposed notion of fairness requires that if the distance between two individuals is small, then the predictions of a fair classifier cannot be very different. More formally, for some small constant $\tau \geq 0$, we say a hypothesis $f : \mathcal{X} \to [-1, 1]$ satisfies (d, τ)-metric fairness[1] if the following (approximate) Lipschitz condition holds for all pairs of individuals from the population $\mathcal{X}$.

$$\forall x, x' \in \mathcal{X} \times \mathcal{X} : \quad \left|f(x) - f(x')\right| \leq d(x, x') + \tau \tag{5.1}$$

Subject to these intuitive similarity constraints, the classifier may be chosen to maximize utility. The definition's conceptual simplicity and modularity make fairness through awareness a very appealing framework. Currently, there are many (sometimes contradictory) notions of what it means for a classifier to be fair [KMR17, CG18, FPCG16, HPS16, HKRR18], and there is much debate on which definitions should be applied in a given context. Basing fairness on a similarity metric offers a flexible approach for formalizing a variety of guarantees and protections from discrimination.

The issue with applying this notion in machine learning applications is that requiring individual fairness to hold for all pairs of individuals may be information-theoretically prohibitive. When the universe of individuals is very large, even writing down an appropriate metric could be completely infeasible. In these cases, rather than require the metric value to be specified for all pairs of individuals, we could instead ask a panel of experts to provide similarity scores for a small sample of pairs of individuals. While we can't enforce individual metric fairness from a small sample, inspired by the approach of multi-calibration, we show that we can enforce a multi-group metric fairness guarantee. When a metric-based fairness notion is applicable, but similarity scores are hard to come by, multi-group metric fairness provides a strong, provable notion of fairness that maintains the theoretical appeal and practical modularity of the fairness through awareness framework. As in [DHP$^+$12], we investigate how to learn a classifier that achieves optimal utility under similarity-based fairness constraints, assuming a weaker model of limited access to the metric.

As in multi-calibration, we define our relaxation of metric fairness with respect to a rich class of statistical tests on the pairs of individuals. Let a comparison be any subset of the pairs of $S \subseteq \mathcal{X} \times \mathcal{X}$. The definition is parameterized by a collection of comparisons $\mathcal{C} \subseteq \{0, 1\}^{\mathcal{X} \times \mathcal{X}}$ and requires that a hypothesis appear Lipschitz on average according to all of the statistical tests defined by the comparisons $S \in \mathcal{C}$.

[1] Note the definition given in [DHP$^+$12] is slightly different; in particular, they propose a more general Lipschitz condition, but fix $\tau = 0$.

Definition 5.1 (Multi-metric fairness). *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}}$ be a collection of comparisons and let $d : \mathcal{X} \times \mathcal{X} \to [0,2]$ be a metric. For some constants $\tau \geq 0$, a hypothesis f is* $(\mathcal{C}, d, \tau)$-multi-metric fair *if for all $S \in \mathcal{C}$,*

$$\mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[\left|f(x) - f(x')\right|\right] \leq \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[d(x,x')\right] + \tau. \tag{5.2}$$

To begin, note that multi-metric fairness is indeed a relaxation of metric fairness; if we take the collection $\mathcal{C} = \{\{(x,x')\} : x,x' \in \mathcal{X} \times \mathcal{X}\}$ to be the collection of all pairwise comparisons, then $(\mathcal{C}, d, \tau)$-multi-metric fairness is equivalent to (d,τ)-metric fairness. When we want to learn multi-metric fair predictors from a small sample of data, we will apply the multi-group perspective, taking $\mathcal{C}$ to be an expressive class of large intersecting comparisons between subpopulations. Still, when we take $\mathcal{C}$ to be sufficiently expressive—enough to identify comparisons between the most similar subpopulations—we may recover many of the guarantees of the original individual fairness notion.

Proposition 5.2. *Suppose there is some $S \in \mathcal{C}$, such that $\mathbf{E}_{(x,x')\sim\mathcal{D}_S}[d(x,x')] \leq \varepsilon$. Then if f is $(\mathcal{C}, d, \tau)$-multi-metric fair, then f satisfies $(d, (\varepsilon + \tau)/p)$-metric fairness for at least a $(1-p)$-fraction of the pairs in S.*

That is, if there is some subset $S \in \mathcal{C}$ that identifies a set of pairs whose metric distance is small, then any multi-metric fair hypothesis must also satisfy the stronger individual metric fairness notion on many pairs from S. This effect will compound if many different (possibly overlapping) comparisons are identified that have small average distance. If the class $\mathcal{C}$ is rich enough to correlate well with various comparisons that reveal significant information about the metric, then any multi-metric fair hypothesis will satisfy individual-level metric fairness on a significant fraction of the population. In this sense, multi-metric fairness ensures that a hypothesis will be fair on all comparisons identifiable within the computational bound specified by $\mathcal{C}$, providing "fairness through computationally-bounded awareness."

5.2 Learning a Multi-Metric Fair Hypothesis

Throughout, our goal is to learn a hypothesis from noisy samples from the metric that satisfies multi-metric fairness. Specifically, we assume an algorithm can obtain a small number of independent random metric samples $(x, x', \Delta(x,x')) \in \mathcal{X} \times \mathcal{X} \times [0,2]$ where

$(x, x') \sim \mathcal{M}$ is drawn at random over the distribution of pairs of individuals, and $\Delta(x, x')$ is a random variable of bounded magnitude with $\mathbf{E}[\Delta(x, x')] = d(x, x')$.

When we learn linear families, our goal will be to learn from a sample of labeled examples. We assume the algorithm can ask for independent random samples $x, y \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. To evaluate the utility guarantees of our learned predictions, we take a comparative approach. Suppose $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}}$ is a collection of comparisons. For $\varepsilon \geq 0$, we say a hypothesis f is $(\mathcal{H}, \varepsilon)$-*optimal* with respect to $\mathcal{F}$, if

$$\mathop{\mathbf{E}}_{x,y\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}[L(f(x), y)] \leq \mathop{\mathbf{E}}_{x,y\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}[L(f^*(x), y)] + \varepsilon \tag{5.3}$$

where $f^* \in \mathcal{F}$ is an optimal $(\mathcal{H}, d, 0)$-multi-metric fair hypothesis.

A convex program. As in [DHP$^+$12], we formulate the problem of learning a fair set of predictions as a convex program. Our objective is to minimize the expected loss $\mathbf{E}_{x,y\sim\mathcal{D}}[L(f(x), y)]$, subject to the multi-metric fairness constraints defined by $\mathcal{C}$.[2] Specifically, we show that a simple variant of stochastic gradient descent due to [Nes09] learns such linear families efficiently.

Theorem 5.3. *Suppose $\gamma, \tau, \delta > 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}}$ is γ-large. With probability at least $1 - \delta$, stochastic switching subgradient descent learns a hypothesis $w \in \mathcal{F}$ that is $(\mathcal{C}, d, \tau)$-multi-metric fair and $(\mathcal{C}, O(\tau))$-optimal with respect to $\mathcal{F}$ in $O\left(\frac{B^2 n^2 \log(n/\delta)}{\tau^2}\right)$ iterations from $m = \tilde{O}\left(\frac{\log(|\mathcal{C}|/\delta)}{\gamma\tau^2}\right)$ metric samples.*

We give a description of the switching subgradient method in Algorithm 8. Intuitively, the algorithm follows the same framework as Algorithm 1 for multi-calibration, but works in a constrained optimization framework. In each iteration, the procedure checks to see if any constraint defined by $\mathcal{C}$ is significantly violated. If it finds a violation, it takes a (stochastic) step towards feasibility. Otherwise, it steps according a stochastic subgradient for the objective.

For convenience of analysis, we define the residual on the constraint defined by S as follows.

$$R_S(w) = \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[\left|f_w(x) - f_w(x')\right|\right] - \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[d(x, x')\right] \tag{5.4}$$

[2] For the sake of presentation, throughout the theorem statements, we will assume that L is $O(1)$-Lipschitz on the domain of legal predictions/labels to guarantee bounded error; our results are proved more generally.

Algorithm 8 Switching-Multi-Metric-Descent

```
Given:
  $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}}$                // Collection of comparisons
  $\tau > 0$                                  // approximation threshold
  $T \in \mathbb{N}$                                  // number of iterations
  $\hat{R}_S$                                     // residual oracle

Initialize:
  $w_0 \in \mathcal{F} = [-B, B]^n$                     // initial parameters
  $W = \varnothing$                                  // feasible iterates

Repeat: for $t = 0, 1, \ldots, T$
  if $\exists S \in \mathcal{C}$ such that $\hat{R}_S(w_k) > 4\tau/5$ then
    $S_k \leftarrow$ any $S \in \mathcal{C}$ such that $\hat{R}_S(w_k) > 4\tau/5$      // some constraint violated
    $w_{k+1} \leftarrow w_k - \frac{\tau}{M^2}\nabla R_{S_k}(w_k)$              // step according to constraint
  else
    $W \leftarrow W \cup \{w_k\}$                      // update set of feasible iterates
    $w_{k+1} \leftarrow w_k - \frac{\tau}{GM}\nabla L(w_k)$                // step according to objective
  end if
return $\bar{w} = \frac{1}{|W|} \cdot \sum_{w\in W} w$              // output average of feasible iterates
```

Note that $R_S(w)$ is convex in the predctions $f_w(x)$ and thus, for linear families is convex in w. We describe the algorithm assuming access to the following estimators, which we can implement efficiently (in terms of time and samples). First, we assume we can estimate the residual $\hat{R}_S(w)$ on each $S \in \mathcal{C}$ with tolerance τ such that for all $w \in \mathcal{F}$, $\left|R_S(w) - \hat{R}_S(w)\right| \leq \tau$. Next, we assume access to a stochastic subgradient oracle for the constraints and the objective. For a function $\phi(w)$, let $\partial\phi(w)$ denote the set of subgradients of ϕ at w. We abuse notation, and let $\nabla\phi(w)$ refer to a vector-valued random variable where $\mathbf{E}[\nabla\phi(w)|w] \in \partial\phi(w)$. We assume access to stochastic subgradients for $\partial R_S(w)$ for all $S \in \mathcal{C}$ and $\partial L(w)$.

Overview of analysis. Here, we give a high-level overview of the analysis of Algorithm 8. We refer to K_f as the set of "feasible iterations" where we step according to the objective; that is,

$$K_f = \left\{k \in [T] : \hat{R}_{S_k}(w_k) \leq 4\tau/5\right\}$$

Fairness analysis. We begin by showing that the hypothesis $\bar{w}$ that Algorithm 8 returns satisfies multi-metric fairness.

Lemma 5.4. *Suppose for all $S \in \mathcal{C}$, the residual oracle $\hat{R}_S$ has tolerance $\tau/5$. Then, $\bar{w}$ is $(\mathcal{C}, d, \tau)$-multi-metric fair.*

Proof. We choose our final hypothesis $\bar{w}$ to be the weighted average of the feasible iterates. Note that the update rules for K_f and W imply that $\bar{w}$ is a convex combination of hypotheses where no constraint appears significantly violated, $\bar{w} = \frac{1}{|K_f|} \cdot \sum_{k \in K_f} w_k$. By convexity of R_S we have the following inequality for all $S \in \mathcal{C}$.

$$R_S(\bar{w}) = R_S\left(\frac{1}{|K_f|}\sum_{k \in K_f} w_k\right) \le \frac{1}{|K_f|}\sum_{k \in K_f} R_S(w_k)$$

Further, for all $S \in \mathcal{C}$ and all $k \in [T]$, by the assumed tolerance of R_S, we know that

$$\left| R_S(w_k) - \hat{R}_S(w_k) \right| \le \tau/5.$$

Given that for all $k \in K_f$, $\hat{R}_{S_k}(w_k) \le 4\tau/5$, then applying the triangle inequality, we conclude that for each comparison $S \in \mathcal{C}$,

$$\mathop{\mathbf{E}}_{(x,x') \sim \mathcal{D}_S}\left[\left|f_{\bar{w}}(x) - f_{\bar{w}}(x')\right| - d(x,x')\right] = R_S(\bar{w}) \le \tau.$$

Hence, $\bar{w}$ is $(\mathcal{C}, d, \tau)$-multi-metric fair. □

Utility and runtime analysis. We analyze the utility of Algorithm 8 using a duality argument. For notational convenience, denote $L(w) = \mathbf{E}_{x_i \sim \mathcal{D}}[L(f_w(x_i), y_i)]$. In addition to the assumptions in the main body, throughout, we assume the following bounds on the subgradients for all $w \in \mathcal{F}$.

$$\forall S \in \mathcal{C}: \; \|\nabla R_S(w)\|_\infty \le m \qquad\qquad \|\nabla L(w)\|_\infty \le g$$

Assuming an ℓ_∞ bound implies a bound on the corresponding second moments of the stochastic subgradients; specifically, we use the notation $\|\nabla R_S(w)\|_2^2 \le M^2 = m^2 n$ and $\|\nabla L(w)\|_2^2 \le G^2 = g^2 n$.

Consider the Lagrangian of the program $\mathcal{L} : \mathcal{F} \times \mathcal{R}_+^{|\mathcal{C}|} \to \mathcal{R}$.

$$\mathcal{L}(w, \lambda) = L(w) + \sum_{S \in \mathcal{C}} \lambda_S R_S(w)$$

Let $w_* \in \mathcal{F}$ be an optimal feasible hypothesis; that is, w_* is a $(\mathcal{C}, d, 0)$-multi-metric fair hypothesis such that $L(w_*) \le L(w)$ for all other $(\mathcal{C}, d, 0)$-multi-metric fair hypotheses $w \in \mathcal{F}$.[3] By its optimality and feasibility, we know that w_* achieves objective value $L(w_*) = \inf_{w \in F} \sup_{\lambda \in \mathcal{R}_+^{|\mathcal{C}|}} \mathcal{L}(w, \lambda)$. Recall, the dual objective is given as $D(\lambda) = \inf_{w \in \mathcal{F}} \mathcal{L}(w, \lambda)$. Weak duality tells us that the dual objective value is upper bounded by the primal objective value.

$$\sup_{\lambda \in \mathcal{R}_+^{|\mathcal{C}|}} D(\lambda) \le L(w_*)$$

As there is a feasible point and the convex constraints induce a polytope, Slater's condition is satisfied and strong duality holds. To analyze the utility of $\bar{w}$, we choose a setting of dual multipliers $\bar{\lambda} \in \mathcal{R}_+^{|C|}$ such that the duality gap $\gamma(w, \lambda) = L(w) - D(\lambda)$ is bounded (with high probability over the random choice of stochastic subgradients). Exhibiting such a setting of $\bar{\lambda}$ demonstrates the near optimality of $\bar{w}$.

Lemma 5.5. *Let $\tau, \delta > 0$ and $\mathcal{F} = [-B, B]^n$. After running Algorithm 8 for $T > \frac{30^2 M^2 B^2 n \log(n/\delta)}{\tau^2}$ iterations, then with probability at least $1 - 8\delta$ (over the stochastic subgradients)*

$$L(\bar{w}) \le L(w_*) + \frac{3G}{5M}\tau.$$

We give the full proof of Lemma 5.5 subsequently.

Residual queries. Next, we describe how to answer residual queries $R_S(w)$ efficiently, in terms of time and samples. We will use the following proposition which follows by Chernoff's inequality.

Proposition 5.6. *Suppose $\mathcal{C}$ is γ-large. Then with probability at least $1 - \delta$, for all $S \in \mathcal{C}$, the empirical estimate for $\mathbf{E}_{\mathcal{D}_S}[|f(x) - f(x')|]$ of n samples $(x, x') \sim \mathcal{M}$ deviates from the true expected value by at most τ provided*

$$n \ge \tilde{\Omega}\left(\frac{B^2 \log(|C| / \delta)}{\gamma \cdot \tau^2}\right).$$

[3] Such a w^* exists, as $w = 0 \in \mathcal{R}^n$ always trivially satisfies all the fairness constraints.

Lemma 5.7. *For $\tau, \gamma > 0$, for a γ-large collection of comparisons $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}}$, with probability $1-\delta$, given access to n metric samples, every residual query $R_S(w)$ can be answered correctly with tolerance τ provided*

$$n \geq \tilde{\Omega}\left(\frac{\log(|C|/\delta)}{\gamma \cdot \tau^2}\right).$$

Each residual query $R_S(w)$ can be answered after $\tilde{O}\left(\frac{\log(T\cdot|\mathcal{C}|/\delta)}{\gamma\cdot\tau^2}\right)$ evaluations of the current hypothesis.

Proof. Recall the definition of $R_S(w)$.

$$R_S(w) = \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[\left|f_w(x) - f_w(x')\right|\right] - \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}\left[d(x,x')\right]$$

Proposition 5.6 shows that $\mathbf{E}_{\mathcal{D}_S}[d(x,x')]$ can be estimated for all $S \in \mathcal{C}$ from a small number of metric samples. The proof follows a standard Chernoff plus union bound argument. Thus, Lemma 5.7 follows by showing that at each iteration $\mathbf{E}_S[|f_w(x) - f_w(x')|]$ can be estimated from a small number of evaluations of the current hypothesis f_w.

We can estimate the expected value of the deviation on f over $S \in \mathcal{C}$ with a small set of unlabeled samples from $\mathcal{X} \times \mathcal{X}$; we will evaluate the hypothesis f for each of these samples. Using an identical argument as in the case of the expected metric value, which shows the lemma. □

Subgradient queries. Next, we argue that the subgradient oracles can be implemented efficiently without accessing any metric samples. First, suppose we want to take a step according to $R_S(w)$; while $R_S(w)$ is not differentiable, we can compute a legal subgradient defined by partial subderivatives given as follows.

$$\frac{\partial R_S(w)}{\partial w_l} = \mathop{\mathbf{E}}_{(x,x')\sim\mathcal{D}_S}[\mathrm{sgn}(\langle w, x - x'\rangle)\cdot(x_l - x'_l)]$$

The subgradient does not depend on d, so no samples from the metric are necessary. Further, Algorithm 8 only assumes access to stochastic subgradient oracle with bounded entries. If we sample a single $(x_i, x_j) \sim \mathcal{M}$, then $\mathrm{sgn}(\langle w, x_i - x_j\rangle)\cdot(x_{il} - x_{jl})$ will be an unbiased estimate of a subgradient of $R_S(w)$; we claim, the entries will also be bounded. In particular, assuming $\|x_i\|_1 \leq 1$ implies each partial is bounded by 2, so that we can take $M^2 = 4n$.

Detailed utility analysis. Here, we give a full proof of Lemma 5.5. We defer the proof of certain technical lemmas for the sake of presentation.

Proof of Lemma 5.5 Let $\tau, \delta > 0$ and $\mathcal{F} = \{w \in \mathcal{R}^n : \|w\|_\infty \leq B\}$. After running Algorithm 8 for $T > \frac{30^2 M^2 B^2 n \log(n/\delta)}{\tau^2}$ iterations, then

$$L(\bar{w}) \leq L(w_*) + \frac{3G}{5M}\tau$$

with probability at least $1 - 8\delta$ over the randomness of the algorithm.

Proof. As before, we refer to $K_f \subseteq [T]$ as the set of feasible iterations, where we step according to the objective, and $[T] \setminus K_f$ as the set of infeasible iterations, where we step according to the violated constraints. Recall, we denote the set of subgradients of a function L (or R) at w by $\partial L(w)$ and denote by $\nabla L(w)$ a stochastic subgradient, where $\mathbf{E}[\nabla L(w)|w] \in \partial L(w)$.

When we do not step according to the objective, we step according to the subgradient of some violated comparison constraint. In fact, we show that stepping according to any convex combination of such subgradients suffices to guarantee progress in the duality gap. In the case wher $t \notin K_f$, we assume that we can find some convex combination $\sum_{S \in \mathcal{C}} \alpha_{k,S} \hat{R}_S(w_k) > 4\tau/5$ where for all $S \in \mathcal{C}$, $\alpha_{k,S} \in \Delta_{|\mathcal{C}|-1}$. We show that if we step according to the corresponding combination of the subgradients of $R_S(w_k)$, we can bound the duality gap. Specifically, for $k \notin K_f$, let the algorithm's step be given by

$$\sum_{S \in \mathcal{C}} \alpha_{k,S} \nabla R_S(w_k)$$

where for each $S \in \mathcal{C}$, we have $\mathbf{E}\left[\nabla R_S(w_k)|w_k\right] \in \partial R_S(w_k)$. Let $\eta_L = \frac{\tau}{GM}$ and $\eta_R = \frac{\tau}{M^2}$ denote the step size for the objective and residual steps, respectively. Then, consider the following choice of dual multipliers for each $S \in \mathcal{C}$.

$$\bar{\lambda}_S = \frac{\eta_R}{\eta_L |K_f|} \sum_{k \notin K_f} \alpha_{k,S}$$

Expanding the definition of $\bar{w}$ and applying convexity, we can bound the duality gap as follows

$$
\begin{aligned}
\gamma(\bar{w},\bar{\lambda}) &= L(\bar{w}) - D(\bar{\lambda}) \\
&\leq \frac{1}{|K_f|}\left(\sum_{k\in K_f} L(w_k)\right) - \inf_{w\in\mathcal{F}}\left\{L(w) + \sum_{S\in\mathcal{C}} \bar{\lambda}_S R_S(w)\right\} \qquad (5.5)\\
&= \sup_{w\in\mathcal{F}}\left\{\frac{1}{|K_f|}\left(\sum_{k\in K_f} L(w_k)\right) - L(w) - \sum_{S\in\mathcal{C}} \bar{\lambda}_S R_S(w)\right\} \\
&= \sup_{w\in\mathcal{F}}\left\{\frac{1}{\eta_L\,|K_f|}\left(\eta_L \sum_{k\in K_f}(L(w_k)-L(w)) - \eta_R \sum_{k\notin K_f}\sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w)\right)\right\} \qquad (5.6)
\end{aligned}
$$

where (5.5) follows from expanding $\bar{w}$ then applying convexity of L and the definition of $d(\bar{\lambda})$ and (5.6) follows by our choice of $\bar{\lambda}_S$ for each $S \in \mathcal{C}$.

With the duality gap expanded into one sum over the feasible iterates and one sum over the infeasible iterates, we can analyze these iterates separately. The following lemmas show how to track the contribution of each term to the duality gap in terms of a potential function u_k defined as

$$u_k(w) = \frac{1}{2}\left\|w - w_k\right\|^2.$$

For notational convenience, for each $k \in K_f$, let $e(w_k) = \mathbf{E}[\nabla L(w_k)|w_k] - \nabla L(w_k)$ be the noise in the subgradient computation.

Lemma 5.8. *For all $w \in \mathcal{F}$ and for all $k \in K_f$,*

$$\eta_L \cdot (L(w_k) - L(w)) \leq u_k(w) - u_{k+1}(w) + \frac{\tau^2}{2M^2} + \eta_L\langle e(w_k), w_k - w\rangle.$$

Again, for notational convenience, for each $k \in [T] \setminus K_f$, let

$$e(w_k) = \sum_{S\in\mathcal{C}} \alpha_{k,S}\left(\mathbf{E}[\nabla R_S(w_k)|w_k] - \nabla R_S(w_k)\right)$$

be the noise in the subgradient computation.

Lemma 5.9. *For all $w \in \mathcal{F}$ and for all $k \in [T] \setminus K_f$,*

$$-\eta_R \sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w) \leq u_k(w) - u_{k+1}(w) - \frac{\tau^2}{10M^2} + \eta_R\langle e(w_k), w_k - w\rangle.$$

We defer the proofs of Lemmas 5.8 and 5.9. Assuming Lemmas 5.8 and 5.9, we bound the duality gap as follows.

$$\sup_{w\in\mathcal{F}}\left\{\frac{1}{\eta_L\left|K_f\right|}\left(\sum_{k=1}^{T}\left[u_{k-1}(w)-u_k(w)\right]+\eta_L\sum_{k\in K_f}\langle e(w_k),w_k-w\rangle\right.\right.$$
$$\left.\left.+\,\eta_R\sum_{k\notin K_f}\langle e(w_k),w_k-w\rangle+\frac{\tau^2}{2M^2}\left|K_f\right|-\frac{\tau^2}{10M^2}(T-\left|K_f\right|)\right)\right\}$$

$$\leq\frac{1}{\eta_L\left|K_f\right|}(*)+\underbrace{\frac{G\tau}{2M}+\frac{G\tau}{10M}}_{(**)}$$

for

$$\underbrace{\left(\sup_{w\in\mathcal{F}}\left\{u_0(w)+\eta_L\sum_{k\in K_f}\langle e(w_k),w_k-w\rangle+\eta_R\sum_{k\notin K_f}\langle e(w_k),w_k-w\rangle\right\}-\frac{\tau^2}{10M^2}T\right)}_{(*)}$$

by rearranging. Noting that $(**)$ can be bounded by $\frac{3G}{5M}\tau$, it remains to bound $(*)$. We show that for a sufficiently large T, then $(*)$ cannot be positive.

Consider the terms in the supremum over $w\in\mathcal{F}$. Note that we can upper bound $\sup\{u_0(w)\}\leq 2B^2n$. Additionally, we upper bound the error incurred due to the objective subgradient noise with the following lemma.

Lemma 5.10. *With probability at least* $1-4\delta$*, the contribution of the noisy subgradient computation to the duality gap can be bounded as follows.*

$$\sup_{w\in\mathcal{F}}\left\{\eta_L\sum_{k\in K_f}\langle e(w_k),w_k-w\rangle+\eta_R\sum_{k\notin K_f}\langle e(w_k),w_k-w\rangle\right\}\leq\frac{\tau B}{M}\sqrt{8Tn\log(n/\delta)}$$

Thus, we can bound $(*)$ as follows.

$$(*)\leq 2B^2n+\frac{\tau B}{M}\sqrt{8Tn\log(n/\delta)}-\frac{\tau^2}{10M^2}T$$

Assuming the lemma and that $T > \frac{30^2 M^2 B^2 n \log(n/\delta)}{\tau^2}$, then, we can bound $(*)$ by splitting the negative term involving T to balance both positive terms.

$$(*) \leq \left(2B^2 n - \frac{\tau^2}{10M^2} \cdot \frac{20T}{30^2}\right) + \left(\frac{\tau B}{M}\sqrt{8n\log(n/\delta)} \cdot \sqrt{T} - \frac{\tau^2}{10M^2} \cdot \frac{(30^2 - 20)T}{30^2}\right)$$

$$\begin{aligned}
&\leq \left(2B^2 n - \frac{\tau^2}{10M^2}\frac{20M^2B^2 n\log(n/\delta)}{\tau^2}\right) \\
&\quad + \left(\frac{\tau B}{M}\sqrt{8n\log(n/\delta)} \cdot \frac{30MB\sqrt{n\log(n/\delta)}}{\tau} - \frac{\tau^2}{10M^2} \cdot \frac{(30^2-20)M^2B^2 n\log(n/\delta)}{\tau^2}\right) \\
&\leq \left(2B^2 n - 2B^2 n\log(n/\delta)\right) + \left(85B^2 n\log(n/\delta) - 88B^2 n\log(n/\delta)\right)
\end{aligned}$$

Thus, the sum of $(*)$ and $(**)$ is at most $\frac{3G}{5M}\tau$. □

Deferred proofs. First, we show a technical lemma that will be useful in analyzing the iterates' contributions to the duality gap. Recall our potential function $u_k : \mathcal{F} \to \mathcal{R}$.

$$u_k(w) = \frac{1}{2}\left\|w_k - w\right\|_2^2$$

We show that the update rule $w_{k+1} \leftarrow \pi_{\mathcal{F}}(w_k - \eta_k g_k)$ implies the following inequality in terms of $\eta_k, g_k, u_k(w)$, and $u_{k+1}(w)$.

Lemma 5.11. *Suppose $w_{k+1} = \pi_{\mathcal{F}}(w_k - \eta_k g_k)$. Then, for all $w \in \mathcal{F}$,*

$$\eta_k\langle g_k, w_k - w\rangle \leq u_k(w) - u_{k+1}(w) + \frac{\eta_k^2}{2}\left\|g_k\right\|_2^2.$$

Proof. Consider the differentiable, convex function $B_k : \mathcal{F} \to \mathcal{R}$.

$$B_k(w) = \eta_k\langle g_k, w - w_k\rangle + \frac{1}{2}\left\|w - w_k\right\|_2^2 \tag{5.7}$$

$$\begin{aligned}
\langle \nabla B_k(w_{k+1}), w - w_{k+1}\rangle &= \langle \eta_k g_k + w_{k+1} - w_k, w - w_{k+1}\rangle && (5.8)\\
&= \langle \pi_{\mathcal{F}}(w_k - \eta_k g_k) - (w_k - \eta_k g_k), w - \pi_{\mathcal{F}}(w_k - \eta_k g_k)\rangle && (5.9)\\
&\geq 0 && (5.10)
\end{aligned}$$

where (5.9) follows by substituting the definition of w_{k+1} twice; and (5.10) follows from the fact that for any closed convex set $\mathcal{F}$ and $w_0 \notin \mathcal{F}$,

$$\langle \pi_{\mathcal{F}}(w_0) - w_0, w - \pi_{\mathcal{F}}(w_0) \rangle \geq 0.$$

Rearranging (5.8) implies the following inequality holds for all $w \in \mathcal{F}$.

$$\begin{aligned} &\langle \eta_k g_k + w_{k+1} - w_k, w - w_{k+1} \rangle \geq 0 \\ \iff &\eta_k \langle g_k, w_{k+1} - w \rangle \leq \langle w_{k+1} - w_k, w - w_{k+1} \rangle \end{aligned} \tag{5.11}$$

We will use the following technical identity to prove the lemma.

Proposition 5.12. *For all $w \in \mathcal{F}$,*

$$\langle w_{k+1} - w_k, w - w_{k+1} \rangle = u_k(w) - u_{k+1}(w) - \frac{1}{2} \|w_{k+1} - w_k\|^2 .$$

Proof.

$$\begin{aligned} u_k(w) - u_{k+1}(w) &= \|w_k - w\|^2 - \|w_{k+1} - w\|^2 \\ &= \|w_k\|^2 + \|w\|^2 - \|w_{k+1}\|^2 - \|w\|^2 + 2\langle w_{k+1} - w_k, w \rangle \\ &= \|w_k\|^2 - \|w_{k+1}\|^2 + 2\langle w_{k+1} - w_k, w \rangle \\ &= \|w_k\|^2 - \|w_{k+1}\|^2 - 2\langle w_k - w_{k+1}, w_{k+1} \rangle + 2\langle w_{k+1} - w_k, w - w_{k+1} \rangle \\ &= \|w_{k+1} - w_k\|^2 + 2\langle w_{k+1} - w_k, w - w_{k+1} \rangle \end{aligned}$$

□

Finally, we can show the inequality stated in the lemma.

$$\begin{aligned} \eta_k \langle g_k, w_k - w \rangle &= \eta_k \langle g_k, (w_{k+1} + \eta g_k) - w \rangle \\ &\leq \langle w_{k+1} - w_k, w - w_{k+1} \rangle + \eta_k^2 \|g_k\|_2^2 &(5.12) \\ &\leq u_k(w) - u_{k+1}(w) - \frac{1}{2} \|w_{k+1} - w_k\|_2^2 + \eta_k^2 \|g_k\|_2^2 &(5.13) \\ &= u_k(w) - u_{k+1}(w) + \frac{\eta_k^2}{2} \|g_k\|_2^2 &(5.14) \end{aligned}$$

where (5.12) follows from (5.11); (5.13) follows by using Proposition 5.12 to write the expression in terms of u_k's; and (5.14) follows by the gradient step $w_{k+1} - w_k = \eta_k g_k$. □

Proof of Lemma 5.8 Here, we bound the contribution to the duality gap of each of the feasible iterations $k \in K_f$ as follows.

$$\eta_L \cdot (L(w_k) - L(w)) \leq u_k(w) - u_{k+1}(w) + \frac{\tau^2}{2M^2} + \eta_L \langle e_L(w_k), w_k - w \rangle$$

Proof. Let $e_L(w_k) = g_L(w_k) - \nabla L(w_k)$ where $g_L(w_k) = \mathbf{E}[\nabla L(w_k)] \in \partial L(w_k)$.

$$\begin{aligned}
\eta_L \cdot (L(w_k) - L(w)) &\leq \eta \langle g_L(w_k), w_k - w \rangle \\
&\leq \eta \langle \nabla L(w_k) + e_L(w_k), w_k - w \rangle && (5.15) \\
&\leq u_k(w) - u_{k+1}(w) + \frac{\eta_L^2}{2} \left\| \nabla L(w_k) \right\|_2^2 + \eta_L \langle e_L(w_k), w_k - w \rangle && (5.16) \\
&\leq u_k(w) - u_{k+1}(w) + \frac{\tau^2}{2M^2} + \eta_L \langle e_L(w_k), w_k - w \rangle && (5.17)
\end{aligned}$$

where (5.15) follows by substituting g_L; (5.16) follows by expanding the inner product and applying Lemma 5.11 to the first term; (5.17) follows by our choice of $\eta_L = \tau / GM$. □

Proof of Lemma 5.9 Here, we bound the contribution to the duality gap of each of the infeasible iterates $k \in [T] \setminus K_f$. We assume $\hat{R}_{S_k}(w_k)$ has tolerance $\tau/5$. Then we show

$$-\eta_R \sum_{S \in \mathcal{C}} \alpha_{k,S} R_S(w) \leq u_k(w) - u_{k+1}(w) - \frac{\tau^2}{10M^2} + \eta_R \langle e_R(w_k), w_k - w \rangle.$$

Proof. Recall, we let $e(w_k) = \sum_{S \in \mathcal{C}} \alpha_{k,S} \left(\mathbf{E}[\nabla R_S(w_k) | w_k] - \nabla R_S(w_k) \right)$. For each $S \in \mathcal{C}$, for any $g_S(w_k) \in \partial R_S(w_k)$, we can rewrite $-R_S(w)$ as follows.

$$\begin{aligned}
-R_S(w) &= R_S(w_k) - R_S(w) - R_S(w_k) \\
&\leq \langle g_S(w_k), w_k - w \rangle - R_S(w_k)
\end{aligned}$$

Multiplying by η_R and taking the convex combination of $S \in \mathcal{C}$ according to α_k, we apply

Lemma 5.11 to obtain the following inequality.

$$\begin{aligned}
-\eta_R \sum_{S\in\mathcal{C}} &\alpha_{k,S} R_S(w) \\
&\leq \eta_R \left\langle \sum_{S\in\mathcal{C}} \alpha_{k,S} \nabla R_S(w_k) + e_R(w_k), w_k - w \right\rangle - \eta_R \sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w_k) \qquad (5.18)\\
&\leq u_k(w) - u_{k+1}(w) + \frac{\eta_R^2}{2} \left\| \sum_{S\in\mathcal{C}} \alpha_{k,S} \nabla R_S(w_k) \right\|_2^2 \\
&\qquad\qquad - \eta_R \sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w_k) + \eta_R \langle e_R(w_k), w_k - w\rangle \qquad (5.19)\\
&\leq u_k(w) - u_{k+1}(w) + \frac{\tau^2}{2M^2} - \frac{\tau}{M^2} \cdot \sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w_k) + \eta_R \langle e_R(w_k), w_k - w\rangle \qquad (5.20)\\
&\leq u_k(w) - u_{k+1}(w) - \frac{\tau^2}{10M^2} + \eta_R \langle e_R(w_k), w_k - w\rangle \qquad (5.21)
\end{aligned}$$

where (5.18) follows by substituting $\nabla R_S(w_k)$ for each $g_S(w_k)$ and the definition of $e_R(w_k)$; (5.19) follows by expanding the inner product and applying Lemma 5.11; (5.20) follows by our choice of $\eta_k = \tau^2/M^2$; (5.21) follows by the fact that when we update according to a constraint, we know $\sum_{S\in\mathcal{C}} \alpha_{k,S} \hat{R}_S(w_k) \geq 4\tau/5$ with tolerance $\tau/5$, so $\sum_{S\in\mathcal{C}} \alpha_{k,S} R_S(w_k) \geq 3\tau/5$. □

Proof of Lemma 5.10 Here, we show that with probability at least $1-4\delta$, the contribution of the noisy subgradient computation to the duality gap can be bounded as follows.

$$\sup_{w\in\mathcal{F}} \left\{ \eta_L \sum_{k\in K_f} \langle e(w_k), w_k - w\rangle + \eta_R \sum_{k\notin K_f} \langle e(w_k), w_k - w\rangle \right\} \leq \frac{\tau B}{M} \sqrt{8Tn\log(n/\delta)}$$

Proof. Let $\varepsilon = \eta_L \cdot g = \eta_R \cdot m = \frac{\tau}{M\sqrt{n}}$. Further, let $\eta_k = \eta_L$ for $k \in K_f$ and η_R for $k \notin K_f$. Then, we can rewrite the expression to bound using η_k and expand as follows.

$$\begin{aligned}
\sup_{w\in\mathcal{F}} \sum_{k\in[T]} \langle \eta_k e(w_k), w_k - w\rangle &= \sup_{w\in\mathcal{F}} \sum_{k\in[T]} \sum_{l=1}^{n} \eta_k e(w_k)_l \cdot (w_k - w)_l \\
&= \sum_{l=1}^{n} (w_k)_l \sum_{k\in[T]} \eta_k e(w_k)_l + \sum_{l=1}^{n} \sup_{(w)_l} \left\{ (w)_l \cdot \sum_{k\in[T]} \eta_k e(w_k)_l \right\}
\end{aligned}$$

Consider the second summation, and consider the summation inside the supremum. Note that this summation is a sum of mean-zero random variables, so it is also mean-zero. Recall, we assume the estimate of the kth subgradient is independent of the prior subgradients, given w_k. Further, by the assumed ℓ_∞ bound on the subgradients, each of these random variables is bounded in magnitude by ε. Using the bounded difference property, we apply Azuma's inequality separately for each $l \in [n]$.

$$\mathbf{Pr}\left[\left|\sum_{k\in[T]} \eta_k e(w_k)_l\right| > Z\right] \leq 2 \cdot \exp\left(-\frac{Z^2}{2T\varepsilon^2}\right)$$

Taking this probability to be at most $2\delta/n$, we can upper bound Z by $\varepsilon\sqrt{2T\log(n/\delta)} = \frac{\tau}{M}\sqrt{2Tn\log(n/\delta)}$. Then, noting that $|(w)_l| < B$ for any $w \in \mathcal{F}$, we can take a union bound to conclude with probability at least $1 - 2\delta$ the following inequalities hold.

$$\begin{aligned}\sum_{l=1}^{n} \sup_{(w)_l}\left\{(w)_l \cdot \sum_{k\in[T]} \eta_k e(w_k)_l\right\} &\leq Bn \cdot Z \\ &= \frac{\tau B}{M}\sqrt{2Tn\log(n/\delta)}\end{aligned}$$

Further, we note that the first summation concentrates at least as quickly as the second, so by union bounding again,

$$\sup_{w\in\mathcal{F}} \sum_{k\in[T]} \eta_k \langle e(w_k), w_k - w\rangle \leq \frac{\tau B}{M}\sqrt{8Tn\log(n/\delta)}$$

with probability at least $1 - 4\delta$. □

Chapter Notes

This chapter is based entirely off of joint work with Omer Reingold and Guy N. Rothblum. Recent years have seen a number of works investigating how to make individual metric fairness of [DHP$^+$12] more practical.

Most pertinent to this chapter is the work of [RY18] which introduces the notion of PACF learning—probably approximately correct and (metric) fair learning—and follow-up work [JKN$^+$19]. Collectively, these works show how to obtain a PAC-style guarantee for metric fairness satisfaction as well as optimality over linear hypotheses, from a small set of individual-outcome and metric samples. Rather than taking the multi-group approach, these works show that for a sufficiently simple class of hypotheses, small metric fairness loss (i.e., degree of individual fairness violations) in sample translates into bounds on the distributional metric fairness violations. The guarantees of PACF are incomparable to multi-metric-fairness: PACF provides a stronger guarantee against violating the metric (by essentially protecting all groups), but multi-metric fairness may obtain better utility.

Two other works have taken interesting steps at making individual fairness notions practically realizable. [Ilv20] demonstrates settings in which it is feasible to learn the fairness metric directly using limited human input. [KRSM19] propose an alternative to metric fairness, focusing instead of "individual error rate parity." To do this, they work with not only a distribution over individuals, but also a distribution over classification instances and show how to equalize each individual's error rates across the distribution of instances.

Chapter 6

Refining Information in the Service of Fairness

While most researchers studying algorithmic fairness can agree on the high-level objectives of the field—to ensure individuals are not mistreated on the basis of protected attributes; to promote social well-being and justice across populations—there is much debate about how to translate these aspirations into a concrete, formal definition of what it means for a prediction system to be fair. Indeed, as this nascent field has progressed, the efforts to promote "fair prediction" have grown increasingly divided, rather than coordinated. Exacerbating the problem, [MPB$^+$18] identifies that each new approach to fairness makes its own set of assumptions, often implicitly, leading to contradictory notions about the right way to approach fairness [Cho17, KMR17]. Complicating matters further, recent works [LDR$^+$18, CG18] have identified shortcomings of many well-established notions of fairness. At the extreme, these works argue that blindly requiring certain statistical fairness conditions may in fact harm the communities they are meant to protect.

In this thesis, our focus has been on multi-calibration as one notion of fairness that ensures high-quality predictions, not just overall but on structured subpopulations. Multi-calibration, however, is not the only proposal for subpopulation fairness. In Chapter 5, we discussed multi-metric fairness based on the metric fairness notion of [DHP$^+$12]. Other works in recent years have proposed multi-group fairness notions based on parity between subgroups [KNRW18, KNRW19, HSNL18]. Often, there are not clear technical directives of when each notion is appropriate, and instead have left it a matter to be decided by practitioners and experts of the application domain. In the hopes of providing more explicit directives and unifying the many directions of research in the area, we take a step back and ask whether there are guiding principles that broadly serve the high-level goals of "fair" prediction, without relying too strongly on any specific notion of fairness.

In this chapter, we argue that understanding the "informativeness" of predictions needs to be part of any sociotechnical conversation about the fairness of a prediction system. We show that multi-calibrated predictors may serve as a principled intermediary in learning a fair classifier, even when our notion of fairness is based on parity. Specifically, rather than enforcing a multi-parity-style notion of fairness on a decision rule directly, we may first learn a multi-calibrated predictor $\tilde{p}$, then select an optimal selection rule (based on $\tilde{p}$) trading off utility, disparity, and impact on groups of interest. In this sense, multi-calibration may provide an effective way to enforce parity-based notions of fairness, running counter to the conventional wisdom that calibration and parity are inherently at odds.

Information in predictors. In Section 6.1, we provide a self-contained exposition of the framework we use to study the *information content* of a predictor. Information content formally quantifies the uncertainty over individuals' outcomes given their predictions. To compare the information content of multiple predictors, we leverage the concept of *refinements* [DF81]. Refinements provide the technical tool for reasoning about how to improve the information content of a predictor.

In Section 6.2, we show a formal connection between refinements and multi-calibration. In particular, we show that the multi-calibration guarantee, parameterized by a class of boolean functions, is essentially equivalent to a strong notion of simultaneous refinement, parameterized by a class of calibrated predictors. In this sense, we can equivalently define multi-calibration in terms of refinement of calibrated predictors.

In Section 6.3, we revisit the question of finding an optimal fair selection rule, in a setting adapted from [LDR$^+$18]. For prominent parity-based fairness desiderata, we show that improving the information content of the underlying predictions via refinements results in a Pareto improvement of the resulting program in terms of utility, disparity, and long-term impact. As one concrete example, if we hold the selection rule's utility constant, then refining the underlying predictions causes the disparity between groups to decrease. Notably, refinements play a key role in the arguments.

6.1 Tracking Information in Binary Predictions

In this section, we give a self-contained exposition of a formal notion of information content in calibrated predictors. These notions have been studied extensively in the forecasting literature (see [GBR07, GR07] and references therein) and can be viewed through the lens of proper scoring rules [Bri50]. For the sake of exposition, throughout the chapter, we will typically discuss exactly calibrated predictors.

Risk distributions. Throughout this chapter, it will be useful to reason not only about predictors, but about their induced *risk distributions.* Given a predictor and a distribution over individuals, the associated risk distribution is the histogram of scores output by the predictor.

Definition 6.1 (Risk distribution)**.** *For a subset $S \subseteq \mathcal{X}$, given a predictor $p : \mathcal{X} \to [0,1]$, the risk distribution over S, $\mathcal{R}_S^p : \mathrm{supp}(p) \to [0,1]$ is defined as*

$$\mathcal{R}_S^p(v) = \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} [\ p(x) = v\ ].$$

Throughout, we will continue to assume that predictors have discrete support and will only consider risk distributions over calibrated predictors. Additionally, we will abuse notation and consider sampling a score from risk distributions where $v \sim \mathcal{R}_S^p$ denotes a sample from the support of p drawn with probability $\mathbf{Pr}_{x \sim \mathcal{D}_S}[p(x) = v]$.

Information content. In the context of binary prediction, a natural way to measure the "informativeness" of a predictor is by the uncertainty in an individual's outcome given their score. Viewing the outcome as a Bernoulli random variable with parameter v, we have perfect information (zero uncertainty) about the outcome if $v = 0$ or $v = 1$, whereas we have no information (maximal uncertainty) in a trial where $v = 1/2$. We quantify this notion of uncertainty through Shannon entropy. The Shannon entropy of a Bernoulli random variable with parameter v is given by the binary entropy function $H_2(v)$ where

$$H_2(v) = -v \cdot \log(v) - (1 - v) \cdot \log(1 - v).$$

Consider a random draw $x, y \sim \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$. Given access to a calibrated predictor p, then the conditional distribution over y given $p(x)$ follows $\mathrm{Ber}(p(x))$. This observation suggests the following definition for the information content of a predictor.

Definition 6.2 (Information content)**.** *Suppose for $S \subseteq \mathcal{X}$, $p : \mathcal{X} \to [0,1]$ is calibrated over S. The information content of p over S is given as*

$$I_S(z) = 1 - \mathop{\mathbf{E}}_{x \sim S}\left[H_2(x)\right].$$

We define information content over subsets of individuals; we use $I(p) = I_{\mathcal{X}}(p)$ to denote the information content of p over the entire domain. At the extremes, a perfect binary classifier has information content 1, whereas a calibrated predictor that always outputs 1/2 has 0 information.

Another motivation for this definition of information content is its connection to the log-likelihood of the predictor. If we view a calibrated predictor p as parameterizing a conditional distribution of outcomes, the likelihood of a sample (x, y) is given as

$$\mathcal{L}\left(p; (x,y)\right) = p(x)^y \cdot \left(1 - p(x)\right)^{1-y}$$

Correspondingly, the log-likelihood is given as

$$\ell(p; (x,y)) = y \cdot \log(p(x)) + (1-y) \cdot \log(1 - p(x)).$$

We observe that information content is simply a shift of the expected log-likelihood over samples drawn from the joint individual-outcome distribution.

Lemma 6.3. *Suppose for $S \subseteq \mathcal{X}$, $p : \mathcal{X} \to [0,1]$ is calibrated over S. Then,*

$$I_S(p) = \mathop{\mathbf{E}}_{(x,y) \sim \mathcal{D}_{S \times \mathcal{Y}}}\left[\ \ell(p; (x,y))\ \right] + 1$$

Proof. We expand the expected log-likelihood as follows.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_{S\times\mathcal{Y}}}\left[\ \ell(p;(x,y))\ \right] \\
&= \mathop{\mathbf{E}}_{(x,y)\sim\mathcal{D}_{S\times\mathcal{Y}}}\left[\ y\cdot\log(p(x)) + (1-y)\cdot\log(1-p(x))\ \right] \\
&= \sum_{v\in\mathrm{supp}(p)} \mathcal{R}_S^p(v)\cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ \mathop{\mathbf{E}}_{y\sim\mathcal{D}_{\mathcal{Y}|\mathcal{X}}}\left[\ y\cdot\log(v) + (1-y)\cdot\log(1-v)\mid x\ \right]\ \middle|\ p(x)=v\ \right] \quad (6.1) \\
&= \sum_{v\in\mathrm{supp}(p)} \mathcal{R}_S^p(v)\cdot \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ v\cdot\log(v) + (1-v)\cdot\log(1-v)\mid p(x)=v\ \right] \quad (6.2) \\
&= -\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ H_2(p(x))\ \right] \quad (6.3) \\
&= I_S(p) - 1
\end{aligned}$$

where (6.1) follows by Bayes rule and the definition of the risk distribution of p; (6.2) follows by the assumption that p is calibrated over S; and (6.3) follows by the definition of Shannon entropy and reversing the equality from (6.1). □

Thus, as the information content increases, the expected likelihood of the predictor also increases; at the extreme, the calibrated predictor that maximizes information content is the max (expected) likelihood predictor.

Refinements. Intuitively, as the information content of a predictor improves, so too should the resulting decisions derived from the predictor in terms of utility and fairness. To make this intuition formal, we formalize the idea of improving a predictor's information content using the notion of *refinement*, first proposed by [DF81]. Specifically, refinements define a partial order over calibrated predictors, where given calibrated $\tilde{p}$ and p, we say that $\tilde{p}$ *refines* p if $\tilde{p}$ incorporates all of the information in p, and possibly additional information. [DF81] defined the notion of a refinement in terms of another concept called a *stochastic transformation.*

Definition 6.4 (Stochastic Transformation [DF81]). *Given predictors* $p : \mathcal{X} \to [0,1]$ *and*

$q : \mathcal{X} \to [0,1]$*, a stochastic transformation* $\tau : \mathrm{supp}(p) \times \mathrm{supp}(q) \to [0,1]$ *is a map such that*

$$\begin{aligned} \tau(u,v) &\geq 0 & \forall u \in \mathrm{supp}(p), v \in \mathrm{supp}(q) \\ \sum_{v \in \mathrm{supp}(q)} \tau(u,v) &= 1 & \forall u \in \mathrm{supp}(p). \end{aligned}$$

A stochastic transformation can be viewed as randomly mapping the risk distribution of p into that of q, where $\tau(u,v)$ controls the conditional "probability" that a unit where $p(x) = u$ is mapped to $q(x) = v$. With this notion in place, we can define refinements formally.

Definition 6.5 (Refinement [DF81])**.** *For a subset* $S \subseteq \mathcal{X}$*, suppose predictors* $\tilde{p} : \mathcal{X} \to [0,1]$ *and* $p : \mathcal{X} \to [0,1]$ *are calibrated over* S*.* $\tilde{p}$ *is a refinement of* p *over* S *if there exists a stochastic transformation* $\tau : \mathrm{supp}(\tilde{p}) \times \mathrm{supp}(p) \to [0,1]$ *such that for all* $v \in \mathrm{supp}(p)$*:*

$$\sum_{u \in \mathrm{supp}(\tilde{p})} \tau(u,v) \cdot \mathcal{R}_S^{\tilde{p}}(u) = \mathcal{R}_S^{p}(v), \textit{ and} \tag{6.4}$$

$$\sum_{u \in \mathrm{supp}(\tilde{p})} \tau(u,v) \cdot \mathcal{R}_S^{\tilde{p}}(u) \cdot u = \mathcal{R}_S^{p}(v) \cdot v. \tag{6.5}$$

In other words, $\tilde{p}$ refines p if the risk distribution of p can be derived from that of $\tilde{p}$ by a stochastic transformation that preserves calibration. Such a map may merge units from multiple risk scores under $\tilde{p}$ into a single score in p. For our subsequent analysis, the following equivalent formulation of refinements will be easier to work with.[1]

Lemma 6.6 (Alternative characterization of refinements)**.** *For a subset* $S \subseteq \mathcal{X}$*, suppose predictors* $\tilde{p} : \mathcal{X} \to [0,1]$ *and* $p : \mathcal{X} \to [0,1]$ *are calibrated over* S*.* $\tilde{p}$ *is a refinement of* p *over* S *if and only if*

$$\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S} [\ \tilde{p}(x) \mid p(x) = v\ ] = v.$$

Proof. We show that assuming that condition (6.4) holds, we can rewrite the expression

[1] [GKR19] takes this formulation to be the definition of refinements.

(6.5) as the desired conditional expectation.

$$\begin{aligned} v &= \frac{\sum_{u \in \mathrm{supp}(\tilde{p})} \tau(u,v) \cdot \mathcal{R}_S^{\tilde{p}}(u) \cdot u}{\mathcal{R}_S^{p}(v)} \\ &= \frac{\sum_{u \in \mathrm{supp}(\tilde{p})} \mathbf{Pr}_{\mathcal{D}_S}\left[\, p(x) = v \mid \tilde{p}(x) = u \,\right] \cdot \mathbf{Pr}_{\mathcal{D}_S}\left[\, \tilde{p}(x) = u \,\right] \cdot u}{\mathbf{Pr}_{\mathcal{D}_S}\left[\, p(x) = v \,\right]} \qquad (6.6) \\ &= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\, \tilde{p}(x) \mid p(x) = v \,\right] \qquad (6.7) \end{aligned}$$

where (6.6) follows by the definition of the risk distribution and the interpretation of the stochastic transformation as a conditional probability; and (6.7) follows by Bayes rule and the definition of the conditional expectation. The reverse direction follows by defining τ in terms of the sets $\mathcal{X}_{uv} = \{x : \tilde{p}(x) = u\} \cap \{x : p(x) = v\}$ to be a stochastic transformation. □

Note that this characterization establishes that p^* refines all other calibrated predictors.

Corollary 6.7. *For a given set of features $\mathcal{X}$ and any calibrated predictor $p : \mathcal{X} \to [0,1]$, the Bayes optimal predictor $p^* : \mathcal{X} \to [0,1]$ is a refinement of p.*

This corollary follows directly by Lemma 6.6 and the restriction to calibrated predictors. In this sense, given a predictor p, any refinement $\tilde{p}$ can be viewed as a candidate optimal predictor. Importantly, this logic depends on refinements, not simply greater information content. To see this, suppose that there is some $v \in \mathrm{supp}(p)$ such that $\mathbf{E}[\tilde{p}(x)|p(x) = v] \neq v$ (even if $I(\tilde{p}) > I(p)$. Because p is calibrated, the inequality suggests that $\mathbf{E}[\tilde{p}(x)|p(x) = v] \neq \mathbf{E}[p^*(x)|p(x) = v]$. This disagreement provides evidence that p has some consistency with the optimal predictor p^* that $\tilde{p}$ lacks.

Refinement distance as information loss. Refinements establish a binary relation over calibrated predictors, which we can use to establish qualitatively that a refined predictor $\tilde{p}$ has more information than p. Next, we define a notion of *refinement distance* which makes such comparisons between risk distributions quantitative. Suppose $\tilde{p}$ is a refinement of p; refinement distance aims to measure how far the predictions according to p will be compared to those of $\tilde{p}$. More formally, given p, imagine sampling a risk score $v \sim \mathcal{R}^p$, then sampling an individual $x \sim \mathcal{D}_{\{x:p(x)=v\}}$. We wish to measure the difference in the belief about the outcome under the distributions $\mathrm{Ber}(p(x))$ and $\mathrm{Ber}(\tilde{p}(x))$. Specifically, we denote by $D_{\mathrm{KL}}(p;q)$ the KL-divergence between two Bernoulli distributions with parameters p and

q, respectively, defined as

$$D_{\mathrm{KL}}(p;q) = p \cdot \log\left(\frac{p}{q}\right) + (1-p)\cdot\log\left(\frac{1-p}{1-q}\right).$$

We define the refinement distance in terms of the expected KL-divergence between a risk distribution and its refinement.

Definition 6.8 (Refinement distance)**.** *For a subset $S \subseteq \mathcal{X}$, suppose that $\tilde{p} : \mathcal{X} \to [0,1]$ refines $p : \mathcal{X} \to [0,1]$ over S. The refinement distance from $\tilde{p}$ to p over S is given as*

$$D_S^{\mathcal{R}}(\tilde{p};p) = \mathop{\mathbf{E}}_{v\sim\mathcal{R}_S^p}\left[\ \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ D_{\mathrm{KL}}\left(\tilde{p}(x);v\right)\ \mid p(x)=v\ ]\ \right]$$

While we define the refinement distance in terms of sampling from the risk distribution, we can express this distance as an *information loss.* Specifically, the refinement distance is equal to the lost information content from $\tilde{p}$ to p.

Lemma 6.9. *For a subset $S \subseteq \mathcal{X}$, suppose that $\tilde{p} : \mathcal{X} \to [0,1]$ refines $p : \mathcal{X} \to [0,1]$ over S. Then,*

$$\mathcal{D}_S^{\mathcal{R}}(\tilde{p},p) = I_S(\tilde{p}) - I_S(p).$$

In this sense, the difference information content between a predictor and its refinement captures a natural notion of distance between their induced risk distributions. We observe that for calibrated predictors, the refinement distance from p^* to p can be used as a progress measure to analyze the convergence of Algorithm 2. In the light of Lemma 6.9, we can understand the progress of each iteration of the algorithm as changes to $D^{\mathcal{R}}(p^*,p)$, and thus, improved information about p^*. We make the connection between information content and multi-calibration explicit in the subsequent section.

Note that because the KL-divergence is a nonnegative quantity, Lemma 6.9 and Corollary 6.7 immediately establish the fact that p^* is the most informative predictor.

Corollary 6.10. *For a given set of features $\mathcal{X}$, the Bayes optimal predictor $p^* : \mathcal{X} \to [0,1]$ maximizes information content.*

$$p^* = \mathop{\mathrm{argmax}}_{p:\mathcal{X}\to[0,1]} \quad I(p)$$

This observation provides another lens into the fact that p^* is the predictor that maximizes the expected likelihood over samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$.

Proof of Lemma 6.9. We expand the refinement distance leveraging the alternative characterization of refinements.

$$D_S^{\mathcal{R}}(\tilde{p}; p) = \mathop{\mathbf{E}}_{v \sim \mathcal{R}_S^p}\left[\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[D_{\mathrm{KL}}\left(\tilde{p}(x); v\right) \mid p(x) = v \right] \right]$$

$$= \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}\left[\tilde{p}(x) \cdot \log\left(\frac{\tilde{p}(x)}{v}\right) + (1 - \tilde{p}(x)) \cdot \log\left(\frac{1 - \tilde{p}(x)}{1 - v}\right) \,\middle|\, p(x) = v \right] \tag{6.8}$$

We split (6.8) into a term that depends only on $\tilde{p}$, that results in its expected entropy.

$$\begin{aligned}
&(6.8) + \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, H_2(\tilde{p}(x)) \,] \\
&= -\sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, \tilde{p}(x) \cdot \log(v) + (1 - \tilde{p}(x)) \cdot \log(1 - v) \mid p(x) = v \,] \\
&= -\sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot \Big(\mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, \tilde{p}(x) \mid p(x) = v \,] \cdot \log(v) \\
&\qquad + (1 - \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, \tilde{p}(x) \mid p(x) = v \,]) \cdot \log(1 - v) \Big) && (6.9) \\
&= \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot H_2(v) && (6.10) \\
&= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, H_2(p(x)) \,]
\end{aligned}$$

where (6.9) follows by linearity of expectation; and (6.10) follows by the assumption that $\tilde{p}$ refines p. In all, this shows that

$$\begin{aligned}
\mathcal{D}_S^{\mathcal{R}}(\tilde{p}; p) &= \mathop{\mathbf{E}}_{x \sim \mathcal{D}_S}[\, H_2(\tilde{p}(x)) - H_2(p(x)) \,] \\
&= I_S(\tilde{p}) - I_S(p).
\end{aligned}$$

□

6.2 Multi-Calibration as Simultaneous Refinement

Corollaries 6.7 and 6.10 demonstrate that p^* is a refinement of every calibrated predictor and maximizes the information content. As we've discussed, however, in many cases p^* may be unattainable due to data and computational limitations. Still, we may ask for a predictor that captures the information contained in a "meaningful" bounded class of calibrated predictors. In this section, we demonstrate that multi-calibration already gives us such a solution concept. We show that the multi-calibration constraints can be characterized as simultaneous refinement over a class of calibrated predictors.

Simultaneous refinement. We start by defining a notion of *multi-refinement* that captures the idea that a predictor $\tilde{p}$ is a refinement of every predictor in a class of calibrated predictors, simultaneously.

Definition 6.11 (Multi-Refinement)**.** *Let $\alpha \geq 0$. Suppose $\mathcal{P} \subseteq \{p : \mathcal{X} \to [0,1]\}$ is a set of predictors α-calibrated over $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. Then, $\tilde{p} : \mathcal{X} \to [0,1]$ is a $(\mathcal{P}, \alpha)$-multi-refinement if $\tilde{p}$ is α-calibrated over $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, and for all $p \in \mathcal{P}$ and all $v \in \mathrm{supp}(p)$*

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

That is, a multi-refinement simultaneously (approximately) refines each predictor in the given class of calibrated predictors. Again, p^* is a $(\mathcal{P}, 0)$-multi-refinement for any class of calibrated predictors $\mathcal{P}$, but as we bound the complexity of $\mathcal{P}$, the set of multi-refinements will grow beyond p^*. Intuitively, the motivation and properties of multi-refinement seem similar to multi-calibration. We show that the notions are closely related; in fact, multi-refinements are equivalent to calibrated multi-accurate predictors, whereas the multi-calibration constraints are even more restrictive.

This characterization of multi-refinements allows us to interpret the guarantees of multi-calibration in terms of the information-theoretic quantities defined in the prior section. Colloquially, a multi-calibrated predictor captures all of the information about p^* captured by $\mathcal{C}$. We begin with the characterization of multi-refinements.

Theorem 6.12 (Characterizing Multi-Refinement)**.** *Let $\alpha \geq 0$.*

(a) For every class of boolean functions $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0,1\}\}$, for any $\varepsilon > 0$, there is an explicit class of ε-calibrated predictors $\mathcal{P}^{\mathcal{C}} \subseteq \{p : \mathcal{X} \to [0,1]\}$ such that for all $\tilde{p} : \mathcal{X} \to [0,1]$

$$\tilde{p} \text{ is a } (\mathcal{P}^{\mathcal{C}}, \alpha)\text{-multi-refinement} \implies \tilde{p} \text{ is } \alpha\text{-calibrated and } (\mathcal{C}, \alpha+\varepsilon)\text{-multi-accurate.}$$

(b) For every class of calibrated predictors $\mathcal{P} \subseteq \{p : \mathcal{X} \to [0,1]\}$, there is an explicit class of boolean functions $\mathcal{C}^{\mathcal{P}} \subseteq \{c : \mathcal{X} \to \{0,1\}\}$ such that for all $\tilde{p} : \mathcal{X} \to [0,1]$

$$\tilde{p} \text{ is } \alpha\text{-calibrated and } (\mathcal{C}^{\mathcal{P}}, \alpha)\text{-multi-accurate} \implies \tilde{p} \text{ is a } (\mathcal{P}, 2\alpha)\text{-multi-refinement.}$$

Proof. (a) Given a boolean function $c : \mathcal{X} \to \{0,1\}$, denote the conditional expectations as follows.

$$v_0^c = \mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ p^*(x) \mid c(x) = 0\ ]$$
$$v_1^c = \mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ p^*(x) \mid c(x) = 1\ ]$$

For each such c, for any arbitrarily small $\varepsilon > 0$, there is an ε-calibrated predictor $p^c : \mathcal{X} \to [0,1]$ defined as

$$p^c(x) = \begin{cases} v_0^c - \varepsilon & \text{if } c(x) = 0 \\ v_1^c + \varepsilon & \text{if } c(x) = 1 \end{cases}$$

Note that the addition and subtraction of ε is used to break ties if $v_0^c = v_1^c$ (where we truncate $p^c(x)$ to $[0,1]$ if $v_0^c \in [0, \varepsilon]$ or $v_1^c \in [1-\varepsilon, \varepsilon]$). In other words, for every $x \in \mathcal{X}$, $p^c(x)$ is roughly the average value amongst x' such that $c(x') = c(x)$. For a class of boolean functions $\mathcal{C} \subseteq \{c : \mathcal{X} \to \{0,1\}\}$, consider the class of ε-calibrated predictors defined as

$$\mathcal{P}^{\mathcal{C}} = \{p^c : c \in \mathcal{C}\}.$$

Suppose $\tilde{p}$ is a $(\mathcal{P}^{\mathcal{C}}, \alpha)$-multi-refinement. First, note that by definition of multi-refinement, $\tilde{p}$ must be α-calibrated overall. Second, note that by the construction of p^c, for each $c \in \mathcal{C}$, we maintain that a bijection where $c(x) = 0$ if and only if $p^c(x) = v_0^c$ (and similarly for $c(x) = 1$). With this observation, we show that $\tilde{p}$ satisfies the multi-accuracy constraint on c.

$$\begin{aligned} \mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ \tilde{p}(x) \mid c(x) = 1\ ] &= \mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ \tilde{p}(x) \mid p^c(x) = v_1^c\ ] && (6.11) \\ &\leq v_1^c + \alpha && (6.12) \\ &\leq \mathop{\mathbf{E}}_{x \sim \mathcal{D}} [\ p^*(x) \mid c(x) = 1\ ] + \alpha + \varepsilon && (6.13) \end{aligned}$$

where (6.11) follows by the bijection established by construction of p^c; (6.12) follows by the assumption that $\tilde{p}$ is a multi-refinement; and (6.13) follows by the construction of v_1^c. The other multi-accuracy inequalities follow similarly. Thus, a $(\mathcal{P}^{\mathcal{C}}, \alpha)$-multi-refinement is α-calibrated and $(\mathcal{C}, \alpha)$-multi-accurate.

(b) For a class of α-calibrated predictors $\mathcal{P}$, consider the following class of boolean functions.

$$\mathcal{C}^{\mathcal{P}} = \{ \ \mathbf{1}\{p(x) = v\} : p \in \mathcal{P}, \ v \in \operatorname{supp}(p) \ \}$$

Suppose that $\tilde{p}$ is $(\mathcal{C}^{\mathcal{P}}, \alpha)$-multi-accurate. We show that $\tilde{p}$ also satisfies the multi-refinement constraints.

$$\begin{aligned} \mathop{\mathbf{E}}_{x \sim \mathcal{D}}[\ \tilde{p}(x) \mid p(x) = v\ ] &\leq \mathop{\mathbf{E}}_{x \sim \mathcal{D}}[\ p^*(x) \mid \mathbf{1}\{p(x) = v\} = 1\ ] + \alpha \qquad (6.14) \\ &\leq v + 2\alpha \qquad (6.15) \end{aligned}$$

where (6.14) follows by the definition of multi-accuracy and construction of $\mathcal{C}^{\mathcal{P}}$; and (6.15) follows by the assumption that every $p \in \mathcal{P}$ is α-calibrated. Thus, if $\tilde{p}$ is α-calibrated and $(\mathcal{C}^{\mathcal{P}}, \alpha)$-multi-accurate, then $\tilde{p}$ is a $(\mathcal{P}, 2\alpha)$-refinement. □

We conclude this section by observing that multi-calibration can be characterized by a stronger form of refinements, which we call *reflexive refinements*. Specifically, we say that $\tilde{p}$ is a reflexive refinement of p if $\tilde{p}$ refines p, not just overall, but also over the sets $\{x : p(x) = v\}$ for all $v \in \operatorname{supp}(p)$. For a class of calibrated predictors $\mathcal{P}$, we say that $\tilde{p}$ is a $\mathcal{P}$-multi-reflexive-refinement if $\tilde{p}$ reflexively refines every $p \in \mathcal{P}$. In fact, this strengthen notion of multi-refinement is equivalent to multi-calibration.

Theorem 6.13 (Characterizing Multi-Reflexive-Refinement, informal)**.** *For every $\mathcal{C}$, there exists a $\mathcal{P}^{\mathcal{C}}$ such that*

$$\tilde{p} \textit{ is a } \mathcal{P}^{\mathcal{C}}\textit{-multi-reflexive-refinement} \implies \tilde{p} \textit{ is } \mathcal{C}\textit{-multi-calibrated.}$$

For every $\mathcal{P}$, there exists a $\mathcal{C}^{\mathcal{P}}$ such that

$$\tilde{p} \textit{ is } \mathcal{C}^{\mathcal{P}}\textit{-multi-calibrated} \implies \tilde{p} \textit{ is a } \mathcal{P}\textit{-multi-reflexive-refinement.}$$

We state this theorem informally and exclude its proof. The formal statement of Theorem 6.13 and its proof exactly follow that of Theorem 6.12 in structure and constructions, applying the definition of reflexive-refinement where necessary.

6.3 The Value of Information in Fair Prediction

In this section, we argue that reasoning about the information content of calibrated predictors provides a lens into understanding how to improve the utility and fairness of predictors, even when the eventual fairness desideratum is based on parity. Our findings demonstrate the importance of maintaining information content—not just overall, but also on sensitive subpopulations—to ensure effective fair decision-making. The conclusions serve as yet another rebuke of the approach of "fairness through blindness" [DHP$^+$12]. Combined with Section 6.2, the findings also suggest that multi-calibrated predictions may play an important role in obtaining fair decisions, even when applying parity-based fairness notions.

We study a prediction setting from [LDR$^+$18] where a lender selects individuals to give loans from a pool of applicants. While we use the language of predicting creditworthiness, the setup is generic and can be applied to diverse prediction tasks. [LDR$^+$18] introduced a notion of "delayed impact" of selection policies, which models the potential negative impact on communities of enforcing parity-based fairness as a constraint.

The concern for delayed impact is based on the following line of reasoning: while giving out more loans to disadvantaged communities might ideally be considered desirable, if many of these loans will result in default, then the increased default rate could end up hurting the community. Indeed, [LDR$^+$18] show that in their model, blindly enforcing parity between subpopulations while maximizing the lender's utility may lead to worse outcomes for the disadvantaged population than under unconstrained utility maximization.

We revisit the question of delayed impact as part of a broader investigation of the role of information in fair prediction. [LDR$^+$18] argues that when there are differences in the risk score distributions between populations, seemingly-natural approaches to ensuring fairness, like enforcing parity amongst groups, may result in a disservice to the underrepresented population. We observe that this delayed impact can arise from differences in the *predicted* risk scores, not only the true risk scores. If the predicted risk distributions are different between populations due to disparate information loss, then enforcing parity-based fairness could still cause harm, even if the populations are identically distributed. We show that counteracting the information disparity by refining the underlying predictions used to choose a selection policy may help to counteract the effect of negative impacts.

6.3.1 Fair prediction setup

We consider a standard prediction setting where a decision maker, who we call the lender, has access to a calibrated predictor $p : \mathcal{X} \to [0,1]$; from the predicted risk score $p(x)$, the decision maker must choose whether to accept or reject the individual $x \in \mathcal{X}$, i.e. whether to give x a loan or not. For simplicity, we assume there are two disjoint subpopulations $A, B \subseteq \mathcal{X}$, and that predictors are calibrated over A and B separately.

When deciding how to select qualified individuals, the lender's goal is to maximize some expected utility. The utility function $u : [0,1] \to [-1,1]$ specifies the lender's expected utility from an individual based on their score and a fixed threshold[2] $\tau_u \in [0,1]$ as given in (6.16). When considering delayed impact, we will measure the expected impact per subpopulation. The impact function $\ell : [0,1] \to [-1,1]$ specifies the expected benefit to an individual from receiving a loan based on their score and a fixed threshold τ_ℓ also given in (6.16).

$$u(v) = v - \tau_u \qquad\qquad \ell(v) = v - \tau_\ell \tag{6.16}$$

[LDR$^+$18] models a risk-averse lender by assuming that $\tau_u > \tau_\ell$; that is, by choosing accepting individuals with $p(x) \in (\tau_\ell, \tau_u)$, the impact on subpopulations may improve beyond the lender's utility-maximizing policy.

To accept/reject individuals, the lender picks a (randomized) group-sensitive selection policy $f : [0,1] \times \{A, B\} \to [0,1]$ that selects individuals on the basis of their predicted score and their sensitive attribute. For every individual $x \in \mathcal{X}$, the probability that x is selected is given as $f(p(x), \mathcal{A}(x))$. As in [LDR$^+$18], we restrict our attention to threshold policies; that is, for sensitive attribute A (resp., B), there is some $\tau_A \in [0,1]$, such that $f(v, A)$ is given as $f(v, A) = 1$ if $v > \tau_A$, $f(v, A) = 0$ if $v < \tau_A$ and $f(v, A) = p_A$ for $v = \tau_A$, where $p_A \in [0,1]$ is a probability used to randomly break ties on the threshold. The motivation for focusing

[2] Assuming such an affine utility function is equivalent in expectation to assuming that the lender receives u_+ from repayments, u_- from defaults, and 0 from individuals that do not receive loans. In this case, the expected utility for score v is $vu_+ + (1-v)u_- = c_u \cdot u(v)$ for some constant c_u. A similar rationale applies to the individuals' impact function.

on threshold policies is their intuitiveness, widespread use, computational efficiency.[3]

Given this setup, we can write the expected utility $U^p(f)$ of a policy f based on a predictor p, that is calibrated on both subpopulations, A and B, as follows.

$$U^p(f) = \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}\left[\, x \in S \,\right] \cdot \left(\sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v,S) \cdot u(v) \right) \tag{6.17}$$

Recall, $\mathcal{R}_S^p(v) = \mathbf{Pr}_{x \sim S}\left[p(x) = v\right]$.

Similarly, the expected impact over the subpopulations $S \in \{A,B\}$ are given as

$$\mathrm{Imp}_S^p(f) = \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v,S) \cdot \ell(v). \tag{6.18}$$

Often, it may make sense to constrain the net impact to each group as defined in (6.18) to be positive, ensuring that the selection policies do not do harm as in [LDR+18].

The following quantities will be of interest to the lender when choosing a selection policy f as a function of p. First, the lender's overall utility $U(f)$ is given as in (6.17). Demographic parity, which serves as our running example, compares the selection rates $\beta_S = \mathbf{Pr}_{x \sim S}[x \text{ selected}]$, given formally as

$$\beta_S^p(f) = \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v,S).$$

Equalized opportunity and related notions compare the true positive rates and false positive rates. Recall, $\mathrm{TPR} = \mathbf{Pr}[x \text{ selected} | y = 1]$ and $\mathrm{FPR} = \mathbf{Pr}[x \text{ selected} | y = 0]$; in this context, we can rewrite these quantities as follows.

$$\mathrm{TPR}_S^p(f) = \frac{1}{r_S} \cdot \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v,S) \cdot v$$
$$\mathrm{FPR}_S^p(f) = \frac{1}{1 - r_S} \cdot \sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v,S) \cdot (1 - v),$$

where r_S represents the base rate of the subpopulation S; that is, $r_S = \mathbf{Pr}_{(x,y) \sim D_S}[y = 1]$.

[3] Indeed, without the restriction to threshold policies, many of the information-theoretic arguments become trivial. As $\tilde{p}$ is a refinement of p, we can always simulate decisions derived from p given $\tilde{p}$, but in general, we cannot do this efficiently.

Another quantity we will track is the positive predictive value, PPV $= \mathbf{Pr}[y = 1|x \text{ selected}]$.

$$\mathrm{PPV}_S^p(f) = \frac{1}{\beta_S^p(f)} \cdot \left(\sum_{v \in \mathrm{supp}(p)} \mathcal{R}_S^p(v) \cdot f(v, S) \cdot v \right)$$

For notational convenience, we may drop the superscript of these quantities when p is clear from the context.

6.3.2 Refinements in the service of fair prediction

Note that the quantities[4] described in Section 6.3.1 can be written as linear functions of $f(v, S)$. As such, we can formulate a generic policy-selection problem as a linear program where p controls many coefficients in the program. Recalling that different contexts may call for different notions of fairness, we consider a number of different linear programs the lender (or regulator) might choose to optimize. At a high-level, the lender can choose to maximize utility, minimize disparity, or maximize positive impact on groups, while also maintaining some guarantees over the other quantities. Across all such programs, we show how refining the predictor p used for determining the selection rule can improve the utility, parity, and impact of the optimal selection rule.

We will consider selection policies given a fixed predictor $p : \mathcal{X} \to [0, 1]$. We refer generically to the fairness quantities by some $h \in \{\beta, \mathrm{TPR}, \mathrm{FPR}\}$; rather than requiring equality, we will consider the disparity $\left|h_A^p(f) - h_B^p(f)\right|$ and in some cases, constrain it to be less than some constant ε. We also use t_i, t_u to denote lower bounds on the desired impact and utility, respectively. For simplicity's sake, we assume that B is the "protected" group, so we only enforce the positive impact constraint for this group; more generally, we could include an impact constraint for each group. Formally, we consider the following constrained optimizations.

Optimization 1

$$\begin{gathered} \max_f \ U^p(f) \\ \text{s.t. } \mathrm{Imp}_B^p(f) \geq t_i \\ \left|h_A^p(f) - h_B^p(f)\right| \leq \varepsilon \end{gathered}$$

(*Utility Maximization*)

Optimization 2

$$\begin{gathered} \min_f \ \left|h_A^p(f) - h_B^p(f)\right| \\ \text{s.t. } \mathrm{Imp}_B^p(f) \geq t_i \\ U^p(f) \geq t_u \end{gathered}$$

(*Disparity minimization*)

Optimization 3

$$\begin{gathered} \max_f \ \mathrm{Imp}_B^p(f) \\ \text{s.t. } U^p(f) \geq t_u \\ \left|h_A^p(f) - h_B^p(f)\right| \leq \epsilon \end{gathered}$$

(*Impact Maximization*)

[4] Excluding positive predictive value.

Lemma 6.14. *Let $h \in \{\beta, \mathrm{TPR}, \mathrm{FPR}\}$. Given a calibrated predictor $p : \mathcal{X} \to [0, 1]$, Optimization 1,2, and 3 are linear programs in the variables $f(v, S)$ for $v \in \mathrm{supp}(p)$ and $S \in \{A, B\}$. Further, for each program, there is an optimal solution f^* that is a threshold policy.*

We sketch the proof of the lemma. The fact that the optimizations are linear programs follows immediately from the observations that each quantity of interest is a linear function in $f(v, S)$. The proof that there is a threshold policy f^* that achieves the optimal value in each program is similar to the proof of Theorem 6.15 given below. Consider an arbitrary (non-threshold) selection policy f_0; let $h_{S,0} = h_S(f_0)$. The key observation is that for the fixed value of $h_{S,0}$, there is some other threshold policy f where $h^p_S(f) = h_{S,0}$ and $U^p(f) \geq U(f_0)$ and $\mathrm{Imp}^p_S(f) \geq \mathrm{Imp}_S(f_0)$. Leveraging this observation, given any non-threshold optimal selection policy, we can construct a threshold policy, which is also optimal.

We remark that our analysis applies even if considering the more general linear maximization:

Optimization 4

$$\max_f \quad \lambda_U \cdot U^p(f) + \lambda_I \cdot \mathrm{Imp}^p_B(f) - \lambda_\beta \cdot \left|h^p_A(f) - h^p_B(f)\right|$$

for any fixed $\lambda_U, \lambda_I, \lambda_\beta \geq 0$.[5] In other words, the arguments hold no matter the relative weighting of the value of utility, disparity, and impact.

Improving the cost of fairness. We argue that in all of these optimizations, increasing information through refinements of the current predictor on both the subpopulations A and B improves this value of the program. We emphasize that this conclusion is true for all of the notions of parity-based fairness we mentioned above. Thus, independent of the exact formulation of fair selection that policy-makers deem appropriate, information content of p is a key factor in determining the properties of the resulting selection rule. We formalize this statement in the following theorem.

Theorem 6.15. *Let $p, \tilde{p} : \mathcal{X} \to [0, 1]$ be two predictors that are calibrated on disjoint subpopulations $A, B \subseteq \mathcal{X}$. For any of the Optimization 1, 2, 3, 4 and their corresponding fixed*

[5] In particular, each of Optimizations 1, 2, and 3 can be expressed as an instance of Optimization 4 by choosing $\lambda_U, \lambda_I, \lambda_\beta$ to be the optimal dual multipliers for each program. We note that the dual formulation actually gives an alternate way to derive results from [LDR$^+$18]. Their main result can be restated as saying that there exist distributions of scores such that the dual multiplier on the positive impact constraint in Optimization 1 is positive; that is, without this constraint, the utility-maximizing policy will result in negative impact to group B.

parameters, let $OPT(p)$ denote their optimal value under predictor p. If $\tilde{p}$ is a refinement of p over A and B, then $OPT(\tilde{p}) \geq OPT(p)$ for Optimization 1, 3, 4 and $OPT(\tilde{p}) \leq OPT(p)$ for Optimization 2.

One way to understand Theorem 6.15 is through a "cost of fairness" analysis. Focusing on the utility maximization setting, let U^* be the maximum unconstrained utility achievable by the lender given the optimal predictions p^*. Let $\mathrm{OPT}(p)$ be the optimal value of Optimization 1 using predictions p; that is, the best utility a lender can achieve under a parity-based fairness constraint ($\varepsilon = 0$) and positive impact constraint ($t_i = 0$). If we take the cost of fairness to be the difference between these optimal utilities, $U^* - \mathrm{OPT}(p)$, then Theorem 6.15 says that by refining p to $\tilde{p}$, the cost of fairness decreases with increasing informativeness; that is, $U^* - \mathrm{OPT}(p) \geq U^* - \mathrm{OPT}(\tilde{p})$. This corollary of Theorem 6.15 corroborates the idea that in some cases the high perceived cost associated with requiring fairness might actually be due to the low informativeness of the predictions in minority populations. No matter what the true p^* is, this cost will decrease as we increase information content by refining subpopulations.

For $S \in \{A, B\}$, we use $\mathrm{TPR}_S^p(\beta)$ to denote the true positive rate of the threshold policy with selection rate β for the subpopulation S while using the predictor p.[6] Similarly, $\mathrm{PPV}_S^p(\beta)$, $\mathrm{FPR}_S^p(\beta)$ are defined. The following lemma, which plays a key role in each proof, shows that refinements broadly improve selection policies across these three statistics of interest.

Lemma 6.16. *If $\tilde{p}$ is a refinement of p on subpopulations A and B, then for $S \in \{A, B\}$, for all $\beta \in [0, 1]$,*

$$\mathrm{TPR}_S^{\tilde{p}}(\beta) \geq \mathrm{TPR}_S^p(\beta), \qquad \mathrm{FPR}_S^{\tilde{p}}(\beta) \leq \mathrm{FPR}_S^p(\beta), \qquad \mathrm{PPV}_S^{\tilde{p}}(\beta) \geq \mathrm{PPV}_S^p(\beta).$$

In particular, the proof of Theorem 6.15 crucially uses the fact that the positive predictive values, true positive rates, and false positive rates improve for *all* selection rates. Leveraging properties of refinements, the improvement across all selection rates guarantees improvement for any fixed objective. As we'll see, the proof actually tells us more: for *any* selection policy using the predictor p, there exists a threshold selection policy that uses the refined predictor $\tilde{p}$ and *simultaneously* has utility, disparity, and impact that are no worse than under p.

[6] Given a predictor, there is a bijection between selection rates and threshold policies.

Before giving the proofs, we observe that, paired with earlier sections of Chapter 6, Lemma 6.16 suggests another do-no-harm property of multi-calibrated predictors. In particular, if we post-process p into a refinement $\tilde{p}$ that satisfies multi-calibration over the class $\mathcal{C}$—and thus, multi-refinement over $\mathcal{P}^{\mathcal{C}}$—then the conditions of the lemma are satisfied for all $S \in \mathcal{C}$. For convenience, we state the corollary in terms of exact multi-calibrated predictors.

Corollary 6.17. *Suppose $\tilde{p}$ is a $\mathcal{C}$-multi-calibrated predictor that refines p. Then for all subpopulations $S \in \mathcal{C}$ and all selection rates $\beta \in [0,1]$,*

$$\mathrm{TPR}_S^{\tilde{p}}(\beta) \geq \mathrm{TPR}_S^{p}(\beta), \qquad \mathrm{FPR}_S^{\tilde{p}}(\beta) \leq \mathrm{FPR}_S^{p}(\beta), \qquad \mathrm{PPV}_S^{\tilde{p}}(\beta) \geq \mathrm{PPV}_S^{p}(\beta).$$

Proofs. Next, we prove Lemma 6.16 and Theorem 6.15.

Proof of Lemma 6.16. Note that for a fixed selection rate β, the quantities PPV, TPR, and FPR are maximized by selecting the top β-fraction of the individuals ranked according to the optimal predictor p^*. Recall, we can interpret a refinement as a "candidate" optimal predictor. Because $\tilde{p}$ refines p over A and B, we know that p is calibrated not only with respect to the true Bayes optimal predictor p^*, but also with respect to the refinement $\tilde{p}$ on both subpopulations. Imagining a world in which $\tilde{p}$ is the Bayes optimal predictor, the PPV, TPR, and FPR must be no worse under a threshold policy derived from $\tilde{p}$ compared to that of p by the initial observation. Thus, the lemma follows. □

Using Lemma 6.16, we are ready to prove Theorem 6.15.

Proof of Theorem 6.15. Let f be any threshold selection policy under the predictor p. Using f, we will construct a selection policy f' that uses the refined score distribution $\tilde{p}$ such that where $U^{\tilde{p}}(f') \geq U^{p}(f)$, $\mathrm{Imp}_B^{\tilde{p}}(f') \geq \mathrm{Imp}_B^{p}(f)$, and $h_A^{\tilde{p}}(f') = h_A^{p}(f)$ and $h_B^{\tilde{p}}(f') = h_B^{p}(f)$. Here, $h \in \{\beta, \mathrm{TPR}, \mathrm{FPR}\}$ specifies the parity-based fairness definition being used. Thus, taking f to be the optimal solution to any of the Optimizations 1, 2, 3, or 4, we see that f' is a feasible solution to the same optimization and has the same or a better objective value compared to f. Therefore, after optimization, objective values can only get better.

We separately construct f' for each fairness notion as follows:

(Demographic Parity) $h = \beta$:

For $S \in \{A, B\}$, let $\beta_S = \beta_S^p(f)$ be the selection rate of f in the population S. Let f' be the threshold policy that uses the predictor $\tilde{p}$ and achieves selection rates β_A and β_B in the subpopulations A and B, respectively. By Lemma 6.16, $\mathrm{PPV}_S^{\tilde{p}}(\beta_S) \geq \mathrm{PPV}_S^{p}(\beta_S)$ for

$S \in \{A, B\}$. The utility of the policy f' can be written as

$$\begin{aligned} U(f') &= \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \left(\sum_{v\in\text{supp}(\tilde{p})} \mathcal{R}_S^{\tilde{p}}(v) \cdot f'(v,S) \cdot v - \sum_{v\in\text{supp}(\tilde{p})} \mathcal{R}_S^{\tilde{p}}(v) \cdot f'(v,S) \cdot \tau_u \right) \\ &= \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \left(\beta_S \cdot (\text{PPV}_S^{\tilde{p}}(\beta_S) - \tau_u) \right) \\ &\geq \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \big(\beta_S \cdot (\text{PPV}_S^{p}(\beta_S) - \tau_u) \big) \\ &= U(f) \end{aligned}$$

Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

(Equalized Opportunity) $h = \text{TPR}$:

Let (β_A, β_B) be the selection rates of policy f on the subpopulations A and B. We know that $\text{TPR}_S^{\tilde{p}}(\beta_S) \geq \text{TPR}_S^{p}(\beta_S)$ $(S \in \{A, B\})$ through Lemma 6.16. Let f' be the threshold selection policy corresponding to a selection rates of $\beta'_S, (S \in \{A, B\})$ such that $\text{TPR}_S^{\tilde{p}}(\beta'_S) = \text{TPR}_S^{p}(\beta_S)$ $(\leq \text{TPR}_S^{\tilde{p}}(\beta_S))$. As the true positive rates increase with increasing selection rate, $\beta'_S \leq \beta_S$. The utility of the policy f' can be written as

$$\begin{aligned} U(f') &= \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \left(\sum_{v\in\text{supp}(\tilde{p})} \mathcal{R}_S^{\tilde{p}}(v) \cdot f'(v,S) \cdot v - \sum_{v\in\text{supp}(\tilde{p})} \mathcal{R}_S^{\tilde{p}}(v) \cdot f'(v,S) \cdot \tau_u \right) \\ &= \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \left(r_S \cdot \text{TPR}_S^{\tilde{p}}(\beta'_S) - \beta'_S \cdot \tau_u \right) \\ &\geq \sum_{S\in\{A,B\}} \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[x \in S] \cdot \big(r_S \cdot \text{TPR}_S^{p}(\beta_S) - \beta_S \cdot \tau_u \big) \\ &= U(f) \end{aligned}$$

Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

(Equalized False Positive Rate) $h = \text{FPR}$:

Let (β_A, β_B) be the selection rates of policy f on the subpopulations A and B. We know that $\text{FPR}_S^{\tilde{p}}(\beta_S) \leq \text{FPR}_S^{p}(\beta_S)$ $(S \in \{A, B\})$ through Lemma 6.16. Let f' be the threshold selection policy corresponding to a selection rates of $\beta'_S, (S \in \{A, B\})$ such that

$\mathrm{FPR}^{\tilde{p}}_S(\beta'_S) = \mathrm{FPR}^{p}_S(\beta_S)$ $(\geq \mathrm{FPR}^{\tilde{p}}_S(\beta_S))$. As the false postive rates increase with increasing selection rate, $\beta'_S \geq \beta_S$. The utility of the policy f' can be written as

$$\begin{aligned}
U(f') &= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}[x \in S] \cdot \left(\sum_{v \in \mathrm{supp}(\tilde{p})} \mathcal{R}^{\tilde{p}}_S(v) \cdot f'(v,S) \cdot v - \sum_{v \in \mathrm{supp}(\tilde{p})} \mathcal{R}^{\tilde{p}}_S(v) \cdot f'(v,S) \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}[x \in S] \cdot \left(\beta'_S - (1 - r_S) \cdot \mathrm{FPR}^{\tilde{p}}_S(\beta'_S) - \beta'_S \cdot \tau_u \right) \\
&= \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}[x \in S] \cdot \left(\beta'_S \cdot (1 - \tau_u) - (1 - r_S) \cdot \mathrm{FPR}^{\tilde{p}}_S(\beta'_S) \right) \\
&\geq \sum_{S \in \{A,B\}} \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}}[x \in S] \cdot \left(\beta_S \cdot (1 - \tau_u) - (1 - r_S) \cdot \mathrm{FPR}^{p}_S(\beta_S) \right) \\
&= U(f)
\end{aligned}$$

Similarly, we can show that the impact on the subpopulation B under f' is at least as good as under f.

This completes the proof of the theorem. □

Chapter Notes

Chapter 6 is based on [GKR19], a joint work with Sumegha Garg and Omer Reingold. The connection between refinement theory and multi-calibration in Section 6.2 is novel to this thesis, but is based on an informal observation made in the Discussions of [GKR19].

The influential work of [DHP+12] provided two observations that are of particular relevance to this chapter. First, [DHP+12] emphasized the pitfalls of hoping to achieve "fairness through blindness" by censoring sensitive information during prediction. Second, the work highlighted how enforcing broad-strokes demographic parity conditions—even if desired or expected in fair outcomes—is insufficient to imply fairness. This chapter provides further evidence for these perspectives.

A few works have (implicitly or explicitly) touched on the relationship between information and fairness. [CJS18] argues that discrimination may arise in prediction systems due to disparity in predictive power; they advocate for addressing discrimination through data collection. Arguably, much of the work on fairness in online prediction [JKMR16] can be seen as a way to gather information while maintaining fairness. From the computational economics literature, [KLMR18] presents a simple planning model that draws similar qualitative conclusions to this work, demonstrating the significance of trustworthy information as a key factor in algorithmic fairness.

Some works frame informativeness differently and arrive at qualitatively different conclusions. Specifically, the original work on delayed impact [LDR+18] suggests that some forms of misestimation (i.e. loss of information) may reduce the potential for harm from applying parity-based fairness notions. If the lender's predictor p is miscalibrated in a way that underestimates the quality of a group S, then increasing the selection rate beyond the global utility-maximizing threshold may be warranted. Other works [KRZ19, ILZ19] have investigated the role of hiding information through strategic signaling. In such settings, it may be strategic for a group to hide information about individuals in order to increase the overall selection rate for the group. These distinctions highlight the fact that understanding exactly the role of information in "fair" prediction is subtle and also depends on the exact environment of decision-making.

The present work can also be viewed as further investigating the tradeoffs between calibration and parity. Inspired by investigative reporting on the "biases" of the COMPAS recidivism prediction system [ALMK16], the incompatibility of calibration and parity-based notions of fairness has received lots of attention in recent years [Cho17, KMR17, PRW+17]. Perhaps counterintuitively, our work shows how to leverage properties of calibrated predictors to improve the disparity of the eventual decisions.

Outside the literature on fair prediction, our notions of information content and refinements are related to other notions from the fields of online forecasting and information theory. In particular, the idea of refinements was first introduced in [DF81]. The concept of information content of calibrated predictions is related to ideas from the forecasting literature [GBR07, GR07], including *sharpness* and *proper scoring rules* [Bri50]. The concept of a refinement of a calibrated predictor can be seen as a special case of Blackwell's informativeness criterion [Bla53, Cré82, DF81].

Chapter 7

Evidence-Based Rankings

In this final chapter, we study the learning-to-rank problem from the perspective of multi-group fairness. As in the rest of the thesis, we continue to assume that the learner has access to samples of individuals and their binary outcomes $(x, y) \sim \mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$. Our goal, however, will be to recover a ranking close to the ordering given by $p^*(x)$, even without access to direct pairwise comparisons, $p^*(x) < p^*(x')$. In this setting, a natural approach to ranking (both in theory and practice) is to start by learning a predictor $p : \mathcal{X} \to [0, 1]$, and then transform it into a ranking $r : \mathcal{X} \to [0, 1]$. When we reduce the problem of ranking to that of prediction, however, we may be concerned that we fail to maintain important properties, like fairness. We may worry that even if we learn a "fair" predictor, the resulting ranking may be unfair, due to the qualitative differences between the objective of ranking (relative ordering of individuals) and that of prediction (absolute loss).

In this chapter, we investigate these concerns, presenting a novel framework for reasoning about fairness in the context of rankings. We begin by formalizing the notion of ranking in our context, and then introduce some natural fairness desiderata in rankings. Our first notion—domination-compatibility—focuses on the relative ranking of groups: consider a pair S, T of disjoint sets, and suppose that a larger fraction of the members of S have positive outcomes than than do the members of T; any ranking that puts all of the individuals from T ahead of those in S seems blatantly discriminatory. Domination-compatibility precludes this form of systematic misrepresentation of groups. Applying the multi-group perspective, we require compatibility simultaneously across a rich class $\mathcal{C}$ of possibly-intersecting subpopulations. The next notion—evidence-consistency—makes the connection to prediction explicit: a ranking r is $\mathcal{C}$-evidence-consistent if it arises as the induced ranking of a $\mathcal{C}$-multi-accurate predictor $\tilde{p}$. In this sense, the consistent predictor $\tilde{p}$ demonstrates consistency with the "evidence" provided by the statistical tests defined by the subpopulations in $\mathcal{C}$. We show that this global notion of evidence-consistency implies the pairwise notion of domination-compatibility.

Despite the appeal of these notions of evidence-based fairness, we show that enforcing domination-compatibility and evidence-consistency over a predefined class of subpopulations $\mathcal{C}$ may allow for insidious forms of discrimination in the ranking. This weakness is similar to the weakness of multi-accuracy in comparison to multi-calibration. Drawing inspiration from multi-calibration, we redefine both notions of evidence-based rankings in a way to protect subpopulations defined by the ranking itself. Somewhat surprisingly, when the notions protect these self-referential subpopulations there is tight equivalence between them: domination-compatibility, evidence-consistency, and multi-calibration all encode essentially the same notion of fairness in rankings.

7.1 Rankings and Predictors

In this chapter, our goal will be to learn rankings of individuals from a discrete universe $\mathcal{X}$ over a fixed, but unknown distribution $\mathcal{D}$. Our goal will be to rank in accordance with the underlying true $p^*(x) = \mathbf{Pr}[y = 1|x]$. Nevertheless, we will continue to assume a learning model where we may access individual-outcome pairs $(x, y) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ from the underlying distribution. Specifically, in this learning from outcomes model, we do not get to see $p^*(x)$, so cannot easily compare $p^*(x)$ and $p^*(x')$ for individuals $x, x' \in \mathcal{X} \times \mathcal{X}$.

In the case where we have a finite collection 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" $\pi^{-1}(n)$. When we want to learn rankings from a small set of samples, however, we need to generalize the idea of a permutation-based ranking to account for unseen individuals. As with predictors, we will consider rankings as functions from $\mathcal{X}$ to $[0, 1]$, with discrete and finite support. Formally, we require the following property to be a ranking function.

Definition 7.1 (Ranking)**.** *A function $r : \mathcal{X} \to [0, 1]$ is a ranking over $\mathcal{D}$ if for all $\tau \in$* supp(r)

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

We denote by $\mathcal{R} \subseteq \{r : \mathcal{X} \to [0,1]\}$ the set of all ranking functions.

A ranking function is effectively a CDF-style function over the individuals (except with a strict inequality). As with the permutation-based ranking, the top-ranked individuals will receive minimum rank value (in this case, $r(x) = 0$), with the magnitude of the rank increasing as we moved down the ranking. 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 $\mathcal{D}$ will have $r(x) \leq \tau$. In other words, $r(x)$ specifies the quantile of x under the ranking r.

Definition 7.2 (Quantiles according to a ranking)**.** *For a subset $S \subseteq \mathcal{X}$, given a ranking $r \in \mathcal{R}$, the quantiles of S according to r partition S as*

$$\mathcal{Q}_S(r) = \{S^r_\tau : \tau \in \mathrm{supp}(r)\}$$

where $S^r_\tau = \{x : r(x) = \tau\}$. We let $\mathcal{Q}(r) = \mathcal{Q}_{\mathcal{X}}(r)$ denote the quantiles of $\mathcal{X}$.

This notion of ranking has the appealing property that it does not require the ranking to distinguish between every pair of individuals when there is not enough information. In particular, the quantiles according to a ranking $r \in \mathcal{R}$ define equivalence classes of individuals according to their rank $r(x)$. Technically, these properties are essential for learning meaningful rankings from a small set of data.

7.1.1 From Predictors to Rankings

Every predictor $p : \mathcal{X} \to [0,1]$ yields a natural ranking $r^p \in R$ that orders individuals in order according to their values under p, defined as follows. Due to our convention—that the lowest magnitude rank is the "top" of the ranking—the induced ranking will give the individuals with the largest p-values the smallest r^p-values.

Definition 7.3 (Induced ranking)**.** *Given a predictor $p : \mathcal{X} \to [0,1]$, the induced ranking $r^p \in \mathcal{R}$ is defined as follows.*

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

Note that a predictor p and its induced ranking r^p can be related by a transformation between their supports. Specifically, the partition of any subset $S \subseteq \mathcal{X}$ according to the level sets of p gives rise to the quantiles of S according to r^p. That is, for each $v \in \mathrm{supp}(p)$, there exists some $\tau_v \in \mathrm{supp}(r^p)$ such that

$$\{x : p(x) = v\} = \{x : r^p(x) = \tau_v\}\,.$$

Given a predictor that we believe to be accurate or fair (e.g., a multi-calibrated $\tilde{p}$), we may want to obtain and understand its induced ranking. While r^p is well-defined for every predictor p, without some access to the distribution $\mathcal{D}$ over individuals, we may not be able to compute r^p. Still, we argue that a small set of unlabeled samples suffice to transform p into r^p. Given access to p and sufficiently many samples $x' \sim \mathcal{D}$, we can implement this comparison oracle and accurately recover the induced ranking r^p, in the following sense.

Proposition 7.4. *Let $\alpha, \beta > 0$. 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 $m \geq \dfrac{2\log(2/\alpha\beta)}{\alpha^2}$ unlabeled samples $x_1, \ldots, x_m \sim \mathcal{D}_{\mathcal{X}}$, produces a ranking $\hat{r}^p : \mathcal{X} \to [0,1]$ such that with probability at least $1-\beta$, for all $x \in \mathcal{X}$,*

$$|r^p(x) - \hat{r}^p(x)| \leq \alpha$$

and for every $\mathcal{X}_\tau \in \mathcal{Q}(r^p)$, there exists some $\mathcal{X}_{\tau'} \in \mathcal{Q}(\hat{r}^p)$, such that $\mathcal{X}_\tau \subseteq \mathcal{X}_{\tau'}$.

The final guarantee says that the approximation will preserve the quantiles of r^p, but possibly merge them; equivalently, if $r^p(x) = r^p(x')$, then $\hat{r}^p(x) = \hat{r}^p(x')$.

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, $\mathbf{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[X_\tau] = \tau$. 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 $T = \left\{0, \alpha/2, \alpha, \ldots, \left\lfloor \frac{2}{\alpha} \right\rfloor \cdot \alpha/2\right\}$ be a set of $2/\alpha$ 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 $\alpha/2$ away from the true expectation.

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

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

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.

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

Note that $\hat{r}^p(x)$ only depends on the value $p(x)$, or equivalently the value of $r^p(x)$; thus, for all $\mathcal{X}_\tau \in \mathcal{Q}(r^p)$, every $x \in \mathcal{X}_\tau$ will be mapped to the same $\hat{r}^p(x)$ value, establishing the final property on the quantiles.

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

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[r^p(x_i) < \tau_-] \quad &\leq \quad \hat{r}^p(x) \quad \leq \quad \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[r^p(x_i) < \tau_+] && (7.1)\\
\mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < \tau_-] - \alpha/2 \quad &\leq \quad \hat{r}^p(x) \quad \leq \quad \mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < \tau_+] + \alpha/2 && (7.2)\\
\mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < r^p(x)] - \alpha \quad &\leq \quad \hat{r}^p(x) \quad \leq \quad \mathop{\mathbf{Pr}}_{x' \sim \mathcal{D}_\mathcal{X}}[r^p(x') < r^p(x)] + \alpha && (7.3)
\end{aligned}$$

where (7.1) follows by the assumption that $r^p(x) \in [\tau_-, \tau_+]$; (7.2) follows by the accuracy of the empirical estimates for all $\tau \in T$; and (7.3) follows by the fact that $\tau_+ - \tau_- = \alpha/2$ for all consecutive $\tau_- < \tau_+ \in T$. Thus, the empirical estimate $\hat{r}^p(x)$ will be within α 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 (including unseen) $x \in \mathcal{X}$. Further, note that all of the arguments used to prove Proposition 7.4 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 7.5. *Suppose $\alpha, \beta, \tau > 0$. Given access to a predictor $p : \mathcal{X} \to [0,1]$, a subset $S \subseteq \mathcal{X}$, and $\tilde{O}\left(\log(1/\beta)/\alpha^2\right)$ 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 $\tau' \in [\tau - \alpha, \tau + \alpha]$ with probability at least $1 - \beta$.*

Determining the top percentiles of individuals within a subpopulation may be particularly useful for affirmative action mechanisms based on selecting the top-qualified individuals based on their standing within a stratum (e.g., high school graduating class, mother's education level, etc.) rather than based on absolute qualification, à la [Roe98].

7.1.2 From Rankings to Consistent Predictors

The previous section demonstrated that given a predictor and an unlabeled set of individuals, we can recover the ranking induced by the predictor. Perhaps less obviously, in this section we show that rankings also give rise to predictors. Specifically, every ranking will have a set of consistent predictors (that respect the ordering of the ranking), and further, given a small set of labeled data, we can recover a predictor that is close to the optimal consistent predictor. Given a ranking r, we can imagine a set of predictors that would respect the ordering given by the ranking.

Definition 7.6 (Consistency with a ranking)**.** *For a ranking $r : \mathcal{X} \to [0,1]$, a predictor $p : \mathcal{X} \to [0,1]$ is r-consistent (or consistent with r) if for all $x, x' \in \mathcal{X} \times \mathcal{X}$:*

- *if $r(x) < r(x')$, then $p(x) \geq p(x')$, 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 r-consistent predictors.

Note that by convention r-consistency allows for the level-sets that arise from the predictor to merge distinct quantiles. For instance, any constant predictor—that makes no distinctions between individuals—is consistent with any ranking $r \in \mathcal{R}$. While there are generally many consistent predictors that form the collection $\mathcal{P}(r)$, one a natural predictor p^r we can derived from r gives the expected value of p^* on each quantile. Naturally, the predictor will be well-calibrated, so we call p^r the calibration of r.

Definition 7.7 (Calibration of a ranking)**.** *For a ranking $r \in \mathcal{R}$, the calibration of r is the predictor $p^r : \mathcal{X} \to [0,1]$ where for each $\mathcal{X}_\tau \in \mathcal{Q}(r)$, for all $x \in \mathcal{X}_\tau$,*

$$p^r(x) = \mathop{\mathbf{E}}_{x' \sim \mathcal{D}_{\mathcal{X}_\tau}}\left[\ p^*(x')\ \right] = \mathop{\mathbf{E}}_{x' \sim \mathcal{D}}\left[\ p^*(x') \mid r(x') = \tau\ \right].$$

Using arguments as in Chapter 3, we can estimate the calibration of a ranking using a small set of labeled samples. In particular, if the quantiles are not too small, then we can obtain a close approximation of p^r for each value $v \in \mathrm{supp}(p^r)$.

Proposition 7.8. *Suppose $r \in \mathcal{R}$ is a ranking such that $\mathbf{Pr}_{\mathcal{D}}[\ r(x) = \tau\ ] \geq \gamma$ for each $\tau \in \mathrm{supp}(r)$. Then, there exists an efficient algorithm that, given oracle access to r and $m \geq \Omega\left(\frac{\log(1/\alpha\beta\gamma)}{\alpha^2\gamma}\right)$ labeled samples $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, produces a predictor $\hat{p}^r : \mathcal{X} \to [0,1]$ such that for all $x \in \mathcal{X}$*

$$|p^r(x) - \hat{p}^r(x)| \leq \alpha$$

with probability at least $1 - \beta$.

The proposition follows directly by an application of Hoeffding's inequality. Amongst all of the possible consistent predictors, we argue that when r is a highly accurate ranking (in a particular sense we define next), the calibration p^r is the most accurate predictor we can obtain using r and labeled samples alone.

Recovering a ranking. Suppose we want to recover a ranking close to the induced ranking p^*. We need to be a bit careful about how we measure "recovery" of the ranking; in particular, very small changes in the underlying Bayes risk may introduce large differences in the resulting numerical values of the induced ranking. Such small changes may not be detectable statistically from a small sample. Intuitively, however, if p^* and $\tilde{p}$ are indistinguishable, then our measure of quality of a ranking should not distinguish between the induced rankings $r^{\tilde{p}}$ and r^{p^*} that arise from these predictors. Information-theoretically, we can measure distinguishability using the statistical distance defined as follows.

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

We introduce a notion of adjacency to capture the idea that a ranking arises as the induced ranking of some predictor that is statistically close to the optimal predictor.

Definition 7.9 (Adjacency)**.** *A ranking r is ε-adjacent to p^* if there exists an r-consistent predictor $p_r \in \mathcal{P}(r)$ such that $\|p_r - p^*\|_1 \leq \varepsilon$.*

Leveraging this notion of closeness, we show that the calibration of a ranking is a good approximation to the best possible consistent predictor. Specifically, given an ε-adjacent ranking r, we show that the calibration p^r is within 2ε of p^*.

Proposition 7.10. *For any $r \in \mathcal{R}$, let p^r denote the calibration of r. If r is ε-adjacent to p^* for some $\varepsilon \geq 0$, then*

$$\|p^* - p^r\|_1 \leq 2\varepsilon.$$

Proposition 7.10 follows as a consequence of a more general lemma.

Lemma 7.11. *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,*

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

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

$$\begin{aligned}\left\|p^{\mathcal{S}} - p_0^{\mathcal{S}}\right\|_1 &= \sum_{i \in [t]} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S_i] \cdot \left| \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{S^i}}[p^*(x)] - v_i \right| \\ &\leq \sum_{i \in [t]} \mathop{\mathbf{Pr}}_{x \sim \mathcal{X}}[x \in S_i] \cdot \mathop{\mathbf{E}}_{x \sim \mathcal{D}_{S_i}}[|p^*(x) - v_i|] \qquad (7.4) \\ &= \left\|p^* - p_0^{\mathcal{S}}\right\|_1\end{aligned}$$

where (7.4) follows by Jensen's inequality. □

With this lemma in place, we can easily prove Proposition 7.10. *Proof of Proposition 7.10.* For a ranking $r \in R$ that is ε-adjacent to p^*, let $p_1 = \operatorname{argmin}_{p \in \mathcal{P}(r)} \|p - p^*\|_1$.

Note that by r-consistency, we know that p_0 is piecewise constant on the quantiles of r. Thus, by applying Lemma 7.11, we can conclude

$$\|p_r - p^*\|_1 \leq 2 \cdot \|p_0 - p^*\|_1 .$$

Thus, $\|p_r - p^*\|_1 \leq 2\varepsilon$. □

We argue that in any learning model that only has access to binary samples $(x, y) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, the factor of 2 loss between the adjacency and the ℓ_1-distance of the recovered predictor is actually optimal.

Observation 7.12 (Informal)**.** *For any $c < 2$, there is an $\varepsilon > 0$ and a distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, such that given access to a ranking $r \in \mathcal{R}$ that is ε-adjacent to p^* and a bounded number of labeled samples $(x, y) \sim \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, no algorithm can produce a predictor p where*

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

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. □

In all, we have shown that a highly accurate ranking plus a set of labeled samples suggest an accurate predictor. Specifically, in the small error regime, the calibration of a ranking achieves the information-theoretic optimal approximate predictor. This formalizes the idea that a good understanding of the relative risk of individuals plus historical evidence may translate into an understanding of the absolute risk of individuals. Next, we introduce fairness notions for rankings, motivated by the common setting where highly-accurate rankings may not be attainable.

7.2 Domination-Compatibility and Evidence-Consistency

In this section, we introduce formally the notions of evidence-based fairness in rankings. The notions are designed with two competing goals in mind: to protect individuals from systematic misrepresentation within a ranking; and to allow for efficient learning from a small set of labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. As with our earlier notions of fairness developed for prediction, we settle on a multi-group perspective that aims to protect significant subpopulations in a way that can be efficiently audited from data through statistical queries.

Domination-Compatibility. 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 the idea that a ranking r gives preference to S over T, which we call domination.

Definition 7.13 (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 "better than" τ is at least as large as the corresponding fraction of individuals in T, up to an additive slack of γ. The notion of domination can be viewed as an approximate form of CDF-domination (again, up to the strict inequality).

Intuitively, there is a natural combinatorial interpretation of the domination condition in terms of matchings. In the special case where S and T are disjoint sets of equal cardinality and the distribution of interest $\mathcal{D}$ 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, $r(x) \leq r(m(x))$. Definition 7.13 generalizes this notion allowing comparison between S and T that are arbitrarily-intersecting subsets over arbitrary discrete probability densities. We argue that domination formally captures the intuition that a ranking strongly prefers one subset over another. 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 7.14. *For any subsets $S, T \subseteq \mathcal{X}$, if S γ-dominates T in r, then for every $p \in \mathcal{P}(r)$,*

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

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

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

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 $\mathbf{E}[X] =$

$\int_{v \geq 0} \mathbf{Pr}[\, X > v \,]\, dv$.

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

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

Lemma 7.14 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 in r for some small γ. Then, Lemma 7.14 show that no r-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.

A baseline notion of fairness for a ranking r would be that r does not exhibit any such blatant reversals for pairs of meaningful sets. Applying the multi-group perspective, we require this domination-compatibility to hold for every pair from a rich collection $\mathcal{C}$ of subpopulations.

Definition 7.15 (Domination-compatibility)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations 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*

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

For $S, T \in \mathcal{C}$, 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 differences in p^*.

Evidence-Consistency. A key reason that rankings r which violate the domination-compatibility constraints seem so objectionable is that there does not exist any r-consistent predictor $p \in \mathcal{P}(r)$ that exhibits the true expectations on the sets $S \in \mathcal{C}$. This observation motivates a notion of fair rankings from the perspective of consistent predictors. Specifically, the notion—evidence-consistency—goes a step further than domination-compatibility and requires that there exists an r-consistent predictor that exhibits the correct expectations for every subset in the collection.

Definition 7.16 (Evidence-Consistency)**.** *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations and $\alpha \geq 0$. A ranking $r \in \mathcal{R}$ is $(\mathcal{C}, \alpha)$-evidence-consistent if there exists an r-consistent predictor $\tilde{p} \in \mathcal{P}(r)$ where for every $S \in \mathcal{C}$,*

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

In other words, a ranking r is evidence-consistent with respect to a class $\mathcal{C}$ if there is an r-consistent predictor $\tilde{p} \in \mathcal{P}(r)$ 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^* (i.e., it respects the *evidence* about p^* in hand).

Essentially, by contrapositive of Lemma 7.14, we can see that evidence-consistency implies domination-compatibility. That is, 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 members 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.

Proposition 7.17. *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 $\tilde{p} \in P(r)$ such that

$$\underset{x \sim \mathcal{D}_S}{\mathbf{E}} [\ p^*(x)\ ] \geq \underset{x \sim \mathcal{D}_S}{\mathbf{E}} [\ \tilde{p}(x)\ ] - \alpha$$

$$\underset{x \sim \mathcal{D}_T}{\mathbf{E}} [\ p^*(x)\ ] \leq \underset{x \sim \mathcal{D}_T}{\mathbf{E}} [\ \tilde{p}(x)\ ] + \alpha$$

Further, by Lemma 7.14, because S γ-dominates T, we know that

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

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 expectations of p^* over S and T satisfy the domination-compatibility requirement with additive 2α. □

In fact, a simple construction shows that the implication is strict, showing that when the sets we wish to protect are predefined by $\mathcal{C}$, then evidence-consistency is a strictly stronger notion than domination-compatibility.

Observation 7.18. *There exist collections of subpopulations $\mathcal{C}$, distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, and approximation parameter $\alpha > 0$, such that the set of $(\mathcal{C},\alpha)$-evidence-consistent rankings is a strict subset of $(\mathcal{C},0)$-domination-compatible rankings.*

Proof Sketch. Suppose we take $\mathcal{C} = \{S, T\}$ such that $S = A\cup B$ and $T = B\cup C$ where A, B, C satisfy the following properties.

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x\in A\ ] = 1-2\varepsilon \qquad \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x\in B\ ] = \varepsilon \qquad \mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}[\ x\in C\ ] = \varepsilon$$

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_A}[\ p^*(x)\ ] = 1.0 \qquad \mathop{\mathbf{E}}_{x\sim\mathcal{D}_B}[\ p^*(x)\ ] = 1.0 \qquad \mathop{\mathbf{E}}_{x\sim\mathcal{D}_C}[\ p^*(x)\ ] = 0.0$$

Thus, we can conclude that

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}[\ p^*(x)\ ] = 1.0 \qquad \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}[\ p^*(x)\ ] = 0.5$$

Consider the ranking that ranks all of A first, then $T = B\cup C$ at the same ranking.

$$r(x) = \begin{cases} 0.0 & x\in A \\ 1-2\varepsilon & x\in T \end{cases}$$

We claim that r is $(\mathcal{C},0)$-domination-compatible but not $(\mathcal{C},\alpha)$-evidence-consistent for any $3\alpha < \varepsilon < 1/9$. Verifying $(\mathcal{C},0)$-domination-compatibility is straightforward, and mainly requires verifying that the fact T γ-dominates S for $\gamma > 1-2\varepsilon$ implies a trivial compatibility constraint.

To see that the ranking does not satisfy $(\mathcal{C}, \alpha)$-evidence consistency, we show that any consistent predictor $\tilde{p}$ must give values to $x \in B$ such that either $\mathbf{E}_{\mathcal{D}_S}[\ \tilde{p}(x)\ ] \ll \mathbf{E}_{\mathcal{D}_S}[\ p^*(x)\ ]$ or $\mathbf{E}_{\mathcal{D}_T}[\ \tilde{p}(x)\ ] \gg \mathbf{E}_{\mathcal{D}_T}[\ p^*(x)\ ]$. The claim follows by contradiction, assuming a $\tilde{p}$ satisfying consistency on S and deriving that $\tilde{p}$ cannot be consistent on T. □

This construction actually highlights the key conceptual difference between the guarantees of domination-compatibility and evidence-consistency. The proposed ranking actually identifies accurately a sizeable portion of the top-ranked individuals, but makes a mistake on a small portion. This small, local mistake (that might be tolerated by most notions of approximate recovery in prediction) actually makes it impossible for the ranking to be globally consistent. In this way, the evidence-consistency constraints—that require a singular global explanation $\tilde{p}$ of the ranking based on the available evidence about p^*—reject rankings that only satisfy the domination-compatibility constraints locally.

7.2.1 Learning Evidence-Based Rankings

With these fairness notions for rankings in place, we turn our attention to learning such rankings. By Proposition 7.17, an algorithm to learn evidence-consistent rankings suffices to obtain domination-compatible rankings. Thus, we focus on evidence-consistency. Towards this goal, we reexamine the definition of evidence-consistency. The notion requires consistency with a predictor $\tilde{p}$ that satisfies the correct expectations over all subpopulations for some $S \in \mathcal{C}$. In other words, the predictor $\tilde{p}$ must be multi-accurate. Specifically, we can characterize evidence-consistency as follows.

Proposition 7.19. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations and $\alpha \geq 0$.*

- *If $\tilde{p} : \mathcal{X} \to [0,1]$ is a $(\mathcal{C}, \alpha)$-multi-accurate predictor, its induced ranking $r^{\tilde{p}}$ is $(\mathcal{C}, \alpha)$-evidence-consistent.*
- *If $r \in \mathcal{R}$ is a $(\mathcal{C}, \alpha)$-evidence-consistent ranking, then it is the induced ranking of a $(\mathcal{C}, \alpha + \varepsilon)$-multi-accurate predictor, for any constant $\varepsilon > 0$.*

Proof. First, we observe that the induced ranking of a $(\mathcal{C}, \alpha)$-multi-accurate predictor is $(\mathcal{C}, \alpha)$-evidence-consisent. This implication follows immediately by the definition of multi-accuracy and the fact that for any predictor p, the predictor is consistent with its induced ranking; i.e., $p \in \mathcal{P}(r^p)$.

Next, suppose $r \in \mathcal{R}$ is $(\mathcal{C}, \alpha)$-evidence-consistent and consider the $(\mathcal{C}, \alpha)$-multi-accurate $\tilde{p} \in \mathcal{P}(r)$ guaranteed by evidence-consistency. We will show that r arises as the induced ranking of some predictor p that is statistically close to $\tilde{p}$. We define p as follows: for each $\tau \in \text{supp}(r)$, designate a tie-breaking value $\xi_\tau \in [-\varepsilon, \varepsilon]$ satisfying the following properties:

- for each $x \in \mathcal{X}$ such that $r(x) = \tau$, let $p(x) = \tilde{p}(x) + \xi_\tau$
- for each $\tau \in \text{supp}(r)$, there exists some unique $v_\tau \in \text{supp}(p)$ such that $\{x \in \mathcal{X} : r(x) = \tau\} = \{x \in \mathcal{X} : p(x) = v_\tau\}$
- for every $x, x' \in \mathcal{X} \times \mathcal{X}$, $r(x) < r(x')$ if and only if $p(x) > p(x')$.

The values of p move by at most ε from those of $\tilde{p}$, so p is $(\mathcal{C}, \alpha+\varepsilon)$-multi-accurate. Further, the p maintains the distinctions in the original ranking r; by construction, the level sets of p are equal to the quantiles according to r. Thus, the induced ranking $r^p = r$. □

As an immediate corollary of Proposition 7.19, we obtain an algorithm for learning evidence-consistent rankings: first, we apply Algorithm 7 to learn a $(\mathcal{C}, \alpha)$-multi-accurate predictor $\tilde{p}$; then, we apply the transformation from Proposition 7.4 on $\tilde{p}$ to obtain the induced ranking $r^{\tilde{p}}$. By the first direction of Proposition 7.19, we know that $r^{\tilde{p}}$ will be $(\mathcal{C}, \alpha)$-evidence-consistent.

7.3 Evidence-Based Rankings and Multi-Calibration

The results of Section 7.2 establish that the strength of evidence-consistency hinges on the expressiveness of $\mathcal{C}$; the richer $\mathcal{C}$ is, the stronger the protections provided by consistency with the actual expectations in sets in $\mathcal{C}$. In this section, we argue that approaches that only protect a *predefined* collection of subpopulations can leave the door open to abuses—including ones that we show can, in fact, be audited from a small set of labeled data. In this section, we build up stronger notions of both domination-compatibility and evidence-consistency. We show tight connections between rankings which satisfy these stronger "reflexive" notions—that incorporate sets to protect based on the quantiles of the ranking itself—and multi-calibrated predictors.

Notation for quantiles. Throughout this section, we will continue to use the notational convention that for a ranking $r \in \mathcal{R}$ and a subset $S \subseteq \mathcal{X}$,

$$S_\tau^r = \{x \in S : r(x) = \tau\}.$$

The ranking $r \in \mathcal{R}$ is typically clear from context. In such cases, we drop the explicit reference to r, and simply denote these quantiles as S_τ. Thus, we can express the quantiles according to a ranking $r \in \mathcal{R}$ as

$$\mathcal{Q}(r) = \{\mathcal{X}_\tau : \tau \in \operatorname{supp}(r)\}.$$

Attacking evidence-consistent rankings. While a seemingly-strong notion, evidence-consistency is vulnerable to subtle forms of manipulation. One such form of manipulation can be developed by revisiting the proof of Proposition 7.19. While it was immediate to see that the induced ranking of a multi-accurate predictor was evidence-consistent, arguing the reverse implication required more care. By definition, we know that if r is $(\mathcal{C}, \alpha)$-evidence-consistent, then there is some multi-accuate $\tilde{p} \in \mathcal{P}(r)$ that is consistent with r; however, this does not mean that $\tilde{p}$ induces the ranking specified by r. Recall that by the definition of r-consistency, the mapping from the quantiles according to r to the level sets of $\tilde{p}$ may not be injective. For instance, in general, for any $x \in \mathcal{X}$,

$$\left\{x' \in \mathcal{X} : r(x') = r(x)\right\} \neq \left\{x' \in \mathcal{X} : \tilde{p}(x') = \tilde{p}(x)\right\}.$$

In other words, evidence-consistency may allow for a ranking r to make distinctions between sets of individuals, even if the "witness" predictor $\tilde{p}$ does not make distinctions.

To highlight the potential weakness of evidence-consistency, consider the following example. Suppose $\mathcal{C}$ has two disjoint, equally-sized subpopulations S and T, and the learner is given access to a multi-accurate predictor $\tilde{p}$ where

$$\tilde{p}(x) = \begin{cases} 0.8 & x \in S \\ 0.5 & x \in T \end{cases}$$

We know that the induced ranking $r^{\tilde{p}}$ will be $(\mathcal{C}, \alpha)$-evidence-consistent. But suppose a manipulative learner wants to promote the ranking of the individuals in T. Consider the following adversarial ranking: the learner splits T in half into $T_0, T_T \subseteq T$ and defines r as

$$r(x) = \begin{cases} 0.0 & x \in T_0 \\ 0.25 & x \in S \\ 0.75 & x \in T_1 \end{cases}$$

This ranking puts half of the individuals in T at the top of the ranking—above all of S—despite the fact that the evidence in $\tilde{p}$ suggests that individuals in T are on-average less qualified than those in S.

Still, we claim that r satisfies $(\mathcal{C}, 0)$-evidence-consistency. In particular, consider the predictor p that gives

$$p(x) = \begin{cases} 1.0 & x \in T_0 \\ 0.8 & x \in S \\ 0.0 & x \in T_1 \end{cases}$$

This predictor p is is r-consistent and agrees in expectation over all of the subpopulations defined by $\mathcal{C}$. Such adversarial manipulation of evidence-consistency is possible regardless of the structure of p^* within S and T. Indeed, this example exploits the fact that the relevant subpopulations T_0 and T_1 are not included in $\mathcal{C}$. Because the protections evidence-consistency requires are predefined by $\mathcal{C}$, it cannot provide guarantees about sets that are defined by the ranking itself. Still, with the ranking in hand, T_0 and T_1 are identifiable; they are quantiles defined by r. 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.

7.3.1 Ordering the quantiles via domination-compatibility

These examples demonstrate that while evidence-consistency provides strong overall protections for the sets in $\mathcal{C}$, it provides limited guarantees to sets defined by r itself, which may intersect nontrivially with the sets in $\mathcal{C}$. This observation motivates enforcing some notion of consistency to ensure the quantiles of r are ordered in accordance with the evidence about their quality. We argue that a ranking that satisfies domination-compatibility over its quantiles satisfies a certain approximate ordering property.

Lemma 7.20. *Let $r \in \mathcal{R}$ be ranking. Suppose $\tau, \tau' \in \operatorname{supp}(r)$ and $\tau \leq \tau'$. Then, for any $S_\tau \subseteq \mathcal{X}_\tau$ and $T_{\tau'} \subseteq \mathcal{X}_{\tau'}$, S_τ 0-dominates $T_{\tau'}$.*

Proof. The proof of the lemma follows immediately from the definition of quantiles and γ-domination. We argue that for all thresholds $\sigma \in [0, 1]$

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

By the fact that the ranking r is constant on each quantile, the statement is equivalent to

$$\mathbf{1}[\ \tau \leq \sigma\ ] \geq \mathbf{1}\left[\ \tau' \leq \sigma\ \right],$$

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

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

Corollary 7.21. *Suppose a ranking $r \in \mathcal{R}$ is $(\mathcal{Q}(r), \alpha)$-domination-compatible. Then, for all $\tau < \tau' \in \operatorname{supp}(r)$,*

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

Note that requiring domination-compatibility with respect to $\mathcal{Q}(r)$ is fundamentally different than requiring it with respect to a fixed, predefined class $\mathcal{C}$. In particular, when we impose constraints defined by $\mathcal{Q}(r)$, these self-referential constraints change as a function of the ranking in question. Note that our motivating examples—highlighting the weakness of evidence-consistency—failed to satisfy domination-compatibility with respect to the quantiles.

With this mind, one way to augment the notions of domination-compatibility and evidence-consistency from Section 7.2 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 Propostion 7.17. Such a notion is strong enough to mitigate the concerns raised in the examples so far, but still may not be enough. Specifically, the attacks we've shown can all be "scaled down" to work, not at the level of $\mathcal{X}$, but within the subpopulations $S \in \mathcal{C}$. With some care, it is possible to construct examples demonstrating that simply adding the quantiles to the set $\mathcal{C}$ may still suffer from undesirable transpositions of subgroups within the sets defined in $\mathcal{C}$. Thus, we turn our attention to even stronger protections.

7.3.2 Incorporating the quantiles into evidence-based notions

In this section, we show a way to incorporate the quantiles to provide a much stronger guarantee. Rather than protecting the union of the set system $\mathcal{C}$ with the quantiles $\mathcal{Q}(r)$, we could protect the intersections of sets $S \in \mathcal{C}$ and each $\mathcal{X}_\tau \in \mathcal{Q}(r)$. Concretely, we define "reflexive" variants of domination-compatibility and evidence-consistency that strengthen the guarantees to include these subpopulations defined by the ranking itself. The protections that arise from such evidence-based fairness in rankings are syntactically similar to the protections of multi-calibration in predictors. In fact, we show that the connection between these notions is quite deep: when we allow the collection of subpopulations to depend on the sets defined by the ranking, the notions of domination-compatibility, evidence-consistency, and multi-calibration are mutually equivalent.

Incorporating the quantiles. 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 7.22. *Let $\alpha \geq 0$ and $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations of $\mathcal{X}$. For a ranking $r \in \mathcal{R}$ with $s = |\mathrm{supp}(r)|$, consider the collection of subpopulations $\mathcal{C}_\alpha(r) \subseteq \{0,1\}^{\mathcal{X}}$ defined as follows.*

$$\mathcal{C}_\alpha(r) = \left\{ S_\tau : \begin{array}{c} S \in \mathcal{C}, \\ \tau \in \mathrm{supp}(r), \\ \underset{x \sim \mathcal{D}_S}{\mathbf{Pr}}\left[\, x \in S_\tau \,\right] \geq \alpha/s \end{array} \right\} \quad \textit{where } S_\tau = \{x \in S : r(x) = \tau\}.$$

Let $Q = \bigcup_{S_\tau \in \mathcal{C}_\alpha(r)} S$; let the effective quantiles $\mathcal{Q}_{\mathcal{C},\alpha}(r)$ be defined as follows.

$$\mathcal{Q}_{\mathcal{C},\alpha}(r) = \{Q_\tau : \tau \in \mathrm{supp}(r)\} \quad \textit{where } Q_\tau = \{x \in Q : r(x) = \tau\}.$$

We make a few remarks about the collection $\mathcal{C}_\alpha(r)$ and the effective quantiles $\mathcal{Q}_{\mathcal{C},\alpha}(r)$. First, as in the definition of multi-calibration, we exclude from $\mathcal{C}_\alpha(r)$ subpopulations that are sufficiently small, anticipating the fact that we wish to learn such rankings from random samples. By scaling the minimum probability considered by the support size, the total fraction of any subpopulation that is excluded from consideration will be at most α. Throughout our subsequent analysis, we will reason about the effective quantiles $\mathcal{Q}_\alpha(r)$ rather than our earlier definition of quantiles. This is a technicality to handle the unnatural case where there is some $\tau \in \text{supp}(r)$ with exceptionally small probability $\mathbf{Pr}_{\mathcal{D}}[\ r(x) = \tau\ ] < \alpha/s$. Note that if all supported elements have considerable probability at least α/s, then $\mathcal{Q}_\alpha(r) = \mathcal{Q}(r)$.

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

Definition 7.23 (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 $(\mathcal{C}_\alpha(r), \alpha)$-domination-compatible.*

Definition 7.24 (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 it is $(\mathcal{C}_\alpha(r), \alpha)$-evidence-consistent.*

Writing the definition of reflexive-evidence-consistency more explicitly reveals the technical connection to multi-calibrated predictions. Specifically, a ranking $r \in \mathcal{R}$ with support $s = |\text{supp}(r)|$ is $(\mathcal{C}, \alpha)$-reflexive-evidence-consistent if there exists some consistent predictor $\tilde{p} \in \mathcal{P}(r)$ such that for all $S \in \mathcal{C}$ and all $\tau \in \text{supp}(r)$ where $\mathbf{Pr}_{\mathcal{D}_S}[\ r(x) = \tau\ ] > \alpha/s$,

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

Recall that without augmenting the class $\mathcal{C}$, domination-compatibility is a strictly weaker notion than evidence-consistency. The main result of this chapter demonstrates that reflexive-domination-compatibility, reflexive-evidence-consistency, and multi-calibration all encode *equivalent* notions of fairness. In other words, any ranking that satisfies domination-compatibility for a rich enough class of sets (informed by the ranking itself) implies the existence of a globally consistent multi-calibrated predictor.

Theorem 7.25 (Equivalence of Evidence-Based Fairness Notions, informal)**.** *Suppose $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations. For $\alpha > 0$ and a ranking $r \in \mathcal{R}$:*

- *r is $(\mathcal{C}, \alpha)$-reflexive-domination-compatible, if and only if*
- *r is $(\mathcal{C}, \Theta(\alpha))$-reflexive-evidence-consistent, if and only if*

- *r is the induced ranking of a $(\mathcal{C}, \Theta(\alpha))$-multi-calibrated predictor.*

The theorem is stated somewhat informally to emphasize the equivalence of the notions. Technically, what we show is that for any of the notions—reflexive-domination-compatibility, reflexive-evidence-consistency, and induced by multi-calibration—a ranking satisfying $(\mathcal{C}, \alpha)$-[*notion*] also satisfies $(\mathcal{C}, 2\alpha)$-[*other notion*].

We will prove the theorem as follows. One direction of implications will follow as corollaries of the results from Section 7.2. Specifically, the induced ranking of a multi-calibrated predictor is reflexive-evidence-consistent; further, a reflexive-evidence-consistent ranking is reflexive-domination-compatible. To complete the equivalence, we show that if a ranking $r \in \mathcal{R}$ is reflexive-domination-compatible, then r must arise as the induced ranking of some multi-calibrated predictor. We present the theorem and proof with a bias towards intuition rather than optimizing numerical constants. Quantitatively tighter connections (in terms of the approximation factor) can be made by translating directly between each notion, but the proofs add little insight beyond that of Theorem 7.28.

From multi-calibration to evidence-based rankings. We begin by demonstrating that, as with the non-reflexive notions, we can go from a multi-group guarantee on a predictor to guarantee evidence-based fairness in rankings. First, we show that the induced ranking of a multi-calibrated predictor is reflexive-evidence-consistent.

Proposition 7.26. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations 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. Recall, for any predictor $p : \mathcal{X} \to [0,1]$, $p \in \mathcal{P}(r^p)$ is consistent with its induced ranking r^p; also, for each $v \in \mathrm{supp}(p)$ there exists some $\tau_v \in \mathrm{supp}(r^p)$ such that for any subset $S \subseteq \mathcal{X}$,

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

Using these facts, we show that any multi-calibrated $\tilde{p} \in \mathcal{P}(r^{\tilde{p}})$ exhibits the correct expectations on the quantiles over subpopulations defined in $\mathcal{C}_\alpha(r^{\tilde{p}})$. Specifically, for each

$S_\tau \in \mathcal{C}_\alpha(r^{\tilde{p}})$,

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

by the multi-calibration guarantee. Thus, $r^{\tilde{p}}$ is $(\mathcal{C}, \alpha)$-reflexive-evidence-consistent. □

By the characterization of the 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 direct corollary of Proposition 7.17.

Corollary 7.27. *Let $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ be a collection of subpopulations 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.*

From domination-compatibility to multi-calibration. As we've seen, domination-compatibility over a predefined set of subpopulations is not sufficient to imply the existence of a globally consistent predictor. Somewhat surprisingly, we show that when we allow the protected subpopulations to depend nontrivially on the quantiles of the ranking, the protections of reflexive domination-compatibility are considerably stronger. Specifically, a reflexivie-domination-compatible ranking must arise as the induced ranking of a multi-calibrated predictor, and thus, must also be reflexive-evidence-consistent.

Theorem 7.28. *Suppose $\mathcal{C} \subseteq \{0,1\}^{\mathcal{X}}$ is a collection of subpopulations and let $\alpha \geq 0$. If $r \in \mathcal{R}$ is a $(\mathcal{C}, \alpha)$-reflexive-domination-compatible ranking, then r is the induced ranking of a $(\mathcal{C}, \alpha + \varepsilon)$-multi-calibrated predictor, for any constant $\varepsilon > 0$.*

At an intuitive level, we will show that if we turn the reflexive domination-compatible ranking into a calibrated predictor by taking expectations over the quantiles, the resulting predictor is multi-calibrated. The proof will be a bit more technical, taking care to work with the effective level sets $\mathcal{Q}_\alpha(r)$ and to break ties between quantiles that may result in the same expected value of p^*. This tie-breaking is used to ensure that subpopulations defined by the quantiles according to r that are deemed too small to be included in $\mathcal{C}_\alpha(r)$—for which we have no guarantees—do not merge to become large enough to be considered in the definition of $(\mathcal{C}, \alpha)$-multi-calibration.

Proof. Suppose $r \in \mathcal{R}$ is $(\mathcal{C}, \alpha)$-reflexive-domination-compatible. Consider the effective quantiles $\mathcal{Q}_\alpha(r)$, and for each $Q_\tau \in \mathcal{Q}_\alpha(r)$, define the following value.

$$p_\tau = \mathop{\mathbf{E}}_{x \sim \mathcal{D}} \left[\, p^*(x) \mid x \in Q_\tau \,\right].$$

We construct a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ based on $\{p_\tau\}$ as follows. Due to the approximate nature of the domination-compatibility constraints, the values of p_τ may not be monotonically decreasing in τ. Thus, we perturb the values of p_τ slightly in our construction of $\tilde{p}$. For each $\tau \in \text{supp}(r)$, for all $x \in \mathcal{X}_\tau$ we define $\tilde{p}(x)$ as

$$\tilde{p}(x) = \min_{\tau' \leq \tau} \{\, p_{\tau'} \,\} + \varepsilon_\tau$$

where $\varepsilon_\tau \in [-\varepsilon, \varepsilon]$ is an arbitrarily small constant to ensure that the effective quantiles are preserved as level sets of $\tilde{p}$. Specifically, for each $\tau \in \text{supp}(r)$, there exists some $v_\tau \in \text{supp}(p)$ such that

$$\{x \in \mathcal{X} : r(x) = \tau\} = \{x \in \mathcal{X} : \tilde{p}(x) = v_\tau\}. \tag{7.5}$$

The construction of $\tilde{p}$ ensures that r arises as the induced ranking of $\tilde{p}$; that is, $r = r^{\tilde{p}}$. Thus, it remains to verify that $\tilde{p}$ is indeed $(\mathcal{C}, \alpha + \varepsilon)$-multi-calibrated.

To see that $\tilde{p}$ is multi-calibrated, we remark that we only have to reason about sets $S_\tau \in \mathcal{C}_\alpha(r)$. Importantly, the bijection from (7.5) preserves the measure of the sub-quantile as a sub-level-set in $\tilde{p}$; that is, $|\text{supp}(r)| = |\text{supp}(p)| = s$ and for all $S_\tau \in \mathcal{C}_\alpha(r)$,

$$\mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} \left[\, r(x) = \tau \,\right] > \alpha/s \iff \mathop{\mathbf{Pr}}_{x \sim \mathcal{D}_S} \left[\, \tilde{p}(x) = v_\tau \,\right] > \alpha/s.$$

As such, we consider the difference in expectations of various sets within $\mathcal{C}_\alpha(r)$. By Lemma 7.20, we know that for any $\tau' \leq \tau$, and any subsets $S, T \subseteq \mathcal{X}$, $T_{\tau'}$ 0-dominates S_τ. By domination-compatibility, for any such $T_{\tau'}, S_\tau \in \mathcal{C}_\alpha(r)$, the expectation over S_τ

cannot exceed that of $T_{\tau'}$ significantly.

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid r(x) = \tau\ \right] - \mathop{\mathbf{E}}_{x\sim\mathcal{D}_T}\left[\ p^*(x) \;\middle|\; r(x) = \tau'\ \right] \leq \alpha \tag{7.6}$$

Because this inequality holds for all sub-quantiles of $T_{\tau'} \subseteq Q_{\tau'}$, by an averaging argument we know that the same inequality holds for the effective quantiles. Rearranging, we obtain the following two inequalities for $\tau' \leq \tau$ and any $S_\tau \in \mathcal{C}_\alpha(r)$:

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid r(x) = \tau\ \right] \leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ p^*(x) \mid x \in Q_{\tau'}\ \right] + \alpha \tag{7.7}$$

$$\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ p^*(x) \mid x \in Q_{\tau}\ \right] \leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid r(x) = \tau\ \right] + \alpha \tag{7.8}$$

where (7.8) follows by letting $\tau' = \tau$, and applying (7.6) in the reverse direction. With these inequalities, we can upper and lower bound the value of $\tilde{p}(x)$ compared to the expectation over sets $S \in \mathcal{C}$ where $\tilde{p}(x) = v_\tau$. Specifically, for some $x \in S_\tau$ for $S_\tau \in \mathcal{C}_\alpha(r)$:

$$\begin{aligned}
\tilde{p}(x) &= \min_{\tau' \leq \tau}\left\{\ p_{\tau'}\ \right\} + \varepsilon_\tau && (7.9)\\
&\leq p_\tau + \varepsilon && (7.10)\\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ p^*(x) \mid x \in Q_\tau\ \right] + \varepsilon && (7.11)\\
&\leq \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid r(x) = \tau\ \right] + \alpha + \varepsilon && (7.12)\\
&= \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid \tilde{p}(x) = v_\tau\ \right] + \alpha + \varepsilon && (7.13)
\end{aligned}$$

where (7.9) follows by definition of $\tilde{p}$; (7.10) follows by the fact that p_τ is feasible for the minimum over $\tau' \leq \tau$; (7.11) follows by definition of p_τ; (7.12) follows by (7.8); and (7.13) follows by (7.5). To establish the other direction:

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid \tilde{p}(x) = v_\tau\ \right] &= \mathop{\mathbf{E}}_{x\sim\mathcal{D}_S}\left[\ p^*(x) \mid r(x) = \tau\ \right] && (7.14)\\
&\leq \min_{\tau' \leq \tau}\left\{\mathop{\mathbf{E}}_{x\sim\mathcal{D}}\left[\ p^*(x) \mid x \in Q_{\tau'}\ \right]\right\} + \alpha && (7.15)\\
&\leq \min_{\tau' \leq \tau}\left\{\ p_{\tau'}\ \right\} + \alpha && (7.16)\\
&\leq \tilde{p}(x) + \alpha + \varepsilon && (7.17)
\end{aligned}$$

where (7.14) follows again by (7.5); (7.15) follows by the fact that (7.7) holds for all $\tau' \leq \tau$

and thus holds for the minimum such τ'; (7.16) follows by definition of p_τ; and (7.17) follows by definition of $\tilde{p}$. In all, we have shown that if a ranking $r \in \mathcal{R}$ is $(\mathcal{C}, \alpha)$-reflexive-domination-compatible, then it must arise as the induced ranking of some $(\mathcal{C}, \alpha + \varepsilon)$-multi-calibrated predictor $\tilde{p}$. □

While the proof of Theorem 7.28 is technical at points, the overall structure is clear: to maintain multi-calibration, it is vital that we (approximately) maintain the ordering of the quantiles according to their p^* values, because by reflexive-domination-compatibility they are already ordered correctly—even on sub-quantiles.

We note that a corollary of this proof and Proposition 7.10 is a way to transform a reflexive domination-compatible ranking into a multi-calibrated predictor. Specifically, given a ranking $r \in \mathcal{R}$, we show that the consistent calibration of a reflexive domination-compatible ranking, as defined in the proof of Theorem 7.28, is a multi-calibrated predictor. Given such a ranking and a small set of labeled samples, we can estimate the consistent calibration efficiently. Thus, provided access to sufficiently many labeled samples from $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, the task of learning reflexive evidence-based rankings is also *computationally* equivalent to learning multi-calibrated predictors.

In all, the results establish the fact that that reflexive domination-compatibility, reflexive evidence-consistency, and multi-calibration are all tightly connected concepts of evidence-based fairness. We can interpret the result from the perspective of ranking or from the perspective of prediction. First, 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 inherit desirable non-transposition properties in terms of the underlying ranking of subpopulations. Ranking is an inherently global task; thus, the characterization supports the intuitive idea that satisfying multi-calibration requires learning accurately across the entire population.

Chapter Notes

Chapter 7 is based on [DKR$^+$19], a joint work with Cynthia Dwork, Omer Reingold, Guy N. Rothblum, and Gal Yona. Our focus on ranking fairly grew out of discussions of fair affirmative action, or equality of opportunity, as proposed by [Roe98]. Roemer's proposal involves stratifying the population according to some criterion, e.g., high school graduating classes (as done in California and Texas) or education level of mother. Subsequently, students within each stratum are ranked by grades (in California and Texas), or hours spent on homework [Roe98], and the top-ranked students from each stratum are admitted. While the individual fairness notion of [DHP$^+$12] makes no explicit use of rankings, the idea of individuals being "similarly situated" can be interpreted through the lens of Roemer, to say that individuals are similarly situated if they are ranked in the same quantile within their respective statra. The work of [KRW17] 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 our work on evidence-based rankings, they assume direct access to the underlying individuals' real-valued outcomes $p^*(x)$.

There is a broad literature on learning to rank; see for instance, [Bur] and [L$^+$09] and the references therein. Much of this work focuses on the "pairwise approach" where the learner receives ordered pairs $(x, x') \in \mathcal{X} \times \mathcal{X}$, indicating that $x \prec x'$. Often, this approach can be applied to aggregate rankings from multiple sources [VLZ12]. Closer to our setting is the "pointwise approach" where individual $x \in \mathcal{X}$ are annotated with either a numerical or ordinal scores. The special case of binary labels is referred to as the bipartite ranking problem and has been in studied in [AGH$^+$05, FISS03]. [NA13] 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. 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 [Joa02] and RankBoost [FISS03]). Recently, [KZ19] 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 [YS17, CSV17, SJ18].

Bibliography

[AGH$^+$05] Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr):393–425, 2005.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[ASB$^+$19] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, Nov 2019.

[BCZ$^+$16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[BDLM01] Shai Ben-David, Philip Long, and Yishay Mansour. Agnostic boosting. In *Computational Learning Theory*, pages 507–516. Springer, 2001.

[BDR$^+$19] Noam Barda, Noa Dagan, Guy N. Rothblum, Gal Yona, Eitan Bachmat, Phil Greenland, and Morton Liebowitz. Improving subpopulation calibration in medical risk prediction. *NeurIPS Workshop on Fair ML for Healthcare*, 2019.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[Bla53] David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, 1953.

[BNS$^+$16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

[BR17] Andrej Bogdanov and Alon Rosen. Pseudorandom functions: Three decades later. In *Tutorials on the Foundations of Cryptography*, pages 79–158. Springer, 2017.

[BRA$^+$20] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levi, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, Guy N. Rothblum, Uri Shalit, Doron Netzer, Ran Balicer, and Noa Dagan. Performing risk stratification for covid-19 when individual level data is not available, the experience of a large healthcare organization. *medRxiv*, 2020.

[Bri50] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950.

[Bur] Christopher JC Burges. From ranknet to lambdarank to lambdamart: An overview.

[CG18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint 1808.00023*, 2018.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 2018.

[CMJ$^+$19] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445, 2019.

[Cré82] Jacques Crémer. A simple proof of blackwell's "comparison of experiments" theorem. *Journal of Economic Theory*, 1982.

[CSV17] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[Daw82] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[DF81] Morris H DeGroot and Stephen E Fienberg. Assessing probability assessors: Calibration and refinement. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS, 1981.

[DFH$^+$15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[DFH$^+$15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[DFH$^+$15c] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015.

[DHP$^+$12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012.

[DIKL17] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613*, 2017.

[DKR$^+$19] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[ES15] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[Fel10] Vitaly Feldman. Distribution-specific agnostic boosting. In *Proceedings of the First Symposium on Innovations in Computer Science'10*, 2010.

[FISS03] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

[FPCG16] Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 2016.

[Fri01] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[FV97] Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.

[FV98] Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[GBR07] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007.

[GGM84] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. In *Foundations of Computer Science, 1984. 25th Annual Symposium on*, pages 464–479. IEEE, 1984.

[GKR19] Sumegha Garg, Michael P. Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 809–824, 2019.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948, 2018.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 61–70. IEEE, 2010.

[HRBLM07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. (07-49), October 2007.

[HSNL18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938, 2018.

[Ilv20] Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[ILZ19] Nicole Immorlica, Katrina Ligett, and Juba Ziani. Access to population-level signaling as a source of inequality. *FAT**, 2019.

[JKMR16] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016.

[JKN$^+$19] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

[JLN$^+$20] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[KAS11] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, pages 643–650, 2011.

[KC11] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, 2011.

[KGZ19] Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Blackbox post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[KK09] Varun Kanade and Adam Kalai. Potential-based agnostic boosting. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 880–888. Curran Associates, Inc., 2009.

[KLMR18] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *AEA Papers and Proceedings*, 2018.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent tradeoffs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[KMV08] Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 629–638. ACM, 2008.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[KNRW19] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

[Koh96] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *KDD*, volume 96, pages 202–207. Citeseer, 1996.

[KRR18] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 2018.

[KRSM19] Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments. *arXiv 1905.10607*, 2019.

[KRW17] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *ICML*, 2017.

[KRZ19] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. *FAT**, 2019.

[KZ19] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *arXiv preprint arXiv:1902.05826*, 2019.

[L$^+$09] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[LDR$^+$18] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 2018.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[MBBF00] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[MCPZ18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393, 2018.

[MPB$^+$18] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions, 2018.

[NA13] Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems*, pages 2913–2921, 2013.

[Nes09] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[OPVM19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[PRT08] Dino Pedreschi, Salvatore Ruggieri, and Franco 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*, pages 560–568, 2008.

[PRW$^+$17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

[Roe98] John E Roemer. *Equality of opportunity.* Number 331.2/R62e. Harvard University Press Cambridge, MA, 1998.

[RY18] Guy Rothblum and Gal Yona. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688, 2018.

[RZ16] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *AISTATS*, 2016.

[SCM20] Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

[SGA$^+$15] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[Sha49] Claude E Shannon. The synthesis of two-terminal switching circuits. *The Bell System Technical Journal*, 28(1):59–98, 1949.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[SJ18] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228. ACM, 2018.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *2009 24th Annual IEEE Conference on Computational Complexity*, pages 126–136. IEEE, 2009.

[Vad17] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

[Val84] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vin18] James Vincent. Amazon reportedly scraps internal ai recruiting tool that was biased against women. *The Verge*, Oct 2018.

[VLZ12] Maksims N Volkovs, Hugo Larochelle, and Richard S Zemel. Learning to rank by aggregating expert preferences. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 843–851, 2012.

[WHT11] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE transactions on pattern analysis and machine intelligence*, 33(10):1978–1990, 2011.

[YS17] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, page 22. ACM, 2017.

[ZWS$^+$13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

Appendix A

Concentration Inequalities

Theorem A.1 (Chernoff's Inequalities)**.** *Suppose for $m \in \mathbb{N}$, $Z_1, \ldots, Z_m$ are independent Bernoulli random variables such that $\mathbf{E}[Z_i] = \mu_i \in [0,1]$ for $i \in [m]$. Let $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and let $\mu = \frac{1}{m}\sum_{i=1}^{m} \mu_i$. Then, for all $\Delta > 0$,*

$$\mathbf{Pr}\left[\ \hat{\mu} \geq (1+\Delta)\cdot\mu\ \right] \leq \exp\left(-\frac{\Delta^2}{2+\Delta}\cdot\mu\cdot m\right),$$

and for all $0 < \Delta < 1$,

$$\mathbf{Pr}\left[\ \hat{\mu} \leq (1-\Delta)\cdot\mu\ \right] \leq \exp\left(-\frac{\Delta^2}{2}\cdot\mu\cdot m\right).$$

Corollary A.2. *Under the conditions of Theorem A.1, for $0 < \Delta < 1$,*

$$\mathbf{Pr}\left[\ |\hat{\mu} - \mu| \geq \Delta\cdot\mu\ \right] \leq 2\cdot\exp\left(-\frac{\Delta^2}{3}\cdot\mu\cdot m\right).$$

Theorem A.3 (Hoeffding's Inequality)**.** *Suppose for $m \in \mathbb{N}$, $Z_1, \ldots, Z_m$ are independent Bernoulli random variables such that $\mathbf{E}[Z_i] = \mu_i \in [0,1]$ for $i \in [m]$. Let $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and let $\mu = \frac{1}{m}\sum_{i=1}^{m} \mu_i$. Then, for all $\Delta > 0$,*

$$\mathbf{Pr}\left[\ |\hat{\mu} - \mu| \geq \Delta\ \right] \leq 2\cdot\exp\left(-2\Delta^2 m\right).$$