THE SOCIETAL IMPACTS OF ALGORITHMIC DECISION-MAKING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Manish Raghavan

August 2021

THE SOCIETAL IMPACTS OF ALGORITHMIC DECISION-MAKING

Manish Raghavan, Ph.D.

Cornell University 2021

Algorithms are used to make decisions in an ever-increasing number of socially consequential domains. From risk assessment tools in the criminal justice system to content moderation tools to assessments in hiring, algorithms play a key role in shaping the lives of people around the world. Algorithms offer many potential benefits: they are consistent, scalable, and can leverage more data than any human could reasonably consume. However, without careful consideration algorithmic decision-making also carries a number of risks, like replicating human biases, creating perverse incentives, and propagating misinformation. This thesis seeks to develop principles for the responsible deployment of algorithms in applications of societal concern, realizing their benefits while addressing their potential harms. What does it mean to make decisions fairly? How do theoretical ideas about societal impacts manifest in practice? How do existing legal protections apply in algorithmic settings, and how can technical insights inform policy?

In this thesis, we explore these questions from a variety of perspectives. Part II leverages theoretical models to surface challenges posed by algorithmic decision-making and potential avenues to overcome them. Part III incorporates models of behavior to better understand the interplay between algorithms and humans decisions. In Part IV, we explore how these insights manifest in practice, studying applications in employment and credit scoring contexts. We conclude in Part V with open directions for future research.

BIOGRAPHICAL SKETCH

Manish Raghavan is a computer scientist who studies how computational tools impact society. He was born in New York and raised in California, where he graduated from UC Berkeley in 2016 with a degree in Electrical Engineering and Computer Science. He began his PhD at Cornell University in 2016 under the supervision of Jon Kleinberg. During his time at Cornell, he completed internships at Microsoft Research, Google Research, and Facebook Responsible AI. In his spare time, he has competed in men's club soccer for both UC Berkeley and Cornell. After completing his PhD, he will be a postdoctoral scholar at the Harvard Center for Research on Computation and Society working with Cynthia Dwork. He will begin as an assistant professor at the MIT Sloan School of Management and Department of Electrical Engineering and Computer Science in Fall 2022.

For my parents, who have always supported me in everything I do.

நன்றிகள்

ACKNOWLEDGEMENTS

The work in this thesis, and getting through graduate school, would not have been possible without the support of so many people. Thanks to Jon Kleinberg for being a truly wonderful mentor and teacher; your kindness, insight, dedication, and well-timed injections of humor have made made me a better researcher and a better person.

I've been incredibly fortunate to get the opportunity to learn from Karen Levy, whose perspective and direction have shaped the course of my time at Cornell. Karen, from AIPP meetings in Ithaca to restaurants in Barcelona, I've always appreciated your warmth, candor, and incisive questions. Thank you for everything.

I'm grateful to Éva Tardos for the insight and wisdom she has shared with me. Éva, your support, feedback, and advice throughout my Ph.D. have been invaluable; thank you.

Working with and learning from Kilian Weinberger has been a wonderful experience. Kilian, I have fond memories of deadline day in your lab followed by Collegetown Bagels—thank you for taking a half-baked class project and encouraging us to keep pushing.

I've also benefited greatly from working with Solon Barocas. Solon, thank you for broadening my research ideas, and for always looking out for me. I'm grateful for all that you've done for me.

Thanks to Alex Slivkins, Jenn Wortman Vaughan, and Steven Wu for teaching me everything I know about online learning and for always making time for me, both during my time at MSR NYC and beyond. To Sreenivas Gollupadi, Ravi Kumar, Manish Purohit, and Andrew Tomkins for a wonderful summer (and numerous Baadal trips) at Google Research. And to Sam Corbett-Davies,

Isabel Kloumann, Hannah Korevaar, and Jonathan Tannen for welcoming me into the Facebook SAIL/RAI family.

I'm grateful to each of my collaborators in the work presented in this thesis: Solon Barocas, Jon Kleinberg, Karen Levy, Sendhil Mullainathan, Geoff Pleiss, Andrew Selbst, Aleksandrs Slivkins, Jennifer Wortman Vaughan, Kilian Weinberger, Felix Wu, and Steven Wu. In addition, I've been fortunate to work with a number of wonderful researchers and mentors throughout graduate school: Rediet Abebe, Ashton Anderson, Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Jessie Finocchiaro, Sreenivas Gollapudi, Melissa Hall, Isabel Kloumann, Michelle Lam, Roland Maio, Faidra Monachou, Sigal Oren, Gourab Patro, Manish Purohit, Joaquin Quiñonero Candela, David Robinson, Joshua Simons, Ana-Andreea Stoica, Jonathan Tannen, Edmund Tong, Stratis Tsirtis, Kate Vredenburgh, and Jiejing Zhao.

The work in this thesis greatly benefited from feedback from Rediet Abebe, Ifeoma Ajunwa, Bilan Ali, Tal Alon, Lewis Baker, Solon Barocas, Larry Blume, Miranda Bogen, Kiel Brennan-Marquez, Heather Bussing, Albert Chang, A. F. Cooper, Fernando Delgado, Magdalen Dobson, Kate Donahue, Madeleine Elish, Dylan Foster, Sorelle Friedler, Stacia Garr, Avital Gertner-Samet, Jim Guszcza, Stephen Hilgartner, Lauren Kilgour, Jon Kleinberg, Katie Van Koevering, Loren Larsen, Karen Levy, Richard Marr, Cassidy McGovern, Helen Nissenbaum, Eric Parsonnet, Samir Passi, David Pedulla, Frida Polli, Sarah Riley, David Robinson, Aaron Roth, Caleb Rottman, John Sumser, Kelly Trindel, Jamie Tucker-Foltz, Briana Vecchione, Suresh Venkatasubramanian, Hanna Wallach, Angela Zhou, Malte Ziewitz, and Lindsey Zuloaga.

During my PhD, I've been fortunate to have been supported by fellowships

from Cornell University, the NSF Graduate Research Fellowships Program, and Microsoft Research.

Making it through grad school would not have been possible without the support and distraction of my friends. Thanks to Ruchie Bhardwaj, Nicolette Canale, Neil Jagtiani, Deepa Manjanatha, Kyle Patel, and Paras Unadkat for traveling with me, letting me sleep on your couches, and giving me things to look forward to. (And, of course, for visiting me in Ithaca—especially you Ruch.) Can't wait for Ski Brew '22.

My soccer teammates have been a constant source of enjoyment over the years, from surf trips to rec leagues to nights out together. Thanks to Warsame Ahmed, Daniel Angell, Callum Gilchrist, Carl Moos, and Eric Parsonnet for always making time, no matter when and where. Daniel and Eric, it's still not too late—just let me know. I'm grateful to both Cal Men's Club Soccer and Cornell Mundial FC for giving me the opportunity to compete, travel around the country, and occasionally win things. I've learned tremendously from my coaches over the years: in particular, Adam Clarke and Artie Cairel, thank you for demanding excellence and helping me take myself less seriously.

Living in Ithaca has been made wonderful by the people I've gotten to spend time with over the years. Thanks to Thodoris Lykouris for the many hours spent perfecting pasta and salads, and to Sebastian Ament and John Ryan for the IBC runs, jam sessions, evenings by the fire, and general faffery. Thanks to Briana Vecchione for making Ithaca a kinder and better place.

Finally, I'm incredibly grateful for my family. Thanks to my parents Srilatha and Prabhakar, for the opportunities and support they've given me, and for the endless entertainment. To my sister Megha, for all the crosswords, coffee, and meeping. To my grandparents, Malini Patti, Meglus Thatha, Amba Patti, and

Gnu Thatha, for always making me smile and bringing out the kid in me. And to my extended family—cousins, aunts, uncles, nieces, and nephews around the world—I couldn't ask for a more loving and caring bunch. Thank you for making me feel at home wherever I go.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part I

Introduction and Background

CHAPTER 1

INTRODUCTION AND SUMMARY OF RESULTS

Algorithms are increasingly used to make decisions in socially consequential domains. Algorithms make or contribute to decisions made in a growing list of contexts including the criminal justice system (Angwin et al., 2016), medicine (Obermeyer and Mullainathan, 2019), employment (Bogen and Rieke, 2018), and many others. The motivation behind the deployment of these algorithmic tools is that they will ultimately lead to more accurate, efficient, and consistent decisions. However, a growing body of research has begun to highlight the potential harms perpetuated by algorithmic decision-making (Barocas and Selbst, 2016; Eubanks, 2018b; Noble, 2018). Algorithms can perpetuate discrimination (Barocas and Selbst, 2016), cause representational harms (Sweeney, 2013; Noble, 2018; Crawford, 2017), and deny opportunity (Eubanks, 2018b).

The research community, as well as society more broadly, have sought to address these harms through a variety of methods, theoretical and empirical, qualitative and quantitative. Taken together, this body of work (informally referred to as the Fairness, Accountability, and Transparency (FAccT) community) demonstrates how a wide range of perspectives can come together to deepen our understanding of the impacts of algorithms and technology on society. This includes empirical efforts to understand performance disparities in commercially developed algorithms (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Koenecke et al., 2020); theoretical characterizations of bias and discrimination in decision-making (Chouldechova, 2017; Kleinberg et al., 2017; Joseph et al., 2016); and legal analyses of algorithmic systems and their impacts (Barocas and Selbst, 2016; Kim, 2016, 2017; Ajunwa, 2021).

In this thesis, we present a number of contributions to this body of literature, drawing on a variety of techniques and tools from computer science, economics, and the law. In Part II, we develop theoretical tools and frameworks to characterize the impacts and limits of decision-making. We further develop these tools in Part III, where we incorporate models of behavior, including behavioral biases, strategic behavior, and competition, to reason about decision-making in complex, multi-agent environments. In Part IV, we leverage these technical insights to deepen our qualitative understanding of algorithmic impacts, particularly with respect to employment and lending. We conclude in Part V with open directions for future work. Part VI contains appendices with supplementary information and proofs.

The work in this thesis has been previously published as Barocas et al. (2020); Kleinberg et al. (2017); Kleinberg and Raghavan (2018, 2020, 2021); Pleiss et al. (2017); Raghavan et al. (2018, 2020).

1.1 Summary of Results

1.1.1 Theoretical Foundations for Fairness in Algorithmic Decision-Making

In Part II, we examine the impacts of algorithmic decision-making from a theoretical perspective. We show fundamental challenges in making fair decisions using tools from theoretical computer science and machine learning.

Fairness in risk assessment. Algorithms are often used to quantitatively assess risk, predicting future outcomes based on historical data. When these predictive algorithms are applied to people, i.e., evaluating the risk of future behavior or outcomes, decision-makers have an obligation to ensure that these predictions are in some sense fair to decision subjects.

In Chapter 3, we formalize three natural definitions of fairness proposed in the context of risk assessment in the criminal justice system. We formally prove that these three definitions cannot be simultaneously satisfied, even approximately, outside of a few highly constrained cases. In particular, we demonstrate a fundamental tension between *calibration* and *equal error rates*. Our results provide a resolution to a debate over algorithms used in recidivism prediction (Angwin et al., 2016; Dieterich et al., 2016; Angwin and Larson, 2016).

We build upon these results in Chapter 4, further exploring the incompatibility between calibration and equal error rates. In particular, we demonstrate the consequences of pursuing equality of error rates while maintaining calibration. We find that such efforts will necessarily come at the expense of predictive accuracy without improving outcomes for anyone. Our results suggest that attempting to remedy disparities found in algorithmic decision-making simply by altering models, as opposed to the data they ingest and the broader context in which they operate, is insufficient.

Externalities in online learning. In supervised settings like risk assessment, we commonly assume the data to be fixed and known before decisions are made. However, in many platforms, decisions are made in dynamic settings where data arrives continuously and past decisions influence the collection of

data to be used in the future. Settings like this are commonly modeled through the online learning framework (Auer, 2002).

Existing techniques from online learning rely on a balance of *exploration* and *exploitation*: a platform may attempt to experiment on some users in order to gain more information, or they may attempt to leverage existing information to make the instantaneously optimal decision for that user. The need to explicitly experiment in order to gather new information results in *externalities*: users' outcomes are influenced by the presence of other users.

Prior work has articulated concerns that exploration may impose an undue burden on small, marginalized communities (Bird et al., 2016). In Chapter 5, we quantify this worry by formally showing how the presence of one population in an online learning setting can result in negative externalities for another population. However, we show conditions under which these externalities may be in a sense unnecessary—if data are sufficiently diverse, we show that a greedy algorithm that maximizes each user's individual utility (and thus does not engage in explicit exploration) is nearly socially optimal. Formally, this involves a smoothed analysis of the Bayesian linear contextual bandits model, showing conditions under which the greedy algorithm can effectively match the performance of any algorithm on any instance.

1.1.2 Models of Behavior

In Part III, we explore the effects that behavioral factors have on decision-making. In particular, we consider the impacts of behavioral biases and strategic behavior.

Implicit bias and selection. Over the past two decades, the notion of implicit bias has come to serve as an important component in our understanding of discrimination in activities such as hiring, promotion, and school admissions (Greenwald and Banaji, 1995; Bertrand and Mullainathan, 2004). Research on implicit bias posits that when people evaluate others – for example, in a hiring context – their unconscious biases about membership in particular groups can have an effect on their decision-making, even when they have no deliberate intention to discriminate against members of these groups. A growing body of experimental work has pointed to the effect that implicit bias can have in producing adverse outcomes.

In Chapter 7, we propose and analyze a theoretical model for studying the effects of implicit bias on selection decisions, and a way of analyzing possible procedural remedies for implicit bias within this model. A canonical situation represented by our model is a hiring setting: a recruiting committee is trying to choose a set of finalists to interview among the applicants for a job, evaluating these applicants based on their future potential, but their estimates of potential are skewed by implicit bias against members of one group. In this model, we show that the Rooney Rule, a commonly imposed requirement that at least one of the finalists be chosen from the affected group, can not only improve the representation of this affected group, but also lead to higher payoffs in absolute terms for the organization performing the recruiting. However, identifying the conditions under which such measures can lead to improved payoffs involves subtle trade-offs between the extent of the bias and the underlying distribution of applicant characteristics, leading to novel theoretical questions about order statistics in the presence of probabilistic side information.

Strategic behavior and evaluation. Algorithms are often used to produce decision-making rules that classify or evaluate individuals. When these individuals have incentives to be classified a certain way, they may behave strategically to influence their outcomes. A growing body of computer science literature models these interactions to construct algorithms that are in some way robust to this behavior (Dalvi et al., 2004; Brückner and Scheffer, 2011; Hardt et al., 2016a; Dong et al., 2018; Hu et al., 2019; Milli et al., 2019). Typically, these models assume that strategic behavior is undesirable—that is, while an agent can manipulate their appearance, they cannot change the underlying property that the decision-maker seeks to determine.

On the other hand, issues of incentives and strategic behavior have long been considered in the economics literature, particularly in the context of principal-agent models (Grossman and Hart, 1983; Holmström and Milgrom, 1987, 1991; Hermalin and Katz, 1991). Unlike the computer science literature, these economic models typically assume that the decision-maker's utility depends directly on an agent's behavior. Thus, in these models, strategic behavior isn't intrinsically good or bad, but decision-makers have preferences over behaviors.

We seek to integrate perspectives from the computer science and economics literatures. We develop a model in Chapter 8 for how strategic agents can invest effort in order to change the outcomes they receive. Drawing from models in the computer science literature, we assume agents are represented by their features, which they have some ability to manipulate. In the style of principal-agent models from economics, we seek to characterize which of these behaviors can be incentivized. We give a tight characterization of when certain behaviors can be incentivized, and we show that whenever any "reasonable" mechanism

can do so, a simple linear mechanism suffices.

Algorithmic monoculture. As algorithms are increasingly applied to screen applicants for high-stakes decisions in employment, lending, and other domains, concerns have been raised about the effects of *algorithmic monoculture*, in which many decision-makers all rely on the same algorithm. This concern invokes analogies to agriculture, where a monocultural system runs the risk of severe harm from unexpected shocks. In Chapter 9, we show that the dangers of algorithmic monoculture run much deeper, in that monocultural convergence on a single algorithm by a group of decision-making agents, even when the algorithm is more accurate for any one agent in isolation, can reduce the overall quality of the decisions being made by the full collection of agents. Unexpected shocks are therefore not needed to expose the risks of monoculture; it can hurt accuracy even under "normal" operations, and even for algorithms that are more accurate when used by only a single decision-maker. Our results rely on minimal assumptions, and involve the development of a probabilistic framework for analyzing systems that use multiple noisy estimates of a set of alternatives.

1.1.3 Application Domains

In Part IV, we explore the consequences of algorithmic decision-making in two applications: algorithmic hiring and explaining credit decisions. Combining perspectives from computer science, sociology, and the law, we analyze the practice of algorithmic decision-making in these contexts.

Algorithmic hiring. There has been rapidly growing interest in the use of algorithms in hiring, especially as a means to address or mitigate bias. However, little is known about how these methods are used in practice. How are algorithmic assessments built, validated, and examined for bias? In Chapter 11, we document and analyze the claims and practices of companies offering algorithms for employment assessment. In particular, we identify vendors of algorithmic pre-employment assessments (i.e., algorithms to screen candidates), document what they have disclosed about their development and validation procedures, and evaluate their practices, focusing particularly on efforts to detect and mitigate bias. Our analysis considers both technical and legal perspectives. Technically, we consider the various choices vendors make regarding data collection and prediction targets, and explore the risks and trade-offs that these choices pose. We also discuss how algorithmic de-biasing techniques interface with, and create challenges for, antidiscrimination law. We conclude with several policy recommendations designed to prevent discrimination in algorithmic employment decisions.

Explanations in credit scoring. Counterfactual explanations are gaining prominence within technical, legal, and business circles as a way to explain the decisions of a machine learning model without disclosing the model itself. These explanations share a trait with the long-established "principal reason" explanations that are required by U.S. credit laws: they both explain a decision by highlighting a set of features deemed most relevant—and withholding others.

These "feature-highlighting explanations" have several desirable properties: They place no constraints on model complexity, do not require model disclosure, provide a justification for a model's decision or instructions for achieving

a different decision, and seem to automate compliance with the law. But they are far more complex and subjective than they appear.

In Chapter 12, we demonstrate that the utility of feature-highlighting explanations relies on a number of easily overlooked assumptions. These assumptions have to do with how features in the model relate to the actions required to change them, the cost of these actions, and the effect of these actions in other domains in people's lives. They also depend on the underlying model having certain properties, without which explanations will fail, and the stability of the explanation over time. We then explore several consequences of acknowledging and attempting to address these assumptions, including a paradox in the way that feature-highlighting explanations aim to respect autonomy, the unchecked power that feature-highlighting explanations grant decision makers, and a tension between making these explanations useful and the need to keep the model confidential.

While new research suggests several ways that feature-highlighting explanations can work around some of the problems that we identify, the disconnect between features in the model and actions in the real world—and the subjective choices necessary to compensate for this—must be understood before these techniques can be usefully implemented.

1.1.4 Open Directions

We conclude with a broad overview of several potential directions for future study of the impacts of algorithmic decision-making on society. Modern algorithmic decision-making systems incorporate machine-learned predictions in

broader pipelines and mechanisms, implicating a number of concerns over accuracy and allocation found in both the **machine learning and mechanism design** literatures. These challenges increasingly require a combination of techniques and insights from the two fields, particularly with respect to how they conceive of and ensure fair decision-making.

Algorithms are often used to make decisions in settings where the law prohibits discrimination. In such contexts, we must be increasingly sensitive to the potential for **algorithmic discrimination**, the potential for algorithmic decisions to be discriminatory. Future work will need to carefully consider both the computational and legal dimensions of this problem.

One avenue to protect consumers from discriminatory or otherwise unfair decisions is to require **transparency**: people should know when and how they are being algorithmically evaluated. Ensuring that such disclosures provide **meaningful explanations** is an emerging field of study that presents a number of open challenges.

Part II

Theoretical Foundations for Fairness in Algorithmic Decision-Making

CHAPTER 2

OVERVIEW OF Part II

In Part II, we apply a theoretical lens to fair algorithmic decision-making. We consider both supervised learning (Chapters 3 and 4) and online learning (Chapter 5).

In Chapter 3, we consider the *risk assessment* setting, where a predictive model is used to determine an individual's risk for a particular behavior or outcome. For example, in the criminal justice system, risk assessment tools are used in pre-trial hearings to determine the risk of *recidivism*, or future arrest. In such settings, there is a public interest in ensuring such predictions are in some sense "fair" to defendants. We show that three natural definitions of fair decision-making in this context are in conflict with one another: outside of two special cases, these criteria cannot be simultaneously satisfied, even approximately.

We build on these results in Chapter 4, showing a tension between *calibration* and efforts to equalize the cost of predictive errors across populations. In particular, our results imply that efforts to simultaneously enforce calibration while equalizing some notion of cost must make predictions worse for some subset of the population without improving outcomes for anyone.

In Chapter 5, we turn to the issue of externalities in online learning. We demonstrate that the inclusion of heterogeneous groups of users into the same online learning algorithm can create negative externalities, leading to worse outcomes for smaller populations than they would have otherwise received if treated separately. However, we develop novel analyses of the greedy algorithm in the Bayesian linear contextual bandit setting, demonstrating conditions un-

der which the greedy algorithm is instance-optimal, meaning it cannot result in negative externalities for any population.

CHAPTER 3

INHERENT TRADE-OFFS IN THE FAIR DETERMINATION OF RISK SCORES

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools mitigate or exacerbate the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants (Angwin et al., 2016; Larson et al., 2016). One of their main contentions was that the tool's errors were asymmet-

ric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool's errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants (Dieterich et al., 2016; Flores et al., 2016; Gong, 2016).

Second, in a very different domain, researchers have begun to analyze the ways in which different genders and racial groups experience advertising and commercial content on the Internet differently (Datta et al., 2015; Sweeney, 2013). We could ask, for example: if a male user and female user are equally interested in a particular product, does it follow that they're equally likely to be shown an ad for it? Sometimes this concern may have broader implications, for example if women in aggregate are shown ads for lower-paying jobs. Other times, it may represent a clash with a user's leisure interests: if a female user interacting with an advertising platform is interested in an activity that tends to have a male-dominated viewership, like professional football, is the platform as likely to show her an ad for football as it is to show such an ad to an interested male user?

A third domain, again quite different from the previous two, is medical testing and diagnosis. Doctors making decisions about a patient's treatment may rely on tests providing probability estimates for different diseases and conditions. Here too we can ask whether such decision-making is being applied uniformly across different groups of patients (Garb, 1997; Williams and Mohammed, 2009), and in particular how medical tests may play a differential role

for conditions that vary widely in frequency between these groups.

Providing guarantees for decision procedures. One can raise analogous questions in many other domains of fundamental importance, including decisions about hiring, lending, or school admissions (Executive Office of the President, 2016), but we will focus on the three examples above for the purposes of this discussion. In these three example domains, a few structural commonalities stand out. First, the algorithmic estimates are often being used as "input" to a larger framework that makes the overall decision — a risk score provided to a human expert in the legal and medical instances, and the output of a machine-learning algorithm provided to a larger advertising platform in the case of Internet ads. Second, the underlying task is generally about classifying whether people possess some relevant property: recidivism, a medical condition, or interest in a product. We will refer to people as being *positive instances* if they truly possess the property, and *negative instances* if they do not. Finally, the algorithmic estimates being provided for these questions are generally not pure yes-no decisions, but instead probability estimates about whether people constitute positive or negative instances.

Let us suppose that we are concerned about how our decision procedure might operate differentially between two groups of interest (such as African-American and white defendants, or male and female users of an advertising system). What sorts of guarantees should we ask for as protection against potential bias?

A first basic goal in this literature is that the probability estimates provided by the algorithm should be *well-calibrated*: if the algorithm identifies a set of

people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances (Crowson et al., 2016; Foster and Vohra, 1998). Moreover, this condition should hold when applied separately in each group as well (Flores et al., 2016). For example, if we are thinking in terms of potential differences between outcomes for men and women, this means requiring that a z fraction of men and a z fraction of women assigned a probability z should possess the property in question.

A second goal focuses on the people who constitute positive instances (even if the algorithm can only imperfectly recognize them): the average score received by people constituting positive instances should be the same in each group. We could think of this as *balance for the positive class*, since a violation of it would mean that people constituting positive instances in one group receive consistently lower probability estimates than people constituting positive instances in another group. In our initial criminal justice example, for instance, one of the concerns raised was that white defendants who went on to commit future crimes were assigned risk scores corresponding to lower probability estimates in aggregate; this is a violation of the condition here. There is a completely analogous property with respect to negative instances, which we could call *balance for the negative class*. These balance conditions can be viewed as generalizations of the notions that both groups should have equal false negative and false positive rates.

It is important to note that balance for the positive and negative classes, as defined here, is distinct in crucial ways from the requirement that the average probability estimate globally over *all* members of the two groups be equal. This latter global requirement is a version of *statistical parity* (Feldman et al., 2015;

Calders and Verwer, 2012; Kamiran and Calders, 2009; Kamishima et al., 2012). In some cases statistical parity is a central goal (and in some it is legally mandated), but the examples considered so far suggest that classification and risk assessment are much broader activities where statistical parity is often neither feasible nor desirable. Balance for the positive and negative classes, however, is a goal that can be discussed independently of statistical parity, since these two balance conditions simply ask that once we condition on the "correct" answer for a person, the chance of making a mistake on them should not depend on which group they belong to.

The present work: Trade-offs among the guarantees. Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal — that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously.

Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to *approximate* versions of the conditions as well.

3.1 A Formal Model of Risk Assessment

In this section we formulate this main result precisely, as a theorem building on a model that makes the discussion thus far more concrete.

3.1.1 Formulating the Goal

Let's start with some basic definitions. As above, we have a collection of people each of whom constitutes either a positive instance or a negative instance of the classification problem. We'll say that the *positive class* consists of the people who constitute positive instances, and the *negative class* consists of the people who constitute negative instances. For example, for criminal defendants, the positive class could consist of those defendants who will be arrested again within some fixed time window, and the negative class could consist of those who will not. The positive and negative classes thus represent the "correct" answer to the classification problem; our decision procedure does not know them, but is trying to estimate them.

Feature vectors. Each person has an associated *feature vector* σ, representing the data that we know about them. Let p_σ denote the fraction of people with feature vector σ who belong to the positive class. Conceptually, we will picture that while there is variation within the set of people who have feature vector σ, this variation is invisible to whatever decision procedure we apply; all people with feature vector σ are indistinguishable to the procedure. Our model will assume that the value p_σ for each σ is known to the procedure.[1]

Groups. Each person also belongs to one of two *groups*, labeled 1 or 2, and we would like our decisions to be unbiased with respect to the members of these

[1] Clearly the case in which the value of p_σ is unknown is an important version of the problem as well; however, since our main results establish strong limitations on what is achievable, these limitations are only stronger because they apply even to the case of known p_σ.

two groups.[2] In our examples, the two groups could correspond to different races or genders, or other cases where we want to look for the possibility of bias between them. The two groups have different distributions over feature vectors: a person of group t has a probability $a_{t\sigma}$ of exhibiting the feature vector σ. However, people of each group have the same probability p_σ of belonging to the positive class provided their feature vector is σ. In this respect, σ contains all the relevant information available to us about the person's future behavior; once we know σ, we do not get any additional information from knowing their group as well.[3]

Risk Assignments. We say that an *instance* of our problem is specified by the parameters above: a feature vector and a group for each person, with a value p_σ for each feature vector, and distributions $\{a_{t\sigma}\}$ giving the frequency of the feature vectors in each group.

Informally, risk assessments are ways of dividing people up into sets based on their feature vectors σ (potentially using randomization), and then assigning each set a probability estimate that the people in this set belong to the positive class. Thus, we define a *risk assignment* to consist of a set of "bins" (the sets), where each bin is labeled with a *score* v_b that we intend to use as the probability for everyone assigned to bin b. We then create a rule for assigning people to bins based on their feature vector σ; we allow the rule to divide people with a fixed feature vector σ across multiple bins (reflecting the possible use of randomiza-

[2] We focus on the case of two groups for simplicity of exposition, but it is straightforward to extend all of our definitions to the case of more than two groups.

[3] As we will discuss in more detail below, the assumption that the group provides no additional information beyond σ does not restrict the generality of the model, since we can always consider instances in which people of different groups never have the same feature vector σ, and hence σ implicitly conveys perfect information about a person's group.

tion). Thus, the rule is specified by values $X_{\sigma b}$: a fraction $X_{\sigma b}$ of all people with feature vector σ are assigned to bin b. Note that the rule does not have access to the group t of the person being considered, only their feature vector σ. (As we will see, this does not mean that the rule is incapable of exhibiting bias between the two groups.) In summary, a risk assignment is specified by a set of bins, a score for each bin, and values $X_{\sigma b}$ that define a mapping from people with feature vectors to bins.

Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

(A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

Why Do These Conditions Correspond to Notions of Fairness? All of these are natural conditions to impose on a risk assignment; and as indicated by the discussion above, all of them have been proposed as versions of fairness. The first one essentially asks that the scores mean what they claim to mean, even when considered separately in each group. In particular, suppose a set of scores lack the first property for some bin b, and these scores are given to a decision-maker; then if people of two different groups both belong to bin b, the decision-maker has a clear incentive to treat them differently, since the lack of calibration within groups on bin b means that these people have different aggregate probabilities of belonging to the positive class. Another way of stating the property of calibration within groups is to say that, conditioned on the bin to which an individual is assigned, the likelihood that the individual is a member of the positive class is independent of the group to which the individual belongs. This means we are justified in treating people with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to.

The second and third ask that if two individuals in different groups exhibit comparable future behavior (negative or positive), they should be treated comparably by the procedure. In other words, a violation of, say, the second condition would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.

We can also interpret some of the prior work around our earlier examples through the lens of these conditions. For example, in the analysis of the COM-

PAS risk tool for criminal defendants, the critique by Angwin et al. focused on the risk tool's violation of conditions (B) and (C); the counter-arguments established that it satisfies condition (A). While it is clearly crucial for a risk tool to satisfy (A), it may still be important to know that it violates (B) and (C). Similarly, to think in terms of the example of Internet advertising, with male and female users as the two groups, condition (A) as before requires that our estimates of ad-click probability mean the same thing in aggregate for men and women. Conditions (B) and (C) are distinct; condition (C), for example, says that a female user who genuinely wants to see a given ad should be assigned the same probability as a male user who wants to see the ad.

3.1.2 Determining What is Achievable: A Characterization Theorem

When can conditions (A), (B), and (C) be simultaneously achieved? We begin with two simple cases where it's possible.

- *Perfect prediction.* Suppose that for each feature vector σ, we have either $p_\sigma = 0$ or $p_\sigma = 1$. This means that we can achieve perfect prediction, since we know each person's class label (positive or negative) for certain. In this case, we can assign all feature vectors σ with $p_\sigma = 0$ to a bin b with score $v_b = 0$, and all σ with $p_\sigma = 1$ to a bin b' with score $v_{b'} = 1$. It is easy to check that all three of the conditions (A), (B), and (C) are satisfied by this risk assignment.
- *Equal base rates.* Suppose, alternately, that the two groups have the same fraction of members in the positive class; that is, the average value of p_σ is

the same for the members of group 1 and group 2. (We can refer to this as the *base rate* of the group with respect to the classification problem.) In this case, we can create a single bin b with score equal to this average value of p_σ, and we can assign everyone to bin b. While this is not a particularly informative risk assignment, it is again easy to check that it satisfies fairness conditions (A), (B), and (C).

Our first main result establishes that these are in fact the only two cases in which a risk assignment can achieve all three fairness guarantees simultaneously.

Theorem 3.1. *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction (with p_σ equal to 0 or 1 for all σ) or have equal base rates.*

Thus, in every instance that is more complex than the two cases noted above, there will be some natural fairness condition that is violated by any risk assignment. Moreover, note that this result applies regardless of how the risk assignment is computed; since our framework considers risk assignments to be arbitrary functions from feature vectors to bins labeled with probability estimates, it applies independently of the method — algorithmic or otherwise — that is used to construct the risk assignment.

The conclusions of the first theorem can be relaxed in a continuous fashion when the fairness conditions are only approximate. In particular, for any $\varepsilon > 0$ we can define ε-approximate versions of each of conditions (A), (B), and (C) (specified precisely in the next section), each of which requires that the corresponding equalities between groups hold only to within an error of ε. For any

$\delta > 0$, we can also define a δ-approximate version of the equal base rates condition (requiring that the base rates of the two groups be within an additive δ of each other) and a δ-approximate version of the perfect prediction condition (requiring that in each group, the average of the expected scores assigned to members of the positive class is at least $1 - \delta$; by the calibration condition, this can be shown to imply a complementary bound on the average of the expected scores assigned to members of the negative class).

In these terms, our approximate version of Theorem 3.1 is the following.

Theorem 3.2. *There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\varepsilon > 0$, and any instance of the problem with a risk assignment satisfying the ε-approximate versions of fairness conditions (A), (B), and (C), the instance must satisfy either the $f(\varepsilon)$-approximate version of perfect prediction or the $f(\varepsilon)$-approximate version of equal base rates.*

Thus, anything that approximately satisfies the fairness constraints must approximately look like one of the two simple cases identified above.

Finally, in connection to Theorem 3.1, we note that when the two groups have equal base rates, then one can ask for the most accurate risk assignment that satisfies all three fairness conditions (A), (B), and (C) simultaneously. Since the risk assignment that gives the same score to everyone satisfies the three conditions, we know that at least one such risk assignment exists; hence, it is natural to seek to optimize over the set of all such assignments. We consider this algorithmic question in Section 3.4.

To reflect a bit further on our main theorems and what they suggest, we note that our intention in the present work isn't to make a recommendation on how

conflicts between different definitions of fairness should be handled. Nor is our intention to analyze which definitions of fairness are violated in particular applications or datasets. Rather, our point is to establish certain unavoidable trade-offs between the definitions, regardless of the specific context and regardless of the method used to compute risk scores. Since each of the definitions reflect (and have been proposed as) natural notions of what it should mean for a risk score to be fair, these trade-offs suggest a striking implication: that outside of narrowly delineated cases, any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias. This is equally true whether the risk score is determined by an algorithm or by a system of human decision-makers.

Special Cases of the Model. Our main results, which place strong restrictions on when the three fairness conditions can be simultaneously satisfied, have more power when the underlying model of the input is more general, since it means that the restrictions implied by the theorems apply in greater generality. However, it is also useful to note certain special cases of our model, obtained by limiting the flexibility of certain parameters in intuitive ways. The point is that our results apply *a fortiori* to these more limited special cases.

First, we have already observed one natural special case of our model: cases in which, for each feature vector σ, only members of one group (but not the other) can exhibit σ. This means that σ contains perfect information about group membership, and so it corresponds to instances in which risk assignments would have the potential to use knowledge of an individual's group membership. Note that we can convert any instance of our problem into a new instance that belongs to this special case as follows. For each feature vector σ,

we create two new feature vectors $\sigma^{(1)}$ and $\sigma^{(2)}$; then, for each member of group 1 who had feature vector σ, we assign them $\sigma^{(1)}$, and for each member of group 2 who had feature vector σ, we assign them $\sigma^{(2)}$. The resulting instance has the property that each feature vector is associated with members of only one group, but it preserves the essential aspects of the original instance in other respects.

Second, we allow risk assignments in our model to split people with a given feature vector σ over several bins. Our results also therefore apply to the natural special case of the model with *integral* risk assignments, in which all people with a given feature σ must go to the same bin.

Third, our model is a generalization of binary classification, which only allows for 2 bins. Note that although binary classification does not explicitly assign scores, we can consider the probability that an individual belongs to the positive class given that they were assigned to a specific bin to be the score for that bin. Thus, our results hold in the traditional binary classification setting as well.

Data-Generating Processes. Finally, there is the question of where the data in an instance of our problem comes from. Our results do not assume any particular process for generating the positive/negative class labels, feature vectors, and group memberships; we simply assume that we are given such a collection of values (regardless of where they came from), and then our results address the existence or non-existence of certain risk assignments for these values.

This increases the generality of our results, since it means that they apply to any process that produces data of the form described by our model. To give an example of a natural generative model that would produce instances with

the structure that we need, one could assume that each individual starts with a "hidden" class label (positive or negative), and a feature vector σ is then probabilistically generated for this individual from a distribution that can depend on their class label and their group membership. (If feature vectors produced for the two groups are disjoint from one another, then the requirement that the value of p_σ is independent of group membership given σ necessarily holds.) Since a process with this structure produces instances from our model, our results apply to data that arises from such a generative process.

It is also interesting to note that the basic set-up of our model, with the population divided across a set of feature vectors for which race provides no additional information, is in fact a very close match to the information one gets from the output of a well-calibrated risk tool. In this sense, one setting for our model would be the problem of applying post-processing to the output of such a risk tool to ensure additional fairness guarantees. Indeed, since much of the recent controversy about fair risk scores has involved risk tools that are well-calibrated but lack the other fairness conditions we consider, such an interpretation of the model could be a useful way to think about how one might work with these tools in the context of a broader system.

Further Related Work. Mounting concern over discrimination in machine learning has led to a large body of new work seeking to better understand and prevent it. Barocas and Selbst (2016) survey a range of ways in which data-analysis algorithms can lead to discriminatory outcomes. We refer the reader to Mehrabi et al. (2019) for a more complete survey of the area.

Broadly speaking, the technical literature considers both *individual* and *group*

definitions of fair decision-making. Individual notions of fairness, beginning with Dwork et al. (2012), are concerned with the notion that similar individuals should be treated similarly (Ilvento, 2020; Bechavod et al., 2020). This line of work has also influenced efforts to define fairness through counterfactuals, comparing an individual's treatment to what it would have been in some hypothetical counterfactual world (Kusner et al., 2017; Russell et al., 2017; Chiappa, 2019).

In contrast, group-based definitions of fairness seek to impose the constraint that some statistical property, such as error rates or selection rates, should not differ across pre-defined subpopulations based on attributes like gender and race (Kamiran and Calders, 2009; Calders and Verwer, 2012; Kamishima et al., 2012; Hajian and Domingo-Ferrer, 2013; Hardt et al., 2016b; Zafar et al., 2017b; Corbett-Davies et al., 2017). The work in this chapter and Chapter 4 closely aligns with this body of literature.

In particular, we study the role of *calibration* in risk assessment, which are often considered necessary for empirical risk analysis tools (Berk et al., 2018; Crowson et al., 2016; Dieterich et al., 2016; Flores et al., 2016). In practical applications, uncalibrated probability estimates can be misleading in the sense that the end user of these estimates has an incentive to mistrust (and therefore potentially misuse) them. There are several post-processing methods for producing calibrated outputs from classification algorithms. For example, Platt Scaling (Platt, 1999) passes outputs through a learned sigmoid function, transforming them into calibrated probabilities. Histogram Binning and Isotonic Regression (Zadrozny and Elkan, 2001) learn a general monotonic function from outputs to probabilities. See Niculescu-Mizil and Caruana (2005); Guo et al. (2017) for

empirical comparisons of these methods.

Beyond this theoretical work, a growing body of studies have sought to empirically characterize bias in real-world algorithmic decision-making systems, including those used for facial recognition (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019), speech recognition (Koenecke et al., 2020), and health risk prediction (Obermeyer and Mullainathan, 2019).

In response to growing interest in theoretical characterizations of fair decision-making critical scholars have raised concerns that this focus on technical aspects of fair decision-making detract from underlying societal inequities (Powles and Nissenbaum, 2018; Green, 2018). We respond to some of these concerns in a recent attempt to articulate positive roles computing scholarship can play in an effort to bring about social change (Abebe et al., 2020).

3.2 The Characterization Theorems

Starting with the notation and definitions from the previous section, we now give a proof of Theorem 3.1.

Informal overview. Let us begin with a brief overview of the proof, before going into a more detailed version of it. For this discussion, let N_t denote the number of people in group t, and μ_t be the number of people in group t who belong to the positive class.

Roughly speaking, the proof proceeds in two steps. First, consider a single bin b. By the calibration condition, the expected total score given to the group-t

people in bin b is equal to the expected number of group-t people in bin b who belong to the positive class. Summing over all bins, we find that the total score given to all people in group t (that is, the sum of the scores received by everyone in group t) is equal to the total number of people in the positive class in group t, which is μ_t.

Now, let x be the average score given to a member of the negative class, and let y be the average score given to a member of the positive class. By the balance conditions for the negative and positive classes, these values of x and y are the same for both groups.

Given the values of x and y, the total number of people in the positive class μ_t, and the total score given out to people in group t — which, as argued above, is also μ_t — we can write the total score as

$$(N - \mu_t)x + \mu_t y = \mu_t.$$

This defines a line for each group t in the two variables x and y, and hence we obtain a system of two linear equations (one for each group) in the unknowns x and y.

If all three conditions — calibration, and balance for the two classes — are to be satisfied, then we must be at a set of parameters that represents a solution to the system of two equations. If the base rates are equal, then $\mu_1 = \mu_2$ and hence the two lines are the same; in this case, the system of equations is satisfied by any choice of x and y. If the base rates are not equal, then the two lines are distinct, and they intersect only at the point $(x, y) = (0, 1)$, which implies perfect prediction — an average score of 0 for members of the negative class and 1 for members of the positive class. Thus, the three conditions can be simultaneously satisfied if and only if we have equal base rates or perfect prediction.

This concludes the overview of the proof; in the remainder of the section we describe the argument at a more detailed level.

Definitions and notation. Recall from our notation in the previous section that an $a_{t\sigma}$ fraction of the people in group t have feature vector σ; we thus write $n_{t\sigma} = a_{t\sigma}N_t$ for the number of people in group t with feature vector σ. Many of the components of the risk assignment and its evaluation can be written in terms of operations on a set of underlying matrices and vectors, which we begin by specifying.

- First, let $|\sigma|$ denote the number of feature vectors in the instance, and let $p \in \mathbb{R}^{|\sigma|}$ be a vector indexed by the possible feature vectors, with the coordinate in position σ equal to p_σ. For group t, let $n_t \in \mathbb{R}^{|\sigma|}$ also be a vector indexed by the possible feature vectors, with the coordinate in position σ equal to $n_{t\sigma}$. Finally, it will be useful to have a representation of p as a diagonal matrix; thus, let P be a $|\sigma| \times |\sigma|$ diagonal matrix with $P_{\sigma\sigma} = p_\sigma$.
- We now specify a risk assignment as follows. The risk assignment involves a set of B bins with associated scores; let $v \in \mathbb{R}^B$ be a vector indexed by the bins, with the coordinate in position b equal to the score v_b of bin b. Let V be a diagonal matrix version of v: it is a $B \times B$ matrix with $V_{bb} = v_b$. Finally, let X be the $|\sigma| \times B$ matrix of $X_{\sigma b}$ values, specifying the fraction of people with feature vector σ who get mapped to bin b under the assignment procedure.

There is an important point to note about the $X_{\sigma b}$ values. If all of them are equal to 0 or 1, this corresponds to a procedure in which all people with the same feature vector σ get assigned to the same bin. When some of the $X_{\sigma b}$ values are

not equal to 0 or 1, the people with vector σ are being divided among multiple bins. In this case, there is an implicit randomization taking place with respect to the positive and negative classes, and with respect to the two groups, which we can think of as follows. Since the procedure cannot distinguish among people with vector σ, in the case that it distributes these people across multiple bins, the subset of people with vector σ who belong to the positive and negative classes, and to the two groups, are divided up randomly across these bins in proportions corresponding to $X_{\sigma b}$. In particular, if there are $n_{t\sigma}$ group-t people with vector σ, the expected number of these people who belong to the positive class and are assigned to bin b is $n_{t\sigma}p_\sigma X_{\sigma b}$.

Let us now proceed with the proof of Theorem 3.1, starting with the assumption that our risk assignment satisfies conditions (A), (B), and (C).

Calibration within groups. We begin by working out some useful expressions in terms of the matrices and vectors defined above. We observe that $n_t^\top P$ is a vector in $\mathbb{R}^{|\sigma|}$ whose coordinate corresponding to feature vector σ equals the number of people in group t who have feature vector σ and belong to the positive class. $n_t^\top X$ is a vector in $\mathbb{R}^B$ whose coordinate corresponding to bin b equals the expected number of people in group t assigned to bin b.

By further multiplying these vectors on the right, we get additional useful quantities. Here are two in particular:

- $n_t^\top XV$ is a vector in $\mathbb{R}^B$ whose coordinate corresponding to bin b equals the expected sum of the scores assigned to all group-t people in bin b. That is, using the subscript b to denote the coordinate corresponding to bin b, we can write $(n_t^\top XV)_b = v_b(n_t^\top X)_b$ by the definition of the diagonal matrix

V.

- $n_t^\top PX$ is a vector in $\mathbb{R}^B$ whose coordinate corresponding to bin b equals the expected number of group-t people in the positive class who are placed in bin b.

Now, condition (A), that the risk assignment is calibrated within groups, implies that the two vectors above are equal coordinate-wise, and so we have the following equation for all t:

$$n_t^\top PX = n_t^\top XV \tag{3.1}$$

Calibration condition (A) also has an implication for the total score received by all people in group t. Suppose we multiply the two sides of (3.1) on the right by the vector $\mathbf{e} \in \mathbb{R}^B$ whose coordinates are all 1, obtaining

$$n_t^\top PX\mathbf{e} = n_t^\top XV\mathbf{e}. \tag{3.2}$$

The left-hand-side is the number of group-t people in the positive class. The right-hand-side, which we can also write as $n_t^\top Xv$, is equal to the sum of the expected scores received by all group-t people. These two quantities are thus the same, and we write their common value as μ_t.

Fairness to the positive and negative classes. We now want to write down vector equations corresponding to the fairness conditions (B) and (C) for the negative and positive classes. First, recall that for the B-dimensional vector $n_t^\top PX$, the coordinate corresponding to bin b equals the expected number of

group-t people in the positive class who are placed in bin b. Thus, to compute the sum of the expected scores received by all group-t people in the positive class, we simply need to take the inner product with the vector v, yielding $n_t^\top PXv$. Since μ_t is the total number of group-t people in the positive class, the average of the expected scores received by a group-t person in the positive class is the ratio $\frac{1}{\mu_t} n_t^\top PXv$. Thus, condition (C), that members of the positive class should receive the same average score in each group, can be written

$$\frac{1}{\mu_1} n_1^\top PXv = \frac{1}{\mu_2} n_2^\top PXv \tag{3.3}$$

Applying strictly analogous reasoning but to the fractions $1 - p_\sigma$ of people in the negative class, we can write condition (B), that members of the negative class should receive the same average score in each group, as

$$\frac{1}{N_1 - \mu_1} n_1^\top (I - P)Xv = \frac{1}{N_2 - \mu_2} n_2^\top (I - P)Xv \tag{3.4}$$

Using (3.1), we can rewrite (3.3) to get

$$\frac{1}{\mu_1} n_1^\top XVv = \frac{1}{\mu_2} n_2^\top XVv \tag{3.5}$$

Similarly, we can rewrite (3.4) as

$$\frac{1}{N_1 - \mu_1} (\mu_1 - n_1^\top XVv) = \frac{1}{N_2 - \mu_2} (\mu_2 - n_2^\top XVv) \tag{3.6}$$

The portion of the score received by the positive class. We think of the ratios on the two sides of (3.3), and equivalently (3.5), as the average of the expected scores received by a member of the positive class in group t: the numerator is the sum of the expected scores received by the members of the positive class, and the denominator is the size of the positive class. Let us denote this fraction by γ_t; we note that this is the quantity y used in the informal overview of the proof

at the start of the section. By (3.2), we can alternately think of the denominator as the sum of the expected scores received by all group-t people. Hence, the two sides of (3.3) and (3.5) can be viewed as representing the ratio of the sum of the expected scores in the positive class of group t to the sum of the expected scores in group t as a whole. (3.3) requires that $\gamma_1 = \gamma_2$; let us denote this common value by γ.

Now, we observe that $\gamma = 1$ corresponds to a case in which the sum of the expected scores in just the positive class of group t is equal to the sum of the expected scores in all of group t. In this case, it must be that all members of the negative class are assigned to bins of score 0. If any members of the positive class were assigned to a bin of score 0, this would violate the calibration condition (A); hence all members of the positive class are assigned to bins of positive score. Moreover, these bins of positive score contain no members of the negative class (since they've all been assigned to bins of score 0), and so again by the calibration condition (A), the members of the positive class are all assigned to bins of score 1. Finally, applying the calibration condition once more, it follows that the members of the negative class all have feature vectors σ with $p_\sigma = 0$ and the members of the positive class all have feature vectors σ with $p_\sigma = 1$. Hence, when $\gamma = 1$ we have perfect prediction.

Finally, we use our definition of γ_t as $\frac{1}{\mu_t} n_t^\top XVv$, and the fact that $\gamma_1 = \gamma_2 = \gamma$ to write (3.6) as

$$\begin{aligned}
\frac{1}{N_1 - \mu_1}(\mu_1 - \gamma\mu_1) &= \frac{1}{N_2 - \mu_2}(\mu_2 - \gamma\mu_2) \\
\frac{1}{N_1 - \mu_1}\mu_1(1 - \gamma) &= \frac{1}{N_2 - \mu_2}\mu_2(1 - \gamma) \\
\frac{\mu_1/N_1}{1 - \mu_1/N_1}(1 - \gamma) &= \frac{\mu_2/N_2}{1 - \mu_2/N_2}(1 - \gamma)
\end{aligned}$$

Now, this last equality implies that one of two things must be the case. Either $1 - \gamma = 0$, in which case $\gamma = 1$ and we have perfect prediction; or

$$\frac{\mu_1/N_1}{1 - \mu_1/N_1} = \frac{\mu_2/N_2}{1 - \mu_2/N_2},$$

in which case $\mu_1/N_1 = \mu_2/N_2$ and we have equal base rates. This completes the proof of Theorem 3.1.

Some Comments on the Connection to Statistical Parity. Earlier we noted that conditions (B) and (C) — the balance conditions for the positive and negative classes — are quite different from the requirement of *statistical parity*, which asserts that the average of the scores over *all* members of each group be the same.

When the two groups have equal base rates, then the risk assignment that gives the same score to everyone in the population achieves statistical parity along with conditions (A), (B), and (C). But when the two groups do not have equal base rates, it is immediate to show that statistical parity is inconsistent with both the calibration condition (A) and with the conjunction of the two balance conditions (B) and (C). To see the inconsistency of statistical parity with the calibration condition, we take Equation (3.1) from the proof above, sum the coordinates of the vectors on both sides, and divide by N_t, the number of people in group t. Statistical parity requires that the right-hand sides of the resulting equation be the same for $t = 1, 2$, while the assumption that the two groups have unequal base rates implies that the left-hand sides of the equation must be different for $t = 1, 2$. To see the inconsistency of statistical parity with the two balance conditions (B) and (C), we simply observe that if the average score assigned to the positive class and to the negative class are the same in the two

groups, then the average of the scores over all members of the two groups cannot be the same provided they do not contain the same proportion of positive-class and negative-class members.

3.3 The Approximate Theorem

In this section we prove Theorem 3.2. First, we must first give a precise specification of the approximate fairness conditions:

$$(1-\varepsilon)[n_t^\top XV]_b \leq [n_t^\top PX]_b \leq (1-\varepsilon)[n_t^\top XV]_b \tag{A'}$$

$$\begin{aligned}(1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv &\leq \left(\frac{1}{N_1-\mu_1}\right) n_t^\top (I-P)Xv \\ &\leq (1+\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv \end{aligned} \tag{B'}$$

$$(1-\varepsilon)\left(\frac{1}{\mu_2}\right) n_t^\top PXv \leq \left(\frac{1}{\mu_1}\right) n_t^\top PXv \leq (1+\varepsilon)\left(\frac{1}{\mu_2}\right) n_t^\top PXv \tag{C'}$$

For (B') and (C'), we also require that these hold when μ_1 and μ_2 are interchanged.

We also specify the approximate versions of perfect prediction and equal base rates in terms of $f(\varepsilon)$, which is a function that goes to 0 as ε goes to 0.

- *Approximate perfect prediction.* $\gamma_1 \geq 1 - f(\varepsilon)$ and $\gamma_2 \geq 1 - f(\varepsilon)$
- *Approximately equal base rates.* $|\mu_1/N_1 - \mu_2/N_2| \leq f(\varepsilon)$

A brief overview of the proof of Theorem 3.2 is as follows. It proceeds by first establishing an approximate form of Equation (3.1) above, which implies that the total expected score assigned in each group is approximately equal to the total size of the positive class. This in turn makes it possible to formulate

approximate forms of Equations (3.3) and (3.4). When the base rates are close together, the approximation is too loose to derive bounds on the predictive power; but this is okay since in this case we have approximately equal base rates. Otherwise, when the base rates differ significantly, we show that most of the expected score must be assigned to the positive class, giving us approximately perfect prediction.

The remainder of this section provides the full details of the proof.

Total scores and the number of people in the positive class. First, we will show that the total score for each group is approximately μ_t, the number of people in the positive class. Define $\hat{\mu}_t = n_t^\top Xv$. Using (A'), we have

$$\begin{aligned}
\hat{\mu}_t &= n_t^\top Xv \\
&= n_t^\top XVe \\
&= \sum_{b=1}^{B} [n_t^\top PX]_b \\
&\leq (1+\varepsilon) \sum_{b=1}^{B} [n_t^\top PX]_b \\
&= (1+\varepsilon) n_t^\top PXe \\
&= (1+\varepsilon)\mu_t
\end{aligned}$$

Similarly, we can lower bound $\hat{\mu}_t$ as

$$\begin{aligned}
\hat{\mu}_t &= \sum_{b=1}^{B} [n_t^\top PX]_b \\
&\geq (1-\varepsilon) \sum_{b=1}^{B} [n_t^\top PX]_b \\
&= (1-\varepsilon)\mu_t
\end{aligned}$$

Combining these, we have

$$(1-\varepsilon)\mu_t \leq \hat{\mu}_t \leq (1+\varepsilon)\mu_t. \tag{3.7}$$

The portion of the score received by the positive class. We can use (C′) to show that $\gamma_1 \approx \gamma_2$. Recall that γ_t, the average of the expected scores assigned to members of the positive class in group t, is defined as $\gamma_t = \frac{1}{\mu_t} n_t P X v$. Then, it follows trivially from (C′) that

$$(1-\varepsilon)\gamma_2 \leq \gamma_1 \leq (1+\varepsilon)\gamma_2. \tag{3.8}$$

The relationship between the base rates. We can apply this to (B′) to relate μ_1 and μ_2, using the observation that the score not received by people of the positive class must fall instead to people of the negative class. Examining the left inequality of (B′), we have

$$\begin{aligned}
(1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right) n_t^\top (I-P)Xv &= (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)(n_t^\top Xv - n_t^\top PXv) \\
&= (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)(\hat{\mu}_2 - \gamma_2\mu_2) \\
&\geq (1-\varepsilon)\left(\frac{1}{N_2-\mu_2}\right)((1-\varepsilon)\mu_2 - \gamma_2\mu_t) \\
&= (1-\varepsilon)\left(\frac{\mu_2}{N_2-\mu_2}\right)(1-\varepsilon-\gamma_2) \\
&\geq (1-\varepsilon)\left(\frac{\mu_2}{N_2-\mu_2}\right)\left(1-\varepsilon-\frac{\gamma_1}{1-\varepsilon}\right) \\
&= (1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right)
\end{aligned}$$

Thus, the left inequality of (B′) becomes

$$(1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right) \leq \left(\frac{1}{N_1-\mu_1}\right) n_t^\top (I-P)Xv \tag{3.9}$$

By definition, $\hat{\mu}_1 = n_t^\top Xv$ and $\gamma_t\mu_t = n_t^\top PXv$, so this becomes

$$(1-2\varepsilon+\varepsilon^2-\gamma_1)\left(\frac{\mu_2}{N_2-\mu_2}\right) \leq \left(\frac{1}{N_1-\mu_1}\right)(\hat{\mu}_1 - \gamma_1\mu_1) \tag{3.10}$$

If the base rates differ. Let ρ_1 and ρ_2 be the respective base rates, i.e. $\rho_1 = \mu_1/N_1$ and $\rho_2 = \mu_2/N_2$. Assume that $\rho_1 \leq \rho_2$ (otherwise we can switch μ_1 and μ_2 in the above analysis), and assume towards contradiction that the base rates differ by at least $\sqrt{\varepsilon}$, meaning $\rho_1 + \sqrt{\varepsilon} < \rho_2$. Using (3.10),

$$\begin{aligned}
\frac{\rho_1 + \sqrt{\varepsilon}}{1 - \rho_1 - \sqrt{\varepsilon}} \leq \frac{\rho_2}{1 - \rho_2} &\leq \left(\frac{1 + \varepsilon - \gamma_1}{1 - 2\varepsilon + \varepsilon^2 - \gamma_1}\right)\left(\frac{\rho_1}{1 - \rho_1}\right) \\
(\rho_1 + \sqrt{\varepsilon})(1 - \rho_1)(1 - 2\varepsilon + \varepsilon^2 - \gamma_1) &\leq \rho_1(1 - \rho_1 - \sqrt{\varepsilon})(1 + \varepsilon - \gamma_1) \\
(\rho_1 + \sqrt{\varepsilon})(1 - \rho_1)(1 - 2\varepsilon) - \rho_1(1 - \rho_1 - \sqrt{\varepsilon})(1 + \varepsilon) &\\
\leq \gamma_1\left[(\rho_1 + \sqrt{\varepsilon})(1 - \rho_1) - \rho_1(1 - \rho_1 - \sqrt{\varepsilon})\right] &\\
\rho_1[(1 - \rho_1)(1 - 2\varepsilon) - (1 - \rho_1 - \sqrt{\varepsilon})(1 + \varepsilon)] + \sqrt{\varepsilon}(1 - \rho_1)(1 - 2\varepsilon) &\\
\leq \gamma_1[\sqrt{\varepsilon}(1 - \rho_1) + \sqrt{\varepsilon}\rho_1] &\\
\rho_1(-2\varepsilon + 2\varepsilon\rho_1 - \varepsilon + \varepsilon\rho_1 + \sqrt{\varepsilon} + \varepsilon\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1 - 2\varepsilon - \rho_1 + 2\varepsilon\rho_1) &\leq \gamma_1\sqrt{\varepsilon} \\
\rho_1(-3\varepsilon + 3\varepsilon\rho_1 + \sqrt{\varepsilon} + \varepsilon\sqrt{\varepsilon} - \sqrt{\varepsilon} + 2\varepsilon\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1 - 2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
\varepsilon\rho_1(-3 + 3\rho_1 + 3\sqrt{\varepsilon}) + \sqrt{\varepsilon}(1 - 2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
3\varepsilon\rho_1(-1 + \rho_1) + \sqrt{\varepsilon}(1 - 2\varepsilon) &\leq \gamma_1\sqrt{\varepsilon} \\
1 - 2\varepsilon - 3\sqrt{\varepsilon}\rho_1(1 - \rho_1) &\leq \gamma_1 \\
1 - \sqrt{\varepsilon}\left(2\sqrt{\varepsilon} + \frac{3}{4}\right) &\leq \gamma_1
\end{aligned}$$

Recall that $\gamma_2 \geq \gamma_1(1 - \varepsilon)$, so

$$\begin{aligned}
\gamma_2 &\geq (1 - \varepsilon)\gamma_1 \\
&\geq (1 - \varepsilon)\left(1 - \sqrt{\varepsilon}\left(2\sqrt{\varepsilon} + \frac{3}{4}\right)\right) \\
&\geq 1 - \varepsilon - \sqrt{\varepsilon}\left(2\sqrt{\varepsilon} + \frac{3}{4}\right) \\
&= 1 - \sqrt{\varepsilon}\left(3\sqrt{\varepsilon} + \frac{3}{4}\right)
\end{aligned}$$

Let $f(\varepsilon) = \sqrt{\varepsilon}\max(1, 3\sqrt{\varepsilon} + 3/4)$. Note that we assumed that ρ_1 and ρ_2 differ by an additive $\sqrt{\varepsilon} \leq f(\varepsilon)$. Therefore if the ε-fairness conditions are met and the

base rates are not within an additive $f(\varepsilon)$, then $\gamma_1 \geq 1 - f(\varepsilon)$ and $\gamma_2 \geq 1 - f(\varepsilon)$. This completes the proof of Theorem 3.2.

3.4 Reducing Loss with Equal Base Rates

In a risk assignment, we would like as much of the score as possible to be assigned to members of the positive class. With this in mind, if an individual receives a score of v, we define their *individual loss* to be v if they belong to the negative class, and $1 - v$ if they belong to the positive class. The loss of the risk assignment in group t is then the sum of the expected individual losses to each member of group t. In terms of the matrix-vector products used in the proof of Theorem 3.1, one can show that the loss for group t may be written as

$$\begin{aligned} \ell_t(X) &= n_t^\top (I - P)Xv + (\mu_t - n_t^\top PXv) \\ &= 2(\mu_t - n_t^\top PXv), \end{aligned}$$

and the total loss is just the weighted sum of the losses for each group.

Now, let us say that a *fair assignment* is one that satisfies our three conditions (A), (B), and (C). As noted above, when the base rates in the two groups are equal, the set of fair assignments is non-empty, since the calibrated risk assignment that places everyone in a single bin is fair. We can therefore ask, in the case of equal base rates, whether there exists a fair assignment whose loss is strictly less than that of the trivial one-bin assignment. It is not hard to show that this is possible if and only if there is any assignment using more than one bin; we will call such an assignment a *non-trivial assignment*.

Note that the assignment that minimizes loss is simply the one that assigns

each σ to a separate bin with a score of p_σ, meaning X is the identity matrix. While this assignment, which we refer to as the identity assignment I, is well-calibrated, it may violate fairness conditions (B) and (C). It is not hard to show that the loss for any other assignment is strictly greater than the loss for I. As a result, unless the identity assignment happens to be fair, every fair assignment must have larger loss than that of I, forcing a tradeoff between performance and fairness.

3.4.1 Characterization of Well-Calibrated Solutions

To better understand the space of feasible solutions, suppose we drop the fairness conditions (B) and (C) for now and study risk assignments that are simply well-calibrated, satisfying (A). As in the proof of Theorem 3.1, we write γ_t for the average of the expected scores assigned to members of the positive class in group t, and we define the *fairness difference* to be $\gamma_1 - \gamma_2$. If this is nonnegative, we say the risk assignment *weakly favors* group 1; if it is nonpositive, it weakly favors group 2. Since a risk assignment is fair if and only if $\gamma_1 = \gamma_2$, it is fair if and only if the fairness difference is 0.

We wish to characterize when non-trivial fair risk assignments are possible. First, we observe that without the fairness requirements, the set of possible fairness differences under well-calibrated assignments is an interval.

Lemma 3.3. *If group 1 and group 2 have equal base rates, then for any two non-trivial well-calibrated risk assignments with fairness differences d_1 and d_2 and for any $d_3 \in [d_1, d_2]$, there exists a non-trivial well-calibrated risk assignment with fairness difference d_3.*

Proof. The basic idea is that we can effectively take convex combinations of well-calibrated assignments to produce any well-calibrated assignment "in between" them. We carry this out as follows.

Let $X^{(1)}$ and $X^{(2)}$ be the allocation matrices for assignments with fairness differences d_1 and d_2 respectively, where $d_1 < d_2$. Choose λ such that $\lambda d_1 + (1 - \lambda)d_2 = d_3$, meaning $\lambda = (d_2 - d_3)/(d_2 - d_1)$. Then, $X^{(3)} = [\lambda X^{(1)} \quad (1-\lambda)X^{(2)}]$ is a nontrivial well-calibrated assignment with fairness difference d_3.

First, we observe that $X^{(3)}$ is a valid assignment because each row sums to 1 (meaning everyone from every σ gets assigned to a bin), since each row of $\lambda X^{(1)}$ sums to λ and each row of $(1-\lambda)X^{(2)}$ sums to $(1-\lambda)$. Moreover, it is nontrivial because every nonempty bin created by $X^{(1)}$ and $X^{(2)}$ is a nonempty bin under $X^{(3)}$.

Let $v^{(1)}$ and $v^{(2)}$ be the respective bin labels for assignments $X^{(1)}$ and $X^{(2)}$. Define $v^{(3)} = \begin{bmatrix} v^{(1)} \\ v^{(2)} \end{bmatrix}$.

Finally, let $V^{(3)} = \text{diag}(v^{(3)})$. Define $V^{(1)}$ and $V^{(2)}$ analogously. Note that $V^{(3)} = \begin{bmatrix} V^{(1)} & 0 \\ 0 & V^{(2)} \end{bmatrix}$.

We observe that $X^{(3)}$ is calibrated because

$$\begin{aligned} n_t^\top P X^{(3)} &= n_t^\top P[\lambda X^{(1)} \quad (1-\lambda)X^{(2)}] \\ &= [\lambda n_t^\top P X^{(1)} \quad (1-\lambda) n_t^\top P X^{(2)}] \\ &= [\lambda n_t^\top X^{(1)} V^{(1)} \quad (1-\lambda) n_t^\top X^{(2)} V^{(2)}] \\ &= n_t^\top [\lambda X^{(1)} \quad (1-\lambda)X^{(2)}] V^{(3)} \\ &= n_t^\top X^{(3)} V^{(3)} \end{aligned}$$

Finally, we show that the fairness difference is d_3. Let $\gamma_1^{(1)}$ and $\gamma_2^{(1)}$ be the portions of the total expected score received by the positive class from each group respectively. Define $\gamma_1^{(2)}, \gamma_2^{(2)}, \gamma_1^{(3)}, \gamma_2^{(3)}$ similarly.

$$\begin{aligned} \gamma_1^{(3)} - \gamma_2^{(3)} &= \frac{1}{\mu} n_1^\top P X^{(3)} v^{(3)} - \frac{1}{\mu} n_2^\top P X^{(3)} v^{(3)} \\ &= \frac{1}{\mu}(n_1^\top - n_2^\top) P X^{(3)} v^{(3)} \\ &= \frac{1}{\mu}(n_1^\top - n_2^\top) P [\lambda X^{(1)} v^{(1)} \quad (1-\lambda) X^{(2)} v^{(2)}] \\ &= \frac{1}{\mu}(\lambda (n_1^\top - n_2^\top) P X^{(1)} v^{(1)} + (1-\lambda)(n_1^\top - n_2^\top) X^{(2)} v^{(2)}]) \\ &= \lambda(\gamma_1^{(1)} - \gamma_2^{(1)}) + (1-\lambda)(\gamma_1^{(2)} - \gamma_2^{(2)}) \\ &= \lambda d_1 + (1-\lambda) d_2 \\ &= d_3 \end{aligned}$$

□

Corollary 3.4. *There exists a non-trivial fair assignment if and only if there exist non-trivial well-calibrated assignments $X^{(1)}$ and $X^{(2)}$ such that $X^{(1)}$ weakly favors group 1 and $X^{(2)}$ weakly favors group 2.*

Proof. If there is a non-trivial fair assignment, then it weakly favors both group 1 and group 2, proving one direction.

To prove the other direction, observe that the fairness differences d_1 and d_2 of $X^{(1)}$ and $X^{(2)}$ are nonnegative and nonpositive respectively. Since the set of fairness differences achievable by non-trivial well-calibrated assignments is an interval by Lemma 3.3, there exists a non-trivial well-calibrated assignment with fairness difference 0, meaning there exists a non-trivial fair assignment. □

It is an open question whether there is a polynomial-time algorithm to find a fair assignment of minimum loss, or even to determine whether a non-trivial fair solution exists.

3.4.2 NP-Completeness of Non-Trivial Integral Fair Risk Assignments

As discussed in the introduction, risk assignments in our model are allowed to split people with a given feature vector σ over several bins; however, it is also of interest to consider the special case of *integral* risk assignments, in which all people with a given feature σ must go to the same bin. For the case of equal base rates, we can show that determining whether there is a non-trivial integral fair assignment is NP-complete. The proof uses a reduction from the Subset Sum problem and is given Appendix A.

The basic idea of the reduction is as follows. We have an instance of Subset Sum with numbers $w_1, \ldots, w_m$ and a target number T; the question is whether there is a subset of the w_i's that sums to T. As before, γ_t denotes the average of the expected scores received by members of the positive class in group t. We first ensure that there is exactly one non-trivial way to allocate the people of group

1, allowing us to control γ_1. The fairness conditions then require that $\gamma_2 = \gamma_1$, which we can use to encode the target value in the instance of Subset Sum. For every input number w_i in the Subset Sum instance, we create $p_{\sigma_{2i-1}}$ and $p_{\sigma_{2i}}$, close to each other in value and far from all other p_σ values, such that grouping σ_{2i-1} and σ_{2i} together into a bin corresponds to choosing w_i for the subset, while not grouping them corresponds to not taking w_i. This ensures that group 2 can be assigned with the correct value of γ_2 if and only if there is a solution to the Subset Sum instance.

3.5 Conclusion

In this chapter we have formalized three fundamental conditions for risk assignments to individuals, each of which has been proposed as a basic measure of what it means for the risk assignment to be fair. Our main results show that except in highly constrained special cases, it is not possible to satisfy these three constraints simultaneously; and moreover, a version of this fact holds in an approximate sense as well.

Since these results hold regardless of the method used to compute the risk assignment, it can be phrased in fairly clean terms in a number of domains where the trade-offs among these conditions do not appear to be well-understood. To take one simple example, suppose we want to determine the risk that a person is a carrier for a disease X, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of X, at least one of the following undesirable properties must hold: (a) the test's probability estimates are systematically

skewed upward or downward for at least one gender; or (b) the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or (c) the test assigns a higher average risk estimate to carriers of the disease in one gender than the other. The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups.

Finally, we note that our results suggest a number of interesting directions for further work. First, when the base rates between the two underlying groups are equal, our results do not resolve the computational tractability of finding the most accurate risk assignment, subject to our three fairness conditions, when the people with a given feature vector can be split across multiple bins. (Our NP-completeness result applies only to the case in which everyone with a given feature vector must be assigned to the same bin.) Second, there may be a number of settings in which the cost (social or otherwise) of false positives may differ greatly from the cost of false negatives. In such cases, we could imagine searching for risk assignments that satisfy the calibration condition together with only one of the two balance conditions, corresponding to the class for whom errors are more costly. Determining when two of our three conditions can be simultaneously satisfied in this way is an interesting open question. More broadly, determining how the trade-offs discussed here can be incorporated into broader families of proposed fairness conditions suggests interesting avenues for future research.

CHAPTER 4

ON FAIRNESS AND CALIBRATION

Recently, there has been growing concern about errors of machine learning algorithms in sensitive domains – including criminal justice, online advertising, and medical testing (Executive Office of the President, 2016) – which may systematically discriminate against particular groups of people (Barocas and Selbst, 2016; Berk et al., 2018; Chouldechova, 2017).

A recent high-profile example of these concerns was raised by the news organization ProPublica, who studied a risk-assessment tool that is widely used in the criminal justice system. This tool assigns to each criminal defendant an estimated probability that they will commit a future crime. ProPublica found that the risk estimates assigned to defendants who did not commit future crimes were on average higher among African-American defendants than Caucasian defendants (Angwin et al., 2016). This is a form of false-positive error, and in this case it disproportionately affected African-American defendants. To mitigate issues such as these, the machine learning community has proposed different frameworks that attempt to quantify fairness in classification (Barocas and Selbst, 2016; Berk et al., 2018; Chouldechova, 2017; Hardt et al., 2016b; Kleinberg et al., 2017; Woodworth et al., 2017; Zafar et al., 2017a). A recent and particularly noteworthy framework is Equalized Odds (Hardt et al., 2016b) (also referred to as Disparate Mistreatment (Zafar et al., 2017a)),[1] which constrains classification algorithms such that no error type (false-positive or false-negative) disproportionately affects any population subgroup. This notion of non-discrimination is feasible in many settings, and researchers have developed tractable algorithms

[1]For the remainder of the chapter, we will use *Equalized Odds* to refer to this notion of non-discrimination.

for achieving it (Hardt et al., 2016b; Goh et al., 2016; Zafar et al., 2017a; Woodworth et al., 2017).

When risk tools are used in practice, a key goal is that they are *calibrated*: if we look at the set of people who receive a predicted probability of p, we would like a p fraction of the members of this set to be positive instances of the classification problem (Dawid, 1982). Moreover, if we are concerned about fairness between two groups G_1 and G_2 (e.g. African-American defendants and white defendants) then we would like this calibration condition to hold simultaneously for the set of people within each of these groups as well (Flores et al., 2016).

Calibration is a crucial condition for risk tools in many settings. If a risk tool for evaluating defendants were not calibrated with respect to groups defined by race, for example, then a probability estimate of p could carry different meaning for African-American and white defendants, and hence the tool would have the unintended and highly undesirable consequence of incentivizing judges to take race into account when interpreting its predictions.

Despite the importance of calibration as a property, our understanding of how it interacts with other fairness properties is limited. We know from recent work that, except in the most constrained cases, it is impossible to achieve calibration while also satisfying Equalized Odds (Kleinberg et al., 2017; Chouldechova, 2017). However, we do not know how best to achieve relaxations of these guarantees that are feasible in practice.

Our goal is to further investigate the relationship between calibration and error rates. We show that even if the Equalized Odds conditions are relaxed

substantially – requiring only that weighted sums of the group error rates match – it is still problematic to also enforce calibration. We provide necessary and sufficient conditions under which this calibrated relaxation is feasible. When feasible, it has a unique optimal solution that can be achieved through post-processing of existing classifiers. Moreover, we provide a simple post-processing algorithm to find this solution: withholding predictive information for randomly chosen inputs to achieve parity and preserve calibration. However, this simple post-processing method is fundamentally unsatisfactory: although the post-processed predictions of our information-withholding algorithm are "fair" in expectation, most practitioners would object to the fact that a non-trivial portion of the individual predictions are withheld as a result of coin tosses – especially in sensitive settings such as health care or criminal justice. The optimality of this algorithm thus has troubling implications and shows that calibration and error-rate fairness are inherently at odds (even beyond the initial results by Chouldechova (2017) and Kleinberg et al. (2017)).

Finally, we evaluate these theoretical findings empirically, comparing calibrated notions of non-discrimination against the (uncalibrated) Equalized Odds framework on several datasets. These experiments further support our conclusion that calibration and error-rate constraints are in most cases mutually incompatible goals. In practical settings, it may be advisable to choose only one of these goals rather than attempting to achieve some relaxed notion of both.

4.1 Problem Setup

The setup of our framework most follows the *Equalized Odds* framework (Hardt et al., 2016b; Zafar et al., 2017a); however, we extend their framework for use

with probabilistic classifiers. Let $P \subset \mathbb{R}^k \times \{0,1\}$ be the input space of a binary classification task. In our criminal justice example, $(\mathbf{x}, y) \sim P$ represents a person, with $\mathbf{x}$ representing the individual's history and y representing whether or not the person will commit another crime. Additionally, we assume the presence of two groups $G_1, G_2 \subset P$, which represent disjoint population subsets, such as different races. We assume that the groups have different *base rates* μ_t, or probabilities of belonging to the positive class: $\mu_1 = \Pr_{(\mathbf{x},y)\sim G_1}[y=1] \neq \Pr_{(\mathbf{x},y)\sim G_2}[y=1] = \mu_2$.

Finally, let $h_1, h_2 : \mathbb{R}^k \to [0,1]$ be binary classifiers, where h_1 classifies samples from G_1 and h_2 classifies samples from G_2.[2] Each classifier outputs the probability that a given sample $\mathbf{x}$ belongs to the positive class. The notion of Equalized Odds non-discrimination is based on the false-positive and false-negative rates for each group, which we generalize here for use with probabilistic classifiers:

Definition 4.1. *The* generalized false-positive rate *of classifier h_t for group G_t is* $c_{fp}(h_t) = \mathbb{E}_{(\mathbf{x},y)\sim G_t}\big[h_t(\mathbf{x}) \mid y=0\big]$. *Similarly, the* generalized false-negative rate *of classifier h_t is* $c_{fn}(h_t) = \mathbb{E}_{(\mathbf{x},y)\sim G_t}\big[(1-h_t(\mathbf{x})) \mid y=1\big]$.

If the classifier were to output either 0 or 1, this represents the standard notions of false-positive and false-negative rates. We now define the Equalized Odds framework (generalized for probabilistic classifiers), which aims to ensure that errors of a given type are not biased against any group.

Definition 4.2 (Probabilistic Equalized Odds)**.** *Classifiers h_1 and h_2 exhibit Equalized Odds for groups G_1 and G_2 if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$.*

[2] In practice, h_1 and h_2 can be trained jointly (i.e. they are the same classifier).

Calibration Constraints. As stated in the introduction, these two conditions do not necessarily prevent discrimination if the classifier predictions do not represent well-calibrated probabilities. Recall that calibration intuitively says that probabilities should carry semantic meaning: if there are 100 people in G_1 for whom $h_1(\mathbf{x}) = 0.6$, then we expect 60 of them to belong to the positive class.

Definition 4.3. *A classifier h_t is* perfectly calibrated *if* $\forall p \in [0, 1]$, $\Pr_{(\mathbf{x},y)\sim G_t}\big[y{=}1 \mid h_t(\mathbf{x}){=}p\big] = p$.

It is commonly accepted amongst practitioners that both classifiers h_1 and h_2 should be calibrated *with respect to groups* G_1 and G_2 to prevent discrimination (Berk et al., 2018; Crowson et al., 2016; Dieterich et al., 2016; Flores et al., 2016). Intuitively, this prevents the probability scores from carrying group-specific information. Unfortunately, Kleinberg et al. (2017) (as well as Chouldechova (2017), in a binary setting) prove that a classifier cannot achieve both calibration and Equalized Odds, even in an approximate sense, except in the most trivial of cases.

4.1.1 Geometric Characterization of Constraints

We now will characterize the calibration and error-rate constraints with simple geometric intuitions. Throughout the rest of this chapter, all of our results can be easily derived from this interpretation. We begin by defining the region of classifiers which are *trivial*, or those that output a constant value for all inputs (i.e. $h^c(\mathbf{x}) = c$, where $0 \leq c \leq 1$ is a constant). We can visualize these classifiers on a graph with generalized false-positive rates on one axis and generalized false-negatives on the other. It follows from the definitions

of generalized false-positive/false-negative rates and calibration that all trivial classifiers h lie on the diagonal defined by $c_{fp}(h) + c_{fn}(h) = 1$ (Figure 4.1a). Therefore, all classifiers that are "better than random" must lie below this diagonal in false-positive/false-negative space (the gray triangle in the figure). Any classifier that lies above the diagonal performs "worse than random," as we can find a point on the trivial classifier diagonal with lower false-positive and false-negative rates.

Now we will characterize the set of calibrated classifiers for groups G_1 and G_2, which we denote as $\mathcal{H}_1^*$ and $\mathcal{H}_2^*$. Kleinberg et al. show that the generalized false-positive and false-negative rates of a calibrated classifier are linearly related by the base rate of the group:[3]

$$c_{fn}(h_t) = (1 - \mu_t)/\mu_t \, c_{fp}(h_t). \tag{4.1}$$

In other words, h_1 lies on a line with slope $(1 - \mu_1)/\mu_1$ and h_2 lies on a line with slope $(1 - \mu_2)/\mu_2$ (Figure 4.1a). The lower endpoint of each line is the *perfect classifier*, which assigns the correct prediction with complete certainty to every input. The upper endpoint is a trivial classifier, as no calibrated classifier can perform "worse than random" (see Lemma B.5 in Appendix B.2). The only trivial classifier that satisfies the calibration condition for a group G_t is the one that outputs the base rate μ_t. We will refer to h^{μ_1} and h^{μ_2} as the trivial classifiers, calibrated for groups G_1 and G_2 respectively. It follows from the definitions that $c_{fp}(h^{\mu_1}) = \mu_1$ and $c_{fn}(h^{\mu_1}) = 1 - \mu_1$, and likewise for h^{μ_2}.

Finally, it is worth noting that for calibrated classifiers, a lower false-positive rate necessarily corresponds to a lower false-negative rate and vice-versa. In

[3]Throughout this chapter we will treat the calibration constraint as holding exactly; however, our results generalize to approximate settings as well. See Appendix B for more details.

other words, for a given base rate, a "better" calibrated classifier lies closer to the origin on the line of calibrated classifiers.

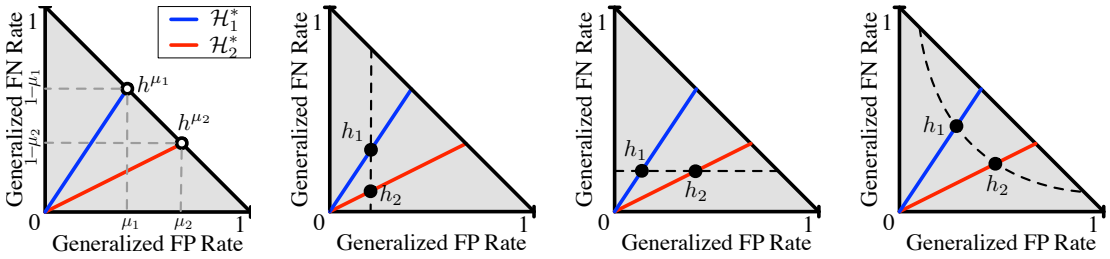
Impossibility of Equalized Odds with Calibration. With this geometric intuition, we can provide a simplified proof of the main impossibility result from Kleinberg et al. (2017):

Theorem (Impossibility Result (Kleinberg et al., 2017))**.** *Let h_1 and h_2 be classifiers for groups G_1 and G_2 with $\mu_1 \neq \mu_2$. h_1 and h_2 satisfy the Equalized Odds and calibration conditions if and only if h_1 and h_2 are perfect predictors.*

Intuitively, the three conditions define a set of classifiers which is overconstrained. Equalized Odds stipulates that the classifiers h_1 and h_2 must lie on the same coordinate in the false-positive/false-negative plane. As h_1 must lie on the blue line of calibrated classifiers for $\mathcal{H}_1^*$ and h_2 on the red line $\mathcal{H}_2^*$ they can only satisfy EO at the unique intersection point — the origin (and location of the perfect classifier). This implies that unless the two classifiers achieve perfect accuracy, we must relax the Equalized Odds conditions if we want to maintain calibration.

4.2 Relaxing Equalized Odds to Preserve Calibration

In this section, we show that a substantially simplified notion of Equalized Odds is compatible with calibration. We introduce a general relaxation that seeks to satisfy a *single equal-cost constraint* while maintaining calibration for each group G_t. We begin with the observation that Equalized Odds sets constraints to equalize false-positives $c_{fp}(h_t)$ and false-negatives $c_{fn}(h_t)$. To capture and generalize



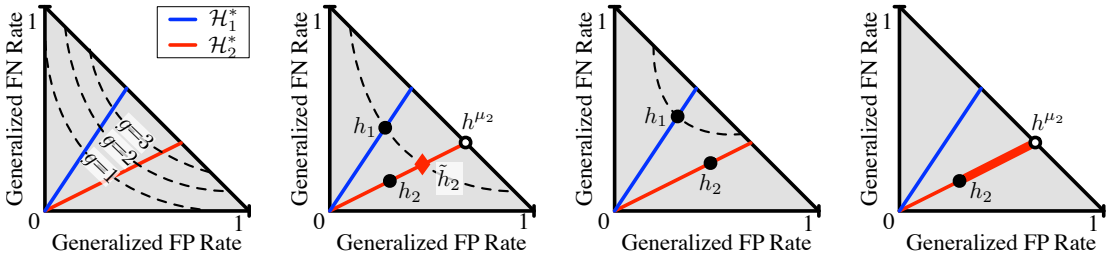
(a) Possible cal. classifiers $\mathcal{H}_1^*, \mathcal{H}_2^*$ (blue/red).

(b) Satisfying cal. and equal F.P. rates.

(c) Satisfying cal. and equal F.N. rates.

(d) Satisfying cal. and a general constraint.

Figure 4.1: Calibration, trivial classifiers, and equal-cost constraints – plotted in the false-pos./false-neg. plane. $\mathcal{H}_1^*, \mathcal{H}_2^*$ are the set of cal. classifiers for the two groups, and h^{μ_1}, h^{μ_2} are trivial classifiers.



(a) Level-order curves of cost. Low cost implies low error rates.

(b) Usually, there is a calibrated classifier $\tilde{h}_2$ with the same cost of h_1.

(c) Cal. and equal-cost are incompatible if h_1 has high error.

(d) Possible cal. classifiers for G_2 (bold red) by mixing h_2 and h^{μ_2}.

Figure 4.2: Calibration-Preserving Parity through interpolation.

this, we define a *cost function* g_t to be a linear function in $c_{fp}(h_t)$ and $c_{fn}(h_t)$ with arbitrary dependence on the group's base rate μ_t. More formally, a cost function for group G_t is

$$g_t(h_t) = a_t c_{fp}(h_t) + b_t c_{fn}(h_t) \tag{4.2}$$

where a_t and b_t are non-negative constants that are specific to each group (and thus may depend on μ_t): see Figure 4.1d. We also make the assumption that for any μ_t, at least one of a_t and b_t is nonzero, meaning $g_t(h_t) = 0$ if and only if $c_{fp}(h_t) = c_{fn}(h_t) = 0$.[4] This class of cost functions encompasses a variety of sce-

[4]By calibration, we cannot have one of $c_{fp}(h_t) = 0$ or $c_{fn}(h_t) = 0$ without the other, see Figure 4.1a.

narios. As an example, imagine an application in which the equal false-positive condition is essential but not the false-negative condition. Such a scenario may arise in our recidivism-prediction example, if we require that non-repeat offenders of any race are not disproportionately labeled as high risk. If we plot the set of calibrated classifiers $\mathcal{H}_1^*$ and $\mathcal{H}_2^*$ on the false-positive/false-negative plane, we can see that ensuring the false-positive condition requires finding classifiers $h_1 \in \mathcal{H}_1^*$ and $h_2 \in \mathcal{H}_2^*$ that fall on the same vertical line (Figure 4.1b). Conversely, if we instead choose to satisfy only the false-negative condition, we would find classifiers h_1 and h_2 that fall on the same horizontal (Figure 4.1c). Finally, if both false-positive and false-negative errors incur a negative cost on the individual, we may choose to equalize a weighted combination of the error rates (Berk, 2016; Berk et al., 2018; Chouldechova, 2017), which can be graphically described by the classifiers lying on a convex and negatively-sloped level set (Figure 4.1d). With these definitions, we can formally define our relaxation:

Definition 4.4 (Relaxed Equalized Odds with Calibration)**.** *Given a cost function g_t of the form in* (4.2)*, classifiers h_1 and h_2 achieve* Relaxed Equalized Odds with Calibration *for groups G_1 and G_2 if both classifiers are calibrated and satisfy the constraint $g_1(h_1) = g_2(h_2)$.*

It is worth noting that, for calibrated classifiers, an increase in cost strictly corresponds to an increase in both the false-negative and false-positive rate. This can be interpreted graphically, as the level-order cost curves lie further away from the origin as cost increases (Figure 4.2a). In other words, the cost function can always be used as a proxy for either error rate.[5]

[5]This holds even for approximately calibrated classifiers — see Appendix B.3.

Feasibility. It is easy to see that Definition 4.4 is always satisfiable – in Figures 4.1b, 4.1c, and 4.1d we see that there are many such solutions that would lie on a given level-order cost curve while maintaining calibration, including the case in which both classifiers are perfect. In practice, however, not all classifiers are achievable. For the rest of the chapter, we will assume that we have access to "optimal" (but possibly discriminatory) calibrated classifiers h_1 and h_2 such that, due to whatever limitations there are on the predictability of the task, we are unable to find other classifiers that have lower cost with respect to g_t. We allow h_1 and h_2 to be learned in any way, as long as they are calibrated. Without loss of generality, for the remainder of the chapter, we will assume that $g_1(h_1) \geq g_2(h_2)$.

Since by assumption we have no way to find a classifier for G_1 with lower cost than h_1, our goal is therefore to find a classifier $\tilde{h}_2$ with cost equal to h_1. This pair of classifiers would represent the lowest cost (and therefore optimal) set of classifiers that satisfies calibration and the equal cost constraint. For a given base rate μ_t and value of the cost function g_t, a calibrated classifier's position in the generalized false-positive/false-negative plane is uniquely determined (Figure 4.2a). This is because each level-order curve of the cost function g_t has negative slope in this plane, and each level order curve only intersects a group's calibrated classifier line once. In other words, there is a unique solution in the false-positive/false-negative plane for classifier $\tilde{h}_2$ (Figure 4.2b).

Consider the range of values that g_t can take. As noted above, $g_t(h_t) \geq 0$, with equality if and only if h_t is the perfect classifier. On the other hand, the trivial classifier (again, which outputs the constant μ_t for all inputs) is the calibrated classifier that achieves maximum cost for any g_t (see Lemma B.5 in Ap-

pendix B.2). As a result, the cost of a classifier for group G_t is between 0 and $g_t(h^{\mu_t})$. This naturally leads to a characterization of feasibility: Definition 4.4 can be achieved if and only if h_1 incurs less cost than group G_2's trivial classifier h^{μ_2}; i.e. if $g_1(h_1) \leq g_2(h^{\mu_2})$. This can be seen graphically in Figure 4.2c, in which the level-order curve for $g_1(h_1)$ does not intersect the set of calibrated classifiers for G_2. Since, by assumption, we cannot find a calibrated classifier for G_1 with strictly smaller cost than h_1, there is no feasible solution. On the other hand, if h_1 incurs less cost than h^{μ_2}, then we will show feasibility by construction with a simple algorithm.

An Algorithm. While it may be possible to encode the constraints of Definition 4.4 into the training procedure of h_1 and h_2, it is not immediately obvious how to do so. Even naturally probabilistic algorithms, such as logistic regression, can become uncalibrated in the presence of optimization constraints (as is the case in Zafar et al. (2017a)). It is not straightforward to encode the calibration constraint if the probabilities are assumed to be continuous, and post-processing calibration methods (Platt, 1999; Zadrozny and Elkan, 2001) would break equal-cost constraints by modifying classifier scores. Therefore, we look to achieve the calibrated Equalized Odds relaxation by post-processing existing calibrated classifiers.

Again, given h_1 and h_2 with $g_1(h_1) \geq g_2(h_2)$, we want to arrive at a calibrated classifier $\tilde{h}_2$ for group G_2 such that $g_1(h_1) = g_2(\tilde{h}_2)$. Recall that, under our assumptions, this would be the best possible solution with respect to classifier cost. We show that this cost constraint can be achieved by withholding predictive information for a randomly chosen subset of group G_2. In other words, rather than always returning $h_2(\mathbf{x})$ for all samples, we will occasionally return

the group's mean probability (i.e. the output of the trivial classifier h^{μ_2}). In Lemma B.6 in Appendix B.2, we show that if

$$\tilde{h}_2(\mathbf{x}) = \begin{cases} h^{\mu_2}(\mathbf{x}) = \mu_2 & \text{with probability } \alpha \\ h_2(\mathbf{x}) & \text{with probability } 1-\alpha \end{cases} \tag{4.3}$$

then the cost of $\tilde{h}_2$ is a linear interpolation between the costs of h_2 and h^{μ_2} (Figure 4.2d). More formally, we have that $g_2(\tilde{h}_2) = (1-\alpha)g_2(h_2) + \alpha g_2(h^{\mu_2}))$, and thus setting $\alpha = \frac{g_1(h_1)-g_2(h_2)}{g_2(h^{\mu_2})-g_2(h_2)}$ ensures that $g_2(\tilde{h}_2) = g_1(h_1)$ as desired (Figure 4.2b). Moreover, this randomization preserves calibration (see Appendix B.4). Algorithm 1 summarizes this method.

Algorithm 1 Achieving Calibration and an Equal-Cost Constraint via Information Withholding

Input: classifiers h_1 and h_2 s.t. $g_2(h_2) \leq g_1(h_1) \leq g_2(h^{\mu_2})$, holdout set P_{valid}.

- Determine base rate μ_2 of G_2 (using P_{valid}) to produce trivial classifier h^{μ_2}.
- Construct $\tilde{h}_2$ using with $\alpha = \frac{g_1(h_1)-g_2(h_2)}{g_2(h^{\mu_2})-g_2(h_2)}$, where α is the interpolation parameter.

return h_1, $\tilde{h}_2$ — which are calibrated and satisfy $g_1(h_1) = g_2(\tilde{h}_2)$.

Implications. In a certain sense, Algorithm 1 is an "optimal" method because it arrives at the unique false-negative/false-positive solution for $\tilde{h}_2$, where $\tilde{h}_2$ is calibrated and has cost equal to h_1. Therefore (by our assumptions) we can find no better classifiers that satisfy Definition 4.4. This simple result has strong consequences, as the tradeoffs to satisfy both calibration and the equal-cost constraint are often unsatisfactory — both intuitively and experimentally (as we will show in Section 4.3).

We find two primary objections to this solution. First, it equalizes costs

simply by making a classifier strictly worse for one of the groups. Second, it achieves this cost increase by withholding information on a randomly chosen population subset, making the outcome inequitable within the group (as measured by a standard measure of inequality like the Gini coefficient). Due to the optimality of the algorithm, the former of these issues is unavoidable in *any* solution that satisfies Definition 4.4. The latter, however, is slightly more subtle, and brings up the question of *individual fairness* (what guarantees we would like an algorithm to make with respect to each individual) and how it interacts with *group fairness* (population-level guarantees). While this certainly is an important issue for future work, in this particular setting, even if one could find another algorithm that distributes the burden of additional cost more equitably, any algorithm will make at least as many false-positive/false-negative errors as Algorithm 1, and these misclassifications will always be tragic to the individuals whom they affect. The performance loss across the entire group is often significant enough to make this combination of constraints somewhat worrying to use in practice, regardless of the algorithm.

Impossibility of Satisfying Multiple Equal-Cost Constraints. It is natural to argue there might be multiple cost functions that we would like to equalize across groups. However, satisfying more than one distinct equal-cost constraint (i.e. different curves in the F.P./F.N. plane) is infeasible.

Theorem 4.5 (Generalized impossibility result)**.** *Let h_1 and h_2 be calibrated classifiers for G_1 and G_2 with equal cost with respect to g_t. If $\mu_1 \neq \mu_2$, and if h_1 and h_2 also have equal cost with respect to a different cost function g'_t, then h_1 and h_2 must be perfect classifiers.*

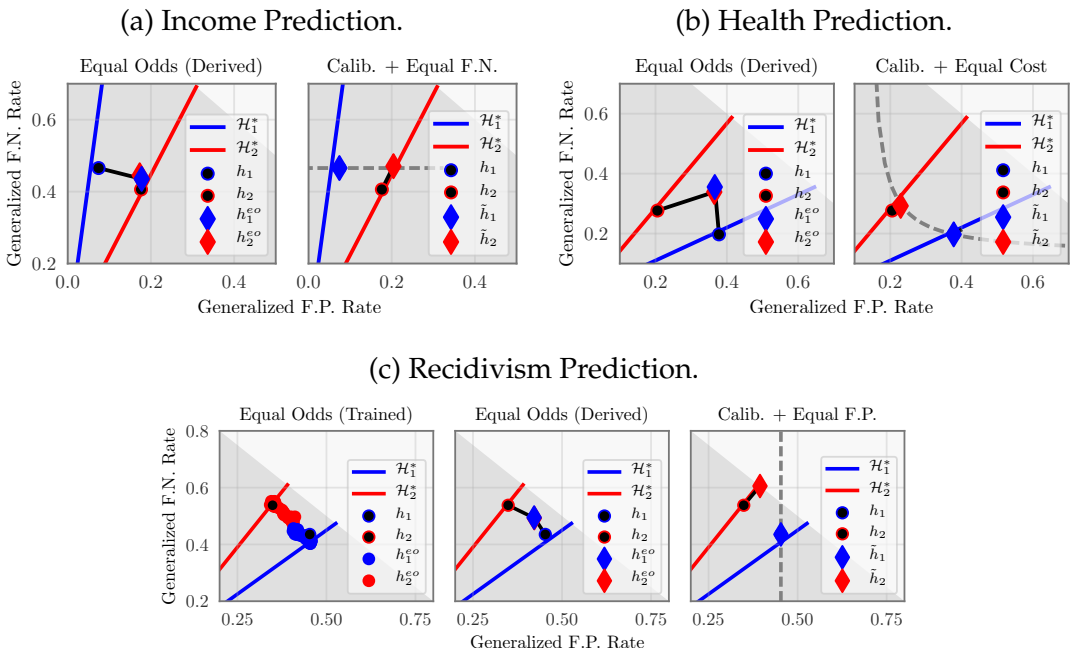


Figure 4.3: Generalized F.P. and F.N. rates for two groups under Equalized Odds and the calibrated relaxation. Diamonds represent post-processed classifiers. Points on the Equalized Odds (trained) graph represent classifiers achieved by modifying constraint hyperparameters.

(Proof in Appendix B.5). Note that this is a generalization of the impossibility result of Kleinberg et al. (2017). Furthermore, we show in Theorem B.11 (in Appendix B.5) that this holds in an approximate sense: if calibration and multiple distinct equal-cost constraints are approximately achieved by some classifier, then that classifier must have approximately zero generalized false-positive and false-negative rates.

4.3 Experiments

In light of these findings, our goal is to understand the impact of imposing calibration and an equal-cost constraint on real-world datasets. We will empirically show that, in many cases, this will result in performance degradation,

while simultaneously increasing other notions of disparity. We perform experiments on three datasets: an income-prediction, a health-prediction, and a criminal recidivism dataset. For each task, we choose a cost function within our framework that is appropriate for the given scenario. We begin with two calibrated classifiers h_1 and h_2 for groups G_1 and G_2. We assume that these classifiers cannot be significantly improved without more training data or features. We then derive $\tilde{h}_2$ to equalize the costs while maintaining calibration. The original classifiers are trained on a portion of the data, and then the new classifiers are derived using a separate holdout set. To compare against the (uncalibrated) Equalized Odds framework, we derive F.P./F.N. matching classifiers using the post-processing method of Hardt et al. (2016b) (**EO-Derived**). On the criminal recidivism dataset, we additionally learn classifiers that directly encode the Equalized Odds constraints, using the methods of Zafar et al. (2017a) (**EO-Trained**). (See Appendix B.6 for detailed training and post-processing procedures.) We visualize model error rates on the generalized F.P. and F.N. plane. Additionally, we plot the calibrated classifier lines for G_1 and G_2 to visualize model calibration.

Income Prediction. The Adult Dataset from UCI Machine Learning Repository (Lichman, 2013) contains 14 demographic and occupational features for various people, with the goal of predicting whether a person's income is above $\$50,000$. In this scenario, we seek to achieve predictions with equalized cost across genders (G_1 represents women and G_2 represents men). We model a scenario where the primary concern is ensuring equal generalized F.N. rates across genders, which would, for example, help job recruiters prevent gender discrimination in the form of underestimated salaries. Thus, we choose our cost

constraint to require equal generalized F.N. rates across groups. In Figure 4.3a, we see that the original classifiers h_1 and h_2 approximately lie on the line of calibrated classifiers. In the left plot (EO-Derived), we see that it is possible to (approximately) match both error rates of the classifiers at the cost of h_1^{eo} deviating from the set of calibrated classifiers. In the right plot, we see that it is feasible to equalize the generalized F.N. rates while maintaining calibration. h_1 and $\tilde{h}_2$ lie on the same level-order curve of g_t (represented by the dashed-gray line), and simultaneously remain on the "line" of calibrated classifiers. It is worth noting that achieving either notion of non-discrimination requires some cost to at least one of the groups. However, maintaining calibration further increases the difference in F.P. rates between groups. In some sense, the calibrated framework trades off one notion of disparity for another while simultaneously increasing the overall error rates.

Health Prediction. The Heart Dataset from the UCI Machine Learning Repository contains 14 processed features from 906 adults in 4 geographical locations. The goal of this dataset is to accurately predict whether or not an individual has a heart condition. In this scenario, we would like to reduce disparity between middle-aged adults (G_1) and seniors (G_2). In this scenario, we consider F.P. and F.N. to both be undesirable. A false prediction of a heart condition could result in unnecessary medical attention, while false negatives incur cost from delayed treatment. We therefore utilize the following cost function $g_t(h_t) = r_{fp} h_t(\mathbf{x})\,(1-y) + r_{fn}\,(1-h_t(\mathbf{x}))\,y$, which essentially assigns a weight to both F.N. and F.P. predictions. In our experiments, we set $r_{fp} = 1$ and $r_{fn} = 3$. In the right plot of Figure 4.3b, we can see that the level-order curves of the cost function form a curved line in the generalized F.P./F.N. plane. Because our orig-

inal classifiers lie approximately on the same level-order curve, little change is required to equalize the costs of h_1 and $\tilde{h}_2$ while maintaining calibration. This is the only experiment in which the calibrated framework incurs little additional cost, and therefore could be considered a viable option. However, it is worth noting that, in this example, the equal-cost constraint does not explicitly match either of the error types, and therefore the two groups will in expectation experience different types of errors. In the left plot of Figure 4.3b (EO-Derived), we see that it is alternatively feasible to explicitly match both the F.P. and F.N. rates while sacrificing calibration.

Criminal Recidivism Prediction. Finally, we examine the frameworks in the context of our motivating example: criminal recidivism. As mentioned in the introduction, African Americans (G_1) receive a disproportionate number of F.P. predictions as compared with Caucasians (G_2) when automated risk tools are used in practice. Therefore, we aim to equalize the generalized F.P. rate. In this experiment, we modify the predictions made by the COMPAS tool (Dieterich et al., 2016), a risk-assessment tool used in practice by the American legal system. Additionally, we also see if it is possible to improve the classifiers with training-time Equalized Odds constraints using the methods of Zafar et al. (2017a) (EO-Trained). In Figure 4.3c, we first observe that the original classifiers h_1 and h_2 have large generalized F.P. and F.N. rates. Both methods of achieving Equalized Odds — training constraints (left plot) and post-processing (middle plot) match the error rates while sacrificing calibration. However, we observe that, assuming h_1 and h_2 cannot be improved, it is infeasible to achieve the calibrated relaxation (Figure 4.3c right). This is an example where matching the F.P. rate of h_1 would require a classifier worse than the trivial classifier h^{μ_2}. This

example therefore represents an instance in which calibration is completely incompatible with any error-rate constraints. If the primary concern of criminal justice practitioners is calibration (Dieterich et al., 2016; Flores et al., 2016), then there will inherently be discrimination in the form of F.P. and F.N. rates. However, if the Equalized Odds framework is adopted, the miscalibrated risk scores inherently cause discrimination to one group, as argued in the introduction. Therefore, the most meaningful change in such a setting would be an improvement to h_2 (the classifier for African Americans) either through the collection of more data or the use of more salient features. A reduction in overall error to the group with higher cost will naturally lead to less error-rate disparity.

4.4 Discussion and Conclusion

We have observed cases in which calibration and relaxed Equalized Odds are compatible and cases where they are not. When it is feasible, the penalty of equalizing cost is amplified if the base rates between groups differ significantly. This is expected, as base rate differences are what give rise to cost-disparity in the calibrated setting. Seeking equality with respect to a single error rate (e.g. false-negatives, as in the income prediction experiment) will necessarily increase disparity with respect to the other error. This may be tolerable (in the income prediction case, some employees will end up over-paid) but could also be highly problematic (e.g. in criminal justice settings). Finally, we have observed that the calibrated relaxation is infeasible when the best (discriminatory) classifiers are not far from the trivial classifiers (leaving little room for interpolation). In such settings, we see that calibration is completely incompatible with an equalized error constraint.

In summary, we conclude that maintaining cost parity *and* calibration is desirable yet often difficult in practice. Although we provide an algorithm to effectively find the unique feasible solution to both constraints, it is inherently based on randomly exchanging the predictions of the better classifier with the trivial base rate. Even if fairness is reached in expectation, for an individual case, it may be hard to accept that occasionally consequential decisions are made by randomly withholding predictive information, irrespective of a particular person's feature representation. In this chapter we argue that, as long as calibration is required, no lower-error solution can be achieved.

CHAPTER 5

THE EXTERNALITIES OF EXPLORATION AND HOW DATA DIVERSITY HELPS EXPLOITATION

Online learning algorithms are a key tool in web search and content optimization, adaptively learning what users want to see. In a typical application, each time a user arrives, the algorithm chooses among various content presentation options (e.g., news articles to display), the chosen content is presented to the user, and an outcome (e.g., a click) is observed. Such algorithms must balance *exploration* (making potentially suboptimal decisions now for the sake of acquiring information that will improve decisions in the future) and *exploitation* (using information collected in the past to make better decisions now). Exploration could degrade the experience of a current user, but improves user experience in the long run. This exploration-exploitation tradeoff is commonly studied in the online learning framework of *multi-armed bandits* (Bubeck and Cesa-Bianchi, 2012).

Concerns have been raised about whether exploration in such scenarios could be unfair, in the sense that some individuals or groups may experience too much of the downside of exploration without sufficient upside (Bird et al., 2016). We formally study these concerns in the *linear contextual bandits* model (Li et al., 2010; Chu et al., 2011), a standard variant of multi-armed bandits appropriate for content personalization scenarios. We focus on *externalities* arising due to exploration, that is, undesirable side effects that the presence of one party may impose on another.

We first examine the effects of exploration at a group level. We introduce the notion of a *group externality* in an online learning system, quantifying how

much the presence of one population (which we dub the majority) impacts the rewards of another (the minority). We show that this impact can be negative, and that, in a particular precise sense, no algorithm can avoid it. This cannot be explained by the absence of suitably good policies since our adoption of the linear contextual bandits framework implies the existence of a feasible policy that is simultaneously optimal for everyone. Instead, the problem is inherent to the process of exploration. We come to a surprising conclusion that more data can sometimes lead to worse outcomes for the users of an explore-exploit-based system.

We next turn to the effect of exploration at an individual level. We interpret exploration as a potential externality imposed on the current user by future users of the system. Indeed, it is only for the sake of the future users that the algorithm would forego the action that currently looks optimal. To avoid this externality, one may use the greedy algorithm that always chooses the action that appears optimal according to current estimates of the problem parameters. While this greedy algorithm performs poorly in the worst case, it tends to work well in many applications and experiments.[1]

In a new line of work, Bastani et al. (2020) and Kannan et al. (2018) analyzed conditions under which inherent diversity in the data makes explicit exploration unnecessary. Kannan et al. (2018) proved that the greedy algorithm achieves a regret rate of $\tilde{O}(\sqrt{T})$ in expectation over small perturbations of the context vectors (which ensure sufficient data diversity). This is the best rate that can be achieved in the worst case (i.e., for all problem instances, without data

[1]Both positive and negative findings are folklore. One way to precisely state the negative result is that the greedy algorithm incurs constant per-round regret with constant probability; while results of this form have likely been known for decades, Mansour et al. (2018, Corollary A.2(b)) proved this for a wide variety of scenarios. Very recently, the good empirical performance has been confirmed by state-of-art experiments in Bietti et al. (2018).

diversity assumptions), but it leaves open the possibilities that (i) another algorithm may perform much better than the greedy algorithm on some problem instances, or (ii) the greedy algorithm may perform much better than worst case under the diversity conditions. We expand on this line of work. We prove that under the same diversity conditions, the greedy algorithm almost matches the best possible Bayesian regret rate of *any* algorithm *on the same problem instance*. This could be as low as $\mathrm{polylog}(T)$ for some instances, and, as we prove, at most $\tilde{O}(T^{1/3})$ whenever the diversity conditions hold.

Returning to group-level effects, we show that under the same diversity conditions, the negative group externalities imposed by the majority essentially vanish if one runs the greedy algorithm. Together, our results illustrate a sharp contrast between the high individual and group externalities that exist in the worst case, and the ability to remove all externalities if the data is sufficiently diverse.

Additional motivation. Whether and when explicit exploration is necessary is an important concern in the study of the exploration-exploitation tradeoff. Fairness considerations aside, explicit exploration is expensive. It is wasteful and risky in the short term, it adds a layer of complexity to algorithm design (Langford and Zhang, 2007; Agarwal et al., 2014), and its adoption at scale tends to require substantial systems support and buy-in from management (Agarwal et al., 2016, 2017). A system based on the greedy algorithm would typically be cheaper to design and deploy.

Further, explicit exploration can run into incentive issues in applications such as recommender systems. Essentially, when it is up to the users which products or experiences to choose and the algorithm can only issue recommen-

dations and ratings, an explore-exploit algorithm needs to provide incentives to explore for the sake of the future users (Kremer et al., 2014; Frazier et al., 2014; Che and Hörner, 2018; Mansour et al., 2015; Papanastasiou et al., 2018). Such incentive guarantees tend to come at the cost of decreased performance, and rely on assumptions about human behavior. The greedy algorithm avoids this problem as it is inherently consistent with the users' incentives.

Additional related work. Our research draws inspiration from the growing body of work on fairness in machine learning (e.g., Dwork et al., 2012; Hardt et al., 2016b; Kleinberg et al., 2017; Chouldechova, 2017). Several other authors have studied fairness in the context of the contextual bandits framework. Our work differs from the line of research on meritocratic fairness in online learning (Kearns et al., 2017; Liu et al., 2017; Joseph et al., 2016), which considers the allocation of limited resources such as bank loans and requires that nobody should be passed over in favor of a less qualified applicant. We study a fundamentally different scenario in which there are no allocation constraints and we would like to serve each user the best content possible. Our work also differs from that of Celis and Vishnoi (2017), who studied an alternative notion of fairness in the context of news recommendations. According to this notion, all users should have approximately the same probability of seeing a particular type of content (e.g., Republican-leaning articles), regardless of their individual preferences, in order to mitigate the possibility of discriminatory personalization.

The data diversity conditions in Kannan et al. (2018) and this chapter are inspired by the smoothed analysis framework of Spielman and Teng (2004), who proved that the expected running time of the simplex algorithm is polynomial for perturbations of any initial problem instance (whereas the worst-case run-

ning time has long been known to be exponential). Such disparity implies that very bad problem instances are brittle. We find a similar disparity for the greedy algorithm in our setting.

Our results on group externalities. A typical goal in online learning is to minimize *regret*, the (expected) difference between the cumulative reward that would have been obtained had the optimal policy been followed at every round and the cumulative reward obtained by the algorithm. We define a corresponding notion of *minority regret*, the portion of the regret experienced by the minority. Since online learning algorithms update their behavior based on the history of their observations, minority regret is influenced by the entire population on which an algorithm is run. If the minority regret is much higher when a particular algorithm is run on the full population than it is when the same algorithm is run on the minority alone, we can view the majority as imposing a negative externality on the minority; the minority population would achieve a higher cumulative reward if the majority were not present. Asking whether this can ever happen amounts to asking whether access to more data points can ever lead an explore-exploit algorithm to make inferior decisions. One might think that more data should always lead to better decisions and therefore better outcomes for the users. Surprisingly, we show that this is not the case, even with a standard algorithm.

Consider LinUCB (Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), a standard algorithm for linear contextual bandits that is based on the principle of "optimism under uncertainty." We provide a specific problem instance on which, after observing T users, LinUCB would have a minority regret of $\Omega(\sqrt{T})$ if run on the full population, but only constant minority regret if run

on the minority alone. While stylized, this example is motivated by the problem of providing driving directions to different populations of users, about which fairness concerns have been raised (Bird et al., 2016). Further, the situation is reversed on a slight variation of this example: LinUCB obtains constant minority regret when run on the full population and $\Omega(\sqrt{T})$ on the minority alone. That is, group externalities can be large and positive in some cases, and large and negative in others.

Although these regret rates are specific to LinUCB, we show that this phenomenon is, in some sense, unavoidable. Consider the minority regret of LinUCB when run on the full population and the minority regret that LinUCB would incur if run on the minority alone. We know that one may be much smaller or larger than the other. We ask whether any algorithm could achieve the minimum of the two on every problem instance. Using a variation of the same problem instance, we prove that this is impossible; in fact, no algorithm could simultaneously approximate both up to any $o(\sqrt{T})$ factor. In other words, an externality-free algorithm would sometimes "leave money on the table."

In terms of techniques, we rely on the special structure of our example, which can be viewed as an instance of the sleeping bandits problem (Kleinberg et al., 2010). This simplifies the behavior and analysis of LinUCB, allowing us to obtain the $O(1)$ upper bounds. The lower bounds are obtained using KL-divergence techniques to show that the two variants of our example are essentially indistinguishable, and an algorithm that performs well on one must obtain $\Omega(\sqrt{T})$ regret on the other.

Our results on the greedy algorithm. We consider a version of linear contextual bandits in which the latent weight vector θ is drawn from a known prior. In each round, an algorithm is presented several actions to choose from, each represented by a *context vector*. The expected reward of an action is a linear product of θ and the corresponding context vector. The tuple of context vectors is drawn independently from a fixed distribution. In the spirit of smoothed analysis, we assume that this distribution has a small amount of jitter. Formally, the tuple of context vectors is drawn from some fixed distribution, and then a small *perturbation*—small-variance Gaussian noise—is added independently to each coordinate of each context vector. This ensures arriving contexts are diverse. We are interested in Bayesian regret, i.e., regret in expectation over the Bayesian prior. Following the literature, we are primarily interested in the dependence on the time horizon T.

We focus on a batched version of the greedy algorithm, in which new data arrives to the algorithm's optimization routine in small batches, rather than every round. This is well-motivated from a practical perspective—in high-volume applications data usually arrives to the "learner" only after a substantial delay (Agarwal et al., 2016, 2017)—and is essential for our analysis.

Our main result is that the greedy algorithm matches the Bayesian regret of any algorithm up to polylogarithmic factors, for each problem instance, fixing the Bayesian prior and the context distribution. We also prove that LinUCB achieves regret $\tilde{O}(T^{1/3})$ for each realization of θ. This implies a worst-case Bayesian regret of $\tilde{O}(T^{1/3})$ for the greedy algorithm under the perturbation assumption.

Our results hold for both natural versions of the batched greedy algorithm,

Bayesian and frequentist, henceforth called BatchBayesGreedy and BatchFreqGreedy. In BatchBayesGreedy, the chosen action maximizes expected reward according to the Bayesian posterior. BatchFreqGreedy estimates θ using ordinary least squares regression and chooses the best action according to this estimate. The results for BatchFreqGreedy come with additive polylogarithmic factors, but are stronger in that the algorithm does not need to know the prior. Further, the $\tilde{O}(T^{1/3})$ regret bound for BatchFreqGreedy is approximately prior-independent, in the sense that it applies even to very concentrated priors such as independent Gaussians with standard deviation on the order of $T^{-2/3}$.

The key insight in our analysis of BatchBayesGreedy is that any (perturbed) data can be used to simulate any other data, with some discount factor. The analysis of BatchFreqGreedy requires an additional layer of complexity. We consider a hypothetical algorithm that receives the same data as BatchFreqGreedy, but chooses actions based on the Bayesian-greedy selection rule. We analyze this hypothetical algorithm using the same technique as BatchBayesGreedy, and then upper bound the difference in Bayesian regret between the hypothetical algorithm and BatchFreqGreedy.

Our analyses extend to group externalities and (Bayesian) minority regret. In particular, we circumvent the impossibility result mentioned above. We prove that both BatchBayesGreedy and BatchFreqGreedy match the Bayesian minority regret of any algorithm run on either the full population or the minority alone, up to polylogarithmic factors

Detailed comparison with prior work. We substantially improve over the $\tilde{O}(\sqrt{T})$ worst-case regret bound from Kannan et al. (2018), at the cost of some

additional assumptions. First, we consider Bayesian regret, whereas their regret bound is for each realization of θ.[2] Second, they allow the context vectors to be chosen by an adversary before the perturbation is applied. Third, they extend their analysis to a somewhat more general model, in which there is a separate latent weight vector for every action (which amounts to a more restrictive model of perturbations). However, this extension relies on the greedy algorithm being initialized with a substantial amount of data. The results of Kannan et al. (2018) do not appear to have implications on group externalities.

Bastani et al. (2020) show that the greedy algorithm achieves logarithmic regret in an alternative linear contextual bandits setting that is incomparable to ours in several important ways. They consider two-action instances where the actions share a common context vector in each round, but are parameterized by different latent vectors. They ensure data diversity via a strong assumption on the context distribution. This assumption does not follow from our perturbation conditions; among other things, it implies that each action is the best action in a constant fraction of rounds. Further, they assume a version of Tsybakov's *margin condition*, which is known to substantially reduce regret rates in bandit problems (e.g., see Rigollet and Zeevi, 2010).

5.1 Preliminaries

We consider the model of *linear contextual bandits* (Li et al., 2010; Chu et al., 2011). Formally, there is a learner who serves a sequence of users over T rounds, where T is the (known) time horizon. For the user who arrives in round t there are (at

[2]Equivalently, they allow point priors, whereas our priors must have variance $T^{-O(1)}$.

most) K actions available, with each action $a \in \{1, \ldots, K\}$ associated with a *context vector* $x_{a,t} \in \mathbb{R}^d$. Each context vector may contain a mix of features of the action, features of the user, and features of both. We assume that the tuple of context vectors for each round t is drawn independently from a fixed distribution. The learner observes the set of contexts and selects an action a_t for the user. The user then experiences a reward r_t which is visible to the learner. We assume that the expected reward is linear in the chosen context vector. More precisely, we let $r_{a,t}$ be the reward of action a if this action is chosen in round t (so that $r_t = r_{a_t,t}$), and assume that there exists an unknown vector $\theta \in \mathbb{R}^d$ such that $\mathbb{E}[r_{a,t} \mid x_{a,t}] = \theta^\top x_{a,t}$ for any round t and action a. Throughout most of the chapter, the realized rewards are either in $\{0, 1\}$ or are the expectation plus independent Gaussian noise of variance at most 1. We sometimes consider a Bayesian version, in which the latent vector θ is initially drawn from some known prior $\mathcal{P}$.

A standard goal for the learner is to maximize the expected total reward over T rounds, or $\sum_{t=1}^{T} \theta^\top x_{a,t}$. This is equivalent to minimizing the learner's *regret*, defined as

$$\text{Regret}(T) = \sum_{t=1}^{T} \theta^\top x_t^* - \theta^\top x_{a_t,t} \tag{5.1}$$

where $x_t^* = \arg\max_{x \in \{x_{1,t}, \cdots, x_{K,t}\}} \theta^\top x$ denotes the context vector associated with the best action at round t. We are mainly interested in *expected regret*, where the expectation is taken over the context vectors, the rewards, and the algorithm's random seed, and *Bayesian regret*, where the expectation is taken over all of the above and the prior over θ.

We introduce some notation in order to describe and analyze algorithms in this model. We write x_t for $x_{a_t,t}$, the context vector chosen at time t. Let $X_t \in \mathbb{R}^{t \times d}$ be the *context matrix* at time t, a matrix whose rows are vectors $x_1, \ldots, x_t \in$

$\mathbb{R}^d$. A $d \times d$ matrix $Z_t := \sum_{\tau=1}^{t} x_\tau x_\tau^\top = X_t^\top X_t$, called the *empirical covariance* matrix, is an important concept in some of the prior work on linear contextual bandits (e.g., Abbasi-Yadkori et al., 2011; Kannan et al., 2018), as well as in this chapter.

Optimism under uncertainty. Optimism under uncertainty is a common paradigm in problems with an explore-exploit tradeoff (Bubeck and Cesa-Bianchi, 2012). The idea is to evaluate each action "optimistically"—assuming the best-case scenario for this action—and then choose an action with the best optimistic evaluation. When applied to the basic multi-armed bandit setting, it leads to a well-known algorithm called UCB1 (Auer et al., 2002), which chooses the action with the highest upper confidence bound (henceforth, UCB) on its mean reward. The UCB is computed as the sample average of the reward for this action plus a term which captures the amount of uncertainty.

Optimism under uncertainty has been extended to linear contextual bandits in the LinUCB algorithm (Chu et al., 2011; Abbasi-Yadkori et al., 2011). The high-level idea is to compute a confidence region $\Theta_t \subset \mathbb{R}^d$ in each round t such that $\theta \in \Theta_t$ with high probability, and choose an action a which maximizes the optimistic reward estimate $\sup_{\theta \in \Theta_t} x_{a,t}^\top \theta$. Concretely, one uses regression to form an empirical estimate $\hat{\theta}_t$ for θ. Concentration techniques lead to high-probability bounds of the form $|x^\top(\theta - \hat{\theta}_t)| \leq f(t)\sqrt{x^\top Z_t^{-1} x}$, where the *interval width function* $f(t)$ may depend on hyperparameters and features of the instance. LinUCB simply chooses an action

$$a_t^{LinUCB} := \arg\max_a x_{a,t}^\top \hat{\theta}_t + f(t)\sqrt{x_{a,t}^\top Z_t^{-1} x_{a,t}}. \tag{5.2}$$

Among other results, Abbasi-Yadkori et al. (2011) use

$$f(t) = S + \sqrt{dc_0 \log(T + tTL^2)}, \tag{5.3}$$

where L and S are known upper bounds on $\|x_{a,t}\|_2$ and $\|\theta\|_2$, respectively, and c_0 is a parameter. For any $c_0 \geq 1$, they obtain regret $\tilde{O}(dS\sqrt{c_0\, K\, T})$, with only a $\mathrm{polylog}$ dependence on TL/d.

5.2 Group Externality of Exploration

In this section, we study the externalities of exploration at a group level, quantifying how much the presence of one population impacts the rewards of another in an online learning system. We consider linear contextual bandits in a setting in which there are two underlying user populations, called the *majority* and the *minority*. The user who arrives at round t is assumed to come from the majority population with some fixed probability and the minority population otherwise, and the population from which the user comes is known to the learner. The tuple of context vectors at time t is then drawn independently from a fixed group-specific distribution.

We assume there is a single hidden vector θ, and that the distribution of rewards conditioned on the chosen context vector is the same for both groups. Only the distribution over tuples of available context vectors differs between groups. This implies that externalities cannot be explained by the absence of a good policy, since there always exists a policy that is simultaneously optimal for everyone. This allows us to focus only on externalities inherent to the process of exploration.

We define the *minority regret* to be the regret experienced by the minority. The *group externality* imposed on the minority by the majority is then the difference between the minority regret of an algorithm run on the minority alone and the minority regret of the same algorithm run on the full population. A negative group externality implies that the minority is worse off due to the presence of the majority. It is generally more meaningful to bound the multiplicative difference between the minority regret obtained with and without the majority present. Several of our results have this form.

We first ask whether large group externalities can exist. We show that on a simple toy example, a large negative group externality arises under LinUCB, while a slight variant of this example leads to a large positive externality. Put another way, more available data can lead to either better or worse outcomes for the users of a system. We show that this general phenomenon is unavoidable. That is, no algorithm can simultaneously approximate the minority regret of LinUCB run on the full population and LinUCB run on the minority alone, up to any $o(\sqrt{T})$ multiplicative factor.

5.2.1 Two-Bridge Instance

We consider a toy example, motivated by a scenario in which a learner is choosing driving routes for two groups of users. Each user starts at point A, B, or C, and wants to get to the same destination, point D, which requires taking one of two bridges, as shown in Figure 5.1. The travel costs for each of the two bridges are unknown. For simplicity, assume all other edges are known to have 0 cost.

Suppose that 95% of users are in the majority group. All of these users start

at point A and have access only to the top bridge. The other 5% are in the minority. Of these users, 95% start at point C, from which they have access only to the bottom bridge. The remaining 5% of the minority users start at point B, and have access to both bridges.

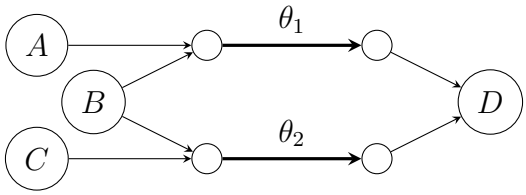


Figure 5.1: Visual illustration of the two-bridge instance.

Consider the behavior of an algorithm that follows the principle of optimism under uncertainty. If run on the full user population, it will quickly collect many observations of the commute time for the top bridge since all users in the majority group must travel over the top bridge. It will collect relatively fewer observations of the commute time over the bottom bridge. Therefore, when the algorithm is faced with a member of the minority population who starts at point B, the algorithm will likely send this user over the bottom bridge in order to collect more data and improve its estimate.

If the same algorithm is instead run on the minority alone, it will quickly collect many more observations of the commute time for the bottom bridge relative to the top. Now when the algorithm is faced with a user who starts at point B, it will likely send her over the top bridge.

Which is better depends on which bridge has the longer commute time. If the top bridge is the better option, then the presence of the majority imposes a negative externality on the minority. If not, then the presence of the majority helps. These two scenarios may be difficult to distinguish.

This toy example can be formalized in the linear contextual bandits framework. There are two underlying actions (the two bridges), but these actions are not always available. To capture this, we define a parameter vector θ in $[0, 1]^2$, with the two coordinates θ_1 and θ_2 representing the expected rewards for taking the top and bottom bridge respectively. (Though we motivated the example in terms of costs, it can be expressed equivalently in terms of rewards.) There are two possible context vectors: $[1\ 0]^\top$ and $[0\ 1]^\top$. A user has available an action with context vector $[1\ 0]^\top$ if and only if she has access to the top bridge. Similarly, she has available an action with context vector $[0\ 1]^\top$ if and only if she has access to the bottom bridge. The instance can then be formalized as follows.

Definition 5.1 (Two-Bridge Instance)**.** *The* two-bridge instance *is an instance of linear contextual bandits. On each round t, the user who arrives is from the majority population with probability 0.95, in which case $x_{1,t} = x_{2,t} = [1\ 0]^\top$. Otherwise, the user is in the minority. In this case, with probability 0.95, $x_{1,t} = x_{2,t} = [0\ 1]^\top$ (based on Figure 5.1, we call these C rounds), while with probability 0.05, $x_{1,t} = [1\ 0]^\top$ and $x_{2,t} = [0\ 1]^\top$ (B rounds). We consider two values for the hidden parameter vector θ, $\theta^{(0)} = [1/2\ \ 1/2 - \varepsilon]^\top$ and $\theta^{(1)} = [1/2 - \varepsilon\ \ 1/2]^\top$ where $\varepsilon = 1/\sqrt{T}$.*

5.2.2 Performance of LinUCB

We start by analyzing the performance of LinUCB on the two-bridge instance. Our main result formalizes the intuition above, showing that when $\theta = \theta^{(0)}$ (that is, the top bridge is better) the majority imposes a large negative group externality on the minority, while the majority imposes a large positive externality

when $\theta = \theta^{(1)}$. We assume rewards are 1-subgaussian.[3]

Theorem 5.2. *Consider LinUCB with any interval width function f satisfying $f(t) \geq 2\sqrt{\log(T)}$.[4] On the two-bridge instance, assuming 1-subgaussian noise on the rewards, when $\theta = \theta^{(0)}$, LinUCB achieves expected minority regret $O(1)$ when run on the minority alone, but $\Omega(\sqrt{T})$ when run on the full population. In contrast, when $\theta = \theta^{(1)}$, LinUCB achieves expected minority regret $O(1)$ when run on the full population, but $\Omega(\sqrt{T})$ when run on the minority alone.*

We omit the proofs of the $\Omega(\sqrt{T})$ lower bounds, which both follow a similar structure to the one used in the proof of the general impossibility result in Section 5.2.3; in fact, both of these lower bounds could be stated as an immediate corollary of Theorem 5.4. Essentially, an argument based on KL-divergence shows that it is difficult to distinguish between the case in which $\theta = \theta^{(0)}$ and the case in which $\theta = \theta^{(1)}$, and therefore LinUCB must choose similar actions in these two cases.

To prove the $O(1)$ upper bounds, we make heavy use of the special structure of the two-bridge instance, which significantly simplifies the analysis of LinUCB. We exploit the fact that the only context vectors available to the learner are the basis vectors $[1\ 0]^\top$ and $[0\ 1]^\top$, which essentially makes this an instance of sleeping bandits (Kleinberg et al., 2010). In this special case, the covariance matrix Z_t is always diagonal, which simplifies Equation (5.2) and leads to LinUCB choosing the ith basis, where i maximizes $(\hat{\theta}_t)_i + f(t)/\sqrt{(Z_t)_{ii}}$ and $(Z_t)_{ii}$ is simply the number of times that this basis vector was already chosen. Additionally,

[3] A random variable X is called σ-subgaussian if $E[e^{\sigma X^2}] < \infty$. A special case is Gaussians with variance σ^2.

[4] For instance, the interval width function in Equation (5.3) satisfies this condition whenever $dc_0 \geq 4$, so one can either set $c_0 \geq 2$, or add two more dimensions to the problem instance (and set $\theta_3 = \theta_4 = 0$).

in this setting $(\hat{\theta}_t)_i$ is just the average reward observed for the ith basis vector, allowing us to bound the difference between each $(\hat{\theta}_t)_i$ and θ_i using standard concentration techniques. Using this, we show that with high probability, after a logarithmic number of rounds—during which the learner can amass at most $O(1)$ regret since the worst-case regret on any round is $\varepsilon = 1/\sqrt{T}$—the probability that LinUCB chooses the wrong action on a B round is small ($O(1/\sqrt{T})$). This leads to constant regret on expectation.

The proof makes use of the following concentration bound:

Lemma 5.3. *Let C_t be the number of C rounds observed in the first t minority rounds in the two-bridge instance. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, $C_t \geq 0.9t$ for all $t \geq 760 \log(T/\delta)$.*

Proof. We apply the following form of the Chernoff bound:

$$\Pr\left[C_t \leq (1-\gamma)\mathbb{E}\left[C_t\right]\right] \leq \exp\left(-\frac{\gamma^2}{2}\mathbb{E}\left[C_t\right]\right).$$

Setting $\gamma = 1/19$, we get

$$\begin{aligned}\Pr\left[C_t \leq \frac{9}{10}t\right] = \Pr\left[C_t \leq \left(1 - \frac{1}{19}\right)\frac{19}{20}t\right] &\leq \exp\left(-\frac{(1/19)^2}{2}\frac{19}{20}t\right) \\ &= \exp\left(-\frac{t}{760}\right) \leq \frac{\delta}{T}\end{aligned}$$

for $t \geq 760 \log(T/\delta)$. Applying a union bound over all T rounds, we have $C_t \geq 0.9t$ for all $t \geq 760 \log(T/\delta)$ with probability at least $1 - \delta$. □

Proof of Theorem 5.2. Now, consider LinUCB run on the minority population alone on the two-bridge instance with $\theta = \theta^{(0)}$. Since we are considering running LinUCB on the minority only, majority rounds are irrelevant, so throughout this proof we abuse notation and use $t \in \{1, \cdots, T_0\}$ for some $T_0 \leq T$ to

index minority rounds. T is still the total number of (minority plus majority) rounds.

This proof heavily exploits the special structure of the two-bridge instance to simplify the analysis of LinUCB. In particular, we exploit the fact that the only contexts ever available are the basis vectors $[1\ 0]^\top$ and $[0\ 1]^\top$. This implies that the covariance matrix Z_t is always diagonal, which greatly simplifies the expression for the chosen action in Equation (5.2). The optimistic estimate of the reward for choosing the ith basis vector is simply

$$UCB_i^t := (\hat{\theta}_t)_i + f(t)/\sqrt{(Z_t)_{ii}}. \tag{5.4}$$

Additionally, in this special case, $(Z_t)_{ii}$ is simply the number of times that the ith basis vector was chosen over the first t minority rounds, and $(\hat{\theta}_t)_i$ is the average reward observed over the $(Z_t)_{ii}$ rounds on which it was chosen.

Using this fact, we can apply concentration bounds to bound the difference between each $(\hat{\theta}_t)_i$ and θ_i. Since rewards were assumed to be 1-subgaussian, Lemma C.8 and a union bound give us that for any δ_1, for any t, with probability at least $1 - 4\delta_1$, for all $i \in \{1, 2\}$,

$$\left|\theta_i - (\hat{\theta}_t)_i\right| \leq \sqrt{2\log(\tfrac{1}{\delta_1})/(Z_t)_{ii}} \tag{5.5}$$

Let B_t and C_t be the number of B and C rounds respectively before round t. By Lemma 5.3, for any δ_2, with probability $1 - \delta_2$, $C_t \geq 9B_t$ when $t \geq 760\log(T/\delta_2)$. Suppose this is the case. Since it is only possible to choose $[1\ 0]$ on B rounds, we have $(Z_t)_{11} \leq B_T$. Similarly, since the algorithm can only choose $[0\ 1]$ on every C round, $(Z_t)_{22} \geq C_T \geq 9B_T$. Fixing $\delta_1 = 1/\sqrt{T}$ and using the assumption that $f(t) \geq 2\sqrt{\log(T)}$, Equations (5.4) and (5.5) then imply that for

any $t \geq 760 \log(T/\delta_2)$, with probability at least $1 - 2\delta_1 = 1 - 2\sqrt{T}$,

$$UCB_1^t \geq \theta_1 - \sqrt{\frac{2\log(\sqrt{T})}{(Z_t)_{11}}} + \frac{f(t)}{\sqrt{(Z_t)_{11}}} \geq \frac{1}{2} + \frac{1}{2}\frac{f(t)}{\sqrt{B_t}},$$

and similarly,

$$UCB_2^t \leq \theta_2 + \sqrt{\frac{2\log(\sqrt{T})}{(Z_t)_{22}}} + \frac{f(t)}{\sqrt{(Z_t)_{22}}} \leq \frac{1}{2} - \varepsilon + \frac{3}{2}\frac{f(t)}{\sqrt{C_T}} \leq \frac{1}{2} + \frac{1}{2}\frac{f(t)}{\sqrt{B_T}} \leq UCB_1^t.$$

This shows that with probability at least $1 - \delta_2$, after the first $760 \log(T/\delta_2)$ rounds, LinUCB picks $[1\ 0]^\top$ on each B round with probability at least $1 - 2\delta_1$, leading to zero regret on that round. To turn this into a bound on expected regret, first note that with at most δ_2 probability, the argument above fails to hold, in which case the minority regret is still bounded by $\varepsilon B_T \leq \varepsilon T$. When the argument above holds, LinUCB may suffer up to ε regret on each of the first $760 \log(T/\delta_2)$ minority rounds. On each additional round, there is a failure probability of $2\delta_1$, and in this case LinUCB again suffers regret of at most ε. Putting this together and setting $\delta_2 = 1/\sqrt{T}$, we get that the expected regret is bounded by $\delta_2 \varepsilon T + 760 \log(T/\delta_2)\varepsilon + 4\delta_1 \varepsilon T = O(1)$. □

5.2.3 An Impossibility Result

It is natural to ask whether it is possible to design an algorithm that can distinguish between the two scenarios analyzed above, obtaining minority regret that is close to the best of LinUCB run on the minority alone and LinUCB run on the full population on any problem instance. In this section, we show that the answer is no. In particular, we prove that on the two-bridge instance, if $\Pr[\theta = \theta^{(0)}] = \Pr[\theta = \theta^{(1)}] = 1/2$, then any algorithm must suffer $\Omega(\sqrt{T})$ regret

on expectation (and therefore $\Omega(\sqrt{T})$ minority regret, since all regret is incurred by minority users).

To prove this result, we begin by formalizing the idea that it is hard to distinguish between the case in which $\theta = \theta^{(0)}$ and the case in which $\theta = \theta^{(1)}$. To do so, we bound the KL-divergence between the joint distributions over the sequences of context vectors, actions taken by the given algorithm, and the given algorithm's rewards that are induced by the two choices of θ. By applying the high-probability Pinsker lemma (Tsybakov, 2009), we show that a low KL-divergence between these distributions implies that the algorithm must be likely either to choose the top bridge on B rounds more than half the time when the bottom bridge is better or to choose the bottom bridge on B rounds more than half the time when the top bridge is better, either of which would lead to high ($\Omega(\sqrt{T})$) regret as long as the number of B rounds is sufficiently large. To finish the proof, we use a simple Chernoff bound to show that the number of B rounds is large with high probability.

To derive the KL-divergence bound, we make use of the assumption that the realized rewards r_t at each round are either 0 or 1. This assumption is not strictly necessary. An analogous argument could be made, for instance, for real-valued rewards with Gaussian noise.

Theorem 5.4. *On the two-bridge instance with realized rewards $r_t \in \{0, 1\}$, any algorithm must incur $\Omega(\sqrt{T})$ minority regret in expectation when $\Pr[\theta = \theta^{(0)}] = \Pr[\theta = \theta^{(1)}] = \frac{1}{2}$.*

Note that "any algorithm" here includes algorithms run on the minority alone, essentially ignoring data from the majority. Theorems 5.2 and 5.4 immediately imply the following corollary.

Corollary 5.5. *No algorithm can simultaneously approximate the minority regret of both LinUCB run on the minority and LinUCB run on the full population up to any* $o(\sqrt{T})$ *multiplicative factor.*

Proof of Theorem 5.4. Fix any algorithm $\mathcal{A}$. We will first derive an $\Omega(\sqrt{T})$ lower bound on the expected regret of $\mathcal{A}$ conditioned on the number of B rounds, B_T, being large. To complete the proof, we then show that B_T is large with high probability.

Let $h_t = \{(x_{1,\tau}, x_{2,\tau}, a_\tau, r_\tau)\}_{\tau=1}^{t}$ be a history of all context vectors, chosen actions, and rewards up to round t, with $h_0 = \emptyset$. Running $\mathcal{A}$ on the two-bridge instance with $\theta = \theta^{(0)}$ induces a distribution over histories h_T. Let P denote the conditional distribution of these histories, conditioned on the event that $B_T \geq T/800$. That is, we define

$$P(h_T) := \Pr\left[h_T \,\middle|\, \theta = \theta^{(0)}, B_T \geq T/800\right].$$

Similarly, we define

$$Q(h_T) := \Pr\left[h_T \,\middle|\, \theta = \theta^{(1)}, B_T \geq T/800\right].$$

We first show that $\mathrm{KL}(P(h_T) \,||\, Q(h_T))$ is upper bounded a constant that does not depend on T. By the chain rule for KL divergences, since r_t is independent of any previous contexts, actions, or rewards conditioned on x_t,

$$\begin{aligned}
&\mathrm{KL}(P(h_T) \,||\, Q(h_T)) \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{h_{t-1}\sim P}[\mathrm{KL}(P((x_{1,t}, x_{2,t}, a_t) \mid h_{t-1}) \,||\, Q((x_{1,t}, x_{2,t}, a_t) \mid h_{t-1}))] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}_{(x_{1,t},x_{2,t},a_t)\sim P}[\mathrm{KL}(P(r_t \mid x_{1,t}, x_{2,t}, a_t) \,||\, Q(r_t \mid (x_{1,t}, x_{2,t}, a_t)))].
\end{aligned}$$

Since the choice of context vectors available at time t is independent of the value of the parameter θ and $\mathcal{A}$ may only base its choices on the observed history and current choice of contexts, it is always the case that $P((x_{1,t}, x_{2,t}, a_t) \mid h_{t-1}) = Q((x_{1,t}, x_{2,t}, a_t) \mid h_{t-1})$, so the first sum in this expression is equal to 0.

To bound the second sum, we make use of the assumption that $r_t \in \{0, 1\}$ for all t.[5] Lemma C.10 then tells us that for any round t, $\mathrm{KL}(P(r_t \mid x_{1,t}, x_{2,t}, a_t) \,||\, Q(r_t \mid x_{1,t}, x_{2,t}, a_t)) \leq 7\varepsilon^2/2$ since the probability of getting reward 1 conditioned on a chosen context is always either $1/2$ or $1/2 - \varepsilon$. Putting this together, we get that

$$\mathrm{KL}(P(h_T) \,||\, Q(h_T)) \leq \frac{7\varepsilon^2 T}{2} = \frac{7}{2}.$$

Now, let E be the event that the algorithm $\mathcal{A}$ chooses arm 2 on at least half of the B rounds, conditioned on $B_T \geq T/800$. If E occurs when $\theta = \theta^{(0)}$, the regret of $\mathcal{A}$ is at least $B_T\varepsilon/2$, which is on the order of $\sqrt{T}$ when $B_T \geq T/800$. If E does not occur (i.e., $\overline{E}$ occurs) when $\theta = \theta^{(1)}$, $\mathcal{A}$ again has regret at least $B_T\varepsilon/2$. We will use the bound on KL divergence to show that one of these bad cases happens with high probability.

By Lemma C.9,

$$P(E) + Q(\overline{E}) \geq \frac{1}{2} e^{-\mathrm{KL}(P(h_T) \,||\, Q(h_T))} \geq \frac{1}{2} e^{-7/2}.$$

[5] If we instead assumed rewards had Gaussian noise with variance σ^2, we would have $\mathrm{KL}(P_t(r_t \mid x_{1,t}, x_{2,t}, a_t) \,||\, Q_t(r_t \mid x_{1,t}, x_{2,t}, a_t)) = \varepsilon^2/(2\sigma^2)$, and the proof would still go through.

Let R be the regret of $\mathcal{A}$. We then have that

$$\begin{aligned}
\mathbb{E}\left[R \,\middle|\, B_T \geq \frac{T}{800}\right] &= \frac{1}{2}\,\mathbb{E}\left[R \,\middle|\, \theta = \theta^{(0)}, B_T \geq \frac{T}{800}\right] + \frac{1}{2}\,\mathbb{E}\left[R \,\middle|\, \theta = \theta^{(1)}, B_T \geq \frac{T}{800}\right] \\
&\geq \frac{1}{2}\Pr\left[E \,\middle|\, \theta = \theta^{(0)}, B_T \geq \frac{T}{800}\right]\mathbb{E}\left[R \,\middle|\, E, \theta = \theta^{(0)}, B_T \geq \frac{T}{800}\right] \\
&\quad + \frac{1}{2}\Pr\left[\overline{E} \,\middle|\, \theta = \theta^{(1)}, B_T \geq \frac{T}{800}\right]\mathbb{E}\left[R \,\middle|\, \overline{E}, \theta = \theta^{(0)}, B_T \geq \frac{T}{800}\right] \\
&\geq \frac{1}{2}\left(P(E) + Q(\overline{E})\right)\frac{\sqrt{T}}{1600} \\
&\geq \frac{\sqrt{T}e^{-\frac{7}{2}}}{6400}.
\end{aligned}$$

It remains to bound the probability that $B_T \geq T/800$. By a Chernoff bound,

$$\Pr\left[B_T \leq \frac{T}{800}\right] = \Pr\left[B_T \leq \frac{\mathbb{E}\left[B_T\right]}{2}\right] \leq \exp\left(-\frac{\mathbb{E}\left[B_T\right]}{8}\right) = \exp\left(-\frac{T}{3200}\right).$$

Thus, for any $\delta \in (0,1)$, if $T \geq 3200\log(1/\delta)$, then with probability at least $1-\delta$, $B_T \geq T/800$. In particular, let $\delta = 1/2$. Then if $T \geq 3200\log 2$, we have

$$\mathbb{E}[R] \geq \Pr\left[B_T \geq \frac{T}{800}\right]\mathbb{E}\left[R \,\middle|\, B_T \geq \frac{T}{800}\right] \geq \left(\frac{1}{2}\right)\left(\frac{\sqrt{T}e^{-\frac{7}{2}}}{6400}\right).$$

This completes the proof that the regret of $\mathcal{A}$ is $\Omega(\sqrt{T})$ on this problem instance.

□

5.3 Greedy Algorithms and LinUCB with Perturbed Contexts

We now turn our attention to externalities at an individual level. We interpret exploration as a potential externality imposed on the current user of a system by future users, since the current user would prefer the learner to take the action that appears optimal. One could avoid such externalities by running the greedy algorithm, which does just that, but it is well known that the greedy algorithm

performs poorly in the worst case. In this section, we build on a recent line of work analyzing the conditions under which inherent data diversity leads the greedy algorithm to perform well.

We analyze the expected performance of the greedy algorithm under small random perturbations of the context vectors. We focus on greedy algorithms that consume new data in batches, rather than every round. We consider both Bayesian and frequentist versions, BatchBayesGreedy and BatchFreqGreedy. Our main result is that for any specific problem instance, both algorithms match the Bayesian regret of any algorithm on that particular instance up to polylogarithmic factors. We also prove that under the same perturbation assumptions, LinUCB achieves regret $\tilde{O}(T^{1/3})$ for each realization of θ, which implies a worst-case Bayesian regret of $\tilde{O}(T^{1/3})$ for the greedy algorithms. Finally, we repurpose our analysis to derive a positive result in the group setting, implying that the impossibility result of Section 5.2.3 breaks down when the data is sufficiently diverse.

Setting and notation. We consider a Bayesian version of linear contextual bandits, with θ drawn from a known multivariate Gaussian prior $\mathcal{P} = \mathcal{N}(\overline{\theta}, \Sigma)$, with $\overline{\theta} \in \mathbb{R}^d$ and invertible $\Sigma \in \mathbb{R}^{d \times d}$.

To capture the idea of data diversity, we assume the context vectors on each round t are generated using the following *perturbed context generation* process: First, a tuple $(\mu_{1,t}, \ \ldots \ , \mu_{K,t})$ of *mean context vectors* is drawn independently from some fixed distribution D_μ over $(\mathbb{R} \cup \{\perp\})^K$, where $\mu_{a,t} = \perp$ means action a is not available. For each available action a, the context vector is then $x_{a,t} = \mu_{a,t} + \varepsilon_{a,t}$, where $\varepsilon_{a,t}$ is a vector of random noise. Each component of $\varepsilon_{a,t}$ is drawn

independently from a zero-mean Gaussian with standard deviation ρ. We refer to ρ as the *perturbation size*. In general, our regret bounds deteriorate if ρ is very small. Together we refer to a distribution D_μ, prior $\mathcal{P}$, and perturbation size ρ as a *problem instance*.

We make several technical assumptions. First, the distribution D_μ is such that each context vector has bounded 2-norm, i.e., $\|\mu_{a,t}\|_2 \leq 1$. It can be arbitrary otherwise. Second, the perturbation size needs to be sufficiently small, $\rho \leq 1/\sqrt{d}$. Third, the realized reward $r_{a,t}$ for each action a and round t is $r_{a,t} = x_{a,t}^\top \theta + \eta_{a,t}$, the mean reward $x_{a,t}^\top \theta$ plus standard Gaussian noise $\eta_{a,t} \sim \mathcal{N}(0,1)$.[6] The history up to round t is denoted by the tuple $h_t = ((x_1, r_1), \ldots, (x_t, r_t))$.

The greedy algorithms. For the batch version of the greedy algorithm, time is divided in batches of Y consecutive rounds each. When forming its estimate of the optimal action at round t, the algorithm may only use the history up to the last round of the previous batch, denoted t_0.

BatchBayesGreedy forms a posterior over θ using prior $\mathcal{P}$ and history h_{t_0}. In round t it chooses the action that maximizes reward in expectation over this posterior. This is equivalent to choosing

$$a_t = \arg\max_a x_{a,t}^\top \theta_t^{\text{bay}}, \quad \text{where } \theta_t^{\text{bay}} := \mathbb{E}[\theta \mid h_{t_0}]\,. \tag{5.6}$$

BatchFreqGreedy does not rely on any knowledge of the prior. It chooses the best action according to the least squares estimate of θ, denoted θ_t^{fre}, computed

[6] Our analysis can be easily extended to handle reward noise of fixed variance, i.e., $\eta_{a,t} \sim \mathcal{N}(0, \sigma^2)$. BatchFreqGreedy would not need to know σ. BatchBayesGreedy would need to know either Σ and σ or just $\Sigma/\,\sigma^2$.

with respect to history h_{t_0}:

$$a_t = \arg\max_a x_{a,t}^\top \theta_t^{\text{fre}}, \quad \text{where } \theta_t^{\text{fre}} = \arg\min_{\theta'} \sum_{\tau=1}^{t_0} ((\theta')^\top x_\tau - r_\tau)^2. \tag{5.7}$$

5.3.1 Main Results

We first state our main results before describing the intuition behind them. We state each theorem in terms of the main relevant parameters T, K, d, Y, and ρ. First, we prove that in expectation over the random perturbations, both greedy algorithms favorably compare to any other algorithm.

Theorem 5.6. *With perturbed context generation, there is some $Y_0 = \text{polylog}(d, T)/\rho^2$ such that with batch duration $Y \geq Y_0$, the following holds. Fix any bandit algorithm, and let $R_0(T)$ be its Bayesian regret on a particular problem instance. Then on that same instance,*

(a) BatchBayesGreedy has Bayesian regret at most $Y \cdot R_0(T/Y) + \tilde{O}(1/T)$,

(b) BatchFreqGreedy has Bayesian regret at most $Y \cdot R_0(T/Y) + \tilde{O}(\sqrt{d}/\rho^2)$.

Our next result asserts that the Bayesian regret for LinUCB and both greedy algorithms is on the order of (at most) $T^{1/3}$. This result requires additional technical assumptions.

Theorem 5.7. *Assume that the maximal eigenvalue of the covariance matrix Σ of the prior $\mathcal{P}$ is at most 1,[7] and the mean vector satisfies $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3 \log T}$. With perturbed context generation,*

(a) With appropriate parameter settings, LinUCB has Bayesian regret $\tilde{O}(d^2\, K^{2/3}\, T^{1/3}/\rho^2)$.

[7] In particular, if $\mathcal{P}$ is independent across the coordinates of θ, then the variance in each coordinate is at most 1.

(b) If $Y \geq Y_0$ as in Theorem 5.6, then both BatchBayesGreedy and BatchFreqGreedy have Bayesian regret at most $\tilde{O}(d^2\, K^{2/3}\, T^{1/3}/\rho^2)$.

The assumption $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3 \log T}$ in Theorem 5.7 can be replaced with $d \geq \log T / \log \log T$. We use Theorem 5.7(b) to derive an "approximately prior-independent" result for BatchFreqGreedy. (For clarity, we state it for independent priors.) The bound in Theorem 5.7(b) deteriorates if $\mathcal{P}$ gets very sharp, but it suffices if $\mathcal{P}$ has standard deviation on the order of (at least) $T^{-2/3}$.

Corollary 5.8. *Assume that the prior $\mathcal{P}$ is independent over the components of θ, with variance $\kappa^2 \leq 1$ in each component. Suppose the mean vector satisfies $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3 \log T}$. With perturbed context generation, if $Y \geq Y_0$ as in Theorem 5.6, then BatchFreqGreedy has Bayesian regret at most $\tilde{O}(d^2\, K^{2/3}\, T^{1/3}/\rho^2)$ as long as $\kappa \geq T^{-2/3}$.*

Finally, we derive a positive result on group externalities. We find that with perturbed context generation, the minority Bayesian regret of the greedy algorithms (i.e., the Bayesian regret incurred on minority rounds) is small compared to the minority Bayesian regret of any algorithm, whether run on the full population *or* on the minority alone. This sidesteps the impossibility result of Section 5.2.3.

Theorem 5.9. *Assume $Y \geq Y_0$ as in Theorem 5.6 and perturbed context generation. Fix any bandit algorithm and instance, and let $R_{\min}(T)$ be the minimum of its minority Bayesian regrets when it is only run over minority rounds or when it is run over the full population. Both greedy algorithms run on the full population achieve minority Bayesian regret at most $Y \cdot R_{\min}(T) + \tilde{O}(\sqrt{d}/\rho^2)$.*

5.3.2 Key Techniques

The key idea behind our approach is to show that, with perturbed context generation, BatchBayesGreedy collects data that is informative enough to "simulate" the history of contexts and rewards from the run of any other algorithm ALG over fewer rounds. This implies that it remains competitive with ALG since it has at least as much information and makes myopically optimal decisions.

We use the same technique to prove a similar simulation result for BatchFreqGreedy. To treat both algorithms at once, we define a template that unifies them. A bandit algorithm is called *batch-greedy-style* if it divides the timeline in batches of Y consecutive rounds each, in each round t chooses some estimate θ_t of θ, based only on the data from the previous batches, and then chooses the best action according to this estimate, so that $a_t = \arg\max_a \theta_t^\top x_{a,t}$. For a batch that starts at round $t_0 + 1$, the *batch history* is the tuple $((x_{t_0+\tau},\, r_{t_0+\tau}) :\ \tau \in [Y])$, and the *batch context matrix* is the matrix X whose rows are vectors $(x_{t_0+\tau} :\ \tau \in [Y])$; here $[Y] = \{1, \cdots, Y\}$. Similarly to the "empirical covariance matrix", we define the *batch covariance matrix* as $X^\top X$.

Let us formulate what we mean by "simulation". We want to use the data collected from a single batch in order to simulate the reward for any one context x. More formally, we are interested in the randomized function that takes a context x and outputs an independent random sample from $\mathcal{N}(\theta^\top x, 1)$. We denote it $\texttt{Rew}_\theta(\cdot)$; this is the realized reward for an action with context vector x.

Definition 5.10. *Consider batch B in the execution of a batch-greedy-style algorithm. Batch history h_B can simulate $\texttt{Rew}_\theta()$ up to radius $R > 0$ if there exists a function $g : \{\text{context vectors}\} \times \{\text{batch histories } h_B\} \to \mathbb{R}$ such that $g(x, h_B)$ is identically*

distributed to $\texttt{Rew}_\theta(x)$ *conditional on the batch context matrix, for all* θ *and all context vectors* $x \in \mathbb{R}^d$ *with* $\|x\|_2 \leq R$.

Let us comment on how it may be possible to simulate $\texttt{Rew}_\theta(x)$. For intuition, suppose that $x = \frac{1}{2}\, x_1 + \frac{1}{2}\, x_2$. Then $(\frac{1}{2}\, r_1 + \frac{1}{2}\, r_2 + \xi)$ is distributed as $\mathcal{N}(\theta^\top x, 1)$ if ξ is drawn independently from $\mathcal{N}(0, \frac{1}{2})$. Thus, we can define $g(x, h) = \frac{1}{2}\, r_1 + \frac{1}{2}\, r_2 + \xi$ in Definition 5.10. We generalize this idea and show that a batch history can simulate $\texttt{Rew}_\theta$ as long as the batch covariance matrix has a sufficiently large minimum eigenvalue, which holds with high probability when the batch size is large.

Lemma 5.11. *With perturbed context generation, there is some* $Y_0 = \text{polylog}(d, T)/\rho^2$ *and* $R = O(\rho\sqrt{d \log(TKd)})$ *such that with probability at least* $1 - T^{-2}$ *any batch history from a batch-greedy-style algorithm can simulate* $\texttt{Rew}_\theta()$ *up to radius* R*, as long as* $Y \geq Y_0$.

If the batch history of an algorithm can simulate $\texttt{Rew}_\theta$, the algorithm has enough information to simulate the outcome of a fresh round of any other algorithm ALG. Eventually, this allows us to use a coupling argument in which we couple a run of BatchBayesGreedy with a slowed-down run of ALG, and prove that the former accumulates at least as much information as the latter, and therefore the Bayesian-greedy action choice is, in expectation, at least as good as that of ALG. This leads to Theorem 5.6(a). We extend this argument to a scenario in which both the greedy algorithm and ALG measure regret over a randomly chosen subset of the rounds, which leads to Theorem 5.9.

To extend these results to BatchFreqGreedy, we consider a hypothetical algorithm that receives the same data as BatchFreqGreedy, but chooses actions based

on the (batched) Bayesian-greedy selection rule. We analyze this hypothetical algorithm using the same technique as above, and then argue that its Bayesian regret cannot be much smaller than that of BatchFreqGreedy. Intuitively, this is because the two algorithms form almost identical estimates of θ, differing only in the fact that the hypothetical algorithm uses the $\mathcal{P}$ as well as the data. We show that this difference amounts to effects on the order of $1/t$, which add up to a maximal difference of $O(\log T)$ in Bayesian regret.

5.4 Analysis: LinUCB with Perturbed Contexts

In this section, we prove Theorem 5.7(a), a Bayesian regret bound for the LinUCB algorithm under perturbed context generation. We focus on a version of LinUCB from Abbasi-Yadkori et al. (2011), as defined in (5.3) on page 80.

Recall that the interval width function in (5.3) is parameterized by numbers L, S, c_0. We use

$$\begin{aligned} L &\geq 1 + \rho\sqrt{2d\log(2T^3Kd)}, \\ S &\geq \|\overline{\theta}\|_2 + \sqrt{3d\log T} \quad \text{(and } S < T\text{)} \\ c_0 &= 1. \end{aligned} \tag{5.8}$$

Recall that ρ denotes perturbation size, and $\overline{\theta} = \mathbb{E}[\theta]$, the prior means of the latent vector θ.

Remark 5.12. *Ideally we would like to set L, S according to* (5.8) *with equalities. We consider a more permissive version with inequalities so as to not require the exact knowledge of ρ and $\|\overline{\theta}\|_2$.*

While the original result in Abbasi-Yadkori et al. (2011) requires $\|x_{a,t}\|_2 \leq L$ and $\|\theta\|_2 \leq S$, in our setting this only happens with high probability.

We prove the following theorem (which implies Theorem 5.7(a)):

Theorem 5.13. *Assume perturbed context generation. Further, suppose that the maximal eigenvalue of the covariance matrix Σ of the prior $\mathcal{P}$ is at most 1, and the mean vector satisfies $\|\bar{\theta}\|_2 \geq 1 + \sqrt{3 \log T}$. The version of LinUCB with interval width function* (5.3) *and parameters given by* (5.8) *has Bayesian regret at most*

$$T^{1/3} \left(d^2 \, S \, (K^2/\rho)^{1/3} \right) \cdot \text{polylog}(TKLd). \tag{5.9}$$

Remark 5.14. *The theorem also holds if the assumption on $\|\bar{\theta}\|_2$ is replaced with $d \geq \frac{\log T}{\log \log T}$. The only change in the analysis is that in the concluding steps (Section 5.4.2), we use Lemma 5.17(b) instead of Lemma 5.17(a).*

On a high level, our analysis proceeds as follows. We massage algorithm's regret so as to elucidate the dependence on the number of rounds with small "gap" between the best and second-best action, call it N. This step does not rely on perturbed context generation, and makes use of the analysis from Abbasi-Yadkori et al. (2011). The crux is that we derive a much stronger upper-bound on $\mathbb{E}[N]$ under perturbed context generation. The analysis relies on some non-trivial technicalities on bounding the deviations from the "high-probability" behavior, which are gathered in Section 5.4.1.

We reuse the analysis in Abbasi-Yadkori et al. (2011) via the following lemma.[8] To state this lemma, define the instantaneous regret at time t as

[8]Lemma 5.15(a) is implicit in the proof of Theorem 3 from Abbasi-Yadkori et al. (2011), and Lemma 5.15(b) is asserted by Abbasi-Yadkori et al. (2011, Lemma 10).

$R_t = \theta^\top x_t^* - \theta^\top x_{a_t,t}$, and let

$$\beta_T = \left(\sqrt{d \log\left(T(1+TL^2)\right)} + S\right)^2.$$

Lemma 5.15 (Abbasi-Yadkori et al. (2011))**.** *Consider a problem instance with reward noise $\mathcal{N}(0,1)$ and a specific realization of latent vector θ and contexts $x_{a,t}$. Consider LinUCB with parameters L, S, c_0 that satisfy $\|x_{a,t}\|_2 \leq L$, $\|\theta\|_2 \leq S$, and $c_0 = 1$. Then*

(a) with probability at least $1 - \frac{1}{T}$ (over the randomness in the rewards) it holds that

$$\sum_{t=1}^{T} R_t^2 \leq 16\beta_T \log(\det(Z_t + I)),$$

where $Z_t = \sum_{\tau=1}^{t} x_\tau x_\tau^\top \in \mathbb{R}^{d\times d}$ is the "empirical covariance matrix" at time t.

(b) $\det(Z_t + I) \leq (1 + tL^2/d)^d$.

The following lemma captures the essence of the proof of Theorem 5.13. From here on, we assume perturbed context generation. In particular, reward noise is $\mathcal{N}(0,1)$.

Lemma 5.16. *Suppose parameter L is set as in* (5.8)*. Consider a problem instance with a specific realization of θ such that $\|\theta\|_2 \leq S$. Then for any $\gamma > 0$,*

$$\mathbb{E}\left[\textit{Regret}(T)\right] \leq \|\theta\|_2^{-1/3} \left(\frac{1}{2\sqrt{\pi}} + 16\beta_T\, d \log(1 + TL^2/d)\right) \left(\frac{TK^2}{\rho}\right)^{1/3} + \tilde{O}\left(1\right).$$

Proof. We will prove that for any $\gamma > 0$,

$$\mathbb{E}\left[\text{Regret}(T)\right] \leq T \cdot \frac{\gamma^2 K^2}{2\rho\|\theta\|_2\sqrt{\pi}} + \frac{1}{\gamma} 16\beta_T\, d \log(1 + TL^2/d) + \tilde{O}(1). \tag{5.10}$$

The Lemma easily follows by setting $\gamma = (TK^2/(\rho\|\theta\|_2))^{-1/3}$.

Fix some $\gamma > 0$. We distinguish between rounds t with $R_t < \gamma$ and those with $R_t \geq \gamma$:

$$\text{Regret}(T) = \sum_{t=1}^{T} R_t \leq \sum_{t \in \mathcal{T}_\gamma} R_t + \sum_{t=1}^{T} \frac{R_t^2}{\gamma} \leq \gamma |\mathcal{T}_\gamma| + \frac{1}{\gamma} \sum_{t=1}^{T} R_t^2, \tag{5.11}$$

where $\mathcal{T}_\gamma = \{t : R_t \in (0, \gamma)\}$.

We use Lemma 5.15 to upper-bound the second summand in (5.11). To this end, we condition on the event that every component of every perturbation $\varepsilon_{a,t}$ has absolute value at most $\sqrt{2 \log 2T^3 Kd}$; denote this event by U. This implies $\|x_{a,t}\|_2 \leq L$ for all actions a and all rounds t. By Lemma C.4, U is a high-probability event: $\Pr[U] \geq 1 - \frac{1}{T^2}$. Now we are ready to apply Lemma 5.15:

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t^2 \mid U\right] \leq 16\, d\, \beta_T\, \log(1 + tL^2/d). \tag{5.12}$$

To plug this into (5.11), we need to account for the low-probability event $\bar{U}$. We need to be careful because R_t could, with low probability, be arbitrarily large. By Lemma 5.18 with $\ell = 0$,

$$\mathbb{E}\left[R_t \mid \bar{U}\right] \leq 2\left[\|\theta\|_2 \left(1 + \rho(1 + \sqrt{2 \log K}) + \sqrt{2 \log(2T^3 Kd)}\right)\right]$$

$$\mathbb{E}\left[\text{Regret}(T) \mid \bar{U}\right] \Pr[\bar{U}] = \sum_{t=1}^{T} \mathbb{E}\left[R_t \mid \bar{U}\right] / T^2 < \tilde{O}(1).$$

$$\mathbb{E}\left[\text{Regret}(T) \mid U\right] \Pr[U] \leq \gamma\, \mathbb{E}\left[\, |\mathcal{T}_\gamma| \,\right] + \tfrac{1}{\gamma} \mathbb{E}\left[\sum_{t=1}^{T} R_t^2 \mid U\right] \qquad \text{(by (5.11))}$$

Putting this together and using (5.12), we obtain:

$$\mathbb{E}\left[\text{Regret}(T)\right] \leq \gamma\, \mathbb{E}\left[\, |\mathcal{T}_\gamma| \,\right] + \frac{16}{\gamma}\, d\, \beta_T\, \log(1 + tL^2/d) + \tilde{O}(1). \tag{5.13}$$

To obtain (5.10), we analyze the first summand in (5.13). Let Δ_t be the "gap" at time t: the difference in expected rewards between the best and second-best actions at time t (where "best" and "second-best" is according to expected rewards). Here, we're taking expectations *after* the perturbations are applied, so

the only randomness comes from the noisy rewards. Consider the set of rounds with small gap, $\mathcal{G}_\gamma := \{t : \Delta_t < \gamma\}$. Notice that $r_t \in (0, \gamma)$ implies $\Delta_t < \gamma$, so $|\mathcal{T}_\gamma| \leq |\mathcal{G}_\gamma|$.

In what follows we prove an upper bound on $\mathbb{E}[|\mathcal{G}_\gamma|]$. This is the step where perturbed context generation is truly used. For any two arms a_1 and a_2, the gap between their expected rewards is

$$\theta^\top(x_{a_1,t} - x_{a_2,t}) = \theta^\top(\mu_{a_1,t} - \mu_{a_2,t}) + \theta^\top(\varepsilon_{a_1,t} - \varepsilon_{a_2,t}).$$

Therefore, the probability that the gap between those arms is smaller than γ is

$$\begin{aligned}
&\Pr\left[|\theta^\top(\mu_{a_1,t} - \mu_{a_2,t}) + \theta^\top(\varepsilon_{a_1,t} - \varepsilon_{a_2,t})| \leq \gamma\right] \\
&\quad = \Pr\left[-\gamma - \theta^\top(\mu_{a_1,t} - \mu_{a_2,t}) \leq \theta^\top(\varepsilon_{a_1,t} - \varepsilon_{a_2,t}) \leq \gamma - \theta^\top(\mu_{a_1,t} - \mu_{a_2,t})\right]
\end{aligned}$$

Since $\theta^\top\varepsilon_{a_1,t}$ and $\theta^\top\varepsilon_{a_2,t}$ are both distributed as $\mathcal{N}(0, \rho^2\|\theta\|_2^2)$, their difference is $\mathcal{N}(0, 2\rho^2\|\theta\|_2^2)$. The maximum value that the Gaussian measure takes is $\frac{1}{2\rho\|\theta\|_2\sqrt{\pi}}$, and the measure in any interval of width 2γ is therefore at most $\frac{\gamma}{\rho\|\theta\|_2\sqrt{\pi}}$. This gives us the bound

$$\Pr\left[|\theta^\top(\mu_{a_1,t} - \mu_{a_2,t}) + \theta^\top(\varepsilon_{a_1,t} - \varepsilon_{a_2,t})| \leq \gamma\right] \leq \frac{\gamma}{\rho\|\theta\|_2\sqrt{\pi}}.$$

Union-bounding over all $\binom{K}{2}$ pairs of actions, we have

$$\begin{aligned}
\Pr[\Delta_t \leq \gamma] &\leq \Pr\left[\bigcup_{a_1,a_2\in[K]} |\theta^\top(x_{a_1,t} - x_{a_2,t})| \leq \gamma\right] \leq \frac{K^2}{2}\frac{\gamma}{\rho\|\theta\|_2\sqrt{\pi}}. \\
\mathbb{E}[\,|\mathcal{G}_\gamma|\,] &= \sum_{t=1}^{T} \Pr[\Delta_t \leq \gamma] \leq T \cdot \frac{K^2}{2}\frac{\gamma}{\rho\|\theta\|_2\sqrt{\pi}}.
\end{aligned}$$

Plugging this into (5.13) (recalling that $|\mathcal{T}_\gamma| \leq |\mathcal{G}_\gamma|$) completes the proof. □

5.4.1 Bounding the Deviations

We make use of two results that bound deviations from the "high-probability" behavior, one on $\|\theta\|_2$ and another on instantaneous regret. First, we prove high-probability upper and lower bounds on $\|\theta\|_2$ under the conditions in Theorem 5.13. Essentially, these bounds allow us to use Lemma 5.16.

Lemma 5.17. *Assume the latent vector θ comes from a multivariate Gaussian, $\theta \sim \mathcal{N}(\overline{\theta}, \Sigma)$, here the covariate matrix Σ satisfies $\lambda_{\max}(\Sigma) \leq 1$.*

(a) If $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3 \log T}$, then for sufficiently large T, with probability at least $1 - \frac{2}{T}$,

$$\frac{1}{2 \log T} \leq \|\theta\|_2 \leq \|\overline{\theta}\|_2 + \sqrt{3d \log T}. \tag{5.14}$$

(b) Same conclusion if $d \geq \frac{\log T}{\log \log T}$.

Proof. We consider two cases, based on whether $d \geq \log T / \log \log T$. We need both cases to prove part (a), and we obtain part (b) as an interesting by-product. We repeatedly use Lemma C.7, a concentration inequality for χ^2 random variables, to show concentration on the Gaussian norm.

Case 1: $d \geq \log T / \log \log T$.

Since the Gaussian measure is decreasing in distance from 0, the $\Pr\left[\|\theta\|_2 \leq c\right] \leq \Pr\left[\|\overline{\theta} - \theta\|_2 \leq c\right]$ for any c. In other words, the norm of a Gaussian is most likely to be small when its mean is 0. Let $X = \Sigma^{-1/2}(\overline{\theta} - \theta)$. Note that X has distribution $\mathcal{N}(0, I)$, and therefore $\|X\|_2^2$ has χ^2 distribution with d degrees of freedom. We

can bound this as

$$\begin{aligned}
\Pr\left[\|\overline{\theta}-\theta\|_2 \le \frac{1}{2\log T}\right] &= \Pr\left[\|\Sigma^{-1/2}X\|_2 \le \frac{1}{2\log T}\right] \\
&\le \Pr\left[\sqrt{\lambda_{\max}(\Sigma)}\|X\|_2 \le \frac{1}{2\log T}\right] \\
&\le \Pr\left[\|X\|_2 \le \frac{1}{2\log T}\right] \\
&= \Pr\left[\|X\|_2^2 \le \frac{1}{4(\log T)^2}\right] \\
&\le \left(\frac{1}{4d(\log T)^2}e^{1-1/((4\log T)^2 d)}\right)^{d/2} && \text{(By Lemma C.7)} \\
&\le \left(\frac{\log\log T}{(\log T)^3}\right)^{\log T/(2\log\log T)} && (d \ge \log T/\log\log T) \\
&= \frac{T^{\log\log\log T/(2\log\log T)}}{T^{3/2}} \\
&\le T^{-1}
\end{aligned}$$

Similarly, we can show

$$\begin{aligned}
\Pr\left[\|\overline{\theta}-\theta\|_2 \ge \sqrt{d\log T}\right] &= \Pr\left[\|\Sigma^{-1/2}X\|_2 \ge \sqrt{d\log T}\right] \\
&\le \Pr\left[\sqrt{\lambda_{\max}(\Sigma)}\|X\|_2 \ge \sqrt{d\log T}\right] \\
&\le \Pr\left[\|X\|_2 \ge \sqrt{d\log T}\right] \\
&= \Pr\left[\|X\|_2^2 \ge d\log T\right] \\
&\le \left(\log T e^{1-\log T}\right)^{d/2} && \text{(By Lemma C.7)} \\
&\le \left(\exp\left(1+\log\log T-\log T\right)\right)^{\log T/(2\log\log T)} \\
& && (d \ge \log T/\log\log T) \\
&= T^{(1+\log\log T-\log T)/(2\log\log T)} \\
&\le T^{-1} && \text{(For } \log T > 1+3\log\log T\text{)}
\end{aligned}$$

By the triangle inequality,

$$\|\overline{\theta}\|_2 - \|\overline{\theta}-\theta\|_2 \le \|\theta\|_2 \le \|\overline{\theta}\|_2 + \|\overline{\theta}-\theta\|_2.$$

Thus, in this case, $\frac{1}{2\log T} \leq \|\theta\|_2 \leq \|\overline{\theta}\|_2 + \sqrt{d\log T}$ with probability at least $1 - 2T^{-1}$.

Case 2: $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3\log T}$ and $d < \log T / \log\log T$.

For this part of the proof, we just need that $d < \log T$, which it is by assumption. Using the triangle inequality, if $\|\overline{\theta}\|_2$ is large, it suffices to show that $\|\overline{\theta} - \theta\|_2$ is small with high probability. Again, let $X = \Sigma^{-1/2}(\overline{\theta} - \theta)$. Then,

$$\begin{aligned}
\Pr\left[\|\overline{\theta} - \theta\|_2 \geq \sqrt{3\log T}\right] &= \Pr\left[\|\Sigma^{1/2}X\|_2 \geq \sqrt{3\log T}\right] \\
&\geq \Pr\left[\sqrt{\lambda_{\max}(\Sigma)}\|X\|_2 \geq \sqrt{3\log T}\right] \\
&= \Pr\left[\|X\|_2 \geq \frac{\sqrt{3\log T}}{\sqrt{\lambda_{\max}(\Sigma)}}\right] \\
&\geq \Pr\left[\|X\|_2 \geq \sqrt{3\log T}\right] \\
&= \Pr\left[\|X\|_2^2 \geq 3\log T\right]
\end{aligned}$$

By Lemma C.7,

$$\begin{aligned}
\Pr\left[\|X\|_2^2 \geq 3\log T\right] &\leq \left(\frac{3\log T}{d}e^{1-\frac{3\log T}{d}}\right)^{d/2} \\
&= \left(T^{-3/d}e\frac{3\log T}{d}\right)^{d/2} \\
&= T^{-1}\left(T^{-1/d}e\frac{3\log T}{d}\right)^{d/2} \\
&\leq T^{-1} \qquad \text{(for sufficiently large } T\text{)}
\end{aligned}$$

Because $\|\overline{\theta}\|_2 \geq 1 + \sqrt{3\log T}$, $1 \leq \|\theta\|_2 \leq \|\overline{\theta}\|_2 + \sqrt{3\log T}$ with probability at least $1 - T^{-1}$. □

Next, we show how to upper-bound expected instantaneous regret in the worst case.[9]

[9]We state and prove this result in a slightly more general version which we use to support Section 5.3. For the sake of this section, a special case of $\ell = 0$ suffices.

Lemma 5.18. *Fix round t and parameter $\ell > 0$. For any θ, conditioned on any history h_{t-1} and the event that $\|\varepsilon_{a,t}\|_\infty \geq \ell$ for each arm a, the expected instantaneous regret of any algorithm at round t is at most*

$$2\,\|\theta\|_2 \left(1 + \rho(2 + \sqrt{2\log K}) + \ell\right).$$

Proof. The expected regret at round t is upper-bounded by the reward difference between the best arm x_t^* and the worst arm $x_t^\dagger$, which is

$$\theta^\top (x_t^* - x_t^\dagger).$$

Note that $x_t^* = \mu_t^* + \varepsilon_t^*$ and $x_t^\dagger = \mu_t^\dagger + \varepsilon_t^\dagger$. Then, this is

$$\begin{aligned}\theta^\top (x_t^* - x_t^\dagger) &= \theta^\top (\mu_t^* - \mu_t^\dagger) + \theta^\top (\varepsilon_t^* - \varepsilon_t^\dagger) \\ &\leq 2\|\theta\|_2 + \theta^\top (\varepsilon_t^* - \varepsilon_t^\dagger)\end{aligned}$$

since $\|\mu_{a,t}\|_2 \leq 1$. Next, note that

$$\theta^\top \varepsilon_t^* \leq \max_a \theta^\top \varepsilon_{a,t}$$

and

$$\theta^\top \varepsilon_t^\dagger \geq \min_a \theta^\top \varepsilon_{a,t}.$$

Since $\varepsilon_{a,t}$ has symmetry about the origin conditioned on the event that at least one component of one of the perturbations has absolute value at least ℓ, i.e. v and $-v$ have equal likelihood, $\max_a \theta^\top \varepsilon_{a,t}$ and $-\min_a \theta^\top \varepsilon_{a,t}$ are identically distributed. Let $E_{\ell,t}$ be the event that at least one of the components of one of the perturbations has absolute value at least ℓ. This means for any choice $\mu_{a,t}$ for all a,

$$\mathbb{E}\left[\theta^\top (x_t^* - x_t^\dagger) \mid E_{\ell,t}\right] \leq 2\|\theta\|_2 + 2\,\mathbb{E}\left[\max_a \theta^\top \varepsilon_{a,t} \mid E_{\ell,t}\right]$$

where the expectation is taken over the perturbations at time t.

Without loss of generality, let $(\varepsilon_{a',t})_j$ be the component such that $|(\varepsilon_{a',t})_j| \geq \ell$. Then, all other components have distribution $\mathcal{N}(0, \rho^2)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\max_a \theta^\top \varepsilon_{a,t} \mid E_{\ell,t}\right] \\
&= \mathbb{E}\left[\max_a \theta^\top \varepsilon_{a,t} \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \\
&= \mathbb{E}\left[\max(\theta^\top \varepsilon_{a',t}, \max_{a \neq a'} \theta^\top \varepsilon_{a,t}) \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \\
&\leq \mathbb{E}\left[\max\left(|\theta_j(\varepsilon_{a',t})_j| + \sum_{i \neq j} \theta_i(\varepsilon_{a',t})_i, \max_{a \neq a'} \theta^\top \varepsilon_{a,t}\right) \mid |(\varepsilon_{a',t})_j| \geq \ell\right]
\end{aligned}$$

Let $(\tilde{\varepsilon}_{a,t})_i = 0$ if $a = a'$ and $i = j$, and $(\varepsilon_{a,t})_i$ otherwise. In other words, we simply zero out the component $(\varepsilon_{a',t})_j$. Then, this is

$$\begin{aligned}
&\mathbb{E}\left[\max\left(|\theta_j(\varepsilon_{a',t})_j| + \theta^\top \tilde{\varepsilon}_{a',t}, \max_{a \neq a'} \theta^\top \tilde{\varepsilon}_{a,t}\right) \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \\
&\leq \mathbb{E}\left[\max_a \left(|\theta_j(\varepsilon_{a',t})_j| + \theta^\top \tilde{\varepsilon}_{a,t}\right) \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \\
&= \mathbb{E}\left[|\theta_j(\varepsilon_{a',t})_j| + \max_a \left(\theta^\top \tilde{\varepsilon}_{a,t}\right) \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \\
&= \mathbb{E}\left[|\theta_j(\varepsilon_{a',t})_j| \mid |(\varepsilon_{a',t})_j| \geq \ell\right] + \mathbb{E}\left[\max_a \left(\theta^\top \tilde{\varepsilon}_{a,t}\right)\right] \\
&\leq \mathbb{E}\left[|\theta_j(\varepsilon_{a',t})_j| \mid |(\varepsilon_{a',t})_j| \geq \ell\right] + \rho\|\theta\|_2\sqrt{2\log K}
\end{aligned}$$

because by Lemma C.5,

$$\mathbb{E}\left[\max_a \theta^\top \tilde{\varepsilon}_{a,t}\right] \leq \mathbb{E}\left[\max_a \theta^\top \varepsilon_{a,t}\right] \leq \rho\|\theta\|_2\sqrt{2\log K}$$

Next, note that by symmetry and since $\theta_j \leq \|\theta\|_2$,

$$\mathbb{E}\left[|\theta_j(\varepsilon_{a',t})_j| \mid |(\varepsilon_{a',t})_j| \geq \ell\right] \leq \|\theta\|_2 \mathbb{E}\left[(\varepsilon_{a',t})_j \mid (\varepsilon_{a',t})_j \geq \ell\right].$$

By Lemma C.1,

$$\mathbb{E}\left[(\varepsilon_{a',t})_j \mid (\varepsilon_{a',t})_j \geq \ell\right] \leq \max(2\rho, \ell + \rho) \leq 2\rho + \ell$$

Putting this all together, the expected instantaneous regret is bounded by

$$2\left(\|\theta\|_2\left(1+\rho(2+\sqrt{2\log K})+\ell\right)\right),$$

proving the lemma. □

5.4.2 Finishing the Proof of Theorem 5.13.

We focus on the "nice event" that (5.14) holds, denote it $\mathcal{E}$ for brevity. In particular, note that it implies $\|\theta\|_2 \leq S$. Lemma 5.16 guarantees that expected regret under this event, $\mathbb{E}\left[\text{Regret}(T) \mid \mathcal{E}\right]$, is upper-bounded by the expression (5.9) in the theorem statement.

In what follows we use Lemma 5.17(a) and Lemma 5.18 guarantee that if $\mathcal{E}$ fails, then the corresponding contribution to expected regret is small. Indeed, Lemma 5.18 with $\ell = 0$ implies that

$$\mathbb{E}\left[R_t \mid \bar{\mathcal{E}}\right] \leq BT\,\|\theta\|_2 \quad \text{for each round } t,$$

where $B = 1+\rho(2+\sqrt{2\log K})$ is the "blow-up factor". Since (5.14) fails with probability at most $\frac{2}{T}$ by Lemma 5.17(a), we have

$$\begin{aligned}
\mathbb{E}\left[\text{Regret}(T) \mid \bar{\mathcal{E}}\right]\ \Pr[\bar{\mathcal{E}}] &\leq \tfrac{2B}{T}\ \mathbb{E}\left[\|\theta\|_2 \mid \bar{\mathcal{E}}\right] \\
&\leq \tfrac{2B}{T}\ \mathbb{E}\left[\|\theta\|_2 \mid \|\theta\|_2 \geq \tfrac{1}{2\log T}\right] \\
&\leq O\left(\tfrac{B}{T}\right)\ \left(\|\overline{\theta}\|_2 + d\log T\right) \\
&\leq O(1).
\end{aligned}$$

The antecedent inequality follows by Lemma C.2 with $\alpha = \frac{1}{2\log T}$, using the assumption that $\lambda_{\max}(\Sigma) \leq 1$. The theorem follows.

5.5 Analysis: Greedy Algorithms with Perturbed Contexts

We present the proofs for our results on greedy algorithms in Section 5.3.[10] This section is structured as follows. In Section 5.5.1, we quantify the diversity of data collected by batch-greedy-style algorithms, assuming perturbed context generation. In Section 5.5.2, we show that a sufficiently "diverse" batch history suffices to simulate the reward for any given context vector, in the sense of Definition 5.10. Jointly, these two subsections imply that batch history generated by a batch-greedy-style algorithm can simulate rewards with high probability, as long as the batch size is sufficiently large. Section 5.5.3 builds on this foundation to derive regret bounds for BatchBayesGreedy. The crux is that the history collected by BatchBayesGreedy suffices to simulate a "slowed-down" run of any other algorithm. This analysis extends to a version of BatchFreqGreedy equipped with a Bayesian-greedy prediction rule (and tracks the performance of the prediction rule). Finally, Section 5.5.4 derives the regret bounds for BatchFreqGreedy, by comparing the prediction-rule version of BatchFreqGreedy with BatchFreqGreedy itself. To derive the results on group externalities, we present all our analysis in Sections 5.5.3 and 5.5.4 in a more general framework in which only the minority rounds are counted for regret.

Preliminaries. We assume perturbed context generation in this section, without further mention.

[10] That is, all results in Section 5.3 except the regret bound for LinUCB (Theorem 5.7(a)), which is proved in Section 5.4.

Throughout, we will use the following parameters as a shorthand:

$$\begin{aligned}\delta_R &= T^{-2}\\ \hat{R} &= \rho\sqrt{2\log(2TKd/\delta_R)}\\ R &= 1+\hat{R}\sqrt{d}.\end{aligned}$$

Recall that ρ denotes perturbation size, and d is the dimension. The meaning of $\hat{R}$ and R is that they are high-probability upper bounds on the perturbations and the contexts, respectively. More formally, by Lemma C.5 we have:

$$\Pr\left[\|\varepsilon_{a,t}\|_\infty \leq \hat{R}: \quad \text{for all arms } a \text{ and all rounds } t\right] \leq \delta_R \tag{5.15}$$

$$\Pr\left[\|x_{a,t}\|_2 \leq R: \quad \text{for all arms } a \text{ and all rounds } t\right] \leq \delta_R \tag{5.16}$$

Let us recap some of the key definitions from Section 5.3.2. We consider batch-greedy-style algorithms, a template that unifies BatchBayesGreedy and BatchFreqGreedy. A bandit algorithm is called *batch-greedy-style* if it divides the timeline in batches of Y consecutive rounds each, in each round t chooses some estimate θ_t of θ, based only on the data from the previous batches, and then chooses the best action according to this estimate, so that $a_t = \arg\max_a \theta_t^\top x_{a,t}$.

For a batch B that starts at round t_0+1, the *batch history* h_B is the tuple $((x_{t_0+\tau},\, r_{t_0+\tau}) : \ \tau \in [Y])$, and the *batch context matrix* X_B is the matrix whose rows are vectors $(x_{t_0+\tau} : \ \tau \in [Y])$. Here and elsewhere, $[Y] = \{1, \cdots, Y\}$. The *batch covariance matrix* is defined as

$$Z_B := X_B^\top X_B = \sum_{t=t_0+1}^{t_0+Y} x_t\, x_t^\top. \tag{5.17}$$

5.5.1 Data Diversity under Perturbations

We are interested in the diversity of data collected by batch-greedy-style algorithms, assuming perturbed context generation. Informally, the observed contexts $x_1, x_2, \ldots$ should cover all directions in order to enable good estimation of the latent vector θ. Following Kannan et al. (2018), we quantify data diversity via the minimal eigenvalue of the empirical covariance matrix Z_t. More precisely, we are interested in proving that $\lambda_{\min}(Z_t)$ is sufficiently large. We adapt some tools from Kannan et al. (2018), and then derive some improvements for batch-greedy-style algorithms.

Tools from Kannan et al. (2018)

Kannan et al. (2018) prove that $\lambda_{\min}(Z_t)$ grows linearly in time t, assuming t is sufficiently large.

Lemma 5.19 (Kannan et al. (2018))**.** *Fix any batch-greedy-style algorithm. Consider round $t \geq \tau_0$, where $\tau_0 = 160\frac{R^2}{\rho^2}\log\frac{2d}{\delta} \cdot \log T$. Then for any realization of θ, with probability $1-\delta$*

$$\lambda_{\min}(Z_t) \geq \frac{\rho^2 t}{32 \log T}.$$

Proof. The claimed conclusion follows from an argument inside the proof of Lemma B.1 from Kannan et al. (2018), plugging in $\lambda_0 = \frac{\rho^2}{2\log T}$. This argument applies for any $t \geq \tau_0'$, where $\tau_0' = \max\left(32\log\frac{2}{\delta}, 160\frac{R^2}{\rho^2}\log\frac{2d}{\delta}\cdot\log T\right)$. We observe that $\tau_0' = \tau_0$ since $R \geq \rho$. □

Recall that Z_t is the sum $Z_t := \sum_{\tau=1}^{t} x_\tau x_\tau^\top$. A key step in the proof of Lemma 5.19 zeroes in on the expected contribution of a single round. We use this tool separately in the proof of Lemma 5.22.

Lemma 5.20 (Kannan et al. (2018))**.** *Fix any batch-greedy-style algorithm, and the latent vector θ. Assume $T \geq 4K$. Condition on the event that all perturbations $\varepsilon_{a,t}$ are upper-bounded by $\hat{R}$, denote it with $\mathcal{E}$. Then with probability at least $\frac{1}{4}$,*

$$\lambda_{\min}\left(\mathbb{E}\left[x_t x_t^\top \mid h_{t-1}, \mathcal{E}\right]\right) \geq \frac{\rho^2}{2\log T}.$$

Proof. The proof is easily assembled from several pieces in the analysis in Kannan et al. (2018). Let $\hat{\theta}_t$ be the algorithm's estimate for θ at time t. As in Kannan et al. (2018), define

$$\hat{c}_{a,t} = \max_{a' \neq a} \hat{\theta}_t^\top x_{a',t},$$

where $\hat{c}_{a,t}$ depends on all perturbations other than the perturbation for $x_{a,t}$. Let us say that $\hat{c}_{a,t}$ is "good" for arm a if

$$\hat{c}_{a,t} \leq \hat{\theta}_t^\top \mu_{a,t} + \rho\sqrt{2\log T}\|\hat{\theta}_t\|_2.$$

First we argue that

$$\Pr\left[\hat{c}_{a,t} \text{ is good for } a \mid a_t = t, \mathcal{E}\right] \geq \tfrac{1}{4}. \tag{5.18}$$

Indeed, in the proof of their Lemma 3.4, Kannan et al. (2018) show that for any round, conditioned on $\mathcal{E}$, if the probability that arm a was chosen over the randomness of the perturbation is at least $2/T$, then the round is good for a with probability at least $\frac{1}{2}$. Let B_t be the set of arms at round t with probability at most $2/T$ of being chosen over the randomness of the perturbation. Then,

$$\Pr_{\varepsilon \sim \mathcal{N}(0,\rho^2 I)}\left[a_t \in B_t\right] \leq \sum_{a \in B_T} \Pr_{\varepsilon \sim \mathcal{N}(0,\rho^2 I)}\left[a_t = a\right] \leq \tfrac{2}{T}|B_t| \leq \tfrac{2K}{T} \leq \tfrac{1}{2}.$$

Since by assumption $T \geq 4K$, (5.18) follows.

Second, we argue that

$$\lambda_{\min}\left(\mathbb{E}\left[x_{a,t}x_{a,t}^{\top} \mid a_t = a, \hat{c}_{a,t} \text{ is good}\right]\right) \geq \frac{\rho^2}{2\log T} \tag{5.19}$$

This is where we use conditioning on the event $\{\varepsilon_{a,t} \leq \hat{R}\}$. We plug in $r = \rho\sqrt{2\log T}$ and $\lambda_0 = \frac{\rho^2}{2\log T}$ into Lemma 3.2 of Kannan et al. (2018). This lemma applies because with these parameters, the perturbed distribution of context arrivals satisfies the $(\rho\sqrt{2\log T}, \rho^2/(2\log T))$-diversity condition from Kannan et al. (2018). The latter is by Lemma 3.6 of Kannan et al. (2018). This completes the proof of (5.19). The lemma follows from (5.18) and (5.19). □

Let θ_t^{fre} be the BatchFreqGreedy estimate for θ at time t, as defined in (5.7). We are interested in quantifying how the quality of this estimate improves over time. Kannan et al. (2018) prove, essentially, that the distance between θ_t^{fre} and θ scales as $\sqrt{t}/\lambda_{\min}(Z_t)$.

Lemma 5.21 (Kannan et al. (2018))**.** *Consider any round t in the execution of Batch-FreqGreedy. Let t_0 be the last round of the previous batch. For any θ and any $\delta > 0$, with probability $1 - \delta$,*

$$\|\theta - \theta_t^{fre}\|_2 \leq \frac{\sqrt{t_0 \cdot 2dR\log\frac{d}{\delta}}}{\lambda_{\min}(Z_{t_0})}.$$

Some improvements We focus on the batch covariance matrix Z_B of a given batch in a batch-greedy-style algorithm. We would like to prove that $\lambda_{\min}(Z_B)$ is sufficiently large with high probability, as long as the batch size Y is large enough. The analysis from Kannan et al. (2018) (a version of Lemma 5.19) would

apply, but only as long as the batch size is least as large as the τ_0 from the statement of Lemma 5.19. We derive a more efficient version, essentially shaving off a factor of 8.[11]

Lemma 5.22. *Fix a batch-greedy-style algorithm and any batch B in the execution of this algorithm. Fix $\delta > 0$ and assume that the batch size Y is at least*

$$Y_0 := (\tfrac{R}{\rho})^2 \tfrac{8e^2}{(e-1)^2} \left(1 + \log \tfrac{2d}{\delta}\right) \log(T) + \tfrac{4e}{e-1} \log \tfrac{2}{\delta}. \tag{5.20}$$

Condition on the event that all perturbations in this batch are upper-bounded by $\hat{R}$, more formally:

$$\mathcal{E}_B = \{\|\varepsilon_{a,t}\|_\infty \leq \hat{R} : \text{ for all arms } a \text{ and all rounds } t \text{ in } B\}.$$

Further, condition on the latent vector θ and the history h before batch B. Then

$$\Pr\left[\, \lambda_{\min}(Z_B) \geq R^2 \mid \mathcal{E}_B, h, \theta \right] \geq 1 - \delta. \tag{5.21}$$

The probability in (5.21) *is over the randomness in context arrivals and rewards in batch B.*

The improvement over Lemma 5.19 comes from two sources: we use a tail bound on the sum of geometric random variables instead of a Chernoff bound on a binomial random variable, and we derive a tighter application of the eigenvalue concentration inequality of Tropp (2012).

Proof. Let t_0 be the last round before batch B. Recalling (5.17), let

$$W_B = \sum_{t=t_0+1}^{t_0+Y} \mathbb{E}\left[x_t x_t^\top \mid h_{t-1}\right]$$

[11] Essentially, the factor of 160 in Lemma 5.19 is replaced with factor $\frac{8e^2}{(e-1)^2} < 20.022$ in (5.20).

be a similar sum over the expected per-round covariance matrices. Assume $Y \geq Y_0$

The proof proceeds in two steps: first we lower-bound $\lambda_{\min}(Z_B)$, and then we show that it implies (5.21). Denoting $m = R^2 \frac{e}{e-1}(1 + \log \frac{2d}{\delta})$, we claim that

$$\Pr\left[\lambda_{\min}(W_B) < m \mid \mathcal{E}_B, h\right] \leq \tfrac{\delta}{2}. \tag{5.22}$$

To prove this, observe that W_B's minimum eigenvalue increases by at least $\lambda_0 = \rho^2/(2\log T)$ with probability at least $1/4$ each round by Lemma 5.20, where the randomness is over the history, i.e., the sequence of (context, reward) pairs. If we want it to go up to m, this should take $4m/\lambda_0$ rounds in expectation. However, we need it to go to m with high probability. Notice that this is dominated by the sum of m/λ_0 geometric random variables with parameter $\frac{1}{4}$. We'll use the following bound from Janson (2018): for $X = \sum_{i=1}^{n} X_i$ where $X_i \sim \text{Geom}(p)$ and any $c \geq 1$,

$$\Pr[X \geq c\mathbb{E}[X]] \leq \exp\left(-n(c - 1 - \log c)\right).$$

Because we want the minimum eigenvalue of W_B to be m, we need $n = m/\lambda_0$, so $\mathbb{E}[X] = 4m/\lambda_0$. Choose $c = \left(1 + \frac{\lambda_0}{m}\log\frac{2}{\delta}\right)\frac{e}{e-1}$. By Corollary C.15,

$$c - 1 - \log c \geq \tfrac{e-1}{e} \cdot c - 1 = \tfrac{\lambda_0}{m}\log\tfrac{2}{\delta}.$$

Therefore,

$$\Pr\left[X \geq c\mathbb{E}[X]\right] \leq \exp\left(-n \cdot \tfrac{\lambda_0}{m}\log\tfrac{2}{\delta}\right) = \left(\tfrac{\delta}{2}\right)^{n\cdot\lambda_0/m} = \tfrac{\delta}{2}$$

Thus, with probability $1 - \frac{\delta}{2}$, $\lambda_{\min}(W_B) \geq m$ as long as the batch size Y is at least

$$\frac{e}{e-1}\left(1 + \frac{\lambda_0}{m}\log\frac{2}{\delta}\right) \cdot \mathbb{E}[X] = \frac{4e}{e-1}\left(\frac{m}{\lambda_0} + \log\frac{2}{\delta}\right) = Y_0.$$

This completes the proof of (5.22).

To derive (5.21) from (5.22), we proceed as follows. Consider the event

$$\mathcal{E} = \left\{ \lambda_{\min}(Z_B) \leq R^2 \text{ and } \lambda_{\min}(W_B) \geq m \right\}.$$

Letting $\alpha = 1 - R^2/m$ and rewriting R^2 as $(1-\alpha)m$, we use a concentration inequality from Tropp (2012) to guarantee that

$$\Pr[\mathcal{E} \mid \mathcal{E}_B, h] \leq d \left(e^{\alpha}(1-\alpha)^{1-\alpha}\right)^{-m/R^2}.$$

Then, using the fact that $x^x \geq e^{-1/e}$ for all $x > 0$, we have

$$\begin{aligned}\Pr[\mathcal{E} \mid \mathcal{E}_B, h] &\leq d \left(e^{1-R^2/m-1/e}\right)^{-m/R^2} = d\, e^{-(m-R^2-m/e)/R^2} \\ &= d \exp\left(-\frac{\left(\frac{e-1}{e}\right) m}{R^2} + 1\right) \leq \tfrac{\delta}{2},\end{aligned}$$

since $m \geq \frac{e}{e-1} R^2 \left(1 + \log \frac{2d}{\delta}\right)$. Finally, observe that, omitting the conditioning on $\mathcal{E}_B, h$, we have:

$$\Pr\left[\lambda_{\min}(Z_B) \leq R^2\right] \leq \Pr\left[\mathcal{E}\right] + \Pr\left[\lambda_{\min}(W_B) < m\right] \leq \tfrac{\delta}{2} + \tfrac{\delta}{2} = \delta.$$

□

5.5.2 Reward Simulation with a Diverse Batch History

We consider reward simulation with a batch history, in the sense of Definition 5.10. We show that a sufficiently "diverse" batch history suffices to simulate the reward for any given context vector. Coupled with the results of Section 5.5.1, it follows that batch history generated by a batch-greedy-style algorithm can simulate rewards as long as the batch size is sufficiently large.

Let us recap the definition of reward simulation (Definition 5.10). Let $\texttt{Rew}_\theta(\cdot)$ be a randomized function that takes a context x and outputs an independent

random sample from $\mathcal{N}(\theta^\top x, 1)$. In other words, this is the realized reward for an action with context vector x.

Definition 5.23. *Consider batch B in the execution of a batch-greedy-style algorithm. Batch history h_B can simulate $\texttt{Rew}_\theta()$ up to radius $R > 0$ if there exists a function $g : \{\text{context vectors}\} \times \{\text{batch histories } h_B\} \to \mathbb{R}$ such that $g(x, h_B)$ is identically distributed to $\texttt{Rew}_\theta(x)$ conditional on the batch context matrix, for all θ and all context vectors $x \in \mathbb{R}^d$ with $\|x\|_2 \leq R$.*

Note that we do not require the function g to be efficiently computable. We do not require algorithms to compute g; a mere existence of such function suffices for our analysis.

The result in this subsection does not rely on the "greedy" property. Instead, it applies to all "batch-style" algorithms, defined as follows: time is divided in batches of Y consecutive rounds each, and the action at each round t only depends on the history up to the previous batch. The data diversity condition is formalized as $\{\lambda_{\min}(Z_B) \geq R^2\}$; recall that it is a high-probability event, in a precise sense defined in Lemma 5.22. The result is stated as follows:

Lemma 5.24. *Fix a batch-style algorithm and any batch B in the execution of this algorithm. Assume the batch covariance matrix Z_B satisfies $\lambda_{\min}(Z_B) \geq R^2$. Then batch history h_B can simulate $\texttt{Rew}_\theta$ up to radius R.*

Proof. Let us construct a suitable function g for Definition 5.23. Fix a context vector $x \in \mathbb{R}^d$ with $\|x\|_2 \leq R$. Let r_B be the vector of realized rewards in batch B, i.e., $r_B = (r_t : \text{rounds } t \text{ in } B) \in \mathbb{R}^Y$. Define

$$g(x, h_B) = w_B^\top r_B + \mathcal{N}\left(0, 1 - \|w_B\|_2^2\right), \text{where } w_B = X_B\, Z_B^{-1}\, x \in \mathbb{R}^Y. \tag{5.23}$$

Recall that the variance of the reward noise is 1. (We can also handle a more general version in which the variance of the reward noise is σ^2. Then the noise variance in (5.23) should be $\sigma^2\left(1-\|w_B\|_2^2\right)$, with essentially no modifications throughout the rest of the proof.)

Note that w_B is well-defined: indeed, Z_B is invertible since $\lambda_{\min}(Z_B) \geq R^2 > 0$. In the rest of the proof we show that g is as needed for Definition 5.23.

First, we will show that for any $x \in \mathbb{R}^d$ such that $\|x\|_2 \leq R$, the weights $w_B \in \mathbb{R}^t$ as defined above satisfy $X_B^\top w_B = x$ and $\|w_B\|_2 \leq 1$. Then, we'll show that if each $r_\tau \sim \mathcal{N}(\theta^\top x_\tau, 1)$, then $r_B^\top w_B + \mathcal{N}(0, 1-\|w_B\|_2^2) \sim \mathcal{N}(\theta^\top x, 1)$.

Trivially, we have

$$X_B^\top w_B = X_B^\top X_B (X_B^\top X_B)^{-1} x = x$$

as desired. We must now show that $\|w_B\|_2^2 \leq 1$. Note that

$$\|w_B\|_2^2 = w_B^\top w_B = w_B^\top X_B Z_B^{-1} x = x^\top Z_B^{-1} x = \|x\|_{Z_B^{-1}}^2$$

where $\|v\|_M^2$ simply denotes $v^\top M v$. Thus, it is sufficient to show that $\|x\|_{Z_B^{-1}}^2 \leq 1$. Since $\|x\|_2 \leq R$ and $\lambda_{\min}(Z_B) \geq R^2$, we have by Lemma C.11

$$Z_B \succeq R^2 I \succeq x x^\top.$$

By Lemma C.12, we have

$$I \succeq Z_B^{-1/2} x x^\top Z_B^{-1/2}.$$

Let $z = Z_B^{-1/2} x$, so $I \succeq z z^\top$. Again by Lemma C.11, $\lambda_{\max}(z z^\top) = z^\top z$. This means that

$$1 \geq z^\top z = (Z_B^{-1/2} x)^\top Z_B^{-1/2} x = x^\top Z_B^{-1} x = \|x\|_{Z_B^{-1}}^2 = \|w_B\|_2^2$$

as desired. Finally, observe that

$$r_B^\top w_B = (X_B\theta + \eta)^\top w_B = \theta^\top X_B^\top w_B + \eta^\top w_B = \theta^\top x + \eta^\top w_B$$

where $\eta \sim \mathcal{N}(0, I)$ is the noise vector. Notice that $\eta^\top w_B \sim \mathcal{N}(0, \|w_B\|_2)$, and therefore, $\eta^\top w_B + \mathcal{N}(0, 1 - \|w_B\|_2^2) \sim \mathcal{N}(0, 1)$. Putting this all together, we have

$$r_B^\top w_B + \mathcal{N}(0, 1 - \|w_B\|_2^2) \sim \mathcal{N}(\theta^\top x, 1)$$

and therefore D can simulate E for any x up to radius R. □

5.5.3 Regret Bounds for BatchBayesGreedy

We apply the tools from Sections 5.5.1 and 5.5.2 to derive regret bounds for BatchBayesGreedy. On a high level, we prove that the history collected by BatchBayesGreedy suffices to simulate a "slowed-down" run of any other algorithm $\texttt{ALG}_0$. Therefore, when it comes to choosing the next action, BatchBayesGreedy has at least as much information as $\texttt{ALG}_0$, so its Bayesian-greedy choice cannot be worse than the choice made by $\texttt{ALG}_0$.

Our analysis extends to a more general scenario which is useful for the analysis of BatchFreqGreedy. We formulate and prove our results for this scenario directly. We consider an extended bandit model which separates data collection and reward collection. Each round t proceeds as follows: the algorithm observes available actions and the context vectors for these actions, then it chooses *two* actions, a_t and a'_t, and observes the reward for the former but not the latter. We refer to a'_t as the "prediction" at round t. We will refer to an algorithm in this model as a bandit algorithm (which chooses actions a_t) with "prediction rule" that chooses the predictions a'_t. More specifically, we will be interested in

an arbitrary batch-greedy-style algorithm with prediction rule given by Batch-BayesGreedy, as per (5.6). We assume this prediction rule henceforth. We are interested in *prediction regret*: a version of regret (5.1) if actions a_t are replaced with predictions a'_t:

$$\texttt{PReg}(T) = \sum_{t=1}^{T} \theta^\top x^*_t - \theta^\top x_{a'_t,t} \tag{5.24}$$

where x^*_t is the context vector of the best action at round t, as in (5.1). More precisely, we are interested in *Bayesian prediction regret*, the expectation of (5.24) over everything: the context vectors, the rewards, the algorithm's random seed, and the prior over θ.

We use essentially the same analysis to derive implications on group externalities. For this purpose, we consider a further generalization in which regret is restricted to rounds that correspond to a particular population. Formally, let $\mathcal{T} \subseteq \mathbb{N}$ be a randomly chosen subset of the rounds where $\Pr[t \in \mathcal{T}]$ is a constant and rounds are chosen to be in $\mathcal{T}$ independently of one another. We allow for the possibility that the underlying context distribution differs for rounds in $\mathcal{T}$ compared to rounds in $[T]\backslash\mathcal{T}$. More precisely, we allow the event $\{t \in \mathcal{T}\}$ be correlated with the context tuple at round t. Similar to the definition of minority regret, we define *$\mathcal{T}$-restricted regret* (resp., prediction regret) in T rounds to be the portion of regret (resp., prediction regret) that corresponds to $\mathcal{T}$-rounds:

$$R^{\mathcal{T}}(T) = \sum_{t\leq T,\ t\in\mathcal{T}} \theta^\top x^*_t - \theta^\top x_{a_t,t}. \tag{5.25}$$

$$\texttt{PReg}^{\mathcal{T}}(T) = \sum_{t\leq T,\ t\in\mathcal{T}} \theta^\top x^*_t - \theta^\top x_{a'_t,t}. \tag{5.26}$$

$\mathcal{T}$-restricted *Bayesian* (prediction) regret is defined as an expectation over everything.

Thus, the main theorem of this subsection is formulated as follows:

Theorem 5.25. *Consider perturbed context generation. Let* ALG *be an arbitrary batch-greedy-style algorithm whose batch size is at least Y_0 from* (5.20)*. Fix any bandit algorithm* $\texttt{ALG}_0$*, and let $R_0^{\mathcal{T}}(T)$ be the $\mathcal{T}$-restricted regret of this algorithm on a particular problem instance $\mathcal{I}$. Then on the same instance,* ALG *has $\mathcal{T}$-restricted Bayesian prediction regret*

$$\mathbb{E}\left[\texttt{PReg}^{\mathcal{T}}(T)\right] \leq Y \cdot \mathbb{E}\left[R_0^{\mathcal{T}}(T/Y)\right] + \tilde{O}(1/T). \tag{5.27}$$

Proof sketch. We use a t-round history of ALG to simulate a (t/Y)-round history of $\texttt{ALG}_0$. More specifically, we use each batch in the history of ALG to simulate one round of $\texttt{ALG}_0$. We prove that the simulated history of $\texttt{ALG}_0$ has exactly the same distribution as the actual history, for any θ. Since ALG predicts the Bayesian-optimal action given the history (up to the previous batch), this action is at least as good (in expectation over the prior) as the one chosen by $\texttt{ALG}_0$ after t/Y rounds. The detailed proof is deferred to Section 5.5.3.

Implications. As a corollary of this theorem, we obtain regret bounds for BatchBayesGreedy in Theorem 5.6 and Theorem 5.7. We take $\mathcal{T}$ to be the set of all rounds, i.e., $\Pr[t \in \mathcal{T}] = 1$, and ALG to be BatchBayesGreedy. For Theorem 5.7(b), we take $\texttt{ALG}_0$ to be LinUCB. Thus:

Corollary 5.26. *In the setting of Theorem 5.25, BatchBayesGreedy has Bayesian regret at most $Y \cdot \mathbb{E}\left[R_0(T/Y)\right] + \tilde{O}(1/T)$ on problem instance $\mathcal{I}$. Further, under the assumptions of Theorem 5.7, BatchBayesGreedy has Bayesian regret at most $\tilde{O}(d^2\, K^{2/3}\, T^{1/3}/\rho^2)$ on all instances.*

We also obtain a similar regret bound on the Bayesian prediction regret of BatchFreqGreedy, which is essential for Section 5.5.4.

Corollary 5.27. *In the setting of Theorem 5.25, BatchFreqGreedy has Bayesian prediction regret* (5.27).

To derive Theorem 5.9 for BatchBayesGreedy, we take $\mathcal{T}$ to be the set of all minority rounds, and apply Theorem 5.25 twice: first when $\texttt{ALG}_0$ is run over the minority rounds only (and can behave arbitrarily on the rest), and then when $\texttt{ALG}_0$ is run over full population.

Proof of Theorem 5.25

We condition on the event that all perturbations are bounded by $\hat{R}$, more precisely, on the event

$$\mathcal{E}_1 = \left\{ \|\varepsilon_{a,t}\|_\infty \leq \hat{R} : \text{ for all arms } a \text{ and all rounds } t \right\}. \tag{5.28}$$

Recall that $\mathcal{E}_1$ is a high-probability event, by (5.15). We also condition on the event

$$\mathcal{E}_2 = \left\{ \lambda_{\min}(Z_B) \geq R^2 : \text{ for each batch } B, \right\}$$

where Z_B is the batch covariance matrix, as usual. Conditioned on $\mathcal{E}_1$, this too is a high-probability event by Lemma 5.22 plugging in δ/T and taking a union bound over all batches.

We will prove that ALG satisfies

$$\mathbb{E}\left[\texttt{PReg}^{\mathcal{T}}(T) \mid \mathcal{E}_1, \mathcal{E}_2\right] \leq Y \cdot \mathbb{E}\left[R_0^{\mathcal{T}}(\lceil T/Y \rceil) \mid \mathcal{E}_1, \mathcal{E}_2\right], \tag{5.29}$$

where the expectation is taken over everything: the context vectors, the rewards, the algorithm's random seed, and the prior over θ. Then we take care of the "failure event" $\overline{\mathcal{E}_1 \cap \mathcal{E}_2}$.

History simulation. Before we prove (5.29), let us argue about using the history of ALG to simulate a (shorter) run of $\texttt{ALG}_0$. Fix round t. We use a t-round history of ALG to simulate a $\lfloor t/Y \rfloor$-round run of $\texttt{ALG}_0$, where Y is the batch size in ALG. Stating this formally requires some notation. Let A_t be the set of actions available in round t, and let $\texttt{con}_t = (x_{a,t} : a \in A_t)$ be the corresponding tuple of contexts. Let CON be the set of all possible context tuples, more precisely, the set of all finite subsets of $\mathbb{R}^d$. Let h_t and h_t^0 denote, resp., the t-round history of ALG and $\texttt{ALG}_0$. Let $\mathcal{H}_t$ denote the set of all possible t-round histories. Note that h_t and h_t^0 are random variables which take values on $\mathcal{H}_t$. We want to use history h_t to simulate history $h^0_{\lfloor t/Y \rfloor}$. Thus, the simulation result is stated as follows:

Lemma 5.28. *Fix round t and let $\sigma = (\texttt{con}_1, \ldots, \texttt{con}_{\lfloor t/Y \rfloor})$ be the sequence of context arrivals up to and including round $\lfloor t/Y \rfloor$. Then there exists a "simulation function"*

$$\texttt{sim} = \texttt{sim}_t : \mathcal{H}_t \times \texttt{CON}_{\lfloor t/Y \rfloor} \to \mathcal{H}_{\lfloor t/Y \rfloor}$$

such that the simulated history $\texttt{sim}(h_t, \sigma)$ is distributed identically to $h^0_{\lfloor t/Y \rfloor}$, conditional on sequence σ, latent vector θ, and events $\mathcal{E}_1, \mathcal{E}_2$.

Proof. Throughout this proof, condition on events $\mathcal{E}_1$ and $\mathcal{E}_2$. Generically, $\texttt{sim}(h_t, \sigma)$ outputs a sequence of pairs $\{(x_\tau, r_\tau)\}_{\tau=1}^{\lfloor t/Y \rfloor}$, where x_τ is a context vector and r_τ is a simulated reward for this context vector. We define $\texttt{sim}(h_t, \sigma)$ by induction on τ with base case $\tau = 0$. Throughout, we maintain a run of algorithm $\texttt{ALG}_0$. For each step $\tau \geq 1$, suppose $\texttt{ALG}_0$ is simulated up to round $\tau - 1$, and

the corresponding history is recorded as $((x_1, r_1), \ldots, (x_{\tau-1}, r_{\tau-1}))$. Simulate the next round in the execution of $\texttt{ALG}_0$ by presenting it with the action set A_τ and the corresponding context tuple $\texttt{con}_\tau$. Let x_τ be the context vector chosen by $\texttt{ALG}_0$. The corresponding reward r_τ is constructed using the τ-th batch in h_t, denote it with B. By Lemmas 5.22 and 5.24, the batch history h_B can simulate a single reward, in the sense of Definition 5.23. In particular, there exists a function $g(x, h_B)$ with the required properties (recall that it is explicitly defined in (5.23)). Thus, we define $r_\tau = g(x_\tau, h_B)$, and return r_τ as a reward to $\texttt{ALG}_0$. This completes the construction of $\texttt{sim}(h_t, \sigma)$. The distribution property of $\texttt{sim}(h_t, \sigma)$ is immediate from the construction. □

Proof of Equation (5.29). We argue for each batch separately, and then aggregate over all batches in the very end. Fix batch B, and let $t_0 = t_0(B)$ be the last round in this batch. Let $\tau = 1 + t_0/Y$, and consider the context vector x^0_τ chosen by $\texttt{ALG}_0$ in round τ. This context vector is a randomized function f of the current context tuple $\texttt{con}_\tau$ and the history $h^0_{\tau-1}$:

$$x^0_\tau = f(\texttt{con}_\tau; h^0_{\tau-1}).$$

By Lemma 5.28, letting $\sigma = (\texttt{con}_1, \ldots, \texttt{con}_{\lfloor t/Y \rfloor})$, it holds that

$$\mathbb{E}\left[x^0_\tau \cdot \theta \mid \sigma, \theta, \mathcal{E}_1, \mathcal{E}_2\right] = \mathbb{E}\left[f(\texttt{con}_\tau; \texttt{sim}(h_{t_0}, \sigma)) \cdot \theta \mid \sigma, \theta, \mathcal{E}_1, \mathcal{E}_2\right] \tag{5.30}$$

Let t be some round in the next batch after B, and let $x'_t = x_{a'_t, t}$, be the context vector predicted by $\texttt{ALG}$ in round t. Recall that x'_t is a Bayesian-greedy choice from the context tuple $\texttt{con}_t$, based on history h_{t_0}. Observe that the Bayesian-greedy action choice from a given context tuple based on history h_{t_0} cannot be worse, in terms of the Bayesian-expected reward, than any other choice from

the same context tuple and based on the same history. Using (5.30), we obtain:

$$\mathbb{E}\left[x'_t \cdot \theta \mid \texttt{con}_t = \texttt{con}, \mathcal{E}_1, \mathcal{E}_2\right] \geq \mathbb{E}\left[x^0_\tau \cdot \theta \mid \texttt{con}_\tau = \texttt{con}, \mathcal{E}_1, \mathcal{E}_2\right], \tag{5.31}$$

for any given context tuple $\texttt{con} \in \texttt{CON}$ that has a non-zero arrival probability given $\mathcal{E}_1 \cap \mathcal{E}_2$.

Given $\texttt{con}_t = \texttt{con}$, the event $t \in \mathcal{T}$ is independent of everything else. Likewise, given $\texttt{con}_\tau = \texttt{con}$, the event $\tau \in \mathcal{T}$ is independent of everything else. It follows that

$$\mathbb{E}\left[x'_t \cdot \theta \mid \texttt{con}_t = \texttt{con}, t \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right] \geq \mathbb{E}\left[x^0_\tau \cdot \theta \mid \texttt{con}_\tau = \texttt{con}, \tau \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right], \tag{5.32}$$

for any given context tuple $\texttt{con} \in \texttt{CON}$ that has a non-zero arrival probability given $\mathcal{E}_1 \cap \mathcal{E}_2$.

Observe that $\texttt{con}_t$ and $\texttt{con}_\tau$ have the same distribution, even conditioned on event $\mathcal{E}_1 \cap \mathcal{E}_2$. (This is because the definitions of $\mathcal{E}_1$ and $\mathcal{E}_2$ treat all rounds in the same batch in exactly the same way.) Therefore, we can integrate (5.32) over the context tuples $\texttt{con}$:

$$\mathbb{E}\left[x'_t \cdot \theta \mid t \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right] \geq \mathbb{E}\left[x^0_\tau \cdot \theta \mid \tau \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right], \tag{5.33}$$

Now, let us sum up (5.33) over all rounds t in the next batch after B, denote it $\texttt{next}(B)$.

$$\sum_{t \in \texttt{next}(B)} \mathbb{E}\left[x'_t \cdot \theta \mid t \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right] \geq Y \cdot \mathbb{E}\left[x^0_\tau \cdot \theta \mid \tau \in \mathcal{T}, \mathcal{E}_1, \mathcal{E}_2\right]. \tag{5.34}$$

Note that the right-hand side of (5.33) stays the same for all t, hence the factor of Y on the right-hand side of (5.34). This completes our analysis of a single batch B.

We obtain (5.29) by over all batches B. Here it is essential that the expectation $\mathbb{E}\left[\mathbf{1}_{\{t\in\mathcal{T}\}}\,\theta^\top x_t^*\right]$ does not depend on round t, and therefore the "regret benchmark" $\theta^\top x_t^*$ cancels out from (5.29). In particular, it is essential that the context tuples $\texttt{con}_t$ are identically distributed across rounds. □

Proof of Theorem 5.25 given Equation (5.29). We must take care of the low-probability failure events $\overline{\mathcal{E}}_1$ and $\overline{\mathcal{E}}_2$. Specifically, we need to upper-bound the expression

$$\mathbb{E}_{\theta\sim P}\left[\texttt{PReg}^{\mathcal{T}}(T)\mid \overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2\right]\cdot\Pr[\overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2].$$

For ease of exposition, we focus on the special case $\Pr\left[t\in\mathcal{T}\right]=1$; the general case is treated similarly. We know that $\Pr[\overline{\mathcal{E}}_1\cup\overline{\mathcal{E}}_2]\leq\delta+\delta_R$. Lemma 5.18 with $\ell=\hat{R}$ gives us that the instantaneous regret of every round is at most

$$\begin{aligned}
&2\,\mathbb{E}_{\theta\sim(\mathcal{P}\mid h_{t-1})}\left[\|\theta\|_2\left(1+\rho(2+\sqrt{2\log K})+\hat{R}\right)\right]\\
&\qquad\leq 2\left[\left(\|\overline{\theta}\|_2+\sqrt{d\lambda_{\max}(\Sigma)}\right)\left(1+\rho(2+\sqrt{2\log K})+\hat{R}\right)\right]
\end{aligned}$$

by Lemma C.6. Letting $\delta=\delta_R=\frac{1}{T^2}$, we verify that our definition of Y means that Lemma 5.22 indeed holds with probability at least $1-T^{-2}$. Using (5.29), the Bayesian prediction regret of ALG is

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathcal{P}}\left[\texttt{PReg}^{\mathcal{T}}(T)\right]\\
&\leq Y\,\mathbb{E}_{\theta\sim\mathcal{P}}\left[R_0^{\mathcal{T}}\left(\tfrac{T}{Y}\right)\right]\\
&\quad+2\,T(\delta+\delta_R)\left[\left(\|\overline{\theta}\|_2+\sqrt{d\lambda_{\max}(\Sigma)}\right)\left(1+\rho(2+\sqrt{2\log K})+\hat{R}\right)\right]\\
&\leq Y\,\mathbb{E}_{\theta\sim\mathcal{P}}\left[R_0^{\mathcal{T}}\left(\tfrac{T}{Y}\right)\right]+\tilde{O}\left(\tfrac{1}{T}\right).
\end{aligned}$$

This completes the proof of Theorem 5.25. □

5.5.4 Regret Bounds for BatchFreqGreedy

To analyze BatchFreqGreedy, we show that its Bayesian regret is not too different from its Bayesian prediction regret, and use Corollary 5.27 to bound the latter. As in the previous subsection, we state this result in more generality for the sake of group externality implications: we consider $\mathcal{T}$-restricted (prediction) regret, exactly as before.

Theorem 5.29. *Assuming perturbed context generation, BatchFreqGreedy satisfies*

$$\left| \mathbb{E}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T)\right] \right| \leq \tilde{O}\left(\frac{\sqrt{d}}{\rho^2}\right)\left(\sqrt{\lambda_{\max}(\Sigma)} + \frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}\right),$$

where Σ is the covariance matrix of the prior and ρ is the perturbation size.

Taking $\mathcal{T}$ to be the set of all contexts, and using Corollary 5.27, we obtain Bayesian regret bounds for BatchFreqGreedy in Theorem 5.6 and Theorem 5.7. To derive Theorem 5.9 for BatchFreqGreedy, we take $\mathcal{T}$ to be the set of all minority rounds.

The remainder of this section is dedicated to proving Theorem 5.29. On a high level, the idea is as follows. As in the proof of Theorem 5.25, we condition on the high-probability event (5.28) that perturbations are bounded. Specifically, we prove that

$$\left| \mathbb{E}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T) \mid \mathcal{E}_1\right] \right| \leq \tilde{O}\left(\frac{\sqrt{d}}{\rho^2}\right)\left(\sqrt{\lambda_{\max}(\Sigma)} + \frac{1}{\sqrt{\lambda_{\min}(\Sigma)}}\right). \quad (5.35)$$

To prove this statement, we fix round t and compare the action a_t taken by BatchFreqGreedy and the predicted action a'_t. We observe that the difference in rewards between these two actions can be upper-bounded in terms of $\theta_t^{\text{bay}} - \theta_t^{\text{fre}}$, the difference in the θ estimates with and without knowledge of the prior.

(Recall (5.6) and (5.7) for definitions.) Specifically, we show that

$$\mathbb{E}\left[\theta^\top(x_{a_t,t} - x_{a'_t,t}) \mid \mathcal{E}_1\right] \leq 2R\,\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2\right]. \tag{5.36}$$

The crux of the proof is to show that the difference $\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2$ is small, namely

$$\mathbb{E}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2 \mid \mathcal{E}_1\right] = \tilde{O}(1/t), \tag{5.37}$$

ignoring other parameters. Thus, summing over all rounds, we get

$$\mathbb{E}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T) \mid \mathcal{E}_1\right] \leq O(\log T) = \tilde{O}(1).$$

Proof of (5.35)**.** Let R^t and $\texttt{PReg}^t$ be, resp., instantaneous regret and instantaneous prediction regret at time t. Then

$$\mathbb{E}_{\theta\sim\mathcal{P}}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T)\right] = \sum_{t\in\mathcal{T}} \mathbb{E}_{\theta\sim\mathcal{P}}\left[R^t - \texttt{PReg}^t\right]. \tag{5.38}$$

Thus, it suffices to bound the differences in instantaneous regret.

Recall that at time t, the chosen action for BatchFreqGreedy and the predicted action are, resp.,

$$a_t = \arg\max_{a\in A} x_{a,t}^\top \theta_t^{\text{fre}}$$
$$a'_t = \arg\max_{a\in A} x_{a,t}^\top \theta_t^{\text{bay}}.$$

Letting $t_0 - 1 = \lfloor t/Y \rfloor$ be the last round in the previous batch, we can formulate θ_t^{fre} and θ_t^{bay} as

$$\theta_t^{\text{fre}} = (Z_{t_0-1})^{-1} X_{t_0-1}^\top \mathbf{r}_{1:t_0-1}$$
$$\theta_t^{\text{bay}} = (Z_{t_0-1} + \Sigma^{-1})^{-1}(X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} + \Sigma^{-1}\overline{\theta}).$$

Therefore, we have

$$\mathbb{E}_{\theta\sim\mathcal{P}\,|\,h_{t-1}}\left[R^t - \texttt{PReg}^t\right] = \mathbb{E}_{\theta\sim\mathcal{P}\,|\,h_{t-1}}\left[(x_{a'_t,t} - x_{a_t,t})^\top \theta_t^{\text{bay}}\right] = (x_{a'_t,t} - x_{a_t,t})^\top \theta_t^{\text{bay}},$$

since the mean of the posterior distribution is exactly θ_t^{bay}, and θ_t^{bay} is deterministic given h_{t-1}. Taking expectation over h_{t-1}, we have

$$\mathbb{E}_{\theta\sim\mathcal{P}}\left[R^t - \texttt{PReg}^t\right] = \mathbb{E}_{\theta\sim\mathcal{P}}\left[(x_{a_t',t} - x_{a_t,t})^\top \theta_t^{\text{bay}}\right].$$

For any fixed θ_t^{bay} and θ_t^{fre}, since BatchFreqGreedy chose a_t over a_t', it must be the case that

$$x_{a_t,t}^\top \theta_t^{\text{fre}} \geq x_{a_t',t}^\top \theta_t^{\text{fre}}. \tag{5.39}$$

Therefore,

$$\begin{aligned}
(x_{a_t',t} - x_{a_t,t})^\top \theta_t^{\text{bay}} &= (x_{a_t',t} - x_{a_t,t})^\top \theta_t^{\text{fre}} + (x_{a_t',t} - x_{a_t,t})^\top (\theta_t^{\text{bay}} - \theta_t^{\text{fre}}) \\
&\leq (x_{a_t',t} - x_{a_t,t})^\top (\theta_t^{\text{bay}} - \theta_t^{\text{fre}}) && \text{(By (5.39))} \\
&\leq (\|x_{a_t',t}\|_2 + \|x_{a_t,t}\|_2)\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2 \\
&\leq 2R\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2
\end{aligned}$$

(5.36) follows.

The crux is to prove (5.37): to bound the expected distance between the Frequentist and Bayesian estimates for θ. By expanding their definitions, we have

$$\begin{aligned}
&\theta_t^{\text{bay}} - \theta_t^{\text{fre}} \\
&= (Z_{t_0-1} + \Sigma^{-1})^{-1}(X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} + \Sigma^{-1}\overline{\theta}) - Z_{t_0-1}^{-1} X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} \\
&= (Z_{t_0-1} + \Sigma^{-1})^{-1}\left[X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} + \Sigma^{-1}\overline{\theta} - (Z_{t_0-1} + \Sigma^{-1}) Z_{t_0-1}^{-1} X_{t_0-1}^\top \mathbf{r}_{1:t_0-1}\right] \\
&= (Z_{t_0-1} + \Sigma^{-1})^{-1}\left[X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} + \Sigma^{-1}\overline{\theta} - X_{t_0-1}^\top \mathbf{r}_{1:t_0-1} - \Sigma^{-1} Z_{t_0-1}^{-1} X_{t_0-1}^\top \mathbf{r}_{1:t_0-1}\right] \\
&= (Z_{t_0-1} + \Sigma^{-1})^{-1}\left[\Sigma^{-1}\overline{\theta} - \Sigma^{-1} Z_{t_0-1}^{-1} X_{t_0-1}^\top \mathbf{r}_{1:t_0-1}\right] \\
&= (Z_{t_0-1} + \Sigma^{-1})^{-1}\Sigma^{-1}\left(\overline{\theta} - \theta_t^{\text{fre}}\right).
\end{aligned}$$

Next, note that

$$\begin{aligned}
&\|(Z_{t_0-1}+\Sigma^{-1})^{-1}\Sigma^{-1}(\overline{\theta}-\theta_t^{\text{fre}})\|_2 \\
&\qquad \le \|(Z_{t_0-1}+\Sigma^{-1})^{-1}\|_2 \, \|\Sigma^{-1}(\overline{\theta}-\theta_t^{\text{fre}})\|_2 \\
&\qquad \le \|(Z_{t_0-1}+\Sigma)^{-1}\|_2 \, \left(\|\Sigma^{-1}(\overline{\theta}-\theta)\|_2 + \|\Sigma^{-1}\|_2 \, \|\theta-\theta_t^{\text{fre}}\|_2\right).
\end{aligned}$$

By Lemma C.13, $\lambda_{\min}\left(Z_{t_0-1}+\Sigma\right) \ge \lambda_{\min}\left(Z_{t_0-1}\right)$. Therefore,

$$\|(Z_{t_0-1}+\Sigma)^{-1}\|_2 \le \frac{1}{\lambda_{\min}\left(Z_{t_0-1}\right)},$$

giving us

$$\begin{aligned}
\|\theta_t^{\text{bay}}-\theta_t^{\text{fre}}\|_2 &\le \frac{\|\Sigma^{-1}(\overline{\theta}-\theta)\|_2 + \|\Sigma^{-1}\|_2 \, \|\theta-\theta_t^{\text{fre}}\|_2}{\lambda_{\min}(Z_{t_0-1})} \\
&\le \frac{\|\Sigma^{-1/2}\|_2\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2 + \|\Sigma^{-1/2}\|_2 \, \|\theta-\theta_t^{\text{fre}}\|_2}{\lambda_{\min}(Z_{t_0-1})} \\
&= \frac{\left(\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2 + \sqrt{\lambda_{\min}(\Sigma)}\|\theta-\theta_t^{\text{fre}}\|_2\right)}{\sqrt{\lambda_{\min}(\Sigma)}\lambda_{\min}(Z_{t_0-1})}.
\end{aligned}$$

Next, recall that for

$$t_0 - 1 \ge t_{\min}(\delta) := 160\tfrac{R^2}{\rho^2}\log\tfrac{2d}{\delta}\cdot\log T$$

the following bounds hold, each with probability at least $1-\delta$:

$$\frac{1}{\lambda_{\min}\left(Z_{t_0-1}\right)} \le \frac{32\log T}{\rho^2(t_0-1)} \qquad \text{(Lemma 5.19)}$$

$$\|\theta-\theta_t^{\text{fre}}\|_2 \le \frac{\sqrt{2dR(t_0-1)\log(d/\delta)}}{\lambda_{\min}(Z_{t_0-1})} \qquad \text{(Lemma 5.21)}$$

Therefore, fixing $t_0 \ge 1 + t_{\min}(\delta/2)$, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\|\theta_t^{\text{bay}}-\theta_t^{\text{fre}}\|_2 \\
&\le \frac{32\log T}{\rho^2(t_0-1)\sqrt{\lambda_{\min}(\Sigma)}}\left(\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2 + \frac{64\sqrt{dR\log(2d/\delta)}\cdot\log T}{\rho^2\sqrt{t_0-1}}\right). \qquad (5.40)
\end{aligned}$$

Note that the high-probability events we need are deterministic given h_{t_0-1}, and therefore are independent of the perturbations at time t. This means that Lemma 5.18 applies, with $\ell = 0$: conditioned on any h_{t_0-1}, the expected regret for round t is upper-bounded by $2\|\theta\|_2(1+\rho(1+\sqrt{2\log K}))$. In particular, this holds for any h_{t_0-1} not satisfying the high probability events from Lemmas 5.19 and 5.21. Therefore, for all $t \geq t_{\min}(\delta)$,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2\right] \\
&\leq \mathbb{E}_{\theta\sim\mathcal{P}}\left[(1-\delta)\frac{32\log T}{\rho^2(t_0-1)\sqrt{\lambda_{\min}(\Sigma)}}\right. \\
&\cdot\left(\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2 + \frac{64\sqrt{dR\log(2d/\delta)}\cdot\log T}{\rho^2\sqrt{t_0-1}}\right) \\
&\qquad\left. + \delta\cdot 2\|\theta\|_2(1+\rho(2+\sqrt{2\log K}))\right] \\
&\leq \frac{32\log T}{\rho^2(t_0-1)\sqrt{\lambda_{\min}(\Sigma)}}\left(\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2\right] + \frac{64\sqrt{dR\log(2d/\delta)}\cdot\log T}{\rho^2\sqrt{t_0-1}}\right) \\
&\quad + \delta\cdot 2(\|\overline{\theta}\|_2 + \mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\overline{\theta}-\theta\|_2\right])(1+\rho(2+\sqrt{2\log K})).
\end{aligned}$$

Because $\theta \sim \mathcal{N}(\overline{\theta}, \Sigma)$, we have $\Sigma^{-1/2}(\overline{\theta}-\theta) \sim \mathcal{N}(0, I)$. By Lemma C.6,

$$\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\Sigma^{-1/2}(\overline{\theta}-\theta)\|_2\right] \leq \sqrt{d} \quad \text{and} \quad \mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\overline{\theta}-\theta\|_2\right] \leq \sqrt{d\lambda_{\max}(\Sigma)}.$$

This means

$$\begin{aligned}
\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2\right] &\leq \frac{32\sqrt{d}\log T}{\rho^2(t_0-1)\sqrt{\lambda_{\min}(\Sigma)}}\left(1 + \frac{64\sqrt{R\log(2d/\delta)}\cdot\log T}{\rho^2\sqrt{t_0-1}}\right) \\
&\quad + \delta\cdot 2(\|\overline{\theta}\|_2 + \sqrt{d\lambda_{\max}(\Sigma)})(1+\rho(2+\sqrt{2\log K})).
\end{aligned}$$

Since $t_0 = \Omega(t)$, for sufficiently small δ, this proves (5.37).

We need to do a careful computation to complete the proof of (5.35). We know from (5.36) that

$$\mathbb{E}_{\theta\sim\mathcal{P}}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T)\right] \leq \sum_{t=1}^{T} 2R\,\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2\right].$$

Choosing $\delta = T^{-2}$, we find that

$$\sum_{t=t_{\min}(T^{-2})}^{T} \delta \cdot 2(\|\overline{\theta}\|_2 + \sqrt{d\lambda_{\max}(\Sigma)})(1 + \rho(2 + \sqrt{2\log K})) = \tilde{O}(1),$$

so this term vanishes. Furthermore,

$$\sum_{t=t_{\min}(T^{-2})}^{T} 2R\frac{32\sqrt{d}\log T}{\rho^2(t_0-1)\sqrt{\lambda_{\min}(\Sigma)}}\left(1 + \frac{64\sqrt{R\log(2d/\delta)}\cdot\log T}{\rho^2\sqrt{t_0-1}}\right)$$
$$= \tilde{O}\left(\frac{R\sqrt{d}}{\rho^2\sqrt{\lambda_{\min}(\Sigma)}}\right)$$

since $t_0 \geq t - Y$, and $\sum_{t=1}^{T} 1/t = O(\log T)$. Using the fact that $R = \tilde{O}(1)$ (since by assumption $\rho \leq d^{-1/2}$), this is simply

$$\tilde{O}\left(\frac{\sqrt{d}}{\rho^2\sqrt{\lambda_{\min}(\Sigma)}}\right).$$

Finally, we note that on the first $t_{\min}(T^{-2}) = \tilde{O}(1/\rho^2)$ rounds, the regret bound from Lemma 5.18 with $\ell = 0$ applies, so the total regret difference is at most

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathcal{P}}\left[R^{\mathcal{T}}(T) - \texttt{PReg}^{\mathcal{T}}(T)\right] \\
&\leq \sum_{t=1}^{t_{\min}(T^{-2})} \mathbb{E}_{\theta\sim\mathcal{P}}\left[R^t - \texttt{PReg}^t\right] + \sum_{t=t_{\min}(T^{-2})}^{T} 2R\,\mathbb{E}_{\theta\sim\mathcal{P}}\left[\|\theta_t^{\text{bay}} - \theta_t^{\text{fre}}\|_2\right], \\
&\leq t_{\min}(T^{-2})\cdot 2(\|\overline{\theta}\|_2 + \sqrt{d\lambda_{\max}(\Sigma)})(1 + \rho(2 + \sqrt{2\log K})) + \tilde{O}\left(\frac{\sqrt{d}}{\rho^2\sqrt{\lambda_{\min}(\Sigma)}}\right) \\
&= \tilde{O}\left(\frac{\sqrt{d\lambda_{\max}(\Sigma)}}{\rho^2}\right) + \tilde{O}\left(\frac{\sqrt{d}}{\rho^2\sqrt{\lambda_{\min}(\Sigma)}}\right),
\end{aligned}$$

which implies (5.35).

Completing the proof of Theorem 5.29 given (5.35)**.** By Theorem 5.29, this holds whenever all perturbations are bounded by $\hat{R}$, which happens with probability at least $1 - \delta_R$. When the bound fail, the total regret is at most

$$2\left[\left(\|\overline{\theta}\|_2 + \sqrt{d\lambda_{\max}(\Sigma)}\right)\left(1 + \rho(2 + \sqrt{2\log K}) + \hat{R}\right)\right]$$

by Lemma 5.18 (with $\ell = \hat{R}$) and Lemma C.6. Since $\delta_R = T^{-2}$, the contribution of regret when the high-probability bound fails is $\tilde{O}(1/T) \leq \tilde{O}(1)$.

Part III

Models of Behavior

CHAPTER 6

OVERVIEW OF Part III

In Part III, we integrate models of behavior into our study of the impacts of algorithmic decision-making. We study the effects of behavioral biases and strategic behavior on theoretical models of decision-making.

In Chapter 7, we analyze a model of selection where a decision-maker discriminates against a subset of the population due to *implicit bias*. Under this model, we consider the effects of the *Rooney Rule*, the requirement that the decision-maker select at least one member from the disadvantaged population. We find that not only does the Rooney Rule increase diversity in selection, for certain parameterizations, it also improves the decision-maker's utility.

Chapter 8 considers the setting where a decision-maker wants to evaluate a decision subject, who can behave strategically in order to receive a more favorable evaluation. We study the types of decision rules that are robust to this strategic behavior. In particular, we assume the decision-maker has preferences over which actions decision subjects take, and we characterize evaluation rules that incentivize the desired actions. We find that linear mechanisms suffice in this setting, and we show that optimizing evaluation rules over data subject to the constraint that they incentivize desirable behaviors is computationally hard.

In Chapter 9, we take a broader view of the competitive effects that arise when multiple decision-makers must simultaneously decide whether or not to deploy an algorithm. We theoretically characterize the problem of *algorithmic monoculture*, where the use of a common algorithm by multiple decision-makers results in worse social welfare than a world where the algorithm does not exist,

even though decision-makers can rationally choose whether or not to deploy the algorithm. Our results suggest caution when standardizing decisions via algorithm, even when it may appear that an algorithm is more accurate than the human decision-makers it replaces.

CHAPTER 7

SELECTION PROBLEMS IN THE PRESENCE OF IMPLICIT BIAS

Over the past two decades, the notion of *implicit bias* (Greenwald and Banaji, 1995) has come to provide on important perspective on the nature of discrimination. Research on implicit bias argues that unconscious attitudes toward members of different demographic groups — for example, defined by gender, race, ethnicity, national origin, sexual orientation, and other characteristics — can have a non-trivial impact on the way in which we evaluate members of these groups; and this in turn may affect outcomes in employment (Bertrand and Mullainathan, 2004; Bohnet et al., 2016; Uhlmann and Cohen, 2005), education (van den Bergh et al., 2010), law (Greenwald and Krieger, 2006; Jolls and Sunstein, 2006), medicine (Green et al., 2007), and other societal institutions.

In the context of a process like hiring, implicit bias thus shifts the question of bias and discrimination to be not just about identifying bad actors who are intentionally discriminating, but also about the tendency of all of us to reach discriminatory conclusions based on the unconscious application of stereotypes. An understanding of these issues also helps inform the design of interventions to mitigate implicit bias — when essentially all of us have a latent tendency toward low-level discrimination, a set of broader practices may be needed to guide the process toward the desired outcome.

A basic mechanism: The Rooney Rule. One of the most basic and widely adopted mechanisms in practice for addressing implicit bias in hiring and selection is the *Rooney Rule* (Collins, 2007), which, roughly speaking, requires that in recruiting for a job opening, one of the candidates interviewed must come from an underrepresented group. The Rooney Rule is named for a protocol adopted

by the National Football League (NFL) in 2002 in response to widespread concern over the low representation of African-Americans in head coaching positions; it required that when a team is searching for a new head coach, at least one minority candidate must be interviewed for the position. Empirical analysis suggests that the Rooney Rule (and policies like it) have positively impacted diversity with little or no negative consequences for performance (DuBois, 2017; DuBois and Schanzenbach, 2017).

Subsequently the Rooney Rule has become a guideline adopted in many areas of business (Cavicchia, 2015); for example, in 2015 then-President Obama exhorted leading tech firms to use the Rooney Rule for hiring executives, and in recent years companies including Amazon, Facebook, Microsoft, and Pinterest have adopted a version of the Rooney Rule requiring that at least one candidate interviewed must be a woman or a member of an underrepresented minority group (Passariello, 2016). In 2017, a much-awaited set of recommendations made by Eric Holder and colleagues to address workplace bias at Uber advocated for the use of the Rooney Rule as one of its key points (Covington and Burling, 2017; Shaban, 2017).

The Rooney Rule is the subject of ongoing debate, and one crucial aspect of this debate is the following tension. On one side is the argument that implicit (or explicit) bias is preventing deserving candidates from underrepresented groups from being fairly considered, and the Rooney Rule is providing a force that counter-balances and partially offsets the consequences of this underlying bias. On the other side is the concern that if a job search process produces a short-list of top candidates all from the majority group, it may be because these are genuinely the strongest candidates despite the underlying bias — particularly

if there is a shortage of available candidates from other groups. In this case, wholesale use of the Rooney Rule may lead firms to consider weaker candidates from underrepresented groups, which works against the elimination of unconscious stereotypes. Of course, there are other reasons to seek diversity in recruiting that may involve broader considerations or longer time horizons than just the specific applicants being evaluated; but even these lines of argument generally incorporate the more local question of the effect on the set of applicants.

Given the widespread consideration of the Rooney Rule from both legal and empirical perspectives (Collins, 2007), it is striking that prior work has not attempted to formalize the inherently mathematical question that forms a crucial ingredient in these debates: given some estimates of the extent of bias and the prevalence of available minority candidates, does the expected quality of the candidates being interviewed by a hiring committee go up or down when the Rooney Rule is implemented? When the bias is large and there are many minority candidates, it is quite possible that a hiring committee's bias has caused it to choose a weaker candidate over a stronger minority one, and the Rooney Rule may be strengthening the pool of interviewees by reversing this decision and swapping the stronger minority candidate in. But when the bias is small or there are few minority candidates, the Rule might be reversing a decision that in fact chose the stronger applicant.

In this chapter, we propose a formalization of this family of questions, via a simplified model of selection with implicit bias, and we give a tight analysis of the consequences of using the Rooney Rule in this setting. In particular, when selecting for a fixed number of slots, we identify a sharp threshold on the effec-

tiveness of the Rooney Rule in our model that depends on three parameters: not just the extent of bias and the prevalence of available minority candidates, but a third quantity as well — essentially, a parameter governing the distribution of candidates' expected future job performance. We emphasize that our model is deliberately stylized, to abstract the trade-offs as cleanly as possible. Moreover, in interpreting these results, we emphasize a point noted above, that there are other reasons to consider using the Rooney Rule beyond the issues that motivate this particular formulation; but an understanding of the trade-offs in our model seems informative in any broader debate about such hiring and selection measures.

7.1 Overview and Summary of Results

We now describe the basic ingredients of our model, followed by a summary of the main results.

7.1.1 A Model of Selection with Implicit Bias

Our model is based on the following scenario. Suppose that a hiring committee is trying to fill an open job position, and it would like to choose the $k \geq 2$ best candidates as *finalists* to interview from among a large set of applicants. We will think of k as a small constant, and indeed most of the subtlety of the question already arises for the case $k = 2$, when just two finalists must be selected.

X-candidates and Y-candidates. The set of all applicants is partitioned into two groups X and Y, where we think of Y as the majority group, and X as a minority group within the domain that may be subject to bias. For some positive real number $\alpha \leq 1$ and a natural number n, there are n applicants from group Y and αn applicants from group X. If a candidate i belongs to X, we will refer to them as an *X-candidate*, and if i belongs to Y, we will refer to them as a *Y-candidate*. (The reader is welcome, for example, to think of the setting of academic hiring, with X as candidates from a group that is underrepresented in the field, but the formulation is general.)

Each candidate i has a (hidden) numerical value that we call their *potential*, representing their future performance over the course of their career. For example, in faculty hiring, we might think of the potential of each applicant in terms of some numerical proxy like their future lifetime citation count (with the caveat that any numerical measure will of course be an imperfect representation). Or in hiring executives, the potential of each applicant could be some measure of the revenue they will bring to the firm.

We assume that there is a common distribution Z that these numerical potentials come from: each potential is an independent draw from Z. (Thus, the applicants can have widely differing values for their numerical potentials; they just arise as draws from a common distribution.) For notational purposes, when i is an X-candidate, we write their potential as X_i, and when j is a Y-candidate, we write their potential as Y_j. We note an important modeling decision in this formulation: we are assuming that all X_i and all Y_j values come from this same distribution Z. While it is also of interest to consider the case in which the numerical potentials of the two groups X and Y are drawn from different group-

specific distributions, we focus on the case of identical distributions for two reasons. First, there are many settings where differences between the underlying distributions for different groups appear to be small compared to the bias-related effects we are seeking to measure; and second, in any formal analysis of bias between groups, the setting in which the groups begin with identical distributions is arguably the first fundamental special case that needs to be understood.

In the domains that we are considering — hiring executives, faculty members, athletes, performers — there is a natural functional form for the distribution Z of potentials, and this is the family of *power laws* (also known as *Pareto distributions*), with $\Pr[Z \geq t] = t^{-(1+\delta)}$ and support $[1, \infty)$ for a fixed $\delta > 0$. Extensive empirical work has argued that the distribution of individual output in a wide range of creative professions can be approximated by power law distributions with small positive values of δ (Clauset et al., 2009). For example, the distribution of lifetime citation counts is well-approximated by a power law, as are the lifetime downloads, views, or sales by performers, authors, and other artists. In the Section 7.3, we also consider the case in which the potentials are drawn from a distribution with bounded support, but for most of the chapter we will focus on power laws.

Selection with Bias. Given the set of applicants, the hiring committee would like to choose k *finalists* to interview. The *utility* achieved by the committee is the sum of the potentials of the k finalists it chooses; the committee's goal is to maximize its utility.[1]

[1]Since our goal is to model processes like the Rooney Rule, which apply to the selection of finalists for interviewing, rather than to the hiring decision itself, we treat the choice of k finalists as the endpoint rather than modeling the interviews that subsequently ensue.

If the committee could exactly evaluate the potential of each applicant, then it would have a straightforward way to maximize the utility of the set of finalists: simply sort all applicants by potential, and choose the top k as finalists. The key feature of the situation we would like to capture, however, is that the committee is biased in its evaluations; we look for a model that incorporates this bias as cleanly as possible.

Empirical work in some of our core motivating settings — such as the evaluation of scientists and faculty candidates — indicates that evaluation committees often systematically downweight female and minority candidates of a given level of achievement, both in head-to-head comparisons and in ranking using numerical scores (Wenneras and Wold, 1997). It is thus natural to model the hiring committee's evaluations as follows: they correctly estimate the potential of a Y-applicant j at the true value Y_j, but they estimate the potential of an X-applicant i at a reduced value $\tilde{X}_i < X_i$. They then rank candidates by these values $\{Y_j\}$ and $\{\tilde{X}_i\}$, and they choose the top k according to this biased ranking.

For most of the chapter, we focus on the case of *multiplicative bias*, in which $\tilde{X}_i = X_i/\beta$ for a bias parameter[2] $\beta > 1$. This is a reasonable approximation to empirical data from human-subject studies (Wenneras and Wold, 1997); and moreover, for power law distributions this multiplicative form is in a strong sense the "right" parametrization of the bias, since biases that grow either faster or slower than multiplicatively have a very simple asymptotic behavior in the power law case.

In this aspect of the model, as in others, we seek the cleanest formulation

[2]When $\beta = 1$, the ranking has no bias.

that exposes the key underlying issues; for example, it would be an interesting extension to consider versions in which the estimates for each individual are perturbed by random noise. A line of previous work (Braverman and Mossel, 2009; Feige et al., 1994; Fu and Lu, 2012) has analyzed models of ranking under noisy perturbations; while our scenario is quite different in that the entities being ranked are partitioned into a fixed set of groups with potentially different levels of bias and noise, it would be natural to see if these techniques could potentially be extended to handle noise in the context of implicit bias.

7.1.2 Main Questions and Results

This then is the basic model in which we analyze interventions with the structure of the Rooney Rule: (i) a set of n Y-applicants and αn X-applicants each have an independent future potential drawn from a power law distribution; (ii) a hiring committee ranks the applicants according to a sorted order in which each X-applicant's potential is divided down by $\beta > 1$, and chooses the top k in this ordering as *finalists*; and (iii) the hiring committee's *utility* is the sum of the potentials of the k finalists.

Qualitatively, the motivation for the Rooney Rule in such settings is that hiring committees are either unwilling or unable to reasonably correct for their bias in performing such rankings, and therefore cannot be relied on to interview X-candidates on their own. The difficulty in removing this skew from such evaluations is a signature aspect of phenomena around implicit bias.

The decision to impose the Rooney Rule is made at the outset, before the actual values of the potentials $\{Y_j\}$ and $\{\tilde{X}_i\}$ are materialized. All that is known

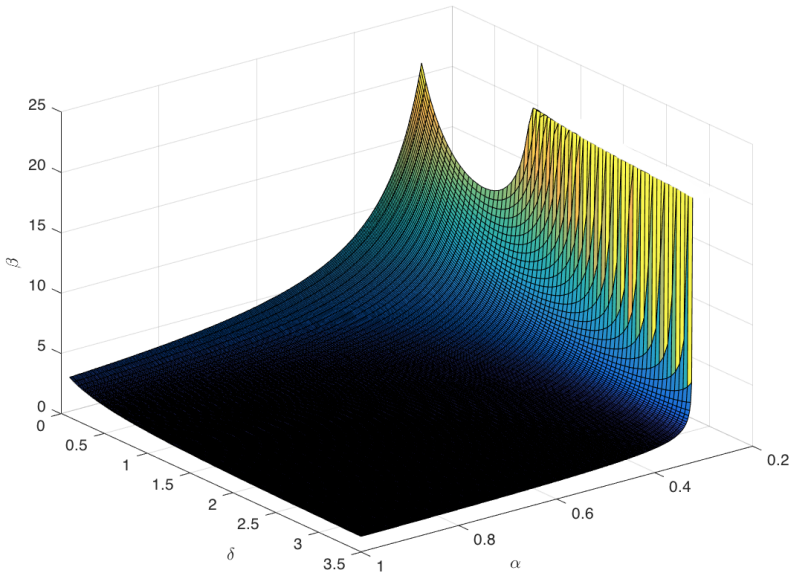


Figure 7.1: Fixing $k = 2$, the (α, β, δ) values for which the Rooney Rule produces a positive expected change for sufficiently large n lie above a surface (depicted in the figure) defined by the function $\phi_2(\alpha, \beta, \delta) = 1$.

at the point of this initial decision to use the Rule or not are the parameters of the domain: the bias β, the relative abundance of X-candidates α, the power law exponent $1 + \delta$, and the number of finalists to be chosen k. The question is: as a function of these parameters, will the use of the Rooney Rule produce a positive or negative expected change in utility, where the expectation is taken over the random draws of applicant values? We note that one could instead ask about the probability that the Rooney Rule produces a positive change in utility as opposed to the expected change; in fact, our techniques naturally extend to characterize not only the expected change, but the probability that this change is positive, as we will show in Section 7.2.

Our model lets us make precise the trade-off in utility that underpins the use of the Rooney Rule. If the committee selects an X-candidate on its own —

even using its biased ranking — then their choice already satisfies the conditions of the Rule. But if all k finalists are Y-candidates, then the Rooney Rule requires that the committee replace the lowest-ranked of these finalists j with the highest-ranked X-candidate i. Because i was not already a finalist, we know that $\tilde{X}_i = X_i/\beta < Y_j$. But to see whether this yields a positive change in utility, we need to understand which of X_i or Y_j has a larger expected value, conditional on the information contained in the committee's decision, that $X_i/\beta < Y_j$.

Our main result is an exact characterization of when the Rooney Rule produces a positive expected change in terms of the four underlying parameters, showing that it non-trivially depends on all four. For the following theorem, and for the remainder of the chapter, we assume $0 < \alpha \leq 1$, $\beta > 1$, and $\delta > 0$. We begin with the case where $k = 2$.

Theorem 7.1. *For $k = 2$ and sufficiently large n, the Rooney Rule produces a positive expected change if and only if $\phi_2(\alpha, \beta, \delta) > 1$ where*

$$\phi_2(\alpha, \beta, \delta) = \frac{\alpha^{1/(1+\delta)} \left[1 - (1 + c^{-1})^{-\delta/(1+\delta)} \left[1 + \frac{\delta}{1+\delta}(1 + c)^{-1}\right]\right]}{\frac{\delta}{1+\delta}(1 + c)^{-1-\delta/(1+\delta)}} \tag{7.1}$$

and $c = \alpha\beta^{-(1+\delta)}$. Moreover, $\phi_2(\alpha, \beta, \delta)$ is increasing in β, so for fixed α and δ there exists β^* such that $\phi_2(\alpha, \beta, \delta) > 1$ if and only if $\beta > \beta^*$.

Thus, we have an explicit characterization for when the Rooney Rule produces positive expected change. The following theorem extends this to larger values of k.

Theorem 7.2. *There is an explicit function $\phi_k(\alpha, \beta, \delta)$ such that the Rooney Rule produces a positive expected change, for n sufficiently large and $k = O(\ln n)$, if and only if $\phi_k(\alpha, \beta, \delta) > 1$.*

Interestingly, even for larger values of k, there are parts of the parameter space for which the Rooney Rule produces a positive expected change and parts for which the Rooney Rule produces a negative expected change, independent of the number of applicants n.

Figure 7.1 depicts a view of the function ϕ_2, by showing the points in three-dimensional (α, β, δ) space for which ϕ takes the value 1. The values for which the Rooney Rule produces a positive expected change for sufficiently large n lie above this surface.

The surface in Figure 7.1 is fairly complex, and it displays unexpected non-monotonic behavior. For example, on certain regions of fixed (α, β), it is non-monotonic in δ, a fact which is not a priori obvious: there are choices of α and β for which the Rooney Rule produces a positive expected change at certain "intermediate" values of δ, but not at values of δ that are sufficiently smaller or sufficiently larger. Moreover, there exist (α, δ) pairs above which the surface does not exist. (One example in Figure 7.1 occurs at $\alpha \approx 0.3$ and $\delta \approx 3$). Characterizing the function ϕ and its level set $\phi = 1$ is challenging, and it is noteworthy that the complexity of this function is arising from our relatively bare-bones formulation of the trade-off in the Rooney Rule; this suggests the function and its properties are capturing something inherent in the process of biased selection.

One monotonocity result we are able to establish for the function ϕ is the following, showing that for fixed (α, β, δ), increasing the number of positions can't make the Rooney Rule go from beneficial to harmful.

Theorem 7.3. *For sufficiently large n and $k = O(\ln n)$, if the Rooney Rule produces a positive expected change at a given number of finalists k, it also produces a positive expected change when there are $k + 1$ finalists (at the same (α, β, δ)).*

We prove these theorems through an analysis of the *order statistics* of the underlying power law distribution. Specifically, if we draw m samples from the power law Z and sort them in ascending order from lowest to highest, then the ℓ^{th} item in the sorted list is a random variable denoted $Z_{(\ell:m)}$. To analyze the effect of the Rooney Rule, we are comparing $Y_{(n-k+1:n)}$ with $X_{(\alpha n:\alpha n)}$. Crucially, we are concerned with their expected values conditional on the fact that the committee chose the k^{th}-ranked Y-candidate over the top-ranked X-candidate, implying as noted above that $X_{(\alpha n:\alpha n)}/\beta < Y_{(n-k+1:n)}$. The crucial comparison is therefore between $\mathbb{E}\left[Y_{(n-k+1:n)}|X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]$ and $\mathbb{E}\left[X_{(\alpha n:\alpha n)}|X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]$. Order statistics conditional on this type of side information turn out to behave in complex ways, and hence the core of the analysis is in dealing with these types of conditional order statistics for power law distributions.

More generally, given the ubiquity of power law distributions (Clauset et al., 2009), we find it surprising how little is known about how their order statistics behave qualitatively. In this respect, the techniques we provide may prove to be independently useful in other applications. For example, we develop a tight asymptotic characterization of the expectations of order statistics from a power law distribution that to our knowledge is novel.

We also note that although our results are expressed for sufficiently large n, the convergence to the asymptotic behavior happens very quickly as n grows; to handle fixed values of n, we need only modify the bounds by correction terms that grow like $\left(1 \pm O\left(\frac{(\ln n)^2}{n}\right)\right)$. In particular, the errors in the asymptotic analysis are small once n reaches 50, which is reasonable for settings in which a job opening receives many applications.

Estimating the level of bias β. The analysis techniques we develop for proving Theorem 7.2 can also be used for related problems in this model. A specific question we are able to address is the problem of estimating the amount of bias from a history of hiring decisions.

In particular, suppose that over m years the hiring committee makes one offer per year; in N of the m years this offer goes to an X-candidate, and in $m - N$ of the m years this offer goes to a Y-candidate. Which value of the bias parameter β maximizes the probability of this sequence of observations?

We provide a tight characterization of the solution to this question, finding again that it depends not only on α (in this case, the sequence of α values for each year), but also on the power law exponent $1 + \delta$. The solution has a qualitatively natural structure, and produces $\beta = 1$ (corresponding to no bias) as the estimate when the fraction of X-candidates hired over the m years is equal to the expected number that would be hired under random selection.

Generalizations to other distributions. Finally, in Section 7.3 we consider how to adapt our approach for classes of distributions other than power laws. A different category of distributions that can be motivated by the considerations discussed here is the set of bounded distributions, which take values only over a finite interval. Just as power laws are characteristic of the performance of employees in certain professions, bounded distributions are appropriate when there are absolute constraints on the maximum effect a single employee can have.

Moreover, bounded distributions are also of interest because they contain the uniform distribution on $[0, 1]$ as a special case. We can think of this special

case as describing an instance in which each candidate is associated with their *quantile* (between 0 and 1) in a ranking of the entire population, and the bias then operates on this quantile value, reducing it in the case of X-candidates.

For bounded distributions, we can handle much more general forms for the bias — essentially, any function that reduces the values X_i strictly below the maximum of the distribution (for instance, a bias that always prefers a Y-candidate to an X-candidate when they are within some ε of each other). When $k = 2$ and there are equal numbers of X-candidates and Y-candidates, we show that for any bounded distribution and any such bias, the Rooney Rule produces a positive expected change in utility for all sufficiently large n.

7.1.3 An Illustrative Special Case: Infinite Bias

To illustrate some of the basic considerations that go into our analysis and its interpretation, we begin with a simple special case that we can think of as "infinite bias" — the committee deterministically ranks every Y-candidate above every X-candidate. This case already exhibits structurally rich behavior, although the complexity is enormously less than the case of general β. We also focus here on $k = 2$. In terms of Figure 7.1, we can visualize the infinite bias case as if we are looking down at the plot from infinitely high up; thus, reasoning about infinite bias amounts to determining which parts of the (α, δ) plane are covered by the surface $\phi_2(\alpha, \beta, \delta) = 1$.

With infinite bias, the committee is guaranteed to choose the two highest-ranked Y-candidates in the absence of an intervention; with the Rooney Rule, the committee will choose the highest-ranked Y-candidate and the highest-

ranked X-candidate. As we discuss in the next section, for power law distributions with exponent $1 + \delta$, if z^* is the expected maximum of n draws from the distribution, then (i) the expected value of the second-largest of the n draws is $\frac{\delta}{(1+\delta)}z^*$; and (ii) the expected maximum of αn draws from the distribution is asymptotically $\alpha^{1/(1+\delta)}z^*$.

This lets us directly evaluate the utility consequences of the intervention. If there is no intervention, the utility of the committee's decision will be $\left(1 + \frac{\delta}{1+\delta}\right) z^*$, and if the Rooney Rule is used, the utility of the committee's decision will be $(1 + \alpha^{1/(1+\delta)})z^*$. Thus, the Rooney Rule produces positive expected change in utility if and only if $\alpha^{1/(1+\delta)} > \frac{\delta}{(1+\delta)}$; that is, if and only if $\alpha > \left(\frac{\delta}{1+\delta}\right)^{1+\delta}$.

In addition to providing a simple closed-form expression for when to use the Rooney Rule in this setting, the condition itself leads to some counter-intuitive consequences. In particular, the closed-form expression for the condition makes it clear that *for every* $\alpha > 0$, there exists a sufficiently small $\delta > 0$ so that when the distribution of applicant potentials is a power law with exponent $1 + \delta$, using the Rooney Rule produces the higher expected utility. In other words, with a power law exponent close to 1, it's a better strategy to commit one of the two offers to the X-candidates, even though they form an extremely small fraction of the population.

This appears to come perilously close to contradicting the following argument. We can arbitrarily divide the Y-candidates into two sets A and B of $n/2$ each; and if $\alpha < 1/2$, each of A and B is larger than the set of all X-candidates. Let a^* be the top candidate in A and b^* be the top candidate in B. Each of a^* and b^* has at least the expected value of the top X-candidate, and moreover, one of them is the top Y-candidate overall. So how can it be that choosing a^* and b^*

fails to improve on the result of using the Rooney Rule?

The resolution is to notice that using the Rooney Rule still involves hiring the *top* Y-candidate. So it's not that the Rooney Rule chooses one of a^* or b^* at random, together with the top X-candidate. Rather, it chooses the *better* of a^* and b^*, along with the top X-candidate. The real point is that power law distributions have so much probability in the tail of the distribution that the best person among a set of αn can easily have a higher expected value than the second-best person among a set of n, even when α is quite small. This is a key property of power law distributions that helps explain what's happening both in this example and in our analysis.

7.1.4 A Non-Monotonicity Effect

As noted above, much of the complexity in the analysis arises from working with expected values of random variables conditioned on the outcomes of certain biased comparisons. One might hope that expected values conditional on these types of comparisons had tractable properties that facilitated the analysis, but this is not the case; in fact, these conditional expectations exhibit some complicated and fairly counter-intuitive behavior. To familiarize the reader with some of these phenomena — both as preparation for the subsequent sections, but also as an interesting end in itself — we offer the following example.

Much of our analysis involves quantities like $\mathbb{E}[X|X > \beta Y]$ — the conditional expectation of X, given that it exceeds some other random variable Y multiplied by a bias parameter. (We will also be analyzing the version in which the inequality goes in the other direction, but we'll focus on the current expres-

sion for now.) If we choose X and Y as independent random variables both drawn from a distribution Z, and then view the conditional expectation as a function just of the bias parameter β, what can we say about the properties of this function $f(\beta) = \mathbb{E}[X|X > \beta Y]$?

Intuitively we'd expect $f(\beta)$ to be monotonically increasing in β — indeed, as β increases, we're putting a stricter lower bound on X, and so this ought to raise the conditional expectation of X.

The surprise is that this is not true in general; we can construct independent random variables X and Y for which $f(\beta)$ is not monotonically increasing. In fact, the random variables are very simple: we can have each of X and Y take values independently and uniformly from the finite set $\{1, 5, 9, 13\}$. Now, the event $X > 2Y$ consists of four possible pairs of (X, Y) values: (5,1), (9,1), (13,1), and (13,5). Thus, $f(2) = \mathbb{E}[X|X > 2Y] = 10$. In contrast, the event $X > 3Y$ consists of three possible pairs of (X, Y) values: (5,1), (9,1), and (13,1). Thus, $f(3) = 9$, which is a smaller value, despite the fact that X is required to be a larger multiple of Y.

The surprising content of this example has a fairly sharp formulation in terms of a story about recruiting. Suppose that two academic departments, Department A and Department B, both engage in hiring each year. In our stylized setting, each interviews one X-candidate and one Y-candidate each year, and hires one of them. Each candidate comes from the uniform distribution on $\{1, 5, 9, 13\}$. Departments A and B are both biased in their hiring: A only hires the X-candidate in a given year if they're more than twice as good as the Y-candidate, while B only hires the X-candidate in a given year if they're more than three times as good as the Y-candidate.

Clearly this bias hurts the average quality of both departments, B more so than A. But you might intuitively expect that at least if you looked at the X-candidates that B has actually hired, they'd be of higher average quality than the X-candidates that A has hired — simply because they had to pass through a stronger filter to get hired. In fact, however, this isn't the case: despite the fact that B imposes a stronger filter, the calculations performed above for this example show that the average quality of the X-candidates B hires is 9, while the average quality of the X-candidates A hires is 10.

This non-monotonicity property shows that the conditional expectations we work with in the analysis can be pathologically behaved for arbitrary (even relatively simple) distributions. However, we will see that with power law distributions we are able — with some work — to avoid these difficulties; and part of our analysis will include a set of explicit monotonicity results.

7.2 Biased Selection with Power Law Distributions

Recall that for a random variable Z, we use $Z_{(\ell:m)}$ to denote the ℓ^{th} order statistic in m draws from Z: the value in position ℓ when we sort m independent draws from Z from lowest to highest. Recall also that when selecting k finalists, the Rooney Rule improves expected utility exactly when

$$\mathbb{E}\left[X_{(\alpha n:\alpha n)} - Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] > 0.$$

Using linearity of expectation and the fact that $\Pr[A|B]\Pr[B] = \Pr\left[A \cdot \mathbf{1}_{\{B\}}\right]$, this is equivalent to

$$\frac{\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]}{\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]} > 1. \tag{7.2}$$

We will show an asymptotically tight characterization of the tuples of parameters $(k, \alpha, \beta, \delta)$ for which this condition holds, up to an error term on the order of $O\left(\frac{(\ln n)^2}{n}\right)$. In order to better understand the terms in (7.2), we begin with some necessary background.

7.2.1 Preliminaries

Fact 7.4. *Let $f_{(p:m)}$ and $F_{(p:m)}$ be, respectively, the probability density function and cumulative distribution function of the p^{th} order statistic out of m draws from the power law distribution with parameter δ. Using definitions from David and Nagaraja (2005),*

$$f_{(p:m)}(x) = (1+\delta)(m-p+1)\binom{m}{p-1}\left(1-x^{-(1+\delta)}\right)^{p-1}\left(x^{-(1+\delta)}\right)^{m-p+1}x^{-1}$$

and

$$F_{(p:m)}(x) = \sum_{j=p}^{m}\binom{m}{j}\left(1-x^{-(1+\delta)}\right)^{j}\left(x^{-(1+\delta)}\right)^{m-j}.$$

Definition 7.5. *We will make extensive use of the Gamma function:*

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\,dt.$$

$\Gamma(\cdot)$ is considered the continuous relaxation of the factorial, and it satisfies

$$\Gamma(a+1) = a\Gamma(a).$$

If a is a positive integer, $\Gamma(a+1) = a!$. Furthermore, $\Gamma(a) > 1$ for $0 < a < 1$ and $\Gamma(a) < 1$ for $1 < a < 2$.

7.2.2 The Case where $k = 2$

For simplicity, we begin with the case where we're selecting $k = 2$ finalists. In this section, we will make several approximations, growing tight with large n,

that we treat formally in Appendices D.1 and D.2. This section is intended to demonstrate the techniques needed to understand the condition (7.2). In the case where $k = 2$, the Rooney Rule increases expected utility if and only if

$$\frac{\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-1:n)}\}}\right]}{\mathbb{E}\left[Y_{(n-1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-1:n)}\}}\right]} > 1. \tag{7.3}$$

Theorems D.1 and D.2 in Appendix D.2 give tight approximations to these quantities; here, we provide an outline for how to find them. For the sake of exposition, we'll only show this for the denominator in this section, which is slightly simpler to approximate. We begin with

$$\mathbb{E}\left[Y_{(n-1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-1:n)}\}}\right] = \int_1^\infty y f_{(n-1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy.$$

Letting $c = \alpha\beta^{-(1+\delta)}$, we can use Lemma D.14 and some manipulation to approximate this by

$$(1+\delta)n(n-1)\int_1^\infty \left(1 - y^{-(1+\delta)}\right)^{n(1+c)-2} \left(y^{-(1+\delta)}\right)^2 dy.$$

Conveniently, the function being integrated is (up to a constant factor) $y \cdot f_{(n(1+c)-1:n(1+c))}(y)$, i.e. y times the probability density function of the second-highest order statistic from $n(1+c)$ samples. Since

$$\begin{aligned}
&\mathbb{E}\left[Z_{(n(1+c)-1:n(1+c))}\right] \\
&= \int_1^\infty z f_{(n(1+c)-1:n(1+c))}(z)\, dz \\
&= (1+\delta)n(1+c)(n(1+c)-1)\int_1^\infty \left(1 - z^{-(1+\delta)}\right)^{n(1+c)-2} \left(z^{-(1+\delta)}\right)^2 dz,
\end{aligned}$$

we have

$$\mathbb{E}\left[Y_{(n-1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-1:n)}\}}\right] \approx \frac{1}{(1+c)^2}\mathbb{E}\left[Z_{(n(1+c)-1:n(1+c))}\right].$$

Then, we can use Lemmas D.22 and D.23 to get $\mathbb{E}\left[Z_{(n(1+c)-1:n(1+c))}\right] \approx (1+c)^{1/(1+\delta)}\mathbb{E}\left[Y_{(n-1:n)}\right]$, meaning that

$$\mathbb{E}\left[Y_{(n-1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-1:n)}\}}\right] \approx (1+c)^{-(1+\delta/(1+\delta))}\mathbb{E}\left[Y_{(n-1:n)}\right]. \tag{7.4}$$

For the numerator of (7.3), a slightly more involved calculation yields

$$\begin{aligned}&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)}<\beta Y_{(n-1:n)}\}}\right] \\ &\approx \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right]\left[1-(1+c^{-1})^{-\delta/(1+\delta)}\left[1+\frac{\delta}{1+\delta}(1+c)^{-1}\right]\right]. \end{aligned} \tag{7.5}$$

By Lemmas D.22 and D.23, $\mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] \approx \Gamma\left(\frac{\delta}{1+\delta}\right)(\alpha n)^{1/(1+\delta)}$ and $\mathbb{E}\left[Y_{(n-1:n)}\right] \approx \Gamma\left(1+\frac{\delta}{1+\delta}\right) n^{1/(1+\delta)}$. Recall that, up to the approximations we made, the Rooney Rule improves utility in expectation if and only if the ratio between (7.5) and (7.4) is larger than 1. Therefore, the following theorem holds:

Theorem 7.6. *For sufficiently large n, the Rooney Rule with $k = 2$ improves utility in expectation if and only if*

$$\frac{\alpha^{1/(1+\delta)}\left[1-(1+c^{-1})^{-\delta/(1+\delta)}\left[1+\frac{\delta}{1+\delta}(1+c)^{-1}\right]\right]}{\frac{\delta}{1+\delta}(1+c)^{-1-\delta/(1+\delta)}} > 1. \tag{7.6}$$

where $c = \alpha\beta^{-(1+\delta)}$.

Note that in the limit as $\beta \to \infty$, $c \to 0$, and the entire expression goes to $\alpha^{1/(1+\delta)}(1+\delta)/\delta$, as noted in Section 7.1.3. Although the full expression in the statement of Theorem 7.6 is fairly complex, it can be directly evaluated, giving a tight characterization of when the Rule yields increased utility in expectation.

With this result, we could ask for a fixed α and δ how to characterize the set of β such that the condition in (7.6) holds. In fact, we can show that this expression is monotonically increasing in β.

Theorem 7.7. *The left hand side of* (7.6) *is decreasing in c and therefore increasing in β. Hence for fixed α and δ there exists β^* such that* (7.6) *holds if and only if $\beta > \beta^*$.*

Non-monotonicity in δ. From Theorem 7.6, we can gain some intuition for the non-monotonicity in δ shown in Figure 7.1. For $\alpha < e^{-1}$, we can show

that even with infinite bias, the Rooney Rule has a negative effect on utility for sufficiently large δ. Intuitively, this is because the condition for positive change with infinite bias is $\alpha > \left(\frac{\delta}{1+\delta}\right)^{1+\delta}$, which can be written as $\alpha > \left(1 - \frac{1}{d}\right)^d$ for $d = 1 + \delta$. Since this converges to e^{-1} from below, for sufficiently large δ and $\alpha < e^{-1}$, we have $\alpha < \left(\frac{\delta}{1+\delta}\right)^{1+\delta}$. On the other hand, as $\delta \to 0$, the Rooney Rule has a more negative effect on utility. For instance, $\phi_2(.3, 10, 1) > 1$ but $\phi_2(.3, 10, .5) < 1$. Intuitively, this non-monotonicity arises from the fact that for large δ and small α, the Rooney Rule always has a negative impact on utility, while for very small δ, samples are very far from each other, meaning that the bias has less effect on the ranking.

7.2.3 The General Case

We can extend these techniques to handle larger values of k. For $k \in [n]$, we define

$$\begin{aligned} r_k(\alpha, \beta, \delta) &= \frac{\mathbb{E}\left[X_{(\alpha n:\alpha n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]}{\mathbb{E}\left[Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]} \\ &= \frac{\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]}{\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]}. \end{aligned}$$

We can see that the Rooney Rule improves expected utility when selecting k candidates if and only if $r_k > 1$. While r_k depends on n, we will show that it is a very weak dependence: for small k, as n increases, r_k converges to a function of $(\alpha, \beta, \delta, k)$ up to a $1 + O((\ln n)^2/n)$ multiplicative factor. To make this precise, we define the following notion of asymptotic equivalence:

Definition 7.8. *For nonnegative functions $f(n)$ and $g(n)$, define*

$$f(n) \approxeq g(n)$$

if and only if there exist $a > 0$ and $n_0 > 0$ such that

$$\frac{f(n)}{g(n)} \leq 1 + \frac{a(\ln n)^2}{n} \quad \textit{and} \quad \frac{g(n)}{f(n)} \leq 1 + \frac{a(\ln n)^2}{n}$$

for all $n \geq n_0$. In other words, $f(n) = g(n)\left(1 \pm O\left(\frac{(\ln n)^2}{n}\right)\right)$. When being explicit about a and n_0, we'll write $f(n) \approxeq_{a;n_0} g(n)$.

Appendix D.3 contains a series of lemmas establishing how to rigorously manipulate equivalences of this form. Now, we formally define a tight approximation to r_k, which serves as an expanded restatement of Theorem 7.2 from the introduction.

Theorem 7.9. *For $k \in [n]$, define*

$$\begin{aligned}&\phi_k(\alpha, \beta, \delta)\\ &= \frac{\alpha^{1/(1+\delta)} c^{\delta/(1+\delta)} (1+c)^{k-1}}{\binom{k-1-\frac{1}{1+\delta}}{k-1}} \left[(1+c^{-1})^{\delta/(1+\delta)} - \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right] \end{aligned} \tag{7.7}$$

where $c = \alpha\beta^{-(1+\delta)}$. Note that ϕ_k does not depend on n. When (α, β, δ) are fixed, we will simply write this as ϕ_k. For $k \leq ((1-c^2)\ln n)/2$, we have

$$r_k \approxeq \phi_k,$$

and therefore the Rooney Rule improves expected utility for sufficiently large n if and only if $\phi_k > 1$.

This condition tightly characterizes when the Rooney Rule improves expected utility, and its asymptotic nature in n becomes accurate even for moderately small n: for example, when $n = 50$, the error between r_k and ϕ_k is around 1% for reasonable choices of (α, β, δ).

Increasing k. Consider the scenario in which we're selecting k candidates, and for the given parameter values, the Rooney Rule improves our expected utility. If we were to instead select $k+1$ candidates, should we still be reserving a spot for an X-candidate? Intuitively, as k increases, the Rule is less likely to change our selections, since we're more likely to have already chosen an X-candidate; however, it is not a priori obvious whether increasing k should make it better for us to use the Rooney Rule (because we have more slots, so we're losing less by reserving one) or worse (because as we take more candidates, we stop needing a reserved slot).

In fact, we can apply Theorem 7.9 to understand how r_k changes with k. The following theorem, proven in Appendix D.2, is an expanded restatement of Theorem 7.3, showing that if the Rooney Rule yields an improvement in expected quality when selecting k candidates, it will do so when selecting $k+1$ candidates as well.

Theorem 7.10. *For $k \leq ((1-c^2)\ln n)/2$, we have $\phi_{k+1} > \phi_k$, and therefore for sufficiently large n, we have $r_{k+1} > r_k$.*

Finally, using these techniques, we can provide a tight characterization of the probability that the Rooney Rule produces a positive change. Specifically, we find the probability that the Rooney Rule has a positive effect conditioned on the event that it changes the outcome.

Theorem 7.11.

$$\Pr\left[X_{(\alpha n:\alpha n)} > Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] \approxeq 1 - \left(\frac{1+\alpha\beta^{-(1+\delta)}}{1+\alpha}\right)^k.$$

To determine whether the Rooney Rule is more likely than not to produce

a positive effect (conditioned on changing the outcome), we can compare the right-hand side to $1/2$.

Note that in the case of infinite bias, the right-hand side becomes $1-(1+\alpha)^{-k}$, and thus, the Rooney Rule produces positive change with probability at least $1/2$ if and only if $\alpha \geq \sqrt[k]{2} - 1$. It is interesting to observe that this means with infinite bias, the condition is independent of δ; in contrast, when considering the effect on the expected value with infinite bias, as we did in Section 7.1.3, the expected change in utility due to the Rooney Rule did depend on δ.

7.2.4 Maximum Likelihood Estimation of β

The techniques established thus far make it possible to answer other related questions, including the following type of question that we consider in this section: "Given some historical data on past selections, can we estimate the bias present in the data?" For example, suppose that for the last m years, a firm has selected one candidate for each year i out of a pool of $\alpha_i n_i$ X-candidates and n_i Y-candidates. If all applicants are assumed to come from the same underlying distribution, then it is easy to see that the expected number of X-selections (in the absence of bias) should be

$$\sum_{i=1}^{m} \frac{\alpha_i}{1+\alpha_i},$$

regardless of what distribution the applicants come from. However, if there is bias in the selection procedure, then this quantity now depends on the bias model and parameters of the distribution. In particular, in our model, we can use Theorem D.3 to get

$$\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n:n)}\right] \approxeq \frac{1}{1+\alpha\beta^{-(1+\delta)}}.$$

This gives us the following approximation for the likelihood of the data $D = (M_1, \ldots, M_m)$ given β, where M_i is 1 if an X-candidate was selected in year i and 0 otherwise:

$$\prod_{i=1}^{m}(1 - M_i) \cdot \frac{1}{1 + \alpha_i\beta^{-(1+\delta)}} + M_i \cdot \frac{\alpha_i\beta^{-(1+\delta)}}{1 + \alpha_i\beta^{-(1+\delta)}}.$$

Taking logarithms, this is

$$\sum_{i:M_i=1} \log(\alpha_i\beta^{-(1+\delta)}) - \sum_{i=1}^{m} \log(1 + \alpha_i\beta^{-(1+\delta)}),$$

and maximizing this is equivalent to maximizing

$$\sum_{i:M_i=1} \log(\beta^{-(1+\delta)}) - \sum_{i=1}^{m} \log(1 + \alpha_i\beta^{-(1+\delta)}) = N \log(\beta^{-(1+\delta)}) - \sum_{i=1}^{m} \log(1 + \alpha_i\beta^{-(1+\delta)})$$

where N is the number of X-candidates selected. Taking the derivative with respect to β, we get

$$-(1 + \delta)N\beta^{-1} + (1 + \delta) \sum_{i=1}^{m} \frac{\alpha_i\beta^{-(2+\delta)}}{1 + \alpha_i\beta^{-(1+\delta)}}.$$

Setting this equal to 0 and canceling common terms, we have

$$\sum_{i=1}^{m} \frac{1}{1 + \alpha_i^{-1}\beta^{1+\delta}} = N$$

Since each $1/(1 + \alpha_i^{-1}\beta^{1+\delta})$ is strictly monotonically decreasing in β, there is a unique $\hat{\beta}$ for which equality holds, meaning that the likelihood is uniquely maximized by $\hat{\beta}$, up to the $1 \pm O((\ln n)^2/n)$ approximation we made for $\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n:n)}\right]$. In the special case where $\alpha_i = \alpha$ for $i = 1, \ldots, m$, then the solution is given by

$$\hat{\beta} = \left(\left(\frac{m}{N} - 1\right)\alpha\right)^{1/(1+\delta)}.$$

7.3 Biased Selection with Bounded Distributions

In this section, we consider a model in which applicants come from a distribution with bounded support. Qualitatively, one would expect different results here from those with power law distributions because in a model with bounded distributions, we expect that for large n, the top order statistics of any distribution will concentrate around the maximum of that distribution. As a result, when there is even a small amount of bias against one population, for large n the probability that *any* of the samples with the highest perceived quality come from that population goes to 0. This means that the Rooney Rule has an effect with high probability, and the effect is positive if the unconditional expectation of the top X-candidate is larger than the unconditional expectation of the Y-candidate that it replaces.

We focus on the case when $\alpha = 1$, meaning we have equal numbers of applicants from both populations. We use the same order statistic notation as before. While all of our previous results have modeled the bias as a multiplicative factor β, we can in fact show that in the bounded distribution setting, for any model of bias $\tilde{X}_{(k:n)} = b(X_{(k:n)})$ such that $b(x) < T$ for $x \geq 0$, where T is strictly less than the maximum of the distribution, the Rooney Rule increases expected utility. Unlike in the previous section the following theorem and analysis are by no means a tight characterization; instead, this is an existence proof that for bounded distributions, there is always a large enough n such that the Rooney Rule improves utility in expectation. We prove our results for continuous distributions with support $[0, 1]$, but a simple scaling argument shows that this extends to any continuous distribution with bounded nonnegative support – specifically, we scale a distribution such that $\inf_{x:f(x)>0} = 0$ and $\sup_{x:f(x)>0} = 1$.

Theorem 7.12. *If f is a continuous probability density function on $[0,1]$ such that $\sup_{x:f(x)>0} = 1$ and $\tilde{X}_{(n:n)} = b(X_{(n:n)})$ is never more than $T < 1$, then for large enough n,*

$$\mathbb{E}\left[X_{(n:n)} - Y_{(n-1:n)} \mid b(X_{(n:n)}) < Y_{(n-1:n)}\right] > 0.$$

While we the defer the full proof to Appendix D.5, the strategy for the proof is as follows:

1. With high probability, $X_{(n:n)}$ and $Y_{(n-1:n)}$ are both large.
2. Whenever $X_{(n:n)}$ and $Y_{(n-1:n)}$ are large, $X_{(n:n)}$ is significantly larger than $Y_{(n-1:n)}$.
3. The gain from switching from $Y_{(n-1:n)}$ to $X_{(n:n)}$ when $X_{(n:n)}$ and $Y_{(n-1:n)}$ are both large outweighs the loss when at least one of them is not large.

7.4 Conclusion

In this work we have presented a model for implicit bias in a selection problem motivated by settings including hiring and admissions, and we analyzed the Rooney Rule, which can improve the quality of the resulting choices. For one of the most natural settings of the problem, when candidates are drawn from a power-law distribution, we found a tight characterization of the conditions under which the Rooney Rule improves the quality of the outcome. In the process, we identified a number of counter-intuitive effects at work, which we believe may also help provide insight into how we can reason about implicit bias. Our techniques also provided a natural solution to an inference problem

in which we estimate parameters of a biased decision-making process. Finally, we performed a similar type of analysis on general bounded distributions.

There are a number of further directions in which these issues could be investigated. One intriguing direction is to consider the possible connections to the theory of optimal delegation (see e.g. Alonso and Matouschek (2008)).[3] In the study of delegation, a *principal* wants a task carried out, but this task can only be performed by an *agent* who may have a utility function that is different from the principal's. In an important family of these models, the principal's only recourse is to impose a restriction on the set of possible actions taken by the agent, creating a more constrained task for the agent to perform, in a way that can potentially improve the quality of the eventual outcome from the principal's perspective. Our analysis of the Rooney Rule can be viewed as taking place from the point of view of a principal who is trying to recruit k candidates, but where the process must be delegated to an agent whose utilities for X-candidates and Y-candidates are different from the principal's, and who is the only party able to evaluate these candidates' potentials. The Rooney Rule, requiring that the agent select at least one X-candidate, is an example of a mechanism that the principal could impose to restrict the agent's set of possible actions, potentially improving the quality of the selected candidates as measured by the principal. More generally, it is interesting to ask whether there are other contexts where such a link between delegation and this type of biased selection provides insight.

Our framework also makes it possible to naturally explore extensions of the basic model. First, the model can be generalized to include noisy observations, potentially with a different level of noise for each group. It would also be inter-

[3] We thank Ilya Segal for suggesting this connection to us.

esting to analyze generalizations of the Rooney Rule; for example, if we were to define the ℓ^{th}-*order Rooney Rule* to be the requirement that at least ℓ of k finalists must be from an underrepresented group, we could ask which ℓ produces the greatest increase in utility for a given set of parameters. Finally, we could benefit from a deeper understanding of the function ϕ that appears in our main theorems. For example, while we showed in Theorem 7.3 that ϕ is monotone in β for $k = 2$, Figure 7.1 shows that ϕ is clearly not monotone in δ. A better understanding of the function ϕ may lead to new insights into our model and into the phenomena it seeks to capture.

CHAPTER 8

HOW DO CLASSIFIERS INDUCE AGENTS TO BEHAVE STRATEGICALLY?

One of the fundamental insights in the economics of information is the way in which assessing people (students, job applicants, employees) can serve two purposes simultaneously: it can identify the strongest performers, and it can also motivate people to invest effort in improving their performance (Spence, 1973). This principle has only grown in importance with the rise in algorithmic methods for predicting individual performance across a wide range of domains, including education, employment, and finance.

A key challenge is that we do not generally have access to the true underlying properties that we need for an assessment; rather, they are encoded by an intermediate layer of *features*, so that the true properties determine the features, and the features then determine our assessment. Standardized testing in education is a canonical example, in which a test score serves as a proxy feature for a student's level of learning, mastery of material, and perhaps other properties we are seeking to evaluate as well. In this case, as in many others, the quantity we wish to measure is unobservable, or at the very least, difficult to accurately measure; the observed feature is a construct interposed between the decision rule and the intended quantity.

This role that features play, as a kind of necessary interface between the underlying attributes and the decisions that depend on them, leads to a number of challenges. In particular, when an individual invests effort to perform better on a measure designed by an evaluator, there is a basic tension between effort invested to raise the true underlying attributes that the evaluator cares about, and

effort that may serve to improve the proxy features without actually improving the underlying attributes. This tension appears in many contexts — it is the problem of *gaming* the evaluation rule, and it underlies the formulation of *Goodhart's Law*, widely known in the economics literature, which states that once a proxy measure becomes a goal in itself, it is no longer a useful measure (Hardt et al., 2016a). This principle also underpins concerns about strategic gaming of evaluations in search engine rankings (Davis, 2006), credit scoring (Bambauer and Zarsky, 2018; Foust and Pressman, 2008), academic paper visibility (Beel et al., 2009), reputation management (Zarsky, 2008), and many other domains. While the results we present here are not unique to the context of classifiers and machine learning, concerns over strategic behavior are especially salient in the context of the algorithmically "scored society" (Citron and Pasquale, 2014).

Incentivizing a designated effort investment. These considerations are at the heart of the following class of design problems, illustrated schematically in Figure 8.1. An *evaluator* creates a decision rule for assessing an *agent* in terms of a set of features, and this leads the agent to make choices about how to invest effort across their actions to improve these features. In many settings, the evaluator views some forms of agent effort as valuable and others as wasteful or undesirable. For example, if the agent is a student and the evaluator is constructing a standardized test, then the evaluator would likely view it as a good outcome if the existence of the test causes the student to study and learn the material, but a bad outcome if the existence of the test causes the student to spend a huge amount of effort learning idiosyncratic test-taking heuristics specific to the format of the test, or to spend effort on cheating. Similarly, a job applicant (the agent) could prepare for a job interview given by a potential employer (the evaluator) either by preparing for and learning material that would

directly improve their job performance (a good outcome for both the agent and the evaluator), or by superficially memorizing answers to questions that they find on-line (a less desirable outcome).

Thus, to view an agent's effort in improving their features as necessarily a form of "gaming" is to miss an important subtlety: some forms of effort correspond intuitively to gaming, while others correspond to self-improvement. If we think of the evaluator as having an opinion on which forms of agent effort they would like to promote, then from the evaluator's point of view, some decision rules work better than others in creating appropriate incentives: they would like to create a decision rule whose incentives lead the agent to invest in forms of effort that the evaluator considers valuable.

These concerns have long been discussed in the education literature surrounding the issue of high-stakes standardized testing. In his book "Measuring Up," Daniel Koretz writes,

> Test preparation has been the focus of intense argument for many years, and all sorts of different terms have been used to describe both good and bad forms…I think it's best to…distinguish between seven different types of test preparation: Working more effectively; Teaching more; Working harder; Reallocation; Alignment; Coaching; Cheating. The first three are what proponents of high-stakes testing want to see (Koretz, 2008).

Because teachers are evaluated based on their students' performance on a test, they change their behavior in order to improve their outcomes. As Koretz notes, this can incentivize the investment of both productive and unproductive forms

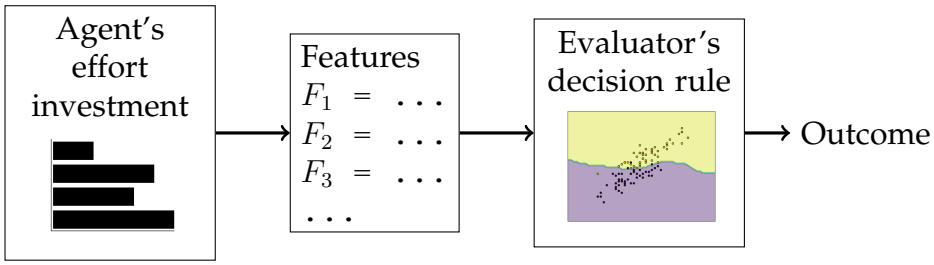


Figure 8.1: The basic framework: an agent chooses how to invest effort to improve the values of certain features, and an evaluator chooses a decision rule that creates indirect incentives favoring certain investments of effort over others.

of effort.

What are the design principles that could help in creating a decision that incentives the kinds of effort that the evaluator wants to promote? Keeping the evaluation rule and the features secret, so as to make them harder to game, is generally not viewed as a robust solution, since information about the evaluation process tends to leak out simply by observing the decisions being made, and secrecy can create inequalities between insiders who know how the system works and outsiders who don't. Nor should the goal be simply to create a decision rule that cannot be affected at all by an agent's behavior; while this eliminates the risk of gaming, it also eliminates the opportunity for the decision rule to incentivize behavior that the evaluator views as valuable.

If there were no intermediate features, and the evaluator could completely observe an agent's choices about how they spread their effort across different actions, then the evaluator could simply reward exactly the actions they want to incentivize. But when the actions taken by an individual are hidden, and can be perceived only through an intermediate layer of proxy features, then the evaluator cannot necessarily tell whether these features are the result of effort they intended to promote (improving the underlying attribute that the feature is

intended to measure) or effort from other actions that also affect the feature. In the presence of these constraints, can one design evaluation rules that nonetheless incentivize the intended set of behaviors?

To return to our stylized example involving students as agents and teachers as evaluators, a teacher can choose among many possible grading schemes to announce to their class; each corresponds to a candidate decision rule, and each could potentially incentivize different forms of effort on the part of the students. For example, the teacher could announce that a certain percentage of the total course grade depends on homework scores, and the remaining percentage depends on exam scores. In this context, the homework and the exam scores are the features that the teacher is able to observe, and the students have various actions at their disposal — studying to learn material, cheating, or other strategies — that can improve these feature values. How does the way in which the teacher balances the percentage weights on the different forms of coursework — producing different possible decision rules — affect the decisions students make about effort? As we will see in the next section, the model we develop here suggests some delicate ways in which choices about a decision rule can in principle have significant effects on agents' decisions about effort allocation.

These effects are not unique to the classroom setting. To take an example from a very different domain, consider a restaurant trying to improve its visibility on a review-based platform (e.g. Yelp). Here we can think of the platform as the evaluator constructing a decision rule and the restaurant as the agent: the platform determines a restaurant's rank based on both the quality of reviews and the number of users who physically visit it, both of which are meant to serve as proxies for its overall quality. The restaurant can individually game ei-

ther of these metrics by paying people to write positive reviews or to physically check in to their location, but improving the quality of the restaurant will ultimately improve both simultaneously. Thus, the platform may wish to consider both metrics, rating and popularity, in some balanced way in order to increase the restaurant's incentive to improve.

Designing evaluation rules. In this chapter, we develop a model for this process of incentivizing effort, when actions can only be measured through intermediate features. We cast our model as an interaction between an *evaluator* who is performing an assessment, and an *agent* who wants to score well on this assessment. An instance of the problem consists of a set of actions in which the agent can invest chosen amounts of *effort*, and a set of functions determining how the levels of effort spent on these actions translate into the values of *features* that are observable to the evaluator.

The evaluator's design task is to create an evaluation rule that takes the feature values as input, and produces a numerical score as output. (Crucially, the evaluation rule is not a function of the agent's level of effort in the actions, only of the feature values.) The agent's goal is to achieve a high score, and to do this, they will optimize how they allocate their effort across actions. The evaluator's goal is to induce a specific *effort profile* from the agent — specifying a level of effort devoted to each action — and the evaluator seeks an evaluation rule that causes the agent to decide on this effort profile. Again, Figure 8.1 gives a basic view of this pipeline of activities.

Our main result is a characterization of the instances for which the evaluator can create an evaluation rule inducing a specified effort profile, and a

polynomial-time algorithm to construct such a rule when it is feasible. As part of our characterization, we find that if there is any evaluation rule, monotone in the feature values, that induces the intended effort profile, then in fact there is one that is linear in the feature values; and we show how to compute a set of coefficients achieving such a rule. Additionally, we provide a tight characterization of which actions can be jointly incentivized.

The crux of our characterization is to consider how an agent is able to "convert" effort from one action to another, or more generally from one set of actions to another set of actions. If it is possible to reallocate effort spent on actions the evaluator is trying to incentivize to actions the evaluator isn't trying to incentivize, in a way that improves the agent's feature values, then it is relatively easy to see that the evaluator won't be able to design a decision rule that incentivizes their desired effort profile: any incentives toward the evaluator's desired effort profile will be undercut by the fact that this effort can be converted away into other undesired forms of effort in a way that improves the agent's outcome. The heart of the result is the converse, providing an if-and-only-if characterization: when such a conversion by the agent isn't possible, then we can use the absence of this conversion to construct an explicit decision rule that incentivizes precisely the effort profile that the evaluator is seeking.

Building on our main result, we consider a set of further questions as well. In particular, we discuss characterizations of the set of all linear evaluation rules that can incentivize a family of allowed effort profiles, identifying tractable structure for this set in special cases, but greater complexity in general. And we consider the problem of choosing an evaluation rule to optimize over a given set of effort profiles, again identifying tractable special cases and computational

hardness in general.

Further Related Work. Our work is most closely related to the principal-agent literature from economics: an evaluator (the principal) wants to set a policy (the evaluation rule) that accounts for the agent's strategic responses. Our main result has some similarities, as well as some key differences, relative to a classical economic formulation in principal-agent models (Grossman and Hart, 1983; Holmström and Milgrom, 1987, 1991; Hermalin and Katz, 1991). We explore this connection in further detail in Section 8.1.4.

A number of more recent models in the economics literature consider settings in which strategic agents are scored to incentivize desirable behavior. These models include elements like intermediaries who perform the scoring (Ball, 2020; Boleslavsky and Kim, 2018), ratings that depend effort exerted over multiple time periods (Holmström, 1999; Hörner and Lambert, 2020; Fong and Li, 2016), binary certification (Zapechelnyuk, 2020; Perez-Richet and Skreta, 2018), and the role of randomness (Rodina and Farragut, 2016). In most of these models, the primary obstacle to the observation of effort is noise; in contrast, uncertainty in our model stems from the confounding of multiple actions in observed features. It would be an interesting subject for future work to combine these sources of uncertainty.

This chapter also ties into two related threads in the economics literature: information design and Bayesian persuasion. Information design concerns the behavior of an agent who controls the information revealed to other agents (Bergemann and Morris, 2019; Taneva, 2019). A special case of information design is Bayesian persuasion (Kamenica and Gentzkow, 2011; Bergemann and Mor-

ris, 2019; Kamenica, 2019), where a principal chooses how to reveal information about an agent or item in order to maximize their utility. Boleslavsky and Kim (2018) extend this model to consider the issue of moral hazard here, where agents under evaluation change their effort investment in response to the scores given by a principal, who is in turn trying to convince a consumer that the agents are high-quality. Because the information scheme is fixed in our model, it lies closer to mechanism design than information design; however, we might extend this model to incorporate elements of information design by allowing the evaluator to modify the mapping from actions to features.

In the computer science literature, a growing body of work seeks to characterize the interaction between a decision-making rule and the strategic agents it governs. This was initially formulated as a zero-sum game (Dalvi et al., 2004), e.g. in the case of spam detection, and more recently in terms of Stackelberg competitions, in which the evaluator publishes a rule and the agent may respond by manipulating their features strategically (Hardt et al., 2016a; Brückner and Scheffer, 2011; Dong et al., 2018; Hu et al., 2019; Milli et al., 2019). This body of work is different from our approach in a crucial respect, in that it tends to assume that all forms of strategic behavior from the agent are undesirable; in our model, on the other hand, we assume that there are certain behaviors that the evaluator wants to incentivize.

There is also work on strategyproof linear regression (Chen et al., 2018; Cummings et al., 2015; Dekel et al., 2010). The setup of these models is also quite different from ours – typically, the strategic agents submit (x, y) pairs where x is fixed and y can be chosen strategically, and the evaluator's goal is to perform linear regression in a way that incentivizes truthful reporting of y. In our set-

ting, on the other hand, agents strategically generate their features x, and the evaluator rewards them in some way based on those features.

Work exploring other aspects of how evaluation rules lead to investment of effort can be found in the economics literature, particularly in the contexts of hiring (Fryer Jr and Loury, 2013; Hu and Chen, 2018) and affirmative action (Coate and Loury, 1993). While these models tend to focus on decisions regarding skill acquisition, they broadly consider the investment incentives created by evaluation. Similar ideas can also be found in the Science and Technology Studies literature (Ziewitz, 2019), considering how organizations respond to guidelines and regulations.

As noted above, principal-agent mechanism design problems in which the principal cannot directly observe the agent's actions have been studied in the economics literature (Arrow, 1963; Pauly, 1968; Arrow, 1968), and include work on the notion of *moral hazard.* Insurance markets are canonical examples in this domain: the agent reduces their liability by purchasing insurance, and this may lead them to act more recklessly and decrease welfare. The principal cannot directly observe how carefully the agent is acting, only whether the agent makes any insurance claims. These models provide some inspiration for ours; in particular, they are often formalized such that the agent's actions are "effort variables" which, at some cost to the agent, increase the agent's level of "production" (Laffont and Martimort, 2009). This could be, for example, acting in more healthy ways or driving more carefully in the cases of health and car insurance respectively. Note, however, that in the insurance case, the agent and the principal have aligned incentives in that both prefer that the agent doesn't — e.g., in the case of car insurance — get into an accident. In our model, we

make no such assumptions: the agent may have no incentive at all to invest in the evaluator's intended forms of effort beyond the utility derived from the mechanism. The types of scenarios considered in insurance markets can be generalized to domains like share-cropping (Cheung, 1969; Stiglitz, 1974), corporate liability (Jensen and Meckling, 1976), and theories of agency (Ross, 1973). Kerr (1975) provides a detailed list of such instances in his classic paper "On the folly of rewarding A, while hoping for B."

Concerns over strategic behavior also manifest in ways that do not necessarily map to intuitive notions of gaming, but instead where the evaluator does not want to incentivize the agent to take actions that might be counter to their interests. For example, Virginia Eubanks (2018a) considers a case of risk assessment in the child welfare system; when a risk tool includes features about a family's history of interaction with public services, including aid such as food stamps and public housing, she argues that it has the potential to incentivize families to avoid such services for fear of being labeled high risk. This too would be a case in which the structure and implementation of an evaluation rule can incentivize potentially undesirable actions in agents, and would be interesting to formalize in the language of our model.

Organization of the remainder of the chapter. Section 8.1 contains all the definitions and technical motivation leading up to the formulation and statement of our two main results, Theorems 8.3 and 8.5. Sections 8.2 and 8.3 contain the proofs of these two results, respectively, and Section 8.4 considers further extensions.

8.1 Model and Overview of Results

8.1.1 A Formal Model of Effort Investment

Here, we develop a formal model of an agent's investment of effort. There are m actions the agent can take, and they must decide to allocate an amount of effort x_j to each action j. We'll assume the agent has some budget B of effort to invest, so $\sum_{j=1}^{m} x_j \leq B$, and we'll call this investment of effort $x = (x_1, x_2, \ldots, x_m)$ an *effort profile*.[1]

The evaluator cannot directly observe the agent's effort profile, but instead observes features $F_1, \ldots, F_n$ derived from the agent's effort profile. The value of each F_i grows monotonically in the effort the agent invests in certain actions according to an *effort conversion function* $f_i(\cdot)$:

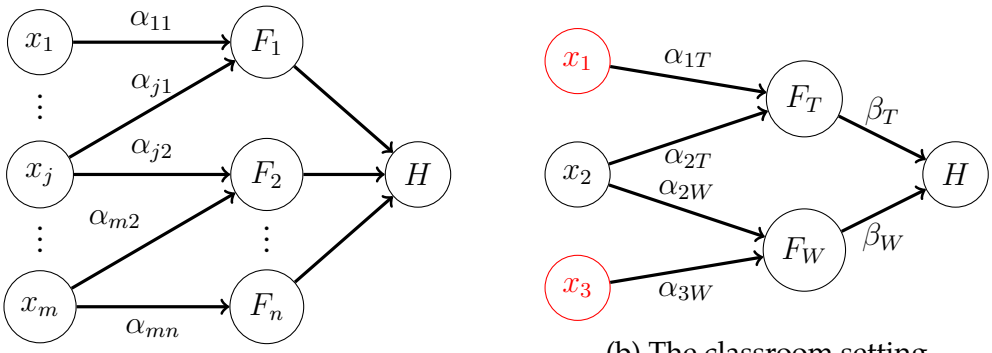
$$F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right), \tag{8.1}$$

where each $f_i(\cdot)$ is nonnegative, smooth, weakly concave (i.e., actions provide diminishing returns), and strictly increasing. We assume $\alpha_{ji} \geq 0$, meaning that effort results in more favorable outcomes.

We might instead model the agent as incurring a fixed cost c per unit effort with no budget. In fact, this formulation is in a sense equivalent: for every cost c, there exists a budget B such that an agent with cost c behaves identically to an agent with fixed budget B (and no cost). For clarity, we will deal only with the budgeted case, but our results will extend to the case where effort comes at a linear cost.

[1] Instead of a fixed budget, we might consider an alternate model in which effort comes at a cost; we discuss the relationship between that model and the one presented here later.

We represent these parameters of the problem using a bipartite graph with the actions $x_1, x_2, \ldots, x_m$ on the left, the features $F_1, \ldots, F_n$ on the right, and an edge of weight α_{ji} whenever $\alpha_{ji} > 0$, so that effort spent on action j contributes to the value of feature F_i. We call this graph, along with the associated parameters (the matrix $\alpha \in \mathbb{R}^{m \times n}$ with entries α_{ji}; functions $f_i : \mathbb{R} \to \mathbb{R}$ for $i \in \{1, ..., n\}$; and a budget B), the *effort graph* G. Figure 8.2 shows some examples of what G might look like.



(a) General model

(b) The classroom setting

Figure 8.2: The conversion of effort to feature values can be represented using a weighted bipartite graph, where effort x_j spent on action j has an edge of weight α_{ji} to feature F_i.

The evaluator combines the features generated by the effort using some mechanism M to produce an output H, which is the agent's utility. M is simply a function of the n feature values. In a classification setting, for example, H may be binary (whether or not the agent is classified as positive or negative), or a continuous value (the probability that the agent receives a positive outcome). Because all features are increasing in the amount of effort invested by the agent — in particular, including the kinds of effort we want to incentivize — we'll restrict our attention to the class of monotone mechanisms, meaning that if agent X has larger values in all features than agent Y, then X's outcome should be at least as good as that of Y. Formally, we write this as follows:

Definition 8.1. *A* monotone mechanism M *on features* F_i *is a mapping* $\mathbb{R}^n \to \mathbb{R}$ *such that for* $F, F' \in \mathbb{R}^n$ *with* $F'_i \geq F_i$ *for all* $i \in \{1, ..., n\}$, $M(F') \geq M(F)$. *Also, for any* F, *there exists* $i \in \{1, ..., n\}$ *such that strictly increasing* F_i *strictly increases* $M(F)$.

The second of these conditions implies that it is strictly optimal for an agent to invest all of its budget. The agent's utility is simply its outcome H. Thus, for a mechanism M, the agent's optimal strategy is to invest effort to maximize $M(F)$ subject to the constraints that $\sum_{j=1}^{m} x_j \leq B$ and $x_j \geq 0$ for all j. (Recall that in this phrasing, the vector F of feature values is determined from the effort value x_i via the functions $F_i = f_i\left(\sum_{j=1}^{m} \alpha_{ji} x_j\right)$.) We can write the agent's search for an optimal strategy succinctly as the following optimization problem:

$$x^* = \underset{x \in \mathbb{R}^m}{\arg\max}\ M(F) \qquad \text{s.t. } \sum_{j=1}^{m} x_j \leq B \qquad x \geq \mathbf{0} \tag{8.2}$$

where each component F_i of F is defined as in (8.1). Throughout this chapter, we'll assume that agents behave rationally and optimally, though it would be an interesting subject for future work to consider extensions of this model where agents suffer from behavioral biases. We also note that this is where we make use of the concavity of the functions f_i, since for arbitrary f_i the agent wouldn't necessarily be able to efficiently solve this optimization problem.

8.1.2 Returning to the classroom example

To illustrate the use of this model, consider the effort graph shown in Figure 8.2b, encoding the classroom example described in the introduction. There

are two pieces of graded work for the class (a test F_T and homework F_W), and the student can study the material (x_2) to improve their scores on both of these. They can also cheat on the test (x_1) and look up homework answers on-line (x_3). Their combined effort $\alpha_{1T}x_1 + \alpha_{2T}x_2$ contributes to their score on the test, and their combined effort $\alpha_{2W}x_2+\alpha_{3W}x_3$ contributes to their score on the homework. To fully specify the effort graph, we would have to provide a budget B and effort conversion functions f_T and f_W; we leave these uninstantiated, as our main conclusions from this example will not depend on them.

From these scores, the teacher must decide on a student's final grade H. For simplicity, we'll assume the grading scheme is simply a linear combination, meaning $H = \beta_T F_T + \beta_W F_W$ for some real numbers $\beta_T, \beta_W \geq 0$.

The teacher's objective is to incentivize the student to learn the material; thus, they want to set β_T and β_W such that the student invests their entire budget into x_2. Of course, this may not be possible. For example, if α_{1T} and α_{3W} are significantly larger than α_{2T} and α_{2W} respectively, so that it is much easier to cheat on the test and copy homework answers than to study, the student would maximize their utility by investing all of their effort into these undesirable activities.

In fact, we can make this precise as follows. For any unit of effort invested in action 2, the student could instead invest $\frac{\alpha_{2T}}{\alpha_{1T}}$ and $\frac{\alpha_{2W}}{\alpha_{3W}}$ units of effort into actions 1 and 3 respectively without changing the values of F_T and F_W. Moreover, if $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} < 1$, then this substitution strictly reduces the sum $x_1 + x_2 + x_3$, leaving additional effort available (relative to the budget constraint) for raising the values of F_T and F_W. It follows that in any solution with $x_2 > 0$, there is a way to strictly improve it through this substitution. Thus, under this condition,

the teacher cannot incentivize the student to only study. This is precisely the type of "conversion" of effort that we discussed briefly in the previous section, from the evaluator's preferred action (2) to other actions (1 and 3)

When $\frac{\alpha_{2T}}{\alpha_{1T}} + \frac{\alpha_{2W}}{\alpha_{3W}} \geq 1$, on the other hand, a consequence of our results is that that no matter what f_T, f_W and B are, there exist some β_T, β_W that the teacher can choose to incentivize the student to invest all their effort into studying. This may be somewhat surprising – for instance, consider the case where $\alpha_{1T} = \alpha_{3W} = 3$ and $\alpha_{2T} = \alpha_{2W} = 2$, meaning that the best way for the student to maximize their score on each piece of graded work individually is to invest undesirable effort instead of studying. Even so, it turns out that the student can still be incentivized to put all of their effort into studying by appropriately balancing the weight placed on the two pieces of graded work.

This example illustrates several points that will be useful in what follows. First, it makes concrete the basic obstacle against incentivizing a particular action: the possibility that effort can be "swapped out" at a favorable exchange rate towards other actions. Second, it shows a particular kind of reason why it might be possible to incentivize a designated action j: if investing effort x_j improves multiple features simultaneously, the agent might select it even if it is not the most efficient way to increase any one feature individually. This notion of activities that "transfer" across different forms of evaluation, versus activities that fail to transfer, arises in the education literature on testing (Koretz et al., 1991), and our model shows how such effects can lead to valuable incentives.

8.1.3 Stating the main results

In our example, it turned out that a linear grading scheme was sufficient for the teacher to incentivize the student to study. We formalize such mechanisms as follows.

Definition 8.2. *A* linear mechanism $M : \mathbb{R}^n \to \mathbb{R}$ *is the mapping* $M(F) = \beta^\top F = \sum_{i=1}^n \beta_i F_i$ *for some* $\beta \in \mathbb{R}^n$ *such that* $\beta_i \geq 0$ *for all* $i \in \{1, ..., n\}$ *and* $\sum_{i=1}^n \beta_i > 0$.

Note that we don't require $\sum_{i=1}^n \beta_i$ to be anything in particular; the agent's optimal behavior is invariant to scaling β, so we can normalize β to sum to any intended quantity without affecting the properties of the mechanism. We rule out the mechanism in which all β_i are equal to 0, as it is not a monotone mechanism.

We will say that a mechanism M *incentivizes* effort profile x if x is an optimal response to M. Ultimately, our main result will be to prove the following theorem, characterizing when a given effort profile can be incentivized. First, we need to define the support of x as

$$\mathcal{S}(x) \triangleq \{j \mid x_j > 0\}. \tag{8.3}$$

With this, we can state the theorem.

Theorem 8.3. *For an effort graph* G *and an effort profile* x^**, the following are equivalent:*

1. *There exists a linear mechanism that incentivizes* x^*.
2. *There exists a monotone mechanism that incentivizes* x^*.
3. *For all* x *such that* $\mathcal{S}(x) \subseteq \mathcal{S}(x^*)$*, there exists a linear mechanism that incentivizes* x.

Furthermore, there is a polynomial time algorithm that decides the incentivizability of x^* and provides a linear mechanism β to incentivize x^* whenever such β exists.

When there exists a monotone mechanism incentivizing x^*, we'll call both x^* and $\mathcal{S}(x^*)$ *incentivizable*.[2] Informally, when x^* is not incentivizable, this algorithm finds a succinct "obstacle" to any solution with support $\mathcal{S}(x^*)$, meaning no x such that $\mathcal{S}(x) = \mathcal{S}(x^*)$ is incentivizable. The following corollary is a direct consequence of Theorem 8.3. (We use the notation $[m]$ to represent $\{1, 2, \ldots, m\}$.)

Corollary 8.4. *For a set $S \subseteq [m]$, some x such that $\mathcal{S}(x) = S$ is incentivizable if and only if all x with $\mathcal{S}(x) = S$ are incentivizable.*

In Section 8.2, we'll prove Theorem 8.3. The proof we give is constructive, and it establishes the algorithmic result.

Optimizing over effort profiles. It may be the case that the evaluator doesn't have a single specific effort profile by the agent that they want to incentivize; instead, they may have an objective function defined on effort profiles, and they would like to maximize this objective function over effort profiles that are incentivizable. In other words, the goal is to choose an evaluation rule so that the resulting effort profile it induces performs as well as possible according to the objective function.

In Section 8.3, we consider the following formulation for such optimization problems. We assume that the evaluator wants to maximize a concave function $g : \mathbb{R}^m \to \mathbb{R}$ over the space of effort profiles, subject to the constraint that the agent only invests effort in a subset $D \subseteq [m]$ of actions. To accomplish this, the

[2] A closely related notion in the principal-agent literature is that of an *implementable* action.

evaluator selects an evaluation rule so as to incentivize an effort profile x^* with $g(x^*)$ as large as possible. This is what we will mean by optimizing g over the space of effort profiles. In this setting, we show the following results, which we prove in Section 8.3.

Theorem 8.5. *Let g be a concave function over the space of effort profiles, and let D be the set of actions in which the evaluator is willing to allow investment by the agent.*

1. *If there exists an x^* such that $\mathcal{S}(x^*) = D$ and x^* is incentivizable, then any concave function g can be maximized over the space of effort profiles in polynomial time.*
2. *If $|D|$ is constant, then any concave function g can be maximized over the space of effort profiles in polynomial time.*
3. *In general, there exist concave functions g that are NP-hard to maximize over the space of effort profiles subject to the incentivizability condition.*

In summary, we establish that it is computationally hard to maximize even concave objectives in general, although as long as the number of distinct actions the evaluator is willing to incentivize is small, concave objectives can be efficiently maximized.

The above results characterize optimization over effort profiles; instead, the evaluator may wish to optimize over the space of mechanisms (e.g., to fit to a dataset). We consider the feasibility of such optimization in Section 8.4, showing that the set of linear mechanisms incentivizing particular actions can be highly nonconvex, making optimization hard in general.

8.1.4 Principal-Agent Models and Linear Contracts

Now that we have specified the formalism, we are in a position to compare our model with well-studied principal-agent models in economics to see how our results and techniques relate to those from prior work. In the standard principal-agent setting, the principal's objective is to incentivize an agent to invest effort in some particular way (Ross, 1973; Grossman and Hart, 1983). Crucially, the principal cannot observe the agent's action – only some outcome that is influenced by the agent's action. Thus, while the principal cannot directly reward the agent based on the action it takes, it can instead provide rewards based on outcomes that are more likely under desirable actions.

To our knowledge, this framework has yet to be applied to settings based on machine-learning classifiers as we do here; and yet, principal-agent models fit quite naturally in this context. A decision-maker wants to evaluate an agent, which it can only observe indirectly through features. These features, in turn, reflect the actions taken by the agent. In this context, the principal offers a "contract" by specifying an evaluation rule, to which the agent responds strategically by investing its effort so as to improve its evaluation. So far, this is in keeping with the abstract principal-agent framework (Ross, 1973; Grossman and Hart, 1983).

Moreover, some of the key results we derive echo known results from previous models, though they also differ in important respects. Linear contracts, in particular, are often necessary or optimal in principal-agent contexts for a variety of reasons. In modeling bidding for government contracts, for example, payment schemes are linear in practice for the sake of simplicity, even though optimal contracts may be nonlinear (McAfee and McMillan, 1986). In other

models, contracts are naturally linear because agents maximize reward in expectation over outcomes generated stochastically from their actions (Grossman and Hart, 1983).

Even when they aren't necessitated by practical considerations or modeling choices, linear contracts have been shown to be optimal in their own right in some principal-agent models. Holmström and Milgrom (1987, 1991) consider the interplay between incentives and risk aversion and characterize optimal mechanisms in this setting, finding that under a particular form of risk aversion (exponential utility), linear contracts optimally elicit desired behavior. Our models do not incorporate a corresponding notion of risk aversion, and the role of linear mechanisms in our work arises for fundamentally different reasons.

Hermalin and Katz (1991) provide a model more similar to ours, in which observations result stochastically from agents' actions. Drawing on basic optimization results that we use here as well (in particular, duality and Farkas' Lemma), they characterize actions as "implementable" based on whether they can be in some sense replaced by other actions at lower cost to the agent. At a high level, we will rely on a similar strategy to prove Theorem 8.3.

There are, however, some further fundamental differences between the principal-agent models arising from the work of Hermalin and Katz and the questions and results we pursue here. In particular, the canonical models of principal-agent interaction in economics typically only have the expressive power to to incentivize a single action, which stochastically produces a single observed outcome. This basic difference leads to a set of important distinctions for the modeling goals we have: because our goal is to incentivize investment over multiple activities given a multi-dimensional feature vector, with the chal-

lenge that different mixtures of activities can deterministically produce the same feature vector, our model cannot be captured by these earlier formalisms.

An important assumption in our model, and in many principal-agent models in general, is that the principal knows how the agent's effort affects observations. Recent work has sought to relax this assumption, finding that linear contracts are optimal even when the principal has incomplete knowledge of the agent's cost structure (Carroll, 2015). It would be an interesting subject for future work to extend our model so that the principal does not know or needs to learn the agent's cost structure.

8.2 Incentivizing Particular Effort Profiles

In this section, we develop a tight characterization of which effort profiles can be incentivized and find linear mechanisms that do so. For simplicity, we'll begin with the special case where the effort profile to be incentivized is $x^* = B \cdot e_j$, with e_j representing the unit vector in coordinate j — that is, the entire budget is invested in effort on action j. Then, we'll apply the insights from this case to the general case.

The special case where $|\mathcal{S}(x^*)| = 1$. Recall that in the classroom example, the tipping point for when the intended effort profile could be incentivized hinged on the question of *substitutability*: the rate at which undesirable effort could be substituted for the intended effort. We'll characterize this rate as the solution to a linear program. In an effort graph G, recall that $\alpha \in \mathbb{R}^{m \times n}$ is the matrix with entries α_{ji}. Let $\tilde{\alpha}_j \in \mathbb{R}^n$ be the jth row of α. Then, we'll define the substitutabil-

ity of x_j to be

$$\kappa_j \triangleq \min_{y \in \mathbb{R}^m} \; y^\top \mathbf{1} \qquad \text{s.t. } \alpha^\top y \geq \tilde{\alpha}_j \qquad y \geq \mathbf{0} \tag{8.4}$$

Intuitively, y is a redistribution of effort out of action j that weakly increases all feature values. Note that $\kappa_j \leq 1$ because the solution $y = e_j$ (the vector with 1 in the jth position and 0 elsewhere) is feasible and has value 1. In Lemma 8.6, we'll use this notion of substitutability to show that whenever $\kappa_j < 1$, the agent will at optimality put no effort into action j. Conversely, in Lemma 8.7, we'll show that the when $\kappa_j = 1$, there exists a linear mechanism β incentivizing the solution $x^* = B \cdot e_j$.

It might seem odd that this characterization depends only on κ_j, which is independent of both the budget B and effort conversion functions f_i; however, the particular mechanisms that incentivize x^* will depend on these. This will also be true in the general case: whether or not a particular effort profile can be incentivized will not depend on B or f_i, but the exact mechanisms that do so will.

Lemma 8.6. *If $\kappa_j < 1$, then in any monotone mechanism M, $x^*_j = 0$.*

Proof. Intuitively, this is an argument formalizing substitution: if $\kappa_j < 1$, replacing each unit of effort on action j with y_k units of effort (where y comes from the optimal solution to (8.4)) on each action k for $k \in [m]$ weakly increases all of the feature values F_i while making the budget constraint slack. Therefore, any solution with $x_j > 0$ cannot be optimal.

In more detail, consider any solution x with $x_j > 0$. We'll begin by showing that the agent's utility is at least as high in the solution x' with $x'_k = x_k + y_k x_j$

for all $k \neq j$ and $x'_j = y_j x_j$, where y is an optimal solution to the linear program in (8.4). Note that $y_j \leq \kappa_j < 1$, so x' is different from x.

We know from the constraint on (8.4) that $\alpha^\top y \geq \tilde{\alpha}_j$, and therefore

$$\sum_{k=1}^{m} \alpha_{ki} y_k \geq \alpha_{ji} \tag{8.5}$$

for all i. Then, by (8.5),

$$f_i\left(\sum_{k=1}^{m} \alpha_{ki} x_k\right) \leq f_i\left(\sum_{k \neq j} \alpha_{ki} x_k + x_j \sum_{k=1}^{m} \alpha_{ki} y_k\right) = f_i\left(\sum_{k=1}^{m} \alpha_{ki} x'_k\right)$$

Thus, the value of each feature weakly increases from x to x', so in any monotone mechanism M, the agent's utility for x' is at least as high as it is for x. Moreover, the budget constraint on x' isn't tight, since

$$\sum_{k=1}^{m} x'_k = \sum_{k \neq j} (x_k + y_k x_j) + y_j x_j = \sum_{k \neq j} x_k + x_j \sum_{k=1}^{m} y_k < \sum_{k=1}^{m} x_k \leq B.$$

By the definition of a monotone mechanism, no solution for which the budget constraint isn't tight can be optimal, meaning x' is not optimal. This implies that x is not optimal. □

Thus, $\kappa_j < 1$ implies that $x_j = 0$ in any optimal solution. All that remains to show in this special case is the converse: if $\kappa_j = 1$, there exists β that incentivizes the effort profile $x^* = B \cdot e_j$. To do so, define $A(x) \in \mathbb{R}^{m \times n}$ to be the matrix with entries $[A(x)]_{ji} = \alpha_{ji} f_i'([\alpha^\top x]_i)$, and define $a_j(x) \in \mathbb{R}^n$ to be the jth row of $A(x)$. Then, we can define the polytope

$$\mathcal{L}_j \triangleq \{\beta \mid A(x^*)\beta \leq \beta^\top a_j(x^*) \cdot \mathbf{1}\}. \tag{8.6}$$

By construction, $\mathcal{L}_j$ is the set of linear mechanisms that incentivize x^*. This

is because for all $k \in [m]$, every $\beta \in \mathcal{L}_j$ satisfies

$$\begin{aligned}
&[A(x^*)\beta]_k \leq \beta^\top a_j(x^*) \\
&\iff \sum_{i=1}^{n} \alpha_{ki}\beta_i f_i'([\alpha^\top x^*]_i) \leq \sum_{i=1}^{n} \alpha_{ji}\beta_i f_i'([\alpha^\top x^*]_i) \iff \left.\frac{\partial H}{\partial x_k}\right|_{x^*} \leq \left.\frac{\partial H}{\partial x_j}\right|_{x^*}
\end{aligned}$$

By Lemma E.1 in Appendix E.1, this implies that x^* is an optimal agent response to any $\beta \in \mathcal{L}_j$. To complete the proof of this special case of Theorem 8.3, it suffices to show that $\mathcal{L}_j$ is non-empty, which we do via linear programming duality.

Lemma 8.7. *If $\kappa_j = 1$, then $\mathcal{L}_j$ is non-empty.*

Proof. Consider the following linear program.

$$\begin{aligned}
\max_{\beta \in \mathbb{R}^n} \ & \beta^\top a_j(x^*) && \text{s.t. } A(x^*)\beta \leq \mathbf{1} \\
& && \beta \geq \mathbf{0}
\end{aligned} \tag{8.7}$$

Clearly, if (8.7) has value at least 1, then $\mathcal{L}_j$ is non-empty because any β achieving the optimum is in $\mathcal{L}_j$ by (8.6). The dual of (8.7) is

$$\begin{aligned}
\min_{y \in \mathbb{R}^m} \ & y^\top \mathbf{1} && \text{s.t. } A(x^*)^\top y \geq a_j(x^*) \\
& && y \geq \mathbf{0}
\end{aligned} \tag{8.8}$$

We can simplify the constraints on (8.8): for all i,

$$\begin{aligned}
&[A(x^*)^\top y]_i \geq [a_j(x^*)]_i \\
&\iff \sum_{k=1}^{m} \alpha_{ki} y_k f_i'([\alpha^\top x^*]_i) \geq \alpha_{ji} f_i'([\alpha^\top x^*]_i) \\
&\iff \sum_{k=1}^{m} \alpha_{ki} y_k \geq \alpha_{ji}
\end{aligned}$$

Thus, (8.8) is equivalent to (8.4), which has value $\kappa_j = 1$ by assumption. By duality, (8.7) also has value $\kappa_j = 1$, meaning $\mathcal{L}_j$ is non-empty. □

We have shown that if $\kappa_j = 1$, then any $\beta \in \mathcal{L}_j$ incentivizes x^*. Otherwise, by Lemma 8.6, there are no monotone mechanisms that incentivize x^*. Next, we'll generalize these ideas to prove Theorem 8.3.

The general case. We'll proceed by defining the analogue of κ_j in the case where the effort profile to be incentivized has support on more than one component. Drawing upon the reasoning in Lemmas 8.6 and 8.7, we'll prove Theorem 8.3.

Consider an arbitrary effort profile x^* such that $\sum_{i=1}^m x_j^* = B$, and let $\mathcal{S}(x^*)$ be the support of x^*. Let α_S be α with the rows not indexed by S zeroed out, i.e., $[\alpha_S]_{ji} = \alpha_{ji}$ if $j \in S$ and 0 otherwise. Let $\mathbf{1}_S$ be the vector with a 1 for every $j \in S$ and 0 everywhere else, so $\mathbf{1}_S = \sum_{j \in S} e_j$. Similarly to how we defined κ_j, define

$$\kappa_S \triangleq \min_{y \in \mathbb{R}^m, z \in \mathbb{R}^m} \; y^\top \mathbf{1} \qquad \text{s.t. } \alpha^\top y \geq \alpha_S^\top z \qquad (8.9)$$
$$z^\top \mathbf{1}_S \geq 1$$
$$y, z \geq \mathbf{0}$$

Intuitively, we can think of the effort given by z as being substituted out and replaced by y. Note that $\kappa_S \leq \min_{j \in S} \kappa_j$, because the special case where $z_j = 1$ yields (8.4). In a generalization of Lemma 8.6, we'll show that $\kappa_S < 1$ implies that no optimal solution has $x_j > 0$ for all $j \in S$. Lemma 8.6 formalized an argument based on substitutability, in which the effort invested in a particular action could be moved to other actions while only improving the agent's utility. We generalize this to the case when effort invested on a subset of the actions can be replaced by moving that effort elsewhere.

Lemma 8.8. *For any $S \subseteq [m]$, if $\kappa_S < 1$, then any effort profile x such that $x_j > 0$ for all $j \in S$ cannot be optimal.*

Proof. The following proof builds on that of Lemma 8.6. Let y and z be optimal solutions to (8.9). We know that for all i,

$$\sum_{j=1}^{m} \alpha_{ji} y_j \geq \sum_{j \in S} \alpha_{ji} z_j \tag{8.10}$$

Let $c \triangleq \min_{j \in S} x_j / z_j$. Note that $c > 0$ because by assumption, $x_j > 0$ for all $j \in S$. It is well-defined because $z^\top \mathbf{1}_S \geq 1$ and $z \geq \mathbf{0}$, so z_j is strictly positive for some $j \in S$. By this definition, $x_j - c z_j \geq 0$ for all $j \in S$.

We'll again define another solution x' with utility at least as high as x, but with the budget constraint slack. For all i,

$$\begin{aligned}
[\alpha^\top x]_i &= \sum_{j=1}^{m} \alpha_{ji} x_j \\
&= \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} x_j \\
&= \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} (x_j - c z_j) + c \sum_{j \in S} \alpha_{ji} z_j \\
&\leq \sum_{j \notin S} \alpha_{ji} x_j + \sum_{j \in S} \alpha_{ji} (x_j - c z_j) + c \sum_{j=1}^{m} \alpha_{ji} y_j && \text{(By (8.10))} \\
&= \sum_{j \notin S} \alpha_{ji} (x_j + c y_j) + \sum_{j \in S} \alpha_{ji} (x_j + c(y_j - z_j)) \\
&\triangleq [\alpha^\top x']_i,
\end{aligned}$$

where we have defined

$$x'_j \triangleq \begin{cases} x_j + c y_j & j \notin S \\ x_j + c(y_j - z_j) & z \in S \end{cases}.$$

Because $x_j - cz_j \geq 0$ for all $j \in S$, x' is a valid effort profile. Since f_i is increasing, $f_i([\alpha^\top x]_i) \leq f_i([\alpha^\top x']_i)$. However,

$$\sum_{i=1}^{m} x'_j = \sum_{j \notin S} x_j + cy_j + \sum_{j \in S} x_j + c(y_j - z_j) = x^\top \mathbf{1} + c(y^\top \mathbf{1} - z^\top \mathbf{1}_S)$$
$$\leq x^\top \mathbf{1} + c(\kappa_S - 1) < B$$

Thus, the budget constraint for x' is not tight, and so for any monotone mechanism, there exists a solution x'' which is strictly better than x' and x, meaning x is not optimal. □

Lemma 8.8 tells us which subsets of variables definitely can't be jointly incentivized. However, given a subset of variables, it doesn't a priori tell us if these variables *can* be jointly incentivized, and if so, which particular effort profiles on these variables are incentivizable. In fact, we'll show that any x^* such that $\kappa_{\mathcal{S}(x^*)} = 1$ is incentivizable.

Lemma 8.9. *Define*

$$\mathcal{L}(x) \triangleq \left\{ \beta \mid A(x)\beta \leq \frac{1}{B} x^\top A(x)\beta \cdot \mathbf{1} \right\} \tag{8.11}$$

If $\kappa_{\mathcal{S}(x^*)} = 1$, then $\mathcal{L}(x^*)$ is the set of linear mechanisms that incentivize x^*, and $\mathcal{L}(x^*)$ is non-empty.

Proof. Let $S = \mathcal{S}(x^*)$. We know that for any z such that $z^\top \mathbf{1}_S \geq 1$,

$$\kappa_S \leq \kappa_S(z) \triangleq \min_{y \in \mathbb{R}^m} \; y^\top \mathbf{1} \qquad \text{s.t. } \alpha^\top y \geq \alpha_S^\top z \qquad y \geq \mathbf{0} \tag{8.12}$$

because we've just written (8.9) without allowing for optimization over z. Therefore, if $\kappa_S = 1$, then $\kappa_S(z) = 1$ for any z. We can write each constraint

$[\alpha^\top y]_i \geq [\alpha_S^\top z]_i$ as

$$\begin{aligned}
[\alpha^\top y]_i \geq [\alpha_S^\top z]_i &\iff \sum_{j=1}^{m} \alpha_{ji} y_j \geq \sum_{j \in S} \alpha_{ji} z_j \\
&\iff \sum_{j=1}^{m} \alpha_{ji} f_i'([\alpha^\top x^*]_i) y_j \geq \sum_{j \in S} \alpha_{ji} f_i'([\alpha^\top x^*]_i) z_j \\
&\iff \sum_{j=1}^{m} [A(x^*)]_{ji}^\top y_j \geq \sum_{j \in S} [A(x^*)]_{ji} z_j
\end{aligned}$$

Thus, (8.12) is equivalent to the following optimization, where similarly to the definition of α_S, we define $A_S(x)$ to be $A(x)$ with all rows $j \notin S$ zeroed out.

$$\begin{aligned}
\kappa_S(z) = \min_{y \in \mathbb{R}^m} \; & y^\top \mathbf{1} && \text{s.t. } A(x^*)^\top y \geq A_S(x^*)^\top z \\
& && y \geq \mathbf{0}
\end{aligned} \tag{8.13}$$

The dual of (8.13) is

$$\begin{aligned}
\eta(z) \triangleq \max_{\beta \in \mathbb{R}^n} \; & \beta^\top (A_S(x^*)^\top z) && \text{s.t. } A(x^*)\beta \leq \mathbf{1} \\
& && \beta \geq \mathbf{0}
\end{aligned} \tag{8.14}$$

Thus, (8.14) has value $\eta(z) = \kappa_S(z) = 1$. Recall that

$$\mathcal{L}(x^*) = \{\beta \mid A(x^*)\beta \leq \frac{1}{B} x^{*\top} A(x^*)\beta \cdot \mathbf{1}\}.$$

Clearly, $\mathcal{L}(x^*)$ is non-empty because plugging in $z = \frac{x^*}{B}$, (8.14) has value $\eta(z) = 1$, meaning there exists β such that for all j,

$$\eta\left(\frac{x^*}{B}\right) = \frac{1}{B}\beta^\top (A_S(x^*)^\top x^*) = 1 \geq [A(x^*)\beta]_j \tag{8.15}$$

We'll show that β incentivizes the agent to invest x^* if and only if $\beta \in \mathcal{L}(x^*)$. Note that (8.15) is true if and only if

$$\left.\frac{\partial H}{\partial x_j}\right|_{x^*} \leq \sum_{k \in S} \frac{x_k^*}{B} \left.\frac{\partial H}{\partial x_k}\right|_{x^*}. \qquad (\forall j \in [m])$$

The right hand side is the convex combination of the partial derivatives of H with respect to each of the $k \in S$. Since this convex combination is at least as large as each partial in the combination, it must be the case that all of these partials on the right hand side are equal to one another. In other words, this is true if and only if $\frac{\partial H}{\partial x_j}\Big|_{x^*} = \frac{\partial H}{\partial x_{j'}}\Big|_{x^*}$ for all $j, j' \in S$.

By Lemma E.1 in Appendix E.1, this is true if and only if x^* is an optimal effort profile, meaning $\mathcal{L}(x^*)$ is exactly the set of linear mechanisms that incentivize x^*. □

Thus, we've shown Theorem 8.3: for any target effort profile x^*, either $\kappa_{\mathcal{S}(x^*)} = 1$, in which case any $\beta \in \mathcal{L}(x^*)$ incentivizes x^*, or $\kappa_{\mathcal{S}(x^*)} < 1$, in which case no monotone mechanism incentivizes x^* by Lemma 8.8.

8.3 Optimizing other Objectives

So far, we have given a tight characterization of which effort profiles can be incentivized. Moreover, we have shown that whenever an effort profile can be incentivized, we can compute a set of linear mechanisms that do so. However, this still leaves room for the evaluator to optimize over other preferences. For instance, perhaps profiles that distribute effort among many activities are more desirable, or perhaps the evaluator has a more complex utility function over the agent's effort profile.

In this section, we consider the feasibility of such optimization subject to the constraints imposed by incentivizability. We show that optimization over effort profiles is possible in particular instances, but in general, it is computationally

hard to optimize even simple objectives over incentivizable effort profiles.

Incentivizing a subset of variables. For the remainder of this section, we will assume that the evaluator has a set of designated actions $D \subseteq [m]$, and they want to incentivize the agent to only invest in actions in D. Recall that a set of actions S is incentivizable if and only if $\kappa_S = 1$, where κ_S is defined in (8.9). We define the set system

$$\mathcal{F}_D = \{S \subseteq D \mid \kappa_S = 1\} \tag{8.16}$$

By Theorem 8.3, $\mathcal{F}_D$ gives the sets of actions that can be jointly incentivized. As we will show, a consequence of our results from Section 8.2 is that $\mathcal{F}_D$ is downward-closed, meaning that if $S \in \mathcal{F}_D$, then $S' \in \mathcal{F}_D$ for any $S' \subseteq S$.

We begin by characterizing when it is feasible to incentivize some x such that $\mathcal{S}(x) \subseteq D$. As the following lemma shows, this can be done if and only if some individual $j \in D$ is incentivizable on its own.

Lemma 8.10. *It is possible to incentivize effort in a subset of a designated set of actions $D \subseteq [m]$ if and only if $\max_{j \in D} \kappa_j = 1$.*

Proof. The set system $\mathcal{F}_D$ is downward closed, since $\kappa_{S \cup \{j\}} = 1$ implies $\kappa_S = 1$ for all S, j. This is because any solution to (8.9) for S is a solution to (8.9) for $S \cup \{j\}$, so $\kappa_S \geq \kappa_{S \cup \{j\}}$. Therefore, if x is such that $\mathcal{S}(x) \subseteq D$ is incentivizable, meaning $\kappa_{\mathcal{S}(x)} = 1$, then $\kappa_j = 1$ for all $j \in \mathcal{S}(x)$. If $\kappa_j < 1$ for all $j \in D$, then no x such that $\mathcal{S}(x) \subseteq D$ is incentivizable. □

Thus, there exists an incentivizable x such that $\mathcal{S}(x) \subseteq D$ if and only if there is some $j \in D$ such that the agent can be incentivized to invest all of its budget into x_j.

Objectives over effort profiles. In the remainder of this section, we prove Theorem 8.5. Lemma 8.10 shows that if the evaluator wants the agent to only invest effort into a subset D of actions, one solution might be to simply incentivize them to invest all of their effort into a single $j \in D$. However, this might not be a satisfactory solution — the evaluator may want the agent to engage in a diverse set of actions, or to invest at least some amount in each designated action. Thus, the evaluator may have some other objective beyond simply incentivizing the designated actions D.

We formalize this as follows: suppose the evaluator has some objective $g : \mathbb{R}^m \to \mathbb{R}$ over the agent's effort profile x, and wants to pick the x that maximizes g subject to the constraint that x is incentivizable and $\mathcal{S}(x) \subseteq D$. Formally, this is

$$\arg\max_{x \in \mathbb{R}^m} \; g(x) \qquad \text{s.t.} \quad \kappa_{\mathcal{S}(x)} = 1 \qquad \mathcal{S}(x) \subseteq D \tag{8.17}$$

To make this more tractable, we assume that g is concave, as it will in general be hard to optimize arbitrary non-concave functions. We will begin by showing that this optimization problem is feasible when $\kappa_D = 1$, or equivalently, when $D \in \mathcal{F}_D$. We will extend this to show that when $|D|$ is small, (8.17) can can be solved. In general, however, we will show that due to the incentivizability constraint, this is computationally hard.

First, we consider the case where $\kappa_D = 1$. Here, it is possible to find a mechanism to maximize $g(x)$ because any x in the simplex $\{x \mid \sum_{j \in D} x_j = B\}$ is incentivizable by Theorem 8.3. Thus, the evaluator could simply maximize g over this simplex to get some effort profile x^* and find a linear mechanism β to

incentivize x^*. Extending this idea, if $\kappa_D < 1$ but $|D|$ is small, the evaluator can simply enumerate all subsets $S \subseteq D$ such that $\kappa_S = 1$, optimize g over each one separately, and pick the optimal x^* out of all these candidates.

However, in general, it is NP-hard to optimize a number of natural objectives over the set of incentivizable effort profiles if $\kappa_D < 1$. From Theorem 8.3, we know that incentivizable effort profiles x can be described by their support $\mathcal{S}(x)$, which must satisfy $\kappa_{\mathcal{S}(x)} = 1$. The following lemma shows that this constraint on x makes it difficult to optimize even simple functions because the family of sets $\mathcal{F}_D = \{S \subseteq D \mid \kappa_S = 1\}$ can be used to encode the set of independent sets of an arbitrary graph. Using this fact, we can show that there exist concave objectives g that are NP-hard to optimize subject to the incentivizability constraint.

Lemma 8.11. *Given a graph $G = (V, E)$, there exists an effort graph G' and a set of designated actions D such that $S \subseteq D$ is an independent set of G if and only if $\kappa_S = 1$ in G'.*

Proof. We construct a designated action for each $v \in V$, so $D = V$. We also construct an undesirable action for each $e \in E$, so the total number of actions is $m = |V| + |E|$. For ease of indexing, we'll refer to the designated actions as x_v for $v \in V$ and the remaining actions as x_e for $e \in E$.

We construct a feature F_v for each vertex $v \in V$. Then, let $\alpha_{v,v} = 3$ for all $v \in V$ and $\alpha_{e,v} = 2$ for all $v \in V$. For each $e \in E$, this creates the gadget shown in Figure 8.3.

First, we'll show that if $(u, v) \in E$, then any $S \subseteq D$ containing both u and v has $\kappa_S < 1$. Recall the definition of κ_S in (8.9). Consider the solution with $z_u = z_v = \frac{1}{2}$ and $y_e = \frac{2}{3}$. This is feasible, so $\kappa_S \leq \frac{2}{3} < 1$. By the contrapositive,

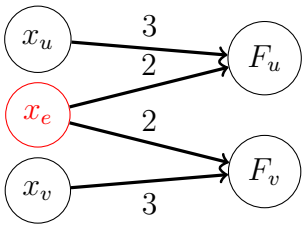


Figure 8.3: Gadget to encode independent sets

if $\kappa_S = 1$ (meaning S is incentivizable), S cannot contain any u, v such that $(u, v) \in E$, meaning S forms an independent set in G.

To show the other direction, consider any independent set S in G. By construction, $S \subseteq D$ because $D = V$. Then, in the optimal solution (y, z) to (8.9), we will show that $y_u = z_u$ for all $u \in S$, meaning $\kappa_S = y^\top \mathbf{1} \geq 1$. To do so, consider the constraint $[\alpha^\top y]_u \geq [\alpha_S^\top z]_u$ for any $u \in S$. This is simply $3y_u + 2\sum_{e=(u,v)\in E} y_e \geq 3z_u$. Because S is an independent set, $z_v = 0$ for any v such that $(u, v) \in E$, so this is the only constraint in which any such y_e appears. Therefore, it is strictly optimal to choose $y_u = z_u$ and $y_e = 0$ for all $e = (u, v) \in E$. As a result, $y_u = z_u$ for all $u \in S$, meaning $\kappa_S \geq \sum_{u\in S} y_u = \sum_{u\in S} z_u \geq 1$ by the constraint $z^\top \mathbf{1}_S \geq 1$. □

Thus, if the evaluator wants to find an incentivizable effort profile x such that $\mathcal{S}(x) \subseteq D$ (the agent only invests effort in designated actions), maximizing an objective like $g(x) = \|x\|_0$ (the number of actions with non-zero effort) is NP-hard, due to a reduction from the maximum independent set problem. Note that $\|x\|_0$ is concave for nonnegative x.

Moreover, other simple and natural objectives are hard to optimize as well. Using a construction similar to the one in Figure 8.3, we can create effort graphs with a set of designated actions D in which $S \subseteq D$ is incentivizable if and only

if $|S| \leq k$, meaning $\|x\|_0 \leq k$. This is known to make optimizing even simple quadratic functions (e.g. $\|\mathcal{A}x - y\|_2$ for some matrix $\mathcal{A}$ and vector y) NP-hard (Natarajan, 1995). In general, then, it is difficult to find the optimal agent effort profile subject to the incentivizability constraint.

8.4 The Structure of the Space of Linear Mechanisms

Thus far, we have seen how to construct linear mechanisms that incentivize particular effort profiles, finding that the mechanisms that do so form a polytope. Suppose that the evaluator doesn't have a particular effort profile that they want to incentivize, but instead wants the agent to only invest effort in a subset of intended actions $D \subseteq [m]$. Generalizing the definition of $\mathcal{L}(x^*)$ as the set of linear mechanisms incentivizing x^*, we define $\mathcal{L}(D)$ to be the set of linear mechanisms incentivizing any x such that $\mathcal{S}(x) \subseteq D$.[3] In the remainder of this section, we give structural results characterizing $\mathcal{L}(D)$, showing that in general it can be highly nonconvex, indicating the richness of the solution space of this problem.

In the simplest case where $|D| = 1$, meaning the evaluator wants to incentivize a single action, we know by (8.6) that $\mathcal{L}(D)$ is simply a polytope. This makes it possible for the evaluator to completely characterize $\mathcal{L}(D)$ and even maximize any concave objective over it.

However, in general, $\mathcal{L}(D)$ can display nonconvexities in several ways. Figure 8.3 gives an example such that if the evaluator only wants to incentivize x_u and x_v, then $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 = 1\}$, meaning β has exactly one nonzero entry. This can be generalized to an example where $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 \leq k\}$ for any k,

[3] With this notation, we could write $\mathcal{L}_j$ as defined in Section 8.2 as $\mathcal{L}(\{j\})$.

which amounts to a nonconvex sparsity constraint.

This form of nonconvexity arises because we're considering mechanisms that incentivize x such that $\mathcal{S}(x) \subseteq D$. In particular, if S and S' are disjoint subsets of D, then we wouldn't necessarily expect the union of $\mathcal{L}(S)$ and $\mathcal{L}(S')$ to be convex. However, we might hope that if each $\mathcal{L}(S)$ for $S \subseteq D$ is convex or can be written as the union of convex sets, then $\mathcal{L}(D)$ could also be written as the union of convex sets.

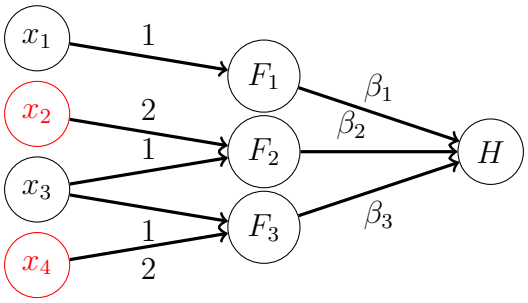


Figure 8.4: Non-convexity of $\mathcal{L}^*(D)$

Unfortunately, this isn't the case. Let $\mathcal{L}^*(D)$ be the set of mechanisms incentivizing x such such that $\mathcal{S}(x) = D$ (as opposed to $\mathcal{S}(x) \subseteq D$). $\mathcal{L}^*(D)$ may still be nonconvex, depending on the particular effort conversion functions $f(\cdot)$. Consider the effort graph shown in Figure 8.4 with $B = 1$, $f_1(y) = f_2(y) = 1 - e^{-y}$ and $f_3(y) = 1 - e^{-2y}$. Let $D = \{1, 3\}$. To incentivize $x_1 > 0$ and $x_3 > 0$ simultaneously with $x_2 = x_4 = 0$, it must be the case that

$$\left.\frac{\partial H}{\partial x_1}\right|_x = \beta_1 f_1'(x_1) = \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x.$$

To incentivize $x_2 = x_4 = 0$, we must also have

$$\left.\frac{\partial H}{\partial x_2}\right|_x = 2\beta_2 f_2'(x_3) \leq \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x$$
$$\left.\frac{\partial H}{\partial x_4}\right|_x = 2\beta_3 f_3'(x_3) \leq \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) = \left.\frac{\partial H}{\partial x_3}\right|_x$$

This is only possible if $\beta_2 f_2'(x_3) = \beta_3 f_3'(x_3)$, meaning β incentivizes x such that $\mathcal{S}(x) = \{1, 3\}$ if and only if

$$\beta_1 f_1'(x_1) = \beta_2 f_2'(x_3) + \beta_3 f_3'(x_3) \tag{8.18}$$

$$\beta_2 f_2'(x_3) = \beta_3 f_3'(x_3) \tag{8.19}$$

Combining (8.18) and (8.19), we get $\beta_1 f_1'(x_1) = 2\beta_2 f_2'(x_3)$, implying

$$\beta_1 f_1'(x_1) = 2\beta_2 f_2'(x_3)$$

$$\beta_1' e^{-x_1} = 2\beta_2 e^{-x_3} \tag{8.20}$$

$$\beta_2 = \frac{\beta_1 e^{x_3 - x_1}}{2} \tag{8.21}$$

Similarly, we can derive

$$\beta_3 = \frac{\beta_1 e^{2x_3 - x_1}}{4} \tag{8.22}$$

We'll show non-convexity by giving two linear mechanisms β and β' that both incentivize an x such that $\mathcal{S}(x) = \{1, 3\}$, but $\beta'' = \frac{1}{2}(\beta + \beta')$ does not incentivize such an x.

Let β and β' incentivize $x = [1/3\ 0\ 2/3\ 0]^\top$ and $x' = [2/3\ 0\ 1/3\ 0]^\top$ respectively. Without loss of generality, we can set $\beta_1 = \beta_1' = 1$. Using (8.21) and (8.22), we get

$$\beta = \begin{bmatrix} 1 & \frac{e^{1/3}}{2} & \frac{e}{4} \end{bmatrix}^\top \qquad\qquad \beta' = \begin{bmatrix} 1 & \frac{e^{-1/3}}{2} & \frac{1}{4} \end{bmatrix}^\top$$

Then, let $\beta'' = \frac{1}{2}(\beta + \beta')$. If β'' incentivizes x^* such that $\mathcal{S}(x^*) = \{1, 3\}$, then

by (8.19), it must be the case that

$$\begin{aligned}
\beta_2'' f_2'(x_3^*) &= \beta_3'' f_3'(x_3^*) \\
\beta_2'' e^{-x_3^*} &= 2\beta_3'' e^{-2x_3^*} \\
e^{x_3^*} &= \frac{2\beta_3''}{\beta_2''} \\
x_3^* &= \log\left(\frac{e+1}{e^{1/3}+e^{-1/3}}\right) \approx 0.566
\end{aligned}$$

On the other hand, by (8.20), we must also have

$$\begin{aligned}
\beta_1'' f_1'(x_1^*) &= 2\beta_2'' e^{-x_3^*} \\
e^{-x_1^*} &= \frac{e^{1/3}+e^{-1/3}}{2} \exp\left(-\log\left(\frac{e+1}{e^{1/3}+e^{-1/3}}\right)\right) \\
e^{-x_1^*} &= \frac{e^{1/3}+e^{-1/3}}{2} \cdot \frac{e^{1/3}+e^{-1/3}}{e+1} \\
x_1^* &= -\log\left(\frac{(e^{1/3}+e^{-1/3})^2}{2(e+1)}\right) \approx 0.511
\end{aligned}$$

Such a solution would fail to respect the budget constraint (since $x_1^* + x_3^* > 1 = B$), meaning β'' cannot incentivize x^* such that $\mathcal{S}(x^*) = \{1, 3\}$. In fact, the above analysis shows that for any x^* incentivized by β'', $\mathcal{S}(x^*)$ must include either 2 or 4 because β'' incentivizes neither $x_1^* = 1$ nor $x_3^* = 1$, meaning the only way to use the entire budget is to set $x_2^* > 0$ or $x_4^* > 0$. Thus, despite the fact that both β and β' incentivize effort profiles with support $\{1, 3\}$, a convex combination of them does not. As a result, the set of linear mechanisms incentivizing a subset of actions may in general exhibit complex structures that don't lend themselves to simple characterization.

We visualize this nonconvexity in Figure 8.5, where for clarity we modify the effort graph in Figure 8.4 by setting $\alpha_{22} = 0$. The yellow region corresponds to (β_2, β_3) values such that $\beta = (1\ \beta_2\ \beta_3)^\top$ incentivizes x such that $\mathcal{S}(x) = \{1, 3\}$. Note that the upper left edge of this region is slightly curved, producing the

non-convexity. As a result, the set of linear mechanisms incentivizing a subset of actions may in general exhibit complex structures that don't lend themselves to simple characterization.

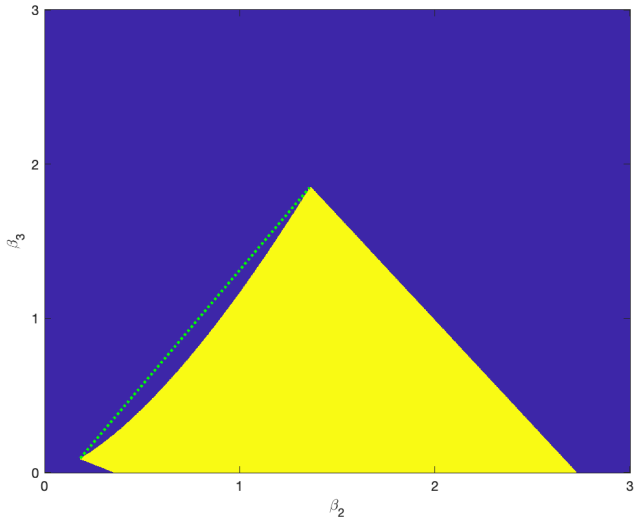


Figure 8.5: Non-convexity in (β_2, β_3) pairs

Implications for optimization. The complexity of $\mathcal{L}(D)$ has immediate hardness implications for optimizing objectives over the space of linear mechanisms. For example, mechanisms that distribute weight on multiple features may be preferable because in practice, measuring multiple distinct features may lead to less noisy evaluations. We might also consider the case where the evaluator has historical data $\mathcal{A} \in \mathbb{R}^{r \times n}$ and $y \in \mathbb{R}^r$, where each row of $\mathcal{A}$ contains the features F of some individual and each entry of y contains their measured outcome of some sort. Then, in the absence of strategic considerations, the evaluator could just choose β that minimizes squared error $\|\mathcal{A}\beta - y\|_2$ between the scores given by the mechanism and the outcomes y in the dataset. As noted above, there are examples for which $\mathcal{L}(D) = \{\beta \mid \|\beta\|_0 \leq k\}$, which is known to make minimizing squared error NP-hard (Natarajan, 1995). However, in the special case

when $D = \{j\}$ (there's only one action the evaluator wants to incentivize), then if $\kappa_j = 1$, the set of linear mechanisms incentivizing $x^* = B \cdot e_j$ is just the convex polytope $\mathcal{L}_j$ defined in (8.6). Thus, it is possible to maximize any concave objective over this set.

8.5 Conclusion

Strategic behavior is a major challenge in designing simple and transparent evaluation mechanisms. In this chapter, we have developed a model in which strategic behavior can be directed toward specified actions through appropriate designs.

Our results leave open a number of interesting questions. All of our analysis has been for the case in which an evaluator designs a mechanism optimized for the parameters of a single agent (or for a group of agents who all have the same parameters). Extending this reasoning to consider the incentives of a heterogeneous group of agents, where the parameters differ across members of the group, is a natural further direction. In addition, we have assumed throughout that agents behave rationally, in that they perfectly optimize their allocation of effort. But it would also be interesting to consider agents with potential biases that reflect human behavioral principles, resulting in sub-optimal behavior that follows certain structured properties. Finally, although we have shown that linear mechanisms suffice whenever a monotone mechanism can incentivize intended behavior, if the output of the mechanisms is constrained in some way (e.g. binary classification), it is an open question to determine what types of mechanisms are appropriate.

CHAPTER 9

ALGORITHMIC MONOCULTURE AND SOCIAL WELFARE

The rise of algorithms used to shape societal choices has been accompanied by concerns over *monoculture*—the notion that choices and preferences will become homogeneous in the face of algorithmic curation. One of many canonical articulations of this concern was expressed in the New York Times by Farhad Manjoo, who wrote, "Despite the barrage of choice, more of us are enjoying more of the same songs, movies and TV shows" (Manjoo, 2019). Because of algorithmic curation, trained on collective social feedback (Salganik et al., 2006), our choices are converging.

When we move from the influence of algorithms on media consumption and entertainment to their influence on high-stakes screening decisions about whom to offer a job or whom to offer a loan, the concerns about algorithmic monoculture become even starker. Even if algorithms are more accurate on a case-by-case basis, a world in which everyone uses the same algorithm is susceptible to correlated failures when the algorithm finds itself in adverse conditions. This type of concern invokes an analogy to agriculture, where monoculture makes crops susceptible to the attack of a single pathogen (Power and Follett, 1987); the analogy has become a mainstay of the computer security literature (Birman and Schneider, 2009), and it has recently become a source of concern about screening decisions for jobs or loans as well. Discussing the post-recession financial system, Citron and Pasquale (2014) write, "Like monocultural-farming technology vulnerable to one unanticipated bug, the converging methods of credit assessment failed spectacularly when macroeconomic conditions changed."

The narrative around algorithmic monoculture thus suggests a trade-off: in

"normal" conditions, a more accurate algorithm will improve the average quality of screening decisions, but when conditions change through an unexpected shock, the results can be dramatically worse. But is this trade-off genuine? In the absence of shocks, does monocultural convergence on a single, more accurate screening algorithm necessarily lead to better average outcomes?

In this chapter, we show that algorithmic monoculture poses risks even in the absence of shocks. We investigate a model involving minimal assumptions, in which two competing firms can either use their own independent heuristics to perform screening decisions or they can use a more accurate algorithm that is accessible to both of them. (Again, we think of screening job applicants or loan applicants as a motivating scenario.) We find that even though it would be rational for each firm in isolation to adopt the algorithm, it is possible for the use of the algorithm by both firms to result in decisions that are *worse* on average. This in turn leads, in the language of game theory, to a type of "Braess' paradox" (Braess, 1968) for screening algorithms: the introduction of a more accurate algorithm can drive the firms into a unique equilibrium that is worse for society than the one that was present before the algorithm existed.

Note that the harm here is to *overall* performance. Another common concern about algorithmic monoculture in screening decisions is the harm it can cause to specific individuals: if all employers or lenders use the same algorithm for their screening decisions, then particular applicants might find themselves locked out of the market when this shared algorithm doesn't like their application for some reason. While this is clearly also a significant concern, our results show that it would be a mistake to view the harm to particular applicants as necessarily balanced against the gains in overall accuracy — rather, it is possi-

ble for algorithmic monoculture to cause harm not just to particular applicants but also to the *average* quality of decisions as well.

Our results thus have a counterintuitive flavor to them: if an algorithm is clearly more accurate than the alternatives when one entity uses it, why does the accuracy become worse than the alternatives when multiple entities use it? The analysis relies on deriving some novel probabilistic properties of rankings, establishing that when we are constructing a ranking from a probability distribution representing a "noisy" version of a true ordering, we can sometimes achieve less error through an incremental construction of the ranking — building it one element at a time — than we can by constructing it in a single draw from the distribution. We now set up the basic model, and then frame the probabilistic questions that underpin its analysis.

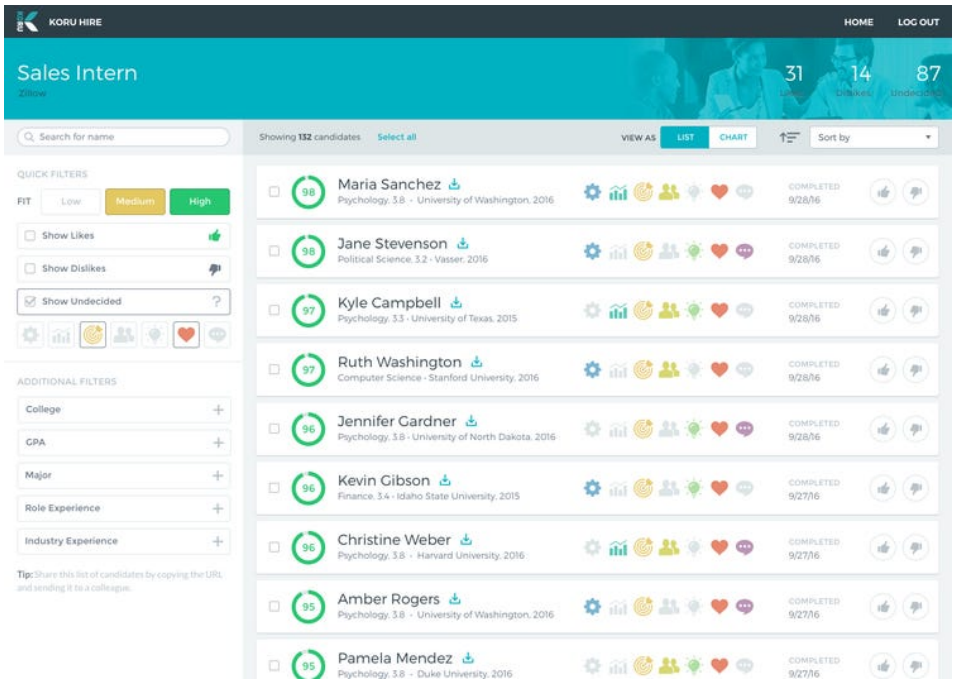


Figure 9.1: Ranking candidates by algorithmically generated scores (Source: https://business.linkedin.com/talent-solutions/blog/recruiting-strategy/2018/the-new-way-companies-are-evaluating-candidates-soft-skills-and-discovering-high-potential-talent)

9.1 Algorithmic hiring as a case study

To instantiate the ideas introduced thus far, we'll focus on the case of algorithmic hiring, where recruiters make decisions based in part on scores or recommendations provided by data-driven algorithms. In this setting, we'll propose and analyze a stylized model of algorithmic hiring with which we can begin to investigate the effects of algorithmic monoculture.

Informally, we can think of a simplified hiring process as follows: rank all of the candidates (see Figure 9.1) and select the first available one. We suppose that each firm has two options to form this ranking: either develop their own, private ranking (which we will refer to as using a "human evaluator"), or use an algorithmically produced ranking. We assume that there is a single vendor of algorithmic rankings, so all firms choosing to use the algorithm receive the same ranking. The firms proceed sequentially, each hiring their favorite remaining candidate according to the ranking they're using—human-generated or algorithmic. Thus, the we can frame the effects of monoculture as follows: are firms better off using the more accurate, common algorithm, or should they instead employ their own less accurate, but private, evaluations?

In what follows, we'll introduce a formal model of evaluation and selection, using it to analyze a setting in which firms seek to hire candidates.

9.1.1 Modeling ranking

More formally, we model the n candidates as having intrinsic values $x_1, \ldots, x_n$, where any employer would derive utility x_i from hiring candidate i. Through-

out the chapter, we assume without loss of generality that $x_1 > x_2 > \cdots > x_n$. These values, however, are unknown to the employer; instead, they must use some noisy procedure to rank the candidates. We model such a procedure as a randomized mechanism $\mathcal{R}$ that takes in the true candidate values and draws a permutation π over those candidates from some distribution. Our main results hold for families of distributions over permutations as defined below:

Definition 9.1 (Noisy permutation family)**.** *A noisy permutation family $\mathcal{F}_\theta$ is a family of distributions over permutations that satisfies the following conditions for any $\theta > 0$ and set of candidates $\mathbf{x}$:*

1. ***(Differentiability)** For any permutation π, $\Pr_{\mathcal{F}_\theta}[\pi]$ is continuous and differentiable in θ.*
2. ***(Asymptotic optimality)** For the true ranking π^*, $\lim_{\theta\to\infty} \Pr_{\mathcal{F}_\theta}[\pi^*] = 1$.*
3. ***(Monotonicity)** For any (possibly empty) $S \subset \mathbf{x}$, let $\pi^{(-S)}$ be the partial ranking produced by removing the items in S from π. Let $\pi_1^{(-S)}$ denote the value of the top-ranked candidate according to $\pi^{(-S)}$. For any $\theta' > \theta$,*

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] \geq \mathbb{E}_{\mathcal{F}_\theta}\left[\pi_1^{(-S)}\right]. \tag{9.1}$$

Moreover, for $S = \emptyset$, (9.1) holds with strict inequality.

θ serves as an "accuracy parameter": for large θ, the noisy ranking converges to the true ranking over candidates. The monotonicity condition states that a higher value of θ leads to a better first choice, even if some of the candidates are removed after ranking. Removal after ranking (as opposed to before) is important because some of the ranking models we will consider later do not satisfy Independence of Irrelevant Alternatives. Examples of noisy permutation families

include Random Utility Models (Thurstone, 1927) and the Mallows Model (Mallows, 1957), both of which we will discuss in detail later.

As an objective function to evaluate the effects of different approaches to ranking and selection, we'll consider each individual employer's utility as well as the sum of employers' utilities. We think of this latter sum as the *social welfare*, since it represents the total quality of the applicants who are hired by any firm. (For example, if all firms deterministically used the correct ranking, then the top applicants would be the ones hired, leading to the highest possible social welfare.)

9.1.2 Modeling selection

Each firm in our model has access to the same underlying pool of n candidates, which they rank using a randomized mechanism $\mathcal{R}$ to get a permutation π as described above. Then, in a random order, each firm hires the highest-ranked remaining candidate according to their ranking. Thus, if two firms both rank candidate i first, only one of them can hire i; the other must hire the next available candidate according to their ranking. In our model, candidates automatically accept the offer they get from a firm. For the sake of simplicity, throughout this chapter, we restrict ourselves to the case where there are two firms hiring one candidate each, although our model readily generalizes to more complex cases.

As described earlier, each firm can choose to use either a private "human evaluator" or an algorithmically generated ranking as its randomized mechanism $\mathcal{R}$. We assume that both candidate mechanisms come from a noisy per-

mutation family F_θ, with differing values of the accuracy parameter θ: human evaluators all have the same accuracy θ_H, and the algorithm has accuracy θ_A. However, while the human evaluator produces a ranking independent of any other firm, the algorithmically generated ranking is identical for all firms who choose to use it. In other words, if two firms choose to use the algorithmically generated ranking, they will both receive the same permutation π.

The choice of which ranking mechanism to use leads to a game-theoretic setting: both firms know the accuracy parameters of the human evaluators (θ_H) and the algorithm (θ_A), and they must decide whether to use a human evaluator or the algorithm. This choice introduces a subtlety: for many ranking models, a firm's rational behavior depends not only on the accuracy of the ranking mechanism, but also on the underlying candidate values $x_1, \ldots, x_n$. Thus, to fully specify a firm's behavior, we assume that $x_1, \ldots, x_n$ are drawn from a known joint distribution $\mathcal{D}$. Our main results will hold for any $\mathcal{D}$, meaning they apply even when the candidate values (but not their identities) are deterministically known.

9.1.3 Stating the main result

Our main result is a pair of intuitive conditions under which a Braess' Paradox-style result occurs—in other words, conditions under which there are accuracy parameters for which both firms rationally choose to use the algorithmic ranking, but social welfare (and each individual firm's utility) would be higher if both firms used independent human evaluators. Recall that the two firms hire in a random order. For a permutation π, let π_i denote the value of the ith-ranked

candidate according to π.

We first state the two conditions, and then the theorem based on them.

Definition 9.2 (Preference for the first position.)**.** *A candidate distribution $\mathcal{D}$ and noisy permutation family $\mathcal{F}_\theta$ exhibits a* preference for the first position *if for all $\theta > 0$, if $\pi, \sigma \sim \mathcal{F}_\theta$,*

$$\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \sigma_1\right] > 0.$$

In other words, for any $\theta > 0$, suppose we draw two permutations π and σ independently from F_θ, and suppose that the first-ranked candidates differ in π and σ. Then the expected value of the first-ranked candidate in π is strictly greater than the expected value of the second-ranked candidate in π.

Definition 9.3 (Preference for weaker competition.)**.** *A candidate distribution $\mathcal{D}$ and noisy permutation family $\mathcal{F}_\theta$, exhibits a* preference for weaker competition *if the following holds: for all $\theta_1 > \theta_2$, $\sigma \sim \mathcal{F}_{\theta_1}$ and $\pi, \tau \sim \mathcal{F}_{\theta_2}$,*

$$\mathbb{E}\left[\pi_1^{(-\{\sigma_1\})}\right] < \mathbb{E}\left[\pi_1^{(-\{\tau_1\})}\right].$$

Intuitively, suppose we have a higher accuracy parameter θ_1 and a lower accuracy parameter $\theta_2 < \theta_1$; we draw a permutation π from $\mathcal{F}_{\theta_2}$; and we then derive two permutations from π: $\pi^{(-\{\sigma_1\})}$ obtained by deleting the first-ranked element of a permutation σ drawn from the more accurate distribution $\mathcal{F}_{\theta_1}$, and $\pi^{(-\{\tau_1\})}$ obtained by deleting the first-ranked element of a permutation τ drawn from the less accurate distribution $\mathcal{F}_{\theta_2}$.

Then the expected value of the first-ranked candidate in $\pi^{(-\{\tau_1\})}$ is strictly greater than the expected value of the first-ranked candidate in $\pi^{(-\{\sigma_1\})}$ — that is, when a random candidate is removed from π, the best remaining candidate is better in expectation when the randomly removed candidate is chosen based on a noisier ranking.

Using these two conditions, we can state our theorem.

Theorem 9.4. *Suppose that a given candidate distribution $\mathcal{D}$ and noisy permutation family $\mathcal{F}_\theta$ satisfy Definition 9.2 (preference for the first position) and Definition 9.3 (preference for weaker competition).*

Then, for any θ_H, there exists $\theta_A > \theta_H$ such that using the algorithmic ranking is a strictly dominant strategy for both firms, but social welfare would be higher if both firms used human evaluators.

9.1.4 A Preference for Independence

Before we prove Theorem 9.4, we provide some intuition for the two conditions in Definitions 9.2 and 9.3. The second condition essentially says that it is better to have a worse competitor: the firm randomly selected to hire second is better off if the firm that hires first uses a less accurate ranking (in this case, a human evaluator instead of the algorithmic ranking).

The first condition states that when two identically distributed permutations disagree on their first element, the first-ranked candidate according to either permutation is still better, in expectation, than the second-ranked candidate according to either permutation. In what follows, we'll demonstrate that this con-

dition implies that firms in our model rationally prefer to make decisions using independent (but equally accurate) rankings.

To do so, we need to introduce some notation. Recall that the two firms hire in a random order. Given a candidate distribution $\mathcal{D}$, let $U_s(\theta_A, \theta_H)$ denote the expected utility of the first firm to hire a candidate when using ranking s, where $s \in \{A, H\}$ is either the algorithmic ranking or the ranking generated by a human evaluator respectively. Similarly, let $U_{s_1 s_2}(\theta_A, \theta_H)$ be the expected utility of the second firm to hire given that the first firm used strategy s_1 and the second firm uses strategy s_2, where again $s_1, s_2 \in \{A, H\}$. Finally, let $\pi, \sigma \sim \mathcal{F}_\theta$.

In what follows, we will show that for any θ,

$$\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \sigma_1\right] > 0 \iff U_{AH}(\theta, \theta) > U_{AA}(\theta, \theta). \tag{9.2}$$

In other words, whenever a ranking model meets Definition 9.2, the firm chosen to select second will prefer to use an *independent* ranking mechanism from it's competitor, given that the ranking mechanisms are equally accurate.

First, we can write

$$\begin{aligned} U_{AH}(\theta_A, \theta_H) &= \mathbb{E}\left[\pi_1 \cdot \mathbf{1}_{\{\pi_1 \neq \sigma_1\}} + \pi_2 \cdot \mathbf{1}_{\{\pi_1 = \sigma_1\}}\right] \\ U_{AA}(\theta_A, \theta_H) &= \mathbb{E}\left[\sigma_2\right] \\ &= \mathbb{E}\left[\sigma_2 \cdot \mathbf{1}_{\{\pi_1 \neq \sigma_1\}} + \sigma_2 \cdot \mathbf{1}_{\{\pi_1 = \sigma_1\}}\right] \end{aligned}$$

Thus,

$$U_{AH}(\theta_A, \theta_H) - U_{AA}(\theta_A, \theta_H) \quad = \mathbb{E}\left[(\pi_1 - \sigma_2) \cdot \mathbf{1}_{\{\pi_1 \neq \sigma_1\}} + (\pi_2 - \sigma_2) \cdot \mathbf{1}_{\{\pi_1 = \sigma_1\}}\right].$$

Conditioned on either $\pi_1 = \sigma_1$ or $\pi_1 \neq \sigma_1$, π_2 and σ_2 are identically distributed and therefore have equal expectations. As a result,

$$U_{AH}(\theta_A, \theta_H) - U_{AA}(\theta_A, \theta_H) = \mathbb{E}\left[(\pi_1 - \pi_2) \cdot \mathbf{1}_{\{\pi_1 \neq \sigma_1\}}\right], \tag{9.3}$$

which implies (9.2). Thus, whenever a ranking model meets Definition 9.2, firms rationally prefer independent assessments, all else equal.

To provide some intuition for what this preference for independence entails, consider a setting where a hiring committee seeks to hire two candidates. They meet, produce a ranking σ, and hire σ_1 (the best candidate according to σ). Suppose they have the option to either hire σ_2 or reconvene the next day to form an independent ranking π and hire the best remaining candidate according to π; which option should they choose? It's not immediately clear why one option should be better than the other. However, whenever Definition 9.2 is met, the committee should prefer to reconvene and make their second hire according to a new ranking π. After proving Theorem 9.4, we will provide natural ranking models that meet Definition 9.2, implying that under these ranking models, independent re-ranking can be beneficial.

9.1.5 Proving Theorem 9.4

With this intuition, we are ready to prove Theorem 9.4.

Proof of Theorem 9.4. For given values of θ_A and θ_H, using the algorithmic ranking is a strictly dominant strategy as long as

$$U_A(\theta_A, \theta_H) + U_{AA}(\theta_A, \theta_H) > U_H(\theta_A, \theta_H) + U_{AH}(\theta_A, \theta_H) \tag{9.4}$$

$$U_A(\theta_A, \theta_H) + U_{HA}(\theta_A, \theta_H) > U_H(\theta_A, \theta_H) + U_{HH}(\theta_A, \theta_H) \tag{9.5}$$

Note that (9.5) is always true for $\theta_A > \theta_H$ by the monotonicity assumption on $\mathcal{F}_\theta$: $U_A(\theta_A, \theta_H) \geq U_H(\theta_A, \theta_H)$ because a more accurate ranking produces a top-ranked candidate with higher expected value, and $U_{HA}(\theta_A, \theta_H) \geq U_{HH}(\theta_A, \theta_H)$

because this holds even conditioned on removing any candidate from the pool (in this case, the candidate randomly selected by the firm that hires first). Crucially, in (9.5), the first firm's random selection is independent from the second firm's selection; the same logic could not be used to argue that (9.4) always holds for $\theta_A \geq \theta_H$. Moreover, when $\theta_A > \theta_H$, $U_A(\theta_A, \theta_H) > U_H(\theta_A, \theta_H)$ by the monotonicity assumption, meaning (9.5) holds.

Let $W_{s_1 s_2}(\theta_A, \theta_H)$ denote social welfare when the two firms employ strategies $s_1, s_2 \in \{A, H\}$. Then, when both firms use the algorithmic ranking, social welfare is

$$W_{AA}(\theta_A, \theta_H) = U_A(\theta_A, \theta_H) + U_{AA}(\theta_A, \theta_H).$$

By (9.2), Definition 9.2 implies that for any θ, $U_{AA}(\theta, \theta) < U_{AH}(\theta, \theta)$, implying

$$U_A(\theta_H, \theta_H) + U_{AA}(\theta_H, \theta_H) < U_H(\theta_H, \theta_H) + U_{AH}(\theta_H, \theta_H).$$

However, by the optimality assumption on $\mathcal{F}_\theta$ in Definition 9.1, for sufficiently large $\hat{\theta}_A$,

$$U_A(\hat{\theta}_A, \theta_H) + U_{AA}(\hat{\theta}_A, \theta_H) > U_H(\hat{\theta}_A, \theta_H) + U_{AH}(\hat{\theta}_A, \theta_H).$$

Note that $U_{s_1}(\theta_A, \theta_H)$ and $U_{s_1 s_2}(\theta_A, \theta_H)$ are continuous with respect to θ_A for any $s_1, s_2 \in \{A, H\}$ since they are expectations over discrete distributions with probabilities that are by assumption differentiable with respect to θ_A. Therefore, by the Differentiability assumption on $\mathcal{F}_\theta$ from Definition 9.1, there is some $\theta_A^* > \theta_H$ such that

$$U_A(\theta_A^*, \theta_H) + U_{AA}(\theta_A^*, \theta_H) = U_H(\theta_A^*, \theta_H) + U_{AH}(\theta_A^*, \theta_H), \tag{9.6}$$

i.e., given that its competitor uses the algorithmic ranking, a firm is indifferent between the two strategies. For such θ_A^*, using the algorithmic ranking is still a

weakly dominant strategy. By definition of W_{AA},

$$W_{AA}(\theta_A^*, \theta_H) = U_H(\theta_A^*, \theta_H) + U_{AH}(\theta_A^*, \theta_H).$$

If both firms had instead used human evaluators, social welfare would be

$$W_{HH}(\theta_A^*, \theta_H) = U_H(\theta_A^*, \theta_H) + U_{HH}(\theta_A^*, \theta_H).$$

By Definition 9.3, for $\sigma \sim \mathcal{F}_{\theta_{A^*}}$ and $\pi, \tau \sim \mathcal{F}_{\theta_H}$,

$$\mathbb{E}\left[\pi_1^{(-\{\sigma_1\})}\right] < \mathbb{E}\left[\pi_1^{(-\{\tau_1\})}\right].$$

Note that

$$\begin{aligned} U_{AH}(\theta_A^*, \theta_H) &= \mathbb{E}\left[\pi_1^{(-\{\sigma_1\})}\right] \\ U_{HH}(\theta_A^*, \theta_H) &= \mathbb{E}\left[\pi_1^{(-\{\tau_1\})}\right] \end{aligned}$$

Thus, Definition 9.3 implies that for $\theta_{A^*} > \theta_H$, $U_{HH}(\theta_A^*, \theta_H) > U_{AH}(\theta_A^*, \theta_H)$. As a result for $\theta_{A^*} > \theta_H$, using the algorithmic ranking is a weakly dominant strategy, but

$$\begin{aligned} W_{HH}(\theta_A^*, \theta_H) &= U_H(\theta_A^*, \theta_H) + U_{HH}(\theta_A^*, \theta_H) \\ &> U_H(\theta_A^*, \theta_H) + U_{AH}(\theta_A^*, \theta_H) \\ &= U_A(\theta_A^*, \theta_H) + U_{AA}(\theta_A^*, \theta_H) \\ &= W_{AA}(\theta_A^*, \theta_H), \end{aligned}$$

meaning social welfare would have been higher had both firms used human evaluators.

We can show that this effect persists for a value θ'_A such that using the algorithmic ranking is a *strictly* dominant strategy. Intuitively, this is simply by

slightly increasing θ_A^* so the algorithmic ranking is strictly dominant. For fixed θ_H, define

$$f(\theta_A) = U_A(\theta_A, \theta_H) + U_{AA}(\theta_A, \theta_H)$$
$$g(\theta_A) = U_H(\theta_A, \theta_H) + U_{AH}(\theta_A, \theta_H)$$
$$h(\theta_A) = U_H(\theta_A, \theta_H) + U_{HH}(\theta_A, \theta_H)$$

Because (9.5) always holds for $\theta_A > \theta_H$, it suffices to show that there exists θ'_A such that $g(\theta'_A) < f(\theta'_A) < h(\theta'_A)$. This is because $g(\theta'_A) < f(\theta'_A)$ is equivalent to (9.4) and $f(\theta'_A) < h(\theta'_A)$ is equivalent to $W_{AA}(\theta'_A, \theta_H) < W_{HH}(\theta'_A, \theta_H)$.

First, note that $h(\theta_A)$ is a constant, and by Definition 9.3, $g(\theta_A) < h(\theta_A)$ for all $\theta_A > \theta_H$. By the optimality assumption of Definition 9.1, there exists sufficiently large $\hat{\theta}_A$ such that $f(\hat{\theta}_A) > g(\hat{\theta}_A)$. Recall that by definition of θ_A^*, $f(\theta_A^*) = g(\theta_A^*)$. Both f and g are continuous by the Differentiability assumption in Definition 9.1. Thus, there must exist some $\theta'_A > \theta_A^*$ such that $g(\theta'_A) < f(\theta'_A) < h(\theta'_A)$. This means that for θ'_A, using the algorithmic ranking is a strictly dominant strategy, but social welfare would still be larger if both firms used human evaluators. □

9.2 Instantiating with Ranking Models

Thus far, we have described a general set of conditions under which algorithmic monoculture can lead to a reduction in social welfare. Under which ranking models do these conditions hold? In the remainder of this chapter, we instantiate the model with two well-studied ranking models: Random Utility Models (RUMs) (Thurstone, 1927) and the Mallows Model (Mallows, 1957). While

RUMs do not always satisfy Definitions 9.2 and 9.3, they do under some realistic parameterizations, regardless of the candidate distribution $\mathcal{D}$. Under the Mallows Model, Definitions 9.2 and 9.3 are always met, meaning that for any candidate distribution $\mathcal{D}$ and human evaluator accuracy θ_H, there exists an accuracy parameter θ_A such that a common algorithmic ranking with accuracy θ_A decreases social welfare.

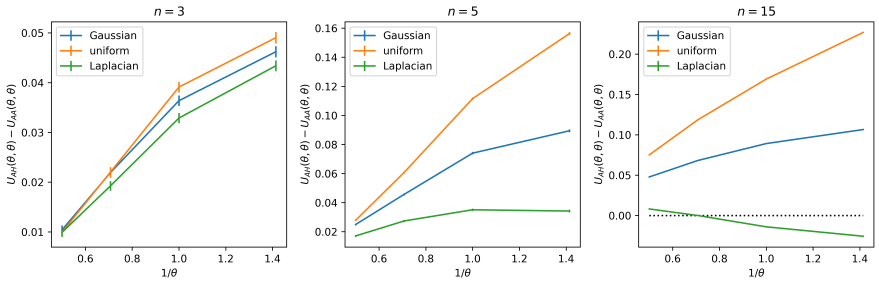


Figure 9.2: $U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta)$ for three noise models with n candidates whose utilities are drawn from a uniform distribution with unit variance for $n = 3$, $n = 5$, and $n = 15$. Note that for $n = 15$, $U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta) < 0$ for Laplacian noise, meaning Definition 9.2 is not met.

9.2.1 Random Utility Models

In Random Utility Models, the underlying candidate values x_i are perturbed by independent and identically distributed noise $\varepsilon_i \sim \mathcal{E}$, and the perturbed values are ranked to produce π. Originally conceived in the psychology literature (Thurstone, 1927), this model has been well-studied over nearly a century, (Daniels, 1950; Block and Marschak, 1960; Joe, 2000; Yellott Jr, 1977; Manski, 1977; Strauss, 1979), including more recently in the computer science and machine learning literature (Azari Soufiani et al., 2012, 2013; Ragain and Ugander, 2016; Zhao et al., 2018; Makhijani and Ugander, 2019).

First, we must define a family of RUMs that satisfies the conditions of Def-

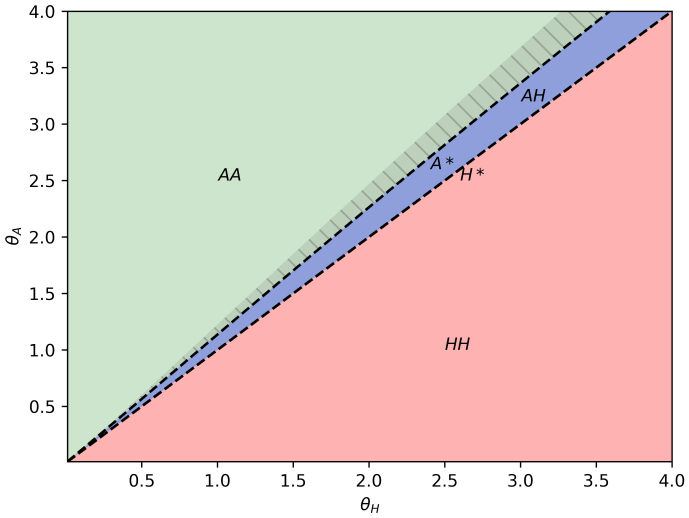


Figure 9.3: Regions for different equilibria. When human evaluators are more accurate than the algorithm, both firms decide to employ humans (HH). When the algorithm is significantly more accurate, both firms use the algorithm (AA). When the algorithm is slightly more accurate than human evaluators, two possible equilibria exist: (1) one firm uses the algorithm and the other employs a human (AH), or (2) both decide whether to use the algorithm with some probability p. The shaded portion of the green AA region depicts where social welfare is smaller at the AA equilibrium than it would be if both firms used human evaluators.

inition 9.1. Assume without loss of generality that the noise distribution $\mathcal{E}$ has unit variance. Then, consider the family of RUMs parameterized by θ in which candidates are ranked according to $x_i + \frac{\varepsilon_i}{\theta}$. By this definition, the standard deviation of the noise for a particular value of θ is simply $1/\theta$. Intuitively, larger values of θ reduce the effect of the noise, making the ranking more accurate. In Theorem F.1 in Appendix F.1, we show as long as the noise distribution $\mathcal{E}$ has positive support on $(-\infty, \infty)$, this definition of $\mathcal{F}_\theta$ meets the differentiability, asymptotic optimality, and monotonicity conditions in Definition 9.1. For distributions with finite support, many of our results can be generalized by relaxing strict inequalities in Definition 9.1 and Theorem 9.4 to weak inequalities.

Because RUMs are notoriously difficult to work with analytically, we restrict ourselves to the case where $n = 3$, i.e., there are 3 candidates. Under this restriction, we can show that for Gaussian and Laplacian noise distributions, Definition 9.2 and 9.3 — the two conditions of Theorem 9.4 — are met, regardless of the candidate distribution $\mathcal{D}$. We defer the proof to Appendix F.3.

Theorem 9.5. *Let $\mathcal{F}_\theta$ be the family of RUMs with either Gaussian or Laplacian noise with standard deviation $1/\theta$. Then, for any candidate distribution $\mathcal{D}$ over 3 candidates, the conditions of Theorem 9.4 are satisfied.*

It might be tempting to generalize Theorem 9.5 to other distributions and more candidates; however, certain noise and candidate distributions violate the conditions of Theorem 9.4. Even for 3-candidate RUMs, there exist distributions for which each of the conditions is violated; we provide such examples in Appendix F.2.

Moreover, while Gaussian and Laplacian distributions provably meet Definitions 9.2 and 9.3 with only 3 candidates, this doesn't necessarily extend to larger candidate sets. Figure 9.2 shows that Definition 9.2 can be violated under a particular candidate distribution $\mathcal{D}$ for Laplacian noise with 15 candidates. This challenges the intuition that independence is preferable—under some conditions, it can actually better in expectation for a firm to use the *same* algorithmic ranking as its competitor, even if an independent human evaluator is equally accurate overall. Unlike Theorem 9.5, which applies for *any* candidate distribution $\mathcal{D}$, certain noise models may violate Definition 9.2 only for particular $\mathcal{D}$. It is an open question as to whether Theorem 9.5 can be extended to larger numbers of candidates under Gaussian noise.

Finally, there exist noise distributions that violate Definition 9.2 for any can-

didate distribution $\mathcal{D}$. In particular, the RUM family defined by the Gumbel distribution is well-known to be equivalent to the Plackett-Luce model of ranking, which is generated by sequentially selecting candidate i with probability

$$\frac{\exp(\theta x_i)}{\sum_{j \in S} \exp(\theta x_j)}, \tag{9.7}$$

where S is the set of remaining candidates (Luce, 1959; Block and Marschak, 1960). Under the Plackett-Luce model, for any θ, $U_{AH}(\theta, \theta) = U_{AA}(\theta, \theta)$. To see this, suppose the firm that hires first selects candidate i^*. Then, the firm that hires second gets each candidate i with probability given by (9.7) with $S = \{1, \ldots, n\} \backslash i^*$. As a result, by (9.3), if $\pi, \sigma \sim \mathcal{F}_\theta$,

$$\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \sigma_1\right] = 0$$

for any candidate distribution $\mathcal{D}$, meaning the Plackett-Luce model never meets Definition 9.2. Thus, under the Plackett-Luce model, monoculture has no effect—the optimal strategy is always to use the best available ranking, regardless of competitors' strategies.

Given the analytic intractability of most RUMs, it might appear that testing the conditions of Theorem 9.4, especially for a particular noise and candidate distributions, may not be possible; however, they can be efficiently tested via simulation: as long as the noise distribution $\mathcal{E}$ and the candidate distribution $\mathcal{D}$ can be sampled from, it is possible to test whether the conditions of Theorem 9.4 are satisfied. Thus, even if the conditions of Theorem 9.4 are not met for *every* candidate distribution $\mathcal{D}$, it is possible to efficiently determine whether they are met for any *particular* $\mathcal{D}$.

It is also interesting to ask about the magnitude of the negative impact produced by monoculture. Our model allows for the qualities of candidates to be ei-

ther positive or negative (capturing the fact that a worker's productivity can be either more or less than their cost to the firm in wages); using this, we can construct instances of the model in which the optimal social welfare is positive but the welfare under the (unique) monocultural equilibrium implied by Theorem 1 is negative. This is a strong type of negative result, in which sub-optimality reverses the sign of the objective function, and it means that in general we cannot compare the optimum and equilibrium by taking a ratio of two non-negative quantities, as is standard in *Price of Anarchy* results. However, as a future direction, it would be interesting to explore such Price of Anarchy bounds in special cases of the problem where structural assumptions on the input are sufficient to guarantee that the welfare at both the social optimum and the equilibrium are non-negative. As one simple example, if the qualities for three candidates are drawn independently from a uniform distribution centered at 0, and the noise distribution is Gaussian, then there exist parameters $\theta_A > \theta_H$ such that expected social welfare at the equilibrium where both firm use the algorithmic ranking is non-negative, and approximately 4% less than it would be had both firms used human evaluators instead.

9.2.2 The Mallows Model

The Mallows Model also appears frequently in the ranking literature (Das and Li, 2014; Lu and Boutilier, 2011), and is much more analytically tractable than RUMs. Under the Mallows Model, the likelihood of a permutation is related to its distance from the true ranking π^*:

$$\Pr[\pi] = \frac{1}{Z}\phi^{-d(\pi,\pi^*)}, \tag{9.8}$$

where Z is a normalizing constant. In this model, $\phi > 1$ is the accuracy parameter: the larger ϕ is, the more likely the ranking procedure is to output a ranking π that is close to the true ranking r. To instantiate this model, we need a notion of distance $d(\cdot, \cdot)$ over permutations. For this, we'll use Kendall tau distance, another standard notion in the literature, which is simply the number of pairs of elements in π that are incorrectly ordered (Kendall, 1938). In Appendix F.4, we verify that the family of distributions $\mathcal{F}_\theta$ given by the Mallows Model satisfies Definition 9.1, defining $\theta = \phi - 1$ (for consistency, so θ is well-defined on $(0, \infty)$).

In contrast to RUMs, the Mallows Model always satisfies the conditions of Theorem 9.4 for any candidate distribution $\mathcal{D}$, which we prove in Appendix F.5.

Theorem 9.6. *Let $\mathcal{F}_\theta$ be the family of Mallows Model distributions with parameter $\theta = \phi - 1$. Then, for any candidate distribution $\mathcal{D}$, the conditions of Theorem 9.4 are satisfied.*

Figure 9.3 characterizes firms' rational behavior at equilibrium in the (θ_H, θ_A) plane under the Mallows Model. The decrease in social welfare found in Theorem 9.6 is depicted by the shaded portion of the green region labeled AA, where social welfare would be higher if both firms used human evaluators.

While the result of Theorem 9.6 is certainly stronger than that of Theorem 9.5, in that it applies to all instances of the Mallows Model without restrictions, it should be interpreted with some caution. The Mallows Model does not depend on the underlying candidate values, so according to this model, monoculture can produce arbitrarily large negative effects. While insensitivity to candidate values may not necessarily be reasonable in practice, our results hold for any candidate distribution $\mathcal{D}$. Thus, to the extent that the Mallows Model can reasonably approximate ranking in particular contexts, our results imply that

monoculture can have negative welfare effects.

9.3 Models with Multiple Firms

Our main focus in this chapter has been on models with two competing firms. However, it is also interesting to consider the case of more than two firms; we will see that the complex and sometimes counterintuitive effects that we found in the two-firm case are further enriched by additional phenomena. Primarily, we will present the result of computational experiments with the model, exposing some fundamental structural properties in the multi-firm problem for which a formal analysis remains an intriguing open problem. For concreteness, we will focus on a model in which rankings are drawn from the Mallows model. As before, each firm must choose to order candidates according to either an independent, human-produced ranking or an algorithmic ranking common to all firms who choose it. These rankings come from instances of the Mallows model with accuracy parameters ϕ_H and ϕ_A respectively as defined in (9.8).

Braess' Paradox for $k > 2$ firms. First, we ask whether the Braess' Paradox effect persists with $k > 2$ firms. We find that it is possible to construct instances of the problem with $k > 2$ for which Braess' Paradox occurs — using an algorithmic evaluation is a dominant strategy, but social welfare would be higher if all firms used human evaluators instead. Under the Mallows Model, suppose $n = 4$, $k = 3$, $\phi_A = 2$, $\phi_H = 1.75$, and candidate qualities are drawn from a uniform distribution on $[0, 1]$. We find via computation that at equilibrium, each firm will rationally decide to use the algorithmic evaluator and experience

utility $\approx .551$, but if all firms instead used human evaluators, they would experience utility $\approx .552$. Thus, Braess' Paradox can occur for $k > 2$ firms. Proving a generalization of Theorem 9.6, to show that Braess' Paradox can occur for any candidate distribution $\mathcal{D}$ and any value of ϕ_H for $k > 2$ firms remains an open question.

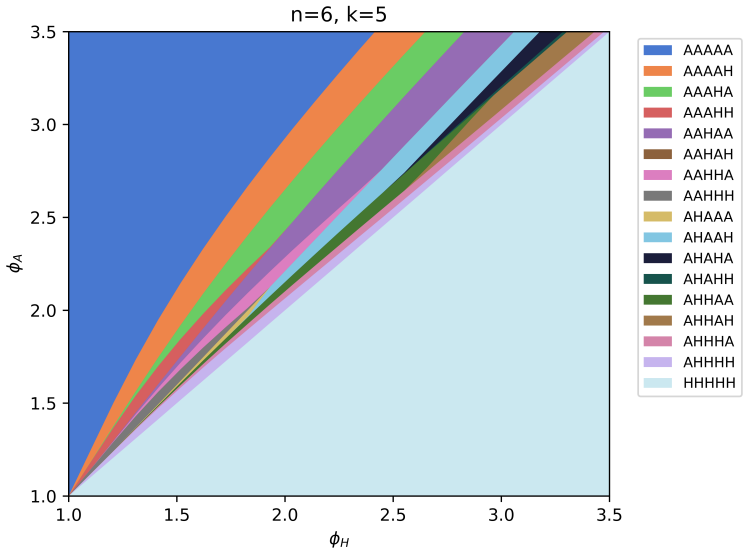


Figure 9.4: Regions for different optimal strategy profiles, where each strategy profile is a sequence of 'A' and 'H' representing the optimal strategies of each firm sequentially. For this plot, there are 5 firms ($k = 5$) and 6 candidates ($n = 6$) whose values are drawn from a uniform distribution. Note that when ϕ_A is much larger than ϕ_H, all firms use the algorithmic ranking, but when ϕ_A is only slightly larger than ϕ_H, only the first firm uses the algorithmic ranking.

Sequential decision-making. Since the equilibrium behaviors we are studying take place in a model where firms make decisions in a random order, a crucial first step is to characterize firms' optimal behavior when making decisions *sequentially*—that is, when firms hire in a fixed, known order as opposed to a random order. In this context, consider the rational behavior of each firm: given a distribution over candidate values, which ranking should each firm use? Clearly, the first firm to make a selection should use the more accurate ranking

mechanism; however, as shown previously, subsequent firms' decisions are less clear-cut. For a fixed number of firms, number of candidates, and distribution over candidate values, we can explore the firms' optimal strategies over the possible space of (ϕ_H, ϕ_A) values.

An optimal choice of strategies for the k firms moving sequentially can be written as a sequence of length k made up of the symbols A and H; the i^{th} term in the sequence is equal to A if the i^{th} firm to move sequentially uses the algorithm as its optimal strategy (given the choices of the previous $i-1$ firms), and it is equal to H if the i^{th} firm uses an independent human evaluation. We can therefore represent the choice of optimal strategies, as the parameters (ϕ_H, ϕ_A) vary, by a labeling of the (ϕ_H, ϕ_A)-plane: we label each point (ϕ_H, ϕ_A) with the length-k sequence that specifies the optimal sequence of strategies.

We can make the following initial formal observation about these optimal sequences:

Theorem 9.7. *When $\phi_H \geq \phi_A$, one optimal sequence is for all firms to choose H. When $\phi_H > \phi_A$, the unique optimal sequence is for all firms to choose H.*

We prove this formally in Appendix F.6.1, but we provide a sketch here. When $\phi_H \geq \phi_A$, the first firm to move in sequence will simply use the more accurate strategy, and hence will choose H. Now, proceeding by induction, suppose that the first i firms have all chosen H, and consider the $(i+1)^{\text{st}}$ firm to move in sequence. Regardless of whether this firm chooses A or H, it will be making a selection that is independent of the previous i selections, and hence it is optimal for it to choose H as well. Hence, by induction, it is an optimal solution for all firms to choose H when $\phi_H \geq \phi_A$. (This argument, slightly adapted, also directly establishes that it is uniquely optimal for all firms to choose H when

$\phi_H > \phi_A$.)

Beyond this observation, if we wish to extend to the case when $\phi_A > \phi_H$, the mathematical analysis of this multi-firm model remains an open question; but it is possible to determine optimal strategies computationally for each choice of (ϕ_H, ϕ_A), and then to look at how these strategies vary over the (ϕ_H, ϕ_A)-plane. Figure 9.4 shows the result of doing this — producing a labeling of the (ϕ_H, ϕ_A)-plane as described above — for $k = 5$ firms and $n = 6$ candidates, with the values of the candidates drawn from a uniform distribution.

We observe a number of interesting phenomena from this labeling of the plane. First, the region where $\phi_H \geq \phi_A$ is labeled with the all-H sequence, reflecting the argument above; for the half-plane $\phi_A > \phi_H$, on the other hand, all optimal sequences begin with A, since it is always optimal for the first firm to use the more accurate method. The labeling of the half-plane $\phi_A > \phi_H$ becomes quite complex; in principle, any sequence over the binary alphabet $\{A, H\}$ that begins with A could be possible, and in fact we see that all $2^4 = 16$ of these sequences appear as labels in some portion of the plane. This means that the sequential choice of optimal strategies for the firms can display arbitrary non-monotonicities in the choice of algorithmic or human decisions, with firms alternating between them; for example, even after the first firm chooses A and the second chooses H, the third may choose A or H depending on the values (ϕ_H, ϕ_A).

The boundaries of the regions labeled by different optimal sequences are similarly complex; some of the regions (such as $AAAHH$) appear to be bounded, while others (such as $AHAHA$ and $AHHAH$) appear to only emerge for sufficiently large values of ϕ_H.

Perhaps the most intriguing observation about the arrangement of regions is the following. Suppose we think of the sequences of symbols over $\{A, H\}$ as binary representations of numbers, with A corresponding to the binary digit 1 and H correspnding to the binary digit 0. (Thus, for example, $AAAHH$ would correspond to the number $16 + 8 + 4 = 28$, while $AHAHA$ would correspond to the number $16 + 4 + 1 = 21$.) The observation is then the following: if we choose any vertical line $\phi_H = x$ (for a fixed x), and we follow it upward in the plane, we encounter regions in increasing order of the numbers corresponding to their labels, in this binary representation. (First $HHHHH$, then $AHHHH$, then $AHHHA$, then $AHHAH$, and so forth.)

We do not know a proof for this fact, or how generally it holds, but we can verify it computationally for the regions of the (ϕ_H, ϕ_A)-plane mapped out in Figure 9.4, as well as similar computational experiments not shown here for other choices of k and n. This binary-counter property suggests a rich body of additional structure to the optimal strategies in the k-firm case, and we leave it as an open question to analyze this structure mathematically.

9.4 Conclusion

Concerns about monoculture in the use of algorithms have focused on the danger of unexpected, correlated shocks, and on the harm to particular individuals who may fare poorly under the algorithm's decision. Our work here shows that concerns about algorithmic monoculture are in a sense more fundamental, in that it is possible for monoculture to cause decisions of globally lower average quality, even in the absence of shocks. In addition to telling us something

about the pervasiveness of the phenomenon, it also suggests that it might be difficult to notice its negative effects even while they're occurring — these effects can persist at low levels even without a shock-like disruption to call our attention to them. Our results also make clear that algorithmic monoculture in decision-making doesn't always lead to adverse outcomes; rather, we given natural conditions under which such outcomes become possible, and show that these conditions hold in a wide range of standard models.

Our results suggest a number of natural directions for further work. To begin with, we have noted earlier in the chapter that it would be interesting to give more comprehensive quantitative bounds on the magnitude of monoculture's possible negative effects in decisions such as hiring — how much worse can the quality of candidates be when selected with an equilibrium strategy involving shared algorithms than with a socially optimal one? In formulating such questions, it will be important to take into account how the noise model for rankings relates to the numerical qualities of the candidates.

We have also focused here on the case of two firms and a single shared algorithm that is available to both. It would be natural to consider generalizations involving more firms and potentially more algorithms as well. With more algorithms, we might see solutions in which firms cluster around different algorithms of varying accuracies, as they balance the level of accuracy and the amount of correlation in their decisions. It would also be interesting to explore the ways in which correlations in firms' decisions can be decomposed into constituent parts, such as the use of standardized tests that form input features for algorithms, and how quantifying these forms of correlation might help firms assess their decisions.

Finally, it will be interesting to consider how these types of results apply to further domains. While the analysis presented here illustrates the consequences of monoculture as applied to algorithmic hiring, our findings have potential implications in a broader range of settings. Algorithmic monoculture not only leads to a lack of heterogeneity in decision-making; by allowing valuable options to slip through the cracks — be they job candidates, potential hit songs, or budding entrepreneurs — it reduces total social welfare, even when the individual decisions are more accurate on a case-by-case basis. These concerns extend beyond the use of algorithms; whenever decision-makers rely on identical or highly correlated evaluations, they miss out on hidden gems, and in this way diminish the overall quality of their decisions.

Part IV

Application Domains

CHAPTER 10

OVERVIEW OF Part IV

In Part IV, we apply some of the conceptual principles derived earlier in this thesis to domains of social interest. We consider algorithmic hiring and explanations in credit scoring.

Chapter 11 considers *algorithmic pre-employment assessments*, which are used by many firms to determine which job applicants to interview. Taking a sample of 18 vendors developing these tools, we empirically determine how they conceive of and mitigate issues of bias and discrimination in their products. Based on these findings, we analyze their practices from both computer science and legal perspectives. We conclude with policy recommendations to promote more robust protections against discrimination going forwards.

In Chapter 12, we explore the use of algorithmically-generated explanations for adverse credit decisions. Traditionally, firms have provided *feature-highlighting* explanations for this purpose. More recently, there have been proposals to use *counterfactual explanations* for this task. We compare and contrast these two styles of explanations, finding that neither can be automated without a firm making a number of subjective choices, each of which transfers power from the decision subject to the decision-maker.

CHAPTER 11

MITIGATING BIAS IN ALGORITHMIC DECISION-MAKING: EVALUATING CLAIMS AND PRACTICES

The study of algorithmic bias and fairness in machine learning has quickly matured into a field of study in its own right, delivering a wide range of formal definitions and quantitative metrics. As industry takes up these tools and accompanying terminology, promises of eliminating algorithmic bias using computational methods have begun to proliferate. In some cases, however, rather than forcing precision and specificity, the existence of formal definitions and metrics has had the paradoxical result of giving undue credence to vague claims about "de-biasing" and "fairness."

In this chapter, we use algorithmic pre-employment assessment as a case study to show how formal definitions of fairness allow us to ask focused questions about the meaning of "fair" and "unbiased" models. The hiring domain makes for an effective case study because of both its prevalence and its long history of bias. We know from decades of audit studies that employers tend to discriminate against women and ethnic minorities (Bertrand and Mullainathan, 2004; Bendick et al., 1997; Bendick and Nunes, 2012; Johnson et al., 2016), and a recent meta-analysis suggests that little has improved over the past 25 years (Quillian et al., 2017). Citing evidence that algorithms may help reduce human biases (Houser, 2019; Kleinberg et al., 2018), advocates argue for the adoption of algorithmic techniques in hiring (Chamorro-Prezumic and Akhtar, 2019; Cowgill, 2018), with a variety of computational metrics proposed to identify and prevent unfair behavior (Feldman et al., 2015). But to date, little is known about how these methods are used in practice.

One of the biggest obstacles to empirically characterizing industry practices is the lack of publicly available information. Much technical work has focused on using computational notions of equity and fairness to evaluate specific models or datasets (Angwin et al., 2016; Buolamwini and Gebru, 2018). Indeed, when these models are available, we can and should investigate them to identify potential problems. But what do we do when we have little or no access to models or the data they produce? Certain models may be completely inaccessible to the public, whether for practical or legal reasons, and attempts to audit these models by examining their training data or outputs might jeopardize users' privacy. With algorithmic pre-employment assessments, we find that this is very much the case: models, much less the sensitive employee data used to construct them, are in general kept private. As such, the only information we can consistently glean about industry practices is limited to what companies publicly disclose. Despite this, one of the key findings of our work is that even without access to models or data, we can still learn a considerable amount by investigating what corporations disclose about their practices for developing, validating, and removing bias from these tools.

Documenting claims and evaluating practices. Following a review of firms offering recruitment technologies, we identify 18 vendors of pre-employment assessments. We document what each company has disclosed about its practices and consider the implications of these claims. In so doing, we develop an understanding of industry attempts to mitigate bias and what critical issues are unaddressed.

Prior work has sought to taxonomize the points at which bias can enter machine learning systems, noting that the choice of target variable or outcome

to predict, the training data used, and labelling of examples are all potential sources of disparities (Barocas and Selbst, 2016; Kleinberg et al., 2019). Following these frameworks, we seek to understand how practitioners handle these key decisions in the machine learning pipeline. In particular, we surface choices and trade-offs vendors face with regard to the collection of data, the ability to validate on representative populations, and the effects of discrimination law on efforts to prevent bias. The heterogeneity we observe in vendors' practices indicates evolving industry norms that are sensitive to concerns of bias but lack clear guidance on how to respond to these worries.

Of course, analyzing publicly available information has its limitations. We are unable, for example, to identify issues that any particular model might raise in practice. Nor can we be sure that vendors aren't doing more behind the scenes to ensure that their models are non-discriminatory. And while other publicly accessible information (e.g., news articles and videos from conferences) might offer further details about vendors' practices, for the sake of consistent comparison, we limit ourselves to statements on vendors' websites. As such, our analysis should not be viewed as exhaustive; however, as we will see, it is still possible to draw meaningful conclusions and characterize industry trends through our methods. One notable limitation we encounter is the lack of information about the validity of these assessments. It is of paramount importance to know the extent to which these tools actually work, but we cannot do so without additional transparency from vendors.

We stress that our analysis is not intended as an exposé of industry practices. Many of the vendors we study exist precisely because they seek to provide a fairer alternative to traditional hiring practices. Our hope is that this

work will paint a realistic picture of the landscape of algorithmic techniques in pre-employment assessment and offer recommendations for their effective and appropriate use.

Organization of the rest of the chapter. In Section 11.2, we systematically review vendors of algorithmic screening tools and empirically characterize their practices based on the claims that they make. We analyze these practices in detail in Sections 11.3 and 11.4 from technical and legal perspectives, examining ambiguities and particular causes for concern. We provide concluding thoughts and recommendations in Section 11.5.

11.1 Background

Pre-employment assessments in the hiring pipeline. Hiring decisions are among the most consequential that individuals face, determining key aspects of their lives, including where they live and how much they earn. These decisions are similarly impactful for employers, who face significant financial pressure to make high-quality hires quickly and efficiently (Mariotti, 2017). As a result, many employers seek tools with which to optimize their hiring processes.

Broadly speaking, there are four distinct stages of the hiring pipeline, though the boundaries between them are not always rigid: sourcing, screening, interviewing, and selection (Bogen and Rieke, 2018). Sourcing consists of building a candidate pool, which is then screened to choose a subset to interview. Finally, after candidates are interviewed, selected candidates receive offers. We will focus on *screening*, and in particular, pre-employment assessments that al-

gorithmically evaluate candidates. This includes, for example, questionnaires and video interviews that are analyzed automatically.

Prior work has considered the rise of algorithmic tools in the context of hiring, highlighting the concerns that they raise for fairness. Bogen and Rieke (2018) provide an overview of the various ways in which algorithms are being introduced into this pipeline, with a focus on their implications for equity. Garr and Jackson (2019) survey a number of platforms designed to promote diversity and inclusion in hiring. Sanchez-Monedero et al. (2020) analyze some of the vendors considered here from the perspective of UK law, addressing concerns over both discrimination and data protection. Broadly considering the use of data science in HR-related activities, Tambe et al. (2019) identify several practical challenges to the use of algorithmic systems in hiring, and propose a framework to help address them. Ajunwa (2020) provides a legal framework to consider the problems algorithmic tools introduce and argues against subjective targets like "cultural fit." Kim (2018, 2020) also raises legal concerns over the use of algorithms in hiring in both advertising and screening contexts.

Scholars in the field of Industrial-Organizational (IO) Psychology have also begun to grapple with the variety of new pre-employment assessment methods and sources of information enabled by algorithms and big data (Guzzo et al., 2015). Chamorro-Premuzic et al. (2016) find that academic research has been unable to keep pace with rapidly evolving technology, allowing vendors to push the boundaries of assessments without rigorous independent research. A 2013 report by the National Research Council summarizes a number of ethical issues that arise in pre-employment assessment, including the role of human intervention, the provision of feedback to candidates, and the goal of hiring for

"fit," especially in light of modern data sources (National Research Council, 2013). And although proponents argue that pre-employment assessments can push back against human biases (Chamorro-Prezumic and Akhtar, 2019), assessments (especially data-driven algorithmic ones) run the risk of codifying inequalities while providing a veneer of objectivity.

A history of equity concerns in assessment. Pre-employment assessments date back to examinations for the Chinese civil service thousands of years ago (Haney, 1982). In the early 1900's, the idea that assessments could reveal innate cognitive abilities gained traction in Western industrial and academic circles, leading to the formation of Industrial Psychology as an academic discipline (Munsterberg, 1998; Gerhardt, 1916; Kemble, 1916). During the two World Wars, the U.S. government turned to these assessments in an attempt to quantify soldiers' abilities, paving the way for their widespread adoption in post-war industry (Baritz, 1960; DuBois, 1970; Dunnette and Borman, 1979). Historically, these assessments were primarily behavioral or cognitive in nature, like the Stanford-Binet IQ test (Terman, 1916), the Myers-Briggs type indicator (Myers, 1962), and the Big Five personality traits (Norman, 1963). IO Psychology remains a prominent component of these modern assessment tools—many vendors we examine employ IO psychologists who work with data scientists to create and validate assessments.

Cognitive assessments have imposed adverse impacts on minority populations since their introduction into mainstream use (Tyler, 1947; Ruda and Albright, 1968; National Research Council, 1989). Critics have long contended that observed group differences in test outcomes indicated flaws in the tests themselves (Cravens, 1978), and a growing consensus has formed around the idea

that while assessments do have some predictive validity, they often disadvantage minorities despite the fact that minority candidates have similar real-world job performance to their white counterparts (National Research Council, 1989).[1]

The American Psychological Association (APA) recognizes these concerns as examples of "predictive bias" (when an assessment systematically over- or under-predicts scores for a particular group) in its Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2018). The APA Principles consider several potential definitions of fairness, and while they encourage practitioners to identify and mitigate predictive bias, they explicitly reject the view that fairness requires equal outcomes (Society for Industrial and Organizational Psychology, 2018). As we will see, this focus on predictive bias over outcome-based definitions of fairness forms interesting connections and contrasts with U.S. employment discrimination law.

A brief overview of U.S. employment discrimination law. Title VII of the Civil Rights Act of 1964 forms the basis of regulatory oversight regarding discrimination in employment. It prohibits discrimination with respect to a number of protected attributes ("race, color, religion, sex and national origin"), establishing the Equal Employment Opportunity Commission (EEOC) to ensure compliance (U.S. Congress, 1964). The EEOC, in turn, issued the Uniform Guidelines on Employment Selection Procedures in 1978 to set standards for how employers can choose their employees.

[1]Disparities in assessment outcomes for minority populations are not limited to pre-employment assessments. In the education literature, the adverse impact of assessments on minorities is well-documented (Madaus and Clarke, 2001). This has led to a decades-long line of literature seeking to measure and mitigate the observed disparities (see Hutchinson and Mitchell (2019) for a survey).

According to the Uniform Guidelines (Equal Employment Opportunity Commission, 1978), the gold standard for pre-employment assessments is *validity*: the outcome of a test should say something meaningful about a candidate's potential as an employee. The EEOC accepts three forms of evidence for validity: criterion, content, and construct. Criterion validity refers to predictive ability: do test scores correlate with meaningful job outcomes (e.g., sales numbers)? An assessment with content validity tests candidates in similar situations to ones that they will encounter on the job (e.g., a coding interview). Finally, assessments demonstrate construct validity if they test for some fundamental characteristic (e.g., grit or leadership) required for good job performance.

When is an assessment legally considered discriminatory? Based on existing precedent, the Uniform Guidelines provide two avenues to challenge an assessment: disparate treatment and disparate impact (Barocas and Selbst, 2016). Disparate treatment is relatively straightforward—it is illegal to explicitly treat candidates differently based on categories protected under Title VII (Equal Employment Opportunity Commission, 1978; U.S. Congress, 1964). Disparate impact is more nuanced, and while we provide an overview of the process here, we refer the reader to Barocas and Selbst (2016) for a more complete discussion.

Under the Uniform Guidelines, the rule of thumb to decide when a disparate impact case can be brought against an employer is the "4/5 rule": if the selection rate for one protected group is less than 4/5 of that of the group with the highest selection rate, the employer may be at risk (Equal Employment Opportunity Commission, 1978). If a significant disparity in selection rates is established, an employer may defend itself by showing that its selection procedures are both valid and necessary from a business perspective (Equal Employment Opportu-

nity Commission, 1978). Even when a business necessity has been established, an employer can be held liable if the plaintiff can produce an alternative selection procedure with less adverse impact that the employer could have used instead with little business cost (Equal Employment Opportunity Commission, 1978).[2] Ultimately, both the APA Principles and the Uniform Guidelines agree that validity is fundamental to a good assessment.[3] And while validity can be used as a defense against disparate selection rates, we will see that the Uniform Guidelines' emphasis on outcome disparities and the 4/5 rule significantly impacts vendors' practices.

11.2 Empirical Findings

11.2.1 Methodology

Identifying companies offering algorithmic pre-employment assessments. In order to get a broad overview of the emerging industry surrounding algorithmic pre-employment assessments, we conducted a systematic review of assessment vendors with English-language websites. To identify relevant companies, we consulted Crunchbase's list of the top 300 start-ups (by funding amount) under its "recruiting" category.[4] Crunchbase offers information on public and

[2]It should be noted that this description is based on a particular (although the most common) interpretation of Title VII. Some legal scholars contend that Title VII offers stronger protections to minorities (Bornstein, 2018; Kim, 2016), and there is disagreement on how (or whether) to operationalize the 4/5 rule through statistical tests (Shoben, 1978; Cohn, 1979b; Shoben, 1979; Cohn, 1979a). In this chapter, we will not consider alternative interpretations of Title VII, nor will we get into the specifics of how exactly to detect violations of the 4/5 rule.

[3]Many psychologists disagree with the specific conception of validity endorsed by the Uniform Guidelines (Mcdaniel et al., 2011; Salas, 2011; Biddle, 2008); however, there is broad agreement that some form of validation is necessary.

[4]https://www.crunchbase.com/hub/recruiting-startups

private companies, providing details on funding and other investment activity. While Crunchbase is not an exhaustive list of all companies working in an industry, it is an often-used resource for tracking developments in start-up companies. Companies can create profiles for themselves, subject to validation.[5] We supplemented this list with an inventory of relevant companies found in recent reports by Upturn (Bogen and Rieke, 2018), a technology research and advocacy firm focused on civil rights, and RedThread Research (Garr and Jackson, 2019), a research and advisory firm specializing in new technologies for human resource management. This resulted in 22 additional companies, for a combined total of 322. There was substantial overlap between the three sources considered.

Thirty-nine of these companies did not have English-language websites, so we excluded them. Recall that the hiring pipeline has four primary stages (sourcing, screening, interviewing, and selection); we ruled out vendors that do not provide assessment services at the screening stage, leaving us with 45 vendors. Note that this excluded companies that merely provide online job boards or marketplaces like Monster.com and Upwork. Twenty-two of the remaining vendors did not obviously use any predictive technology (e.g., coding interview platforms that only evaluated correctness or rule-based screening) or did not offer explicit assessments (e.g., scraping candidate information from other sources), and an additional 5 did not provide enough information for us to make concrete determinations, leaving us with 18 vendors in our sample. With these 18 vendors, in April 2019,[6] we recorded administrative information available on Crunchbase (approximate number of employees, location, and total funding) and undertook a review of their claims and practices, which we explain

[5]https://support.crunchbase.com/hc/en-us/articles/115011823988-Create-a-Crunchbase-Profile

[6]Our empirical findings are specific to this moment in time; practices and documentation may have changed since then.

below.

Documenting vendors' claims and practices. Based on prior frameworks intended to interrogate machine learning pipelines for bias (Barocas and Selbst, 2016; Kleinberg et al., 2019), we ask the following questions of vendors:

- What types of assessments do they provide (e.g., questions, video interviews, or games)? [Features]
- What is the outcome or quality that these assessments aim to predict (e.g., sales revenue, annual review score, or grit)? [Target variable]
- What data are used to develop the assessment (e.g., the client's or the vendor's own data)? [Training data]
- What information do they provide regarding validation processes (e.g., validation studies or whitepapers)? [Validation]
- What claims or guarantees (if any) are made regarding bias or fairness? When applicable, how do they achieve these guarantees? [Fairness]

To answer these questions, we exhaustively searched the websites of each company. This included following all internal links, downloading any reports or whitepapers they provided, and watching webinars found on their websites. Almost all vendors provided an option to request a demo; we avoided doing so since our focus is on accessible and public information. Sometimes, company websites were quite sparse on information, and we were unable to conclusively answer all questions for all vendors.

11.2.2 Findings

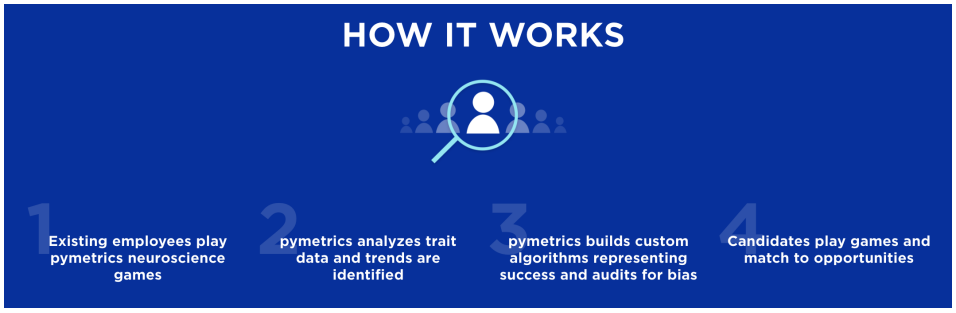


Figure 11.1: Description of the pymetrics process (screenshot from the pymetrics website: https://www.pymetrics.com/employers/)

In our review, we found 18 vendors providing algorithmically driven pre-employment assessments. Those that had available funding information on Crunchbase (16 out of 18) ranged in funding from around $1 million to $93 million. Most vendors (14) had 50 or fewer employees, and half (9) were based in the United States. 15 vendors were present in Crunchbase's "Recruiting Startups" list; the remaining vendors were taken from reports by Upturn (Bogen and Rieke, 2018) and RedThread Research (Garr and Jackson, 2019). Many vendors were present in all of these sources. Table 11.1 summarizes our findings. Table G.1 in Appendix G.1 contains administrative information about the vendors we included.

Assessment types. The types of assessments offered varied by vendor. The most popular assessment types were questions (11 vendors), video interview analysis (6 vendors), and gameplay (e.g., puzzles or video games) (6 vendors). Note that many vendors offered multiple types of assessments. Question-based assessments included personality tests, situational judgment tests, and other formats. For video interviews, candidates were typically either asked to record answers to particular questions or more free-form "video resumes" highlighting

Vendor name	Assessment types [Features]	Custom? [Target & Training data]	Validation info [Validation]	Adverse impact [Fairness]
8 and Above	phone, video	S	–	bias mentioned
ActiView	VR assessment	C	validation claimed	bias mentioned
Assessment Innovation	games, questions	–	–	bias mentioned
Good&Co	questions	C, P	multiple studies	adverse impact
Harver	games, questions	S	–	–
HireVue	games, questions, video	C, P	–	4/5 rule
impress.ai	questions	S	–	–
Knockri	video	S	–	bias mentioned
Koru	questions	S	some description	adverse impact
LaunchPad Recruits	questions, video	–	–	bias mentioned
myInterview	video	–	–	compliance
Plum.io	questions, games	S	validation claimed	bias mentioned
PredictiveHire	questions	C	–	4/5 rule
pymetrics	games	C	small case study	4/5 rule
Scoutible	games	C	–	–
Teamscope	questions	S, P	–	bias mentioned
ThriveMap	questions	C	–	bias mentioned
Yobs	video	C, S	–	adverse impact

Table 11.1: Examining the websites of vendors of algorithmic pre-employment assessments, we answer a number of questions regarding their assessments in relation to questions of fairness and bias. This involves exhaustively searching their websites, downloading whitepapers they provide, and watching webinars they make available. This table presents our findings. The "Assessment types" column gives the types of assessments each vendor offers. In the "Custom?" column, we consider the source of data used to build an assessment: C denotes "custom" (uses employer data), S denotes "semi-custom" (qualitatively tailored to employer without data) and P denotes "pre-built." The "Validation?" column contains information vendors publicly provided about their validation processes. In the "Adverse impact" column, we recorded phrases found on vendors' websites addressing concerns over bias.

their strengths. These videos are then algorithmically analyzed by vendors.

Target variables and training data. Most of the vendors (15) offer custom or customizable assessments, adapting the assessment to the client's particular data or job requirements. In practice, decisions about target variables and training data are made together based on where the data come from. Eight vendors

build assessments based on data from the client's past and current employees (see Figure 11.1). Vendors in general leave it up to clients to determine what outcomes they want to predict, including, for example, performance reviews, sales numbers, and retention time. Other vendors who offer customizable assessments without using client data either use human expertise to determine which of a pre-determined set of competencies are most relevant to the particular job (the vendor's analysis of a job role or a client's knowledge of relevant requirements) or don't explicitly specify their prediction targets. In such cases, the vendor provides an assessment that scores applicants on various competencies, which are then combined into a "fit" score based on a custom formula. Thus, even among vendors who tailor their assessments to a client, they do so in different ways.

Vendors who only offer pre-built assessments typically either provide assessments designed for a particular job role (e.g., salesperson), or provide a sort of "competency report" with scores on a number of cognitive or behavioral traits (e.g., leadership, grit, teamwork). These assessments are closer in spirit to traditional psychometric assessments like the Myers-Briggs Type Indicator or Big Five Personality Test; however, unlike traditional assessments that rely on a small number of questions, modern assessments may build psychographic profiles using machine learning to analyze rich data sources like a video interview or gameplay.

Validation. Generally, vendors' websites do not make clear whether vendors validate their models, what validation methodologies they use, how they select validation data, or how validation procedures might be tailored to the particu-

Vendor	Claim about bias
HireVue	Provide "a highly valid, bias-mitigated assessment"
pymetrics	"…the Pre-Hire assessment does not show bias against women or minority respondents."
PredictiveHire	"AI bias is testable, hence fixable."
Knockri	"Knockri's A.I. is unbiased because of its full spectrum database that ensures there's no benchmark of what the 'ideal candidate' looks like."

Table 11.2: Examples of claims that vendors make about bias, taken from their websites.

lar client. Good & Co.,[7] notably, provides fairly rigorous validation studies of the psychometric component of their assessment, as well as a detailed audit of how the scores differ across demographic groups; however, they do not provide similar documentation justifying the algorithmic techniques they use to recommend candidates based on "culture fit."

Accounting for bias. In total, while 15 of the vendors made at least abstract references to "bias" (sometimes in the context of well-established human bias in hiring), only 7 vendors explicitly discussed compliance or adverse impact with respect to the assessments they offered. Three vendors explicitly mentioned the $4/5$ rule, and an additional 4 advertised "compliance" or claimed to control adverse impact more generally. Several of these vendors claimed to test models for bias, "fixing" it when it appeared. HireVue and pymetrics, in particular, offered a detailed description of their overall approaches to de-biasing, which involves removing features correlated with protected attributes when adverse impact is detected. Other vendors (e.g., Knockri and PredictiveHire) claimed to "fix" adverse impact when it is found without going into further detail.

[7] https://good.co/

Among those that do make concrete claims, all vendors we examined specifically focus on equality of outcomes and compliance with the 4/5 rule. Roughly speaking, there are two ways in which vendors claim to achieve these goals: naturally unbiased assessments and active algorithmic de-biasing. Typically, vendors claiming to provide naturally unbiased assessments seek to measure underlying cognitive or behavioral traits, so their assessments output a small number of scores, one for each competency being measured. In this setting, a naturally unbiased assessment in one that produces similar score distributions across demographic groups. Koru, for instance, measures 7 traits (e.g., "grit" and "presence") and claims that "[i]n all panels since 2015, the Pre-Hire assessment does not show bias against women or minority respondents" (Jarrett and Croft, 2018).

Other vendors actively intervene in their learned models to remove biases. One technique that we have observed across multiple vendors (e.g., HireVue, pymetrics, PredictiveHire) is the following: build a model and test it for adverse impact against various subgroups.[8] As Bogen and Rieke (2018) also observe, if adverse impact is found, the model and/or data are modified to try to remove it, and then the model is tested again for adverse impact. HireVue and pymetrics downweight or remove features found to be highly correlated with the protected attribute in question, noting that this can significantly reduce adverse impact with little effect on the predictive accuracy of the assessment. This is done prior to the model's deployment on actual applicants, though some vendors claim to periodically test and update models. In Section 11.4, we discuss in depth these efforts to define and remove bias.

[8]pymetrics, for instance, open-sources the tests it uses: https://github.com/pymetrics/audit-ai

11.3 Analysis of Technical Concerns

Our findings in Section 11.2 raise several technical challenges for the pre-employment assessment process. In this section, we focus on two areas that are particularly salient in the context of algorithmic hiring: **data choices**, where vendors must decide where to draw data from and what outcomes to predict; and the use of **alternative assessment formats**, like game- or video-based assessments that rely on larger feature sets and more complex machine learning tools than traditional question-based assessments.

11.3.1 Data Choices

Machine learning is often viewed as a process by which we predict a given output from a given set of inputs. In reality, neither the inputs nor outputs are fixed. Where do the data come from? What is the "right" outcome to predict? These and others are crucial decisions in the machine learning pipeline, and can create opportunities for bias to enter the process.

Custom assessments. Consider a hypothetical practitioner building a custom assessment to identify the "best" candidates for her client. As is the case in many domains, translating this to a feasible data-driven task forces our practitioner to make certain compromises (Passi and Barocas, 2019). It quickly becomes clear that she must operationalize "best" in some measurable way. What does the client value? Sales numbers? Cultural fit? Retention? And, crucially, what data does the client have? This is a nontrivial constraint: many companies don't maintain comprehensive and accessible data about employee performance, and

thus, a practitioner may be forced to do the best she can with the limited data that she is given (Tambe et al., 2019). Note that relying on the client's data has already forced the practitioner to only learn from the client's existing employees; at the outset, at least, she has data on how those who *weren't* hired would have performed.

Once a target is identified, the practitioner needs a dataset on which to train a model. Since she has performance data on previous employees, she needs them to take the assessment so she can link their scores to their observed job performance. How many employees' data does she need in order to get an accurate model? What if certain employees don't want to or don't have time to take the assessment? Is the set of employees who respond representative of the larger applicant pool who will ultimately be assessed?

Finally, the practitioner is in a position to actually build a model. Along the way, however, she had to make several key choices, often based on factors (like client data availability) outside her control. The choice of target variable is particularly salient. Proxies like job evaluations, for instance, can exhibit biases against minorities (Sidanius and Crane, 1989; Neumark et al., 1996; Riach and Rich, 2002). Moreover, predicting the success of future employees based on current employees inherently skews the task toward finding candidates resembling those who have already been hired.

Some vendors go beyond trying to identify candidates who are generically good, or even good for a particular client, and explicitly focus on finding candidates who "fit" with an existing employee or team. Both Good & Co. and Teamscope provide these team-specific tools for employers, and Good & Co. further advertises their assessments as a way to "[r]eplicate your top perform-

ers."[9] If models are localized to predict fit with particular teams, any role at any company could in principle have its own tailor-made predictive model. But when models are customized at such a small scale, it can be quite difficult to determine what it means for such a model to be biased or discriminatory. Does each team-specific model need to be audited for bias? How would a vendor go about doing so?

And yet, while it is easy to criticize vendors for their choices, it's not clear that there are better alternatives. In practice, it is impossible to even define, let alone collect data on, an objective measure of a "good" employee. Nor is it always feasible to get completely representative data. Vendors and advocates point out that many of the potentially problematic elements here (subjective evaluations; biased historical samples; emphasis on fit) are equally present, if not more so, in traditional human hiring practices (Chamorro-Prezumic and Akhtar, 2019).

Customizable and pre-built assessments. Instead of building a new custom assessment for each client, it may be tempting to instead offer a pre-built assessment (perhaps specific to a particular job role) that has been validated across data from a variety of clients. This approach has its advantages: it isn't subject to the idiosyncratic data of each client, and it can draw from a diverse range of candidates and employees to learn a broad notion of what a "good" employee looks like. Additionally, pre-built assessments may be attractive to clients with too few existing employees to build a custom assessment.

Some vendors offer assessments that are mostly pre-built but somewhat cus-

[9] https://good.co/pro/

tomizable. Koru and Plum.io, for example, provide pre-built assessments to evaluate a fixed number of competencies. Experts then analyze the job description and role for a particular client and determine which competencies are most important for the client's needs. Thus, these vendors hope to get the best of both worlds: assessments validated on large populations that are still flexible enough to adapt to the specific requirements of each client. As shown in Figure 11.2, the firm 8 and Above profiles over 60 traits based on a video interview, but also reports a single "Elev8" score tailored to the particular client.

Despite these benefits, pre-built assessments do have drawbacks. Individual competencies like "grit" or "openness" are themselves constructs, and attempts to measure them must rely on other psychometric assessments as "ground truth." Given that traits can be measured by multiple tests that don't perfectly correlate with one another (Rodriguez and Maeda, 2006), it may be difficult to create an objective benchmark against which to compare an algorithmic assessment. Furthermore, it is generally considered good practice to build and validate assessments on a representative population for a particular job role (Society for Industrial and Organizational Psychology, 2018), and both underlying candidate pools and job specifics differ across locations, companies, and job descriptions. Pre-built assessments must by nature be general, but as a consequence, they may not adapt well to the client's requirements.

Necessary trade-offs. This leads to an inherently challenging technical problem: on the one hand, more data is usually beneficial in creating and validating an assessment; on the other hand, drawing upon data from related but somewhat different sources may lead to inaccurate conclusions. We can view this as an instance of domain adaptation and the bias-variance tradeoff, well studied in

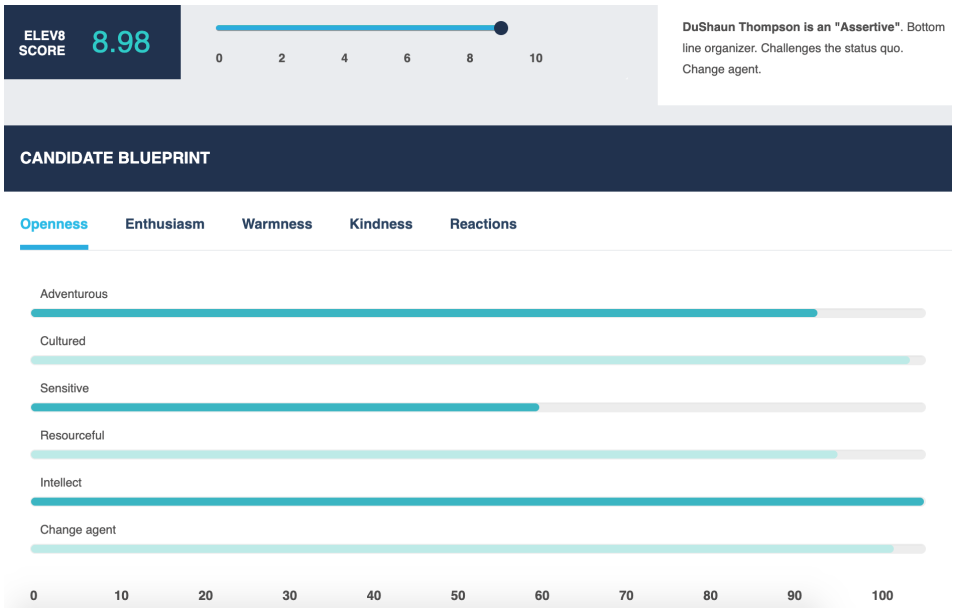


Figure 11.2: Part of a sample candidate profile from 8 and Above, based on a 30-second recorded video cover letter (screenshot from the 8 and Above website: https://www.8andabove.com/p/profile/blueprint/643)

the statistics and machine learning literature (Ben-David et al., 2010; Friedman et al., 2001). Pooling data from multiple companies or geographic locations may reduce variance due to small sample sizes at a particular company, but comes at the cost of biasing the outcomes away from the client's specific needs. There is no obvious answer or clear best practice here, and vendors and clients must carefully consider the pros and cons of various assessment types. Larger clients may be better positioned for vendors to build custom assessments based solely on their data; smaller clients may turn to pre-built assessments, making the assumption that the candidate pool and job role on which the assessment was built is sufficiently similar to warrant generalizing its conclusions.

11.3.2 Alternative Assessment Formats

Once an assessment has been built, it must be validated to verify that it performs as expected. Psychologists have developed extensive standards to guide assessment creators in this process (Society for Industrial and Organizational Psychology, 2018); however, modern assessment vendors are pushing the boundaries of assessment formats far beyond the pen-and-paper tests of old, often with little regulatory oversight (Chamorro-Premuzic et al., 2016). Game- and video-based assessments, in particular, are increasingly common. Vendors point to an emerging line of literature showing that features derived from these modern assessment formats correlate with job outcomes and personality traits (Kramer and Ward, 2010; Grimmett, 2017) as evidence that these assessments contain information that can be predictive of job outcomes, though they rarely release rigorous validation studies of their own.

Technical challenges for alternative assessments. While there is evidence for the predictive validity of alternative assessments, empirical correlation is no substitute for theoretical justification. Historically, IO psychologists have designed assessments based on their research-driven knowledge that certain traits correlate with desirable outcomes. To some extent, machine learning attempts to automate this process by discovering relationships (e.g., between actions in a video game and personality traits) instead of quantifying known relationships. Of course, machine learning can be used to unearth meaningful relationships. But it may also find relationships that experts don't understand. When the expert is unable to explain why, for example, the cadence of a candidate's voice indicates higher job performance, or why reaction time predicts employee reten-

tion, should a vendor rely on these features? From a technical perspective, correlations that cannot be theoretically justified may fail to generalize well or remain stable over time, and, in light of such concerns, the APA Principles caution that a practitioner should "establish a clear rationale for linking the resulting scores to the criterion constructs of interest" (Society for Industrial and Organizational Psychology, 2018). Yet when an algorithm takes in "millions of data points" for each candidate (as advertised by pymetrics[10]), it may not be possible to provide a qualitative justification for the inclusion of each feature.

Moreover, automated discovery of relationships makes it difficult for a critical expert to detect when the model makes indirect use of a proscribed characteristic. Rich sources of data can easily encode properties that are illegal to use in the hiring process. Facial analysis, in particular, has been heavily scrutinized recently. A wave of studies has shown that several commercially available facial analysis techniques suffer from disparities in error rates across gender and racial lines (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Rhue, 2018), and more broadly, evidence suggests that we may not be able to reliably infer emotions from facial expressions, especially cross-culturally (Barrett et al., 2019). Concerns have also been raised over the use of affect and emotion recognition for those with disabilities, particularly in the context of employment (Fruchterman and Melllea, 2018; Guo et al., 2019; Hurley-Hanson and Giannantonio, 2016).

Because it can be quite expensive and technically challenging to build facial analysis software in-house, vendors will often turn to third parties (e.g., Affectiva[11]) who provide facial analysis as a service. As a result, vendors lack

[10]https://perma.cc/3284-WTS8
[11]https://www.affectiva.com/

the ability or resources to thoroughly audit the software they use. With these concerns in mind, U.S. Senators Kamala Harris, Patty Murray, and Elizabeth Warren recently wrote a letter to the EEOC asking for a report on the legality and potential issues with the use of facial analysis in pre-employment assessments (Harris et al., 2018). Even more recently, Illinois passed a law requiring applicants to be notified and provide consent if their video interviews will be analyzed by artificial intelligence (Assembly, 2019), though it's not clear what happens if an applicant refuses to consent.

While heightened publicity regarding racial disparities in facial analysis has prompted many third-party vendors of this technology to respond by improving the performance of their tools on minority populations (Puri, 2018; Roach, 2018), it remains unclear what information facial analysis relies on to draw conclusions about candidates. Facial expressions may contain information about a range of sensitive attributes from obvious ones like ethnicity, gender, and age to more subtle traits like a candidate's mental and physical health (Kramer and Ward, 2010; Zhou et al., 2015).[12] Given the opacity of the deep learning models used for facial analysis, it can be difficult or even impossible to detect if a model inadvertently learns proxies for prohibited features.

11.4 Algorithmic De-Biasing

Under Title VII, employers bear ultimate legal responsibility for their hiring decisions. Employers, then, remain strongly motivated to mitigate their potential liability against disparate impact claims. Vendors, in turn, are incentivized to

[12] As a general matter, the Americans with Disabilities Act prohibits employers from collecting or considering information about candidates' health (U.S. Congress, 1990).

build demonstrably unbiased tools that help employers to avoid such liability.

As we have described, all vendors in our sample who made concrete claims about de-biasing (including the two best-funded firms in our sample) did so with reference to equality of outcomes and compliance with the 4/5 rule. In this section, we explore the effects of this reliance on the stages of a typical disparate impact lawsuit. We then explore technical approaches that have been proposed to control outcome disparities, and their relationship to the law. Finally, we describe some important consequences of the de-biasing strategies favored by vendors.

11.4.1 Algorithmic De-Biasing and Disparate Impact Litigation

Recall the three steps in a disparate impact case. The plaintiff must first establish that the employer's selection procedure generates a disparate impact. Once established, the employer must then defend itself by justifying the disparate impact by reference to some business necessity. In this case, an employer would likely do so by establishing the validity of the model driving its hiring decisions. Finally, the plaintiff may then challenge the proffered justification as faulty or demonstrate that an alternative practice exists that would serve the employer's business objective equally well while reducing the disparate impact in its selection rates.

Note that disparate impact doctrine does not prohibit disparate impact altogether; it renders employers liable for an *unjustified* or *avoidable* disparate impact. Vendors' choice to enforce the 4/5 rule might therefore seem overly cautious: although employers could justify an assessment that has a disparate im-

pact by demonstrating its validity (as we discuss in Section 2), vendors take steps to ensure that employers are not placed in this position, because assessments are prevented from having a disparate impact in the first place. One possible explanation for adopting the 4/5 rule is that vendors might be catering to employers' aversion to legal risk.

As to the second step, the practical effect of vendors' reliance on the 4/5 rule is to obviate the need for an employer to demonstrate business necessity through a legally rigorous validation process. According to the Uniform Guidelines, employers only need to validate their selection procedure if it has a disparate impact. Of course, clients might still expect and even demand validation studies from vendors, given their goal of selecting qualified candidates. As a consequence, the choice of how to validate seems to become a *business* decision rather than a *legal* imperative.

The final step in a disparate impact case raises yet another possible explanation for vendors' decisions to adopt the 4/5 rule as a constraint. Recall that, at this stage, employers bear liability if they failed to adopt an alternative practice that could have minimized any business-justified disparity created by their selection procedure, provided that such practices were not too costly. Employers therefore run significant legal risks if they do not take such steps. In turn, should vendors have some way to minimize disparity without sacrificing the accuracy of their assessments, failing to do so might place their clients in legal jeopardy. A plaintiff could assert that this very possibility reveals that any evident disparate impact—even if justified by a validation study—was avoidable.

While the burden of identifying this alternative business practice rests with the plaintiff, vendors may want to preempt this argument by taking affirmative

steps to explore how to minimize disparate impact without imposing unwelcome costs on the employer. In the past, such exploratory efforts might have been costly and difficult, since discovering an alternative business practice that is equally effective for the firm, while generating less disparity in selection rates, was no easy task. Many modern assessments (e.g., those with a large number of features) make some degree of exploration almost trivial, allowing vendors to find a model that (nearly) maintains maximum accuracy while reducing disparate impact.

In this way, the ready availability of algorithmic techniques might effectively create a legal *imperative* to use them. If the adverse impact of a business-justified model could be reduced through algorithmic de-biasing—without significantly harming predictive ability, and at trivial cost—de-biasing itself might be considered an "alternative business practice," and therefore render the employer liable for not adopting it.

11.4.2 Methods to Control Outcome Disparities

Thus, for legal reasons, a vendor may choose to control outcome disparities in strict adherence to the $4/5$ rule. But this is not the end of the story; multiple techniques exist to control outcome differences. Here, we explore both historical and contemporary approaches in comparison with the de-biasing techniques we observe.

The most straightforward approach to control outcome differences is known as "within-group scoring," under which scores are reported as a percentile with respect to the particular group in question. Employers could then select candi-

dates above a particular threshold for each group (top 10% from Group A, top 10% from Group B, etc.), which would naturally result in equal selection rates. Recall that in the de-biasing reviewed above, vendors achieve (approximately) equal selection rates by systematically removing features from the model that contribute to a disparate impact. In so doing, they may lose useful information contained in these features as well, undermining their ability to maintain an accurate rank order within each group. In contrast, within-group scoring may theoretically be the optimal way to equalize selection rates, since it preserves rank order (Corbett-Davies et al., 2017; Lipton et al., 2018).

In fact, within-group scoring was used for the General Aptitude Test Battery (GATB), a pre-employment assessment developed in the 1940s by the US Employment Service (USES), due to significant differences in score distributions across ethnic groups. In particular, the USES reported results as within-group percentile scores by ethnicity—black, Hispanic, and other (National Research Council, 1989; Schuler et al., 1993). Commissioned to investigate the justification for such a policy, a National Academy of Sciences study recommended the continued use of within-group percentiles because without them, minority applicants would suffer from "higher false-rejection rates" (National Research Council, 1989).

In principle, within-group score reporting (also known as "race-norming") would satisfy the $4/5$ rule; so why don't vendors use it? In fact, within-group reporting would likely be considered illegal today. In 1986 the Department of Justice challenged its legality in the GATB, claiming that it constituted disparate treatment (Schuler et al., 1993), and the practice was prohibited by the Civil Rights Act of 1991 (U.S. Congress, 1991).

This points to a longstanding tension between disparate treatment and disparate impact: some techniques to control outcome disparities require the use of protected attributes, which may be considered disparate treatment. To circumvent this, the vendors we observe engaging in algorithmic de-biasing take into account protected attributes when *building* models, but ultimately produce models that do not take protected attributes as input. In this way, individual decisions do not exhibit disparate treatment, and yet, outcome disparities can still be mitigated.

In fact, these techniques fit into a broader category of methods known as Disparate Learning Processes (DLPs), a family of algorithms designed to produce decision rules that (approximately) equalize outcomes without engaging in disparate treatment at the individual level (Lipton et al., 2018; Pedreshi et al., 2008; Zafar et al., 2017b). There are slight differences between DLPs as found in the computer science literature and vendors' algorithmic de-biasing efforts: DLPs typically work by imposing constraints that prevent outcome disparities on the learning algorithm that produces the model; the algorithmic de-biasing we observe, on the other hand, simply removes features correlated with protected attributes until outcomes are within a tolerable range. In spirit, however, these techniques are ultimately quite related.

Similar connections exist to "fair representation" learning, where an "encoder" is built to process data by removing information about protected attributes, including proxies and correlations (Zemel et al., 2013; Madras et al., 2018; Edwards and Storkey, 2016). Thus, any model built on data processed by the encoder would have approximately equal outcomes, since outputs of the encoder contain very little information about protected attributes. As in

DLPs, protected attributes are used only to create the encoder; after deployment, when the encoder processes any individual's data, it does not have access to protected attributes. We can think of some vendors' practices as analogous to building such an encoder—one that "processes" data by simply discarding features highly correlated with protected attributes.

11.4.3 Limitations of Outcome-Based De-Biasing

Despite the perhaps good reasons vendors have to use the particular form of algorithmic de-biasing discussed above, these techniques face important caveats and consequences worth mentioning.

Outcome-based notions of bias are intimately tied to the datasets on which they are evaluated. As both the EEOC Guidelines and APA Principles clearly articulate, a representative sample is crucial for validation (Equal Employment Opportunity Commission, 1978; Society for Industrial and Organizational Psychology, 2018). The same holds true for claims regarding outcome disparities: they may depend on whether the assessment is taken by recent college grads in Michigan applying for sales positions or high school dropouts in New York applying for jobs stocking warehouses. Thus, when evaluating claims regarding outcome disparities, it is critical to understand how vendors collect and maintain relevant, representative data.

While outcome disparities are important for vendors to consider, especially in light of U.S. regulations, discrimination and the $4/5$ rule should not be conflated. Vendors may find it necessary from a legal or business perspective to build models that satisfy the $4/5$ rule, but this is not a substitute for a critical

analysis into the mechanisms by which bias and harm manifest in an assessment. For example, differential validity, which occurs when an assessment is better at ranking members of one group than another, should be a top-level concern when examining an assessment (Society for Industrial and Organizational Psychology, 2018; Young, 2001). But because of the legal emphasis placed on adverse impact, vendors have little incentive to structure their techniques around it. Furthermore, it can be challenging to identify and mitigate outcome disparities with respect to protected attributes employers typically don't collect (e.g., sexual orientation (Legislature, 1959)). In such cases, vendors may need to consider alternative approaches to prevent discrimination.

More broadly, bias is not limited to the task of predicting outputs from inputs; a critical, holistic examination of the entire assessment development pipeline may surface deeper concerns. Where do inputs and outputs come from, and what justification do they have? Are there features that shouldn't be used? This isn't to say that some vendors are not already asking these questions; however, in the interest of forming industry standards surrounding algorithmic assessments, the legal operationalization of the $4/5$ rule as a definition of bias runs the risk of downplaying the importance of examining a system as a whole.

Both the law and existing techniques focus on assessment outcomes as binary (screened in/out); however, some platforms actually rank candidates (explicitly, or implicitly by assigning numerical scores). While screening decisions can ultimately be viewed as binary (a candidate is either interviewed or not), there are a number of subtleties induced by ranking: a lower-ranked candidate may only be interviewed after higher-ranked candidates, and their lower score could unduly bias future decision-makers against them (Bogen and Rieke, 2018).

There is no clear analog of the 4/5 rule for ranking; in practice, vendors may choose a cut-off score and test for adverse impact via the 4/5 rule (Equal Employment Opportunity Commission, 1978; Baker et al., 2018). In the computer science literature, there are ongoing efforts to define technical constraints on rankings in the spirit of equal representation and the 4/5 rule (Celis et al., 2018; Yang and Stoyanovich, 2017; Zehlike et al., 2017), and LinkedIn has adopted a similar approach to encourage demographic diversity in its search results (Geyik et al., 2019). However, none of these approaches has received any sort of consensus or official endorsement.

From a policy perspective, the EEOC can and should clarify its position on the use of algorithmic de-biasing techniques, which to our knowledge has yet to be challenged in court. Legal scholars have begun to debate the legality of "algorithmic affirmative action" in various contexts (Kroll et al., 2016; Bent, 2020; Hellman, 2020; Kim, 2017; Raub, 2018), but the debate is far from settled. While existing guidelines can be argued to apply to ML-based assessments, the de-biasing techniques described above do present new opportunities and challenges.

11.5 Discussion and Recommendations

In this chapter, we have presented an in-depth analysis into the bias-related practices of vendors of algorithmic pre-employment assessments. Our findings have implications not only for hiring pipelines, but more broadly for investigations into algorithmic and socio-technical systems. Given the proprietary and sensitive nature of models built for actual clients, it is often infeasible for exter-

nal researchers to perform a traditional audit; despite this, we are able to glean valuable information by delving into vendors' publicly available statements. Broadly speaking, models result from the application of a **vendor's practices** to a real-world setting. Thus, by learning about these practices, we can draw conclusions and raise relevant questions about the resultant models. In doing so, we can create a common vocabulary with which we can discuss and compare practices. We found it useful to **limit the scope** of our inquiry in order to be able to ask and answer concrete questions. Even just considering algorithms used in the context of hiring, we found enough heterogeneity (as have previous reports on the subject (Bogen and Rieke, 2018; Garr and Jackson, 2019)) that it was necessary to further refine our focus to those used in pre-employment assessments. While this did lead us to exclude a number of innovative and intriguing hiring technologies (see, e.g., Textio[13] or Jopwell[14]), it allowed us to make specific and direct comparisons between vendors and get a more detailed understanding of the technical challenges specific to assessments.

In analyzing models via practices, we observe that it is crucial to consider technical systems in conjunction with the **context** surrounding their use and deployment. It would be difficult to understand vendors' design decisions without paying attention to the relevant legal, historical, and social influences. Moreover, in order to push beyond hypothetical or anecdotal accounts of algorithmic bias, we need to incorporate empirical evidence from the field.

Based on our findings, we summarize the following policy recommendations discussed throughout this chapter. We refer the reader to Raghavan and

[13]Textio (https://textio.com/) analyzes job descriptions for gender bias and makes suggestions for alternative, gender-neutral framings.

[14]Jopwell (https://www.jopwell.com/) builds and maintains a network of Black, Latinx, and Native American students and connects students these with employers.

Barocas (2019) for further discussion of these recommendations.

1. Transparency is crucial to further our understanding of these systems. While there are some exceptions, vendors in general are not particularly forthcoming about their practices. Additional transparency is necessary to craft effective policy and enable meaningful oversight.
2. Disparate impact is not the only indicator of bias. Vendors should also monitor other metrics like differential validity.
3. Outcome-based measures of bias (including tests for disparate impact and differential validity) are limited in their power. They require representative datasets for particular applicant pools and fail to critically examine the appropriateness of individual predictors. Moreover, they depend on access to protected attributes that are not always available.
4. We may need to reconsider legal standards of validity under the Uniform Guidelines in light of machine learning. Because machine learning may discover relationships that we do not understand, a statistically valid assessment may inadvertently leverage ethically problematic correlations.
5. Algorithmic de-biasing techniques have significant implications for "alternative business practices," since they automate the search for less discriminatory alternatives. Vendors should explore these techniques to reduce disparate impact, and the EEOC should offer clarity about how the law applies.

Our work leads naturally to a range of questions, ranging from those that seem quite technical (What is the effect of algorithmic de-biasing on model outputs? When should data from other sources be incorporated?) to socio-political

(What additional regulatory constraints could improve the use of algorithms in assessment? How can assessments promote the autonomy and dignity of candidates?). Because the systems we examine are shaped by technical, legal, political, and social forces, we believe that an interdisciplinary approach is necessary to get a broader picture of both the problems they face and the potential avenues for improvement.

CHAPTER 12

THE HIDDEN ASSUMPTIONS BEHIND COUNTERFACTUAL EXPLANATIONS AND PRINCIPAL REASONS

Calls for explanations have become a standard part of the push for algorithmic accountability. As algorithmic decision-making becomes ever more complex and proliferates to more domains, explanations are increasingly seen as a way to reconnect those algorithms to their human subjects: to respect the autonomy of people subject to automated decisions, to allow people to navigate the rules that govern their lives, to help people recognize when they should contest decisions or object to the decision-making process, and to facilitate direct oversight and regulation of algorithms (Wachter et al., 2018; Selbst and Barocas, 2018).

In this chapter, we examine two related approaches to explanation: the counterfactual explanations that have been explored in recent computer science research and which are gaining traction in industry, and the "principal reason" approach drawn from United States credit laws. Collectively, we will call these two approaches "feature-highlighting explanations." At a high level, these approaches provide the subject of a decision with a set of factors that "explain" the decision. Though they are distinct in operation and motivation, both methods highlight a certain subset of features that are deemed most deserving of the decision subject's attention. It is this property that underlies most of our observations.

Other approaches to understanding the decisions of a model have focused on building purposefully simple models that lend themselves to direct inspection, approximating complex models in simpler—and thus more practically

inspectable—forms, and accounting for the relative importance of different features in the model overall (rather than in a specific decision) (Selbst and Barocas, 2018). While these are also useful, there are at least five reasons for the growing popularity of feature-highlighting explanations. First, this approach allows practitioners to abandon any constraints on model complexity–a constraint often seen as a barrier to improved model performance. Second, it allows businesses to avoid disclosing models in their entirety, thereby protecting trade secrets and businesses' other proprietary interests, while limiting decision subjects' ability to game the model. Third, the approach promises a concrete justification for a decision or concrete instructions for achieving a different outcome. Fourth, this approach allows firms to automate the difficult task of generating explanations for a model's decisions, then communicating these explanations to consumers. Fifth, this approach appears to generate explanations that comply with legal requirements both in the United States and Europe.

Generating feature-highlighting explanations is far from straightforward, however, and requires decision makers to make many consequential and subjective choices along the way. Adopting methods to automate the process of explaining specific decisions does not relieve decisions makers of the burden and power to decide what they ultimately disclose to decision subjects. Furthermore, useful explanations cannot be generated in isolation from the world in which decision subjects will have to act. While feature-highlighting explanations are typically inward-looking, determined by examining the model alone and the data used to train it, the true difficulty a decision subject will face in changing a feature is inevitably intertwined with her life circumstances, information rarely available to decision makers.

In this chapter, we demonstrate that the promised utility of factor-based explanations rests on five key assumptions, easily overlooked, and sometimes improper: (1) that a change in feature value clearly maps to an action in the real world; (2) that features can be made commensurate by looking only at the underlying distribution of feature values in the training data; (3) that explanations can be offered without regard to decision-making in other areas of people's lives; (4) that the model generating decisions is monotonic and its potential outcomes are binary; and (5) that explanations remain valid over the time period necessary for a decision subject to change feature values.

The chapter then explores three tensions at the heart of feature-highlighting explanations. First, while feature-highlighting explanations are designed to respect or enhance the autonomy of decision subjects, the decision maker is put in the position of having to make determinations about what is best for the decision subject; this is paternalism in the name of autonomy. Furthermore, the only way for a decision maker to be sensitive to decision subjects' needs and preferences is to further intrude into their lives, gathering enough information to respect their autonomy, while comprising the autonomy afforded by privacy in the process. Second, partial disclosure puts the decision subject at the mercy of the decision maker. The choice of what to disclose grants a great deal of power to the decision maker. By granting such power to businesses, we invite them to use that power for interests other than those of the decision subject. Finally, attempts to overcome some of these challenges by providing decision subjects a larger number and more diverse set of explanations or affording them the opportunity to explore the consequences of specific changes will eventually risk revealing the model altogether. If these techniques fail to protect their intellec-

tual property, firms are unlikely to adopt them.

12.1 What are feature-highlighting explanations?

Counterfactual explanations have begun to attract the interest of businesses, regulators, and legal scholars, with many converging on the belief that such explanations is the preferred approach to explaining machine learning models and their decisions. Principal-reason explanations are well established in U.S. credit laws. Various business have well developed procedures for generating and issuing adverse action notices (AANs). Both methods aim to produce explanations of a particular decision by highlighting factors deemed useful or important. This section will describe both approaches, and how they are linked.

12.1.1 Counterfactual explanations

Recent proposals from computer scientists have focused on counterfactual explanations (Martens and Provost, 2014; Wachter et al., 2018; Ustun et al., 2019; Mothilal et al., 2020; Karimi et al., 2020; Hendricks et al., 2018; Grath et al., 2018; Ribeiro et al., 2016; Lou et al., 2012; Dhurandhar et al., 2018). The goal of counterfactual explanations is to provide actionable guidance—to explain how things could have been different and provide a concrete set of steps a consumer might take to achieve a different outcome in the future. Counterfactual explanations are achieved by identifying the features that, if minimally changed, would alter the output of the model.

In particular, an emerging theme in the computer science literature is to

frame the search for such features as an optimization problem, seeking to find the "nearest" hypothetical point that is classified differently from the point currently in question (Wachter et al., 2018; Mothilal et al., 2020; Russell, 2019; Karimi et al., 2020; Ustun et al., 2019). In casting the search for counterfactual explanations as an optimization problem, a key challenge is to define a notion of distance. Different features are rarely directly comparable because they are represented on numerical scales that do not meaningfully map onto one another. We discuss this challenge more in Section 12.2.2.

Wachter et al. (2018) have also argued that counterfactual explanations could satisfy the explanation requirements of the EU's General Data Protection Regulation (GDPR). Over the last several years, lawyers and legal scholars have debated whether certain provisions of the GDPR create a right to an explanation of algorithmic decisions, and if it exists, whether and when it requires an explanation of specific decisions or the model. (Kaminski, 2019; Selbst and Powles, 2017; Brkan, 2019; Wachter et al., 2017; Edwards and Veale, 2017; Casey et al., 2019; Mendoza and Bygrave, 2017; Malgieri and Comandé, 2017). The official interpretation of the Article 29 Working Party—a government body charged with creating official interpretations of European data protection law—has concluded that the GDPR requires, at a minimum, explanations of specific decisions (Article 29 Data Protection Working Party). Thus, part of the rationale to employ counterfactual explanations is to satisfy the legal requirements of the GDPR.

12.1.2 Principal reason explanations

The other type of feature-highlighting explanation is what we call a "principal reason" explanation. The principal reason approach has a long history in the United States, where the Fair Credit Reporting Act (FCRA) (Fair Credit Reporting Act, Public Law 91-508), Equal Credit Opportunity Act (ECOA) (Equal Credit Opportunities Act, Public Law 93-495), and Regulation B (Regulation B) require creditors—and others using credit information—to provide consumers with reasons explaining their adverse decisions (e.g., consumers being given a subprime interest rate, denied credit outright, or denied a job based on credit, etc.) (Selbst and Barocas, 2018). Under ECOA and Regulation B, these decision makers are required to issue adverse action notices (AANs) to such consumers. Under FCRA, consumers are given a list of "key factors" that negatively affect their credit score. These notices must include a statement of no more than four "specific reasons" for the adverse decision (Fair Credit Reporting Act, Public Law 91-508; Regulation B).[1] A Sample Form in the Appendix to the regulation offers a non-exhaustive list of acceptable reasons, such as "income insufficient for amount of credit requested," "unable to verify income," "length of employment," "poor credit performance with us," "bankruptcy," and "no credit file" (Regulation B, Appx. C, (Sample Form)). Under the regulation, "no factor that was a *principal reason* for adverse action may be excluded from disclosure, [and t]he creditor must disclose the *actual reasons* for denial" (Regulation B, § 1002.9(b)(2), emphasis added).

What counts as a principal reason is not well-defined in either the statutes or regulation. The legislative history of ECOA indicates that consumer education

[1]The number four is not a hard limit under Regulation B, as it is under FCRA, but it is observed in practice.

is a primary goal:

> [R]ejected credit applicants will now be able to learn where and how their credit status is deficient and this information should have a pervasive and valuable educational benefit. Instead of being told only that they do not meet a particular creditor's standards, consumers particularly should benefit from knowing, for example, that the reason for the denial is their short residence in the area, or their recent change of employment, or their already over-extended financial situation (Senate Report No. 94-589, p. 4)

This would seem to suggest that counterfactual explanations, as currently conceived, would serve the intended purpose of AANs. And indeed, some scholars have suggested as much (Ustun et al., 2019). But this very ambiguity also demonstrates that principal reasons are satisfied by a broader array of possible feature-highlighting explanations. For example, the Official Staff Interpretation to Regulation B, originally published in 1985 (Federal Register), suggests that two ways creditors can generate principal reasons:

> One method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors achieved by applicants whose total score was at or slightly above the minimum passing score. Another method is to identify the factors for which the applicant's score fell furthest below the average score for each of those factors achieved by all applicants (Regulation B, Supplement I)

Note that neither approach uses the decision boundary as the relevant point of comparison. Instead, they compare the value of applicants' features to the average value of these features for the credit-receiving or general population. When comparing an applicant against the credit-receiving population, the lender might be able to generate a set of principal reasons that give a rough approximation of the changes that the applicant would need to make to obtain credit. Yet, because the point of comparison is *not* the decision boundary, but rather the average value of each feature across the credit-receiving population, there is no guarantee that a rejected applicant would have received the loan even if the value of the proffered features had been equal to or greater than the average value of the population. Indeed, the rejected applicant could have exceeded the average value of the population on the identified feature, while still falling short on other features necessary to obtain credit. This problem would be exacerbated when comparing the applicant to the general population, where the average value of features would be lower, given that it includes failed applicants along with those that had been successful. Though they are written into the regulation, is it not clear that firms actually use these methods to generate principal reasons.

We can also imagine a similar approach that measures an applicant against the decision boundary or the maximum point on the response surface (i.e., "the ideal, most creditworthy possible customer" (Hall et al., 2017))—evaluating how far an applicant falls from either point along each feature. The decision maker could then rank order features by the distance the applicant would need to travel to reach either point, and generate an explanation by highlighting the top four features. In effect, the applicant would learn the four features along which she is most deficient.

12.1.3 Different ways of respecting decision subject autonomy

There is no natural way to choose between these different approaches. Yet rarely is the choice to use one method over another discussed explicitly, or, indeed, even recognized as is a choice in the first place. These methods produce different explanations and serve fundamentally different goals.

Focusing on features that are furthest from the average value of the features in the credit-receiving or general population casts the problem of identifying principal reasons as one of identifying extreme deficiencies that would seem to rule out the applicant completely, rather than near-misses that applicants might readily address before applying for credit again in the future. While the former may strike us as a less attractive or sensible approach to explanation, there may be good reason to favor an explanation that makes clear the features that were held against an applicant. With the latter approach, while the applicant might receive helpful advice, she might not learn that other features were viewed by the model as a crucial mark against her.

Principal-reason explanations treat importance in terms of procedural justice: to respect the autonomy of a decision subject, the decision subject deserves to know which factors dominated the decision (Tyler, 2006). In counterfactual explanations, respect for autonomy means that decision subjects need to know how choices affect outcomes, and thus how they can take actions that will most effectively serve their interests in the future. The former operates more like a justification for a decision—a rationale, with little immediate concern for recourse—whereas the latter serves a more practical purpose—providing explicit guidance for achieving a different decision in the future. In keeping with these differences, the principal reasons offered by creditors tend to be vaguer

("Income insufficient for amount of credit requested"), while counterfactual explanations aim for precision ("Had you earned $5,000 more, your request for credit would have been approved").

12.1.4 A shared focus on subsets of features

Importantly, these methods all have one thing in common. None of them involve disclosing the model in its entirety. They focus, instead, on disclosing a limited set of features that are most deserving of a decision subject's attention. By design, they do not provide an exhaustive inventory of all the features that a model considers. In practice, learned models can often consider a very large set of features, and an explanation that suggests changes to each of those features would be overwhelming. As a result, both the law (in the form of principal reasons) and the emerging technical literature (in the form of counterfactual explanations) seek to produce "sparse" explanations that present the decision subject with only a small subset of features (Wachter et al., 2018). This fact is what binds them together conceptually in a single group, and forms the core point on which the chapter builds its discussion.

12.2 Feature-highlighting explanations in practice

The utility of feature-highlighting explanations to a decision subject relies on several hidden assumptions. In this section, we identify five such assumptions, explain why they might not be valid, and explore the consequences of that realization. Some of these assumptions appear to be necessary to the utility of

feature-highlighting explanations. Others may not be strictly necessary, but still seem to underlie the great majority of work on counterfactual explanations so far.

12.2.1 Features do not clearly map to actions

Feature-highlighting explanations often assume a clear and direct translation from suggested changes in feature values to actions in the real world. In many cases, this is a reasonable assumption: instructing someone to reduce their total lines of credit maps onto the obvious action of cancelling a credit card or fully repaying—and thus dispensing with—a loan. In most of the contexts in which existing scholarship considers the challenge of explaining the decisions of a machine learning model, there is a clear correspondence between the feature values that one is told to change and the actions that one would take to achieve those changes. The mapping can be so obvious as to go unquestioned—as if there is never a situation in which one is unable to translate the necessary changes in feature values into a set of concrete steps in the real world. And yet, in many cases, we are only able to perform this mapping because we have relevant domain knowledge and an implicit causal model in mind that relates specific actions to predictable changes in feature values.

Even when the highlighted feature seems to refer to something rather concrete, the actions a decision subject can take to affect those features may not line up with the features themselves. For example, a recommendation that someone increase his income can lead a person to take one of several actions: he can seek a new job, ask for a raise, or take on more hours. As Figure 12.1 illustrates, these

actions are not as simple as "increase income" or "increase length of employment."

But this example still assumes a relatively direct relationship. To act on explanations that instruct us to change certain feature values, we need to know what causes features to change value in the real world. This might not be obvious; we may struggle to identify the actions that would cause a feature value to change—or change in a predictable way. Consider a model that takes credit score as an input, as in certain insurance pricing models (Kiviat, 2019). A counterfactual explanation for a bad insurance quote might be to raise one's credit score. But figuring out how to raise one's credit is itself famously unclear, so this information is hard to know how to act on. In another example, as data from alternative sources like social media are incorporated into credit scoring models, we may receive explanations for adverse decisions that instruct us to improve our "social media score" without providing meaningful guidance for how to do so.

When we consider the gap between what needs to happen to feature values to get a different decision from the model and what needs to happen in the real world for a person to change these feature values, we can begin to recognize some additional challenges with feature-highlighting explanations.

Consider the recent work on "actionable" explanations, which has focused on avoiding explanations that tell people to make changes that are impossible, placing the burden on decision makers to give advice that is sensitive to the actual steps that decision subjects would need to take to achieve the change in feature values (Ustun et al., 2019; Mothilal et al., 2020). Avoiding these potential explanations is a matter of identifying the lack of any possible causal mecha-

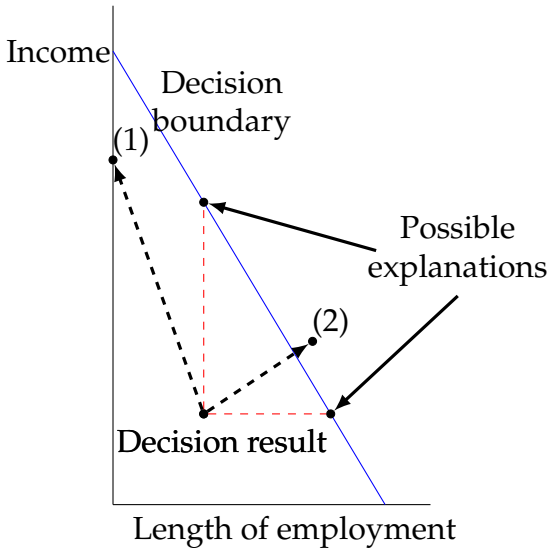


Figure 12.1: A decision based on two features—income and length of employment—will be explained by reference to one of the features, either the shortest or longest distance from the boundary. But the explanations do not map to the decision subject's possible actions that can affect them. Point (1) represents getting a higher-paying job, and point (2) represents waiting for a raise.

nism that would have the necessary effect on the value of some feature. This becomes clear if we compare mutable characteristics under one's control (e.g., income) to (1) mutable characteristics once under one's control that cannot be undone (e.g., declaring bankruptcy), (2) mutable characteristics outside one's control (e.g., age), and (3) immutable characteristics (e.g., race). While it might be helpful to know in advance that one should avoid declaring bankruptcy, given its effects on future credit decisions, being told that you should not have declared bankruptcy after the fact might not be a helpful explanation if the goal is to provide actionable advice.

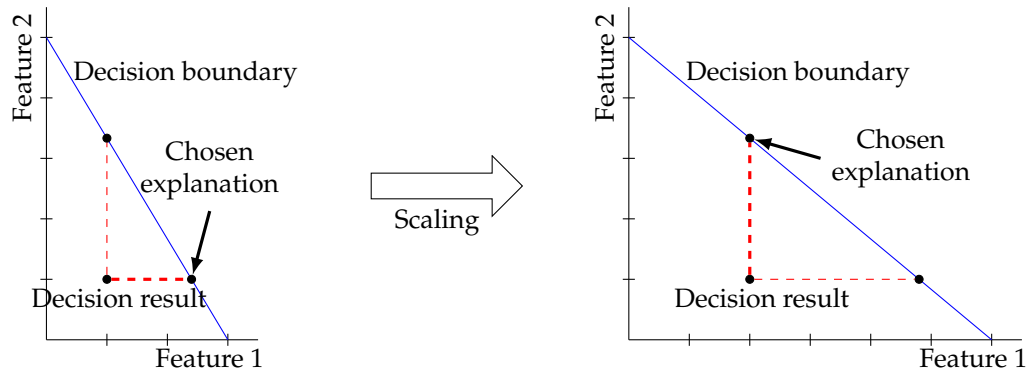
We can also recognize that "gaming" is another case of the disconnect between feature and action. When a decision maker instructs someone to change certain features, the decision maker will often assume that the person will take

a specific desirable sequence of actions because that is the causal mechanism that the decision maker has in mind. But there are often many other ways to change feature values that don't require taking these steps. Opening a new line of credit, for instance, can increase an individual's overall available credit without fundamentally changing their ability to pay off a loan. And yet, in recognizing this gap, we can also appreciate that feature-highlighting explanations leave room—perhaps much needed room—for decision subjects to find strategies to change feature values that decision-makers might not be able to anticipate in advance or on their own.

This insight also alerts us to the possibility that actions may affect multiple features simultaneously. As Figure 12.1 demonstrates, whether one increases his income by finding a higher-paying job or waiting for his performance review to get his raise, the action will affect both income *and* length of employment, a separate feature in the model. In the case of a job change, length of employment will be negatively affected. Thus, increasing income may not be enough to get credit, which is why point (1) is on the left of the decision boundary. In the case of waiting for a raise, a smaller increase in income might be needed than the explanation would say, because length of employment increases at the same time. This is why point (2) is on the right side of the decision boundary, despite not increasing income as much as the explanation suggests. Highlighting certain features as those that need to change to obtain a different decision implicitly relies on the belief that everything else can be held constant while making these changes. In reality, changes in the value of one feature may affect the value of another feature, if the two features interact, thus changing the efficacy of the explanation.

Insisting that explanations exhibit sensitivity to these constraints is analogous to insisting that explanations consider the causal mechanisms that allow decision subjects to alter the value of specific features. Indeed, the only way to ensure that the recommended change is even possible, to prevent gaming, and to account for dependencies between features is to model the outcome of interest using features that directly figure into the causal mechanism. The idea that we can identify all such constraints in advance assumes that the actions necessary to change specific feature values will always be self-evident.

12.2.2 Features cannot be made commensurate by looking only at the distribution of the training data



(a) Before scaling (b) After scaling

Figure 12.2: The counterfactual explanation depends on how axes are scaled. Scaling of the Feature 1 axis by a factor of 2, the closest explanation changes to highlight Feature 2 instead of Feature 1.

All feature-highlighting explanations rely on some notion of a distance between the observed values for various features and some reference point, whether the average person, maximum value, or the decision boundary. Relying on distance requires normalizing features, because there needs to be a

shared scale between features in order to meaningfully compare them. For example, as discussed in Section 12.1.1, an increase in length of employment is not commensurable with an increase in salary. Normalization attempts to capture the fact that salaries may vary on the order of thousands or tens of thousands of dollars, but length of employment varies at a numerically much smaller scale.

Several statistical techniques exist to address this problem, scaling features so as to make them seemingly comparable, and different explanation techniques will use different approaches. Following Wachter et al. (2018), the literature on counterfactual explanations has mostly converged on a heuristic that finds the Median Absolute Deviation (MAD) under an L1 distance norm (Russell, 2019; Mothilal et al., 2020; Grath et al., 2018). Meanwhile, it is entirely unclear what methods principal-reason explanations use—the regulations do not specify, and it is never discussed in practice—but the nature of a distance metric requires that *something* be used.

Functionally, the details of different normalization methods are not terribly important for our purposes, but what they do is measure how spread out the training data are for each feature. Most importantly, the normalization relies on information contained entirely within the data. Different statistical techniques will use ranges, standard deviations, or MAD, but in all methods, the axes are scaled based entirely on the distribution of the data, not some external point of reference.

The choice of how to normalize is extremely important because the ultimate explanation is highly sensitive to scale. Figure 12.2 shows the effect of relative scale. Doubling one axis relative to the other completely changes the explanation on offer.

When examined from the perspective of a decision subject who must take some action in response to these explanations, normalization based simply of the distribution of data is somewhat arbitrary. One decision maker might scale the axes such that increasing income by $5,000 annually is equivalent to an additional year on the job. But if the data were distributed differently, the result could have been that $3,000 of additional income is equivalent to another year. Similarly, a competing lender, using different training data, could conclude that $10,000 of income corresponds to one year of work. Because normalization techniques are inward-looking, they are agnostic as to the meaning and origin of the data, and the scale is just an artifact of the data distribution.

What counts as "important" or "easiest to change" with respect to the data is not necessarily what counts as important or easiest to change with respect to the decision subject. Without an external point of reference to ground our scales, the meaning of the relative difference in feature values is unclear. This is an example of what Selbst et al. (2019) have called the "framing trap," where an attempt to solve an accountability problem ignores interactions with the outside world.

Because we are concerned with decision subject's menu of actions, the most sensible external referent would be something akin to the cost of making the required change, where cost can imply dollars spent, effort, or time. For counterfactual explanations, those features that involve little *cost* to change, even if they involve considerable change along a normalized numeric scale, may be far more useful to highlight than those that would involve considerable cost to change. Thinking this way forces us to think about the changes that would be necessary to realize the recommendations offered by these explanations, and

their difficulty. Telling someone to go find a new job seems like a much more difficult demand than telling her to raise her income by $1,000, but that might be what such a person is being told to do.

If, instead, the preferred approach is in line with generating principal reasons, features that are costly or impossible to change may be precisely the ones that should be highlighted. Knowing that one's application for credit has been rejected due to characteristics outside her control might be paramount if the goal is to ensure some degree of procedural justice or to reveal when to contest the decision (Mendoza and Bygrave, 2017; Wachter et al., 2018; Brkan, 2019). Given its focus on identifying the features easiest to change, counterfactual explanations might give a company engaged in intentional discrimination the opportunity to conceal that its decisions largely hinged on immutable characteristics like race. Worse, such explanations might mislead the subjects of these decisions into believing that their fate was determined by factors under their control.

Thinking about changes in terms of their real-world cost therefore helps to translate numerical changes in feature values to real-world actions. This is true whether we want to point out either what is easiest or most difficult to change. Some works seek to account for the cost of actually manipulating those features in practice by assuming domain knowledge or user input of these costs (Ustun et al., 2019; Grath et al., 2018). Other formulations provide the option to solicit user input (Mothilal et al., 2020), but in general it can be infeasible for the individual to specify all of the relevant real-world constraints that affect the utility of an explanation, and eliciting or estimating each individual's conception of cost is notoriously difficult. Humans struggle to articulate the costs they experience (Kahneman and Tversky, 1977), and the very notion of "cost" often suffers

from difficulties pinning down what it measures (dollars spent, effort, time, etc) and how to properly value intangibles (Frank, 2000).

Worse yet, the cost of making certain changes will not be consistent across different people. Changes that might be rather inexpensive for one person to make might be costly for another person to make. Thus, when we use explanations to identify the easiest or most difficult features for someone to change to achieve a different decision from a model, the explanation must be sensitive not only to how these changes involve different costs, but how these costs vary across the population. Different subsets of features may be appropriate for different people with different life circumstances. This complication cuts against the very desirability of these explanations: the idea that we can automatically determine what is easiest or hardest to change by examining what values are closest or farthest from the decision boundary.

12.2.3 Features may be relevant to decision-making in multiple domains

Feature-highlighting explanations may interact with facts about a person's life that are invisible to the model. The supposition of a counterfactual explanation is that it is offering advice for the kinds of changes that it would be rational for a person to make to achieve better results in future decisions. Some commentators and scholars have cautioned that explanations should never encourage people to take actions that are irrational or harmful (Hall et al., 2017; Eubanks, 2018b). What they mean more specifically is that there may be some recommendations that are indeed rational if the only goal is to obtain a positive decision from

the model, but irrational with respect to other goals in a person's life. It is not possible to determine what would be a rational action in isolation.

An oft-cited common-sense example for this proposition is that an explanation should never recommend that a person seek to make less money (e.g. Lipton, 2018). While we believe it unrealistic that an actual credit model would ever recommend such a thing, the example still holds value. It is self-evident that no one would want to make less money, even if doing so would improve their access to credit. Or consider an example that reverses this dependency: a person contemplating applying to a new job for its superior health insurance is unlikely to remain at his current job because an explanation for a failed credit application told him to do so. In this case, acting on the recommendation would impose an opportunity cost on the consumer by forcing him to forgo benefits in other domains. When other aspects of one's life depend on some of the same features, explanations for how to get the desired outcome in one aspect of your life may conflict with those in another.

We can reason about this the other way around as well. From the point of view of a counterfactual explanation, an applicant might be best off trying to change a number of other features *besides* income. Yet, from the perspective of the applicant, increasing her income might have ancillary benefits in other parts of her life that make this change more attractive—and indeed rational—than those suggested by the explanation. Increasing her income would grant her improved access not only to credit, but to improved quality of life, generally.

In the first case, a change in feature might benefit the decision subject in one domain, while hurting her in others. In the second case, a change in a feature might benefit the decision subject in multiple domains, not just one. These

spillovers—both negative and positive—complicate the process of determining which features would be most useful to highlight in an explanation. Ideally, feature-highlighting explanations would allow decision subjects to avoid negative spillovers and identify opportunities for positive spillover. But a decision maker will lack information about the many other goals that a person might have in her life and the features that are relevant in those domains.

This lack of information goes both ways, as well. By design, feature-highlighting explanations do not offer an exhaustive inventory of all the features that the model considers. Yet withheld features may still matter to the decision to some degree. Decision subjects are only told what needs to change to obtain a different result from the model in the future. But if other features need to remain unchanged for the recommended changes to have their promised effect, this may cause problems.

Thus, due to other life goals, decision subjects may change undisclosed features unless otherwise instructed. For example, if a counterfactual explanation tells someone to increase their income and lower their debt, but fails to mention that they should not reduce their length of employment, the person may have no idea that they should avoid any career change while attempting to address these other issues, stumbling accidentally into point (1) in Figure 12.1. Indeed, the person might not even know that length of employment figured into the credit decision in the first place. Or, as noted in Section 12.2.1, if no action lines up exactly with that feature, she might be forced to take an action that changes multiple features simultaneously.

12.2.4 Models must have certain properties: monotonicity and binary outcomes

The utility of feature-highlighting explanations tends to rely on two aspects of underlying models: monotonicity and binary outcomes. The former is likely necessary to their utility; the latter may not be.

When dealing with continuous variables, feature-highlighting explanations—particularly counterfactual explanations—will often take the form of instructing decision subjects to make changes in feature values in a specific direction and by a specific amount: increase your length of employment by 1 year; decrease your debt by $10,000; etc.

Decision subjects will not necessarily be able to alter the value of these features through some sudden change. Instead, they may have to make incremental changes in the direction of the specified value. And despite their best efforts, decision subjects might struggle to hit the specified feature value; their efforts could move the value of these features in the right direction, but ultimately fail to get the decision subject all the way there. Similarly, decision subjects might lack precise control over the value of a feature, making it difficult to avoid overshooting the mark when they take some action.

This can pose serious problems for feature-highlighting explanations if the features in the underlying model have not been subject to monotonicity constraints. A monotonicity constraint guarantees that as the value of the features move in the recommended direction, the decision subject's chances of success consistently improves. Without monotonicity constraints, a model might learn complex and even counter-intuitive relationships between certain features and

an outcome of interest. For example, a model might learn that people who have spent two to four years at their current job are good candidates for credit, while those who have stayed five or more are not. Likewise, carrying more debt might render applicants less attractive, until they start earning more income, at which point additional debt might make them more attractive.

Unless the model exhibits monotonicity with respect to the highlighted features, the decision subject might find herself in a *worse* position as she moves toward the specified value or if she overshoots the mark. Explanations of models that lack a monotonicity constraint will be brittle. Worse, because monotonicity is intuitive, decision subjects are likely to assume that the property applies, even if they have no formal understanding of the concept.

Separately, in the computer science literature, model explanations often assume that outcomes are binary: did the applicant receive a loan or not? But in reality, the creditor decides not only whether or not a loan is given, but also the loan's interest rate. A counterfactual explanation presented to the applicant must necessarily account for this. Does the decision-maker choose a specific target interest rate when providing an explanation? What if the applicant is only interested in a loan below a certain rate?

As before, we see that in order to provide a useful explanation, the decision-maker needs to have relevant information about a particular individual's life circumstances. Alternative goals will require different explanations, which may have nothing to do with each other. For example, consider a financially responsible borrower who will only accept a loan at a sufficiently low interest rate that she is confident that she would be able to make her monthly payments. If she is told that she could qualify for a high-interest loan by increasing her in-

come without reducing her debt, she learns nothing about how to qualify for a low-interest loan; what it takes to obtain a high- or low-interest loan is not necessarily related. Getting a low-interest loan might additionally require either paying down her debt or opening up and using new credit lines, which in effect raises her debt. There is no way to extrapolate from the counterfactual explanation that gets her to a high-interest loan. Indeed, she may not even know that the counterfactual explanation that tells her how to get a loan is specific to a high-interest loan, instead seeing the interest rate on offer as the only option, and concluding that she cannot get the loan.

Without explicitly accounting for the all the possible types of outcome and the subject's preferences for them, the decision-maker must make assumptions and choices on behalf of the subject that significantly impact the usefulness of the explanation given.

12.2.5 The validity of explanations may not remain stable over time

Explanations occur at a single point in time, even though a person's life circumstances can change over time. Likewise, a model may very well change over time, as decision-makers retrain their models in light of new data—including new behaviors induced by the explanations offered to prior decision subjects.

Individuals cannot in general instantaneously change the features suggested by an explanation. Some features, like length of credit history, are inherently time-dependent. Others changes may take varying degrees of time to imple-

ment. Suppose, for instance, the decision subject could obtain credit either by reducing their debt by $5,000 or by increasing the length of their credit history by 6 months. If it would take 6 months to pay down this debt anyway, it would seem unnecessary for them to do so instead of simply waiting 6 months to qualify. On the other hand, if the subject needs credit immediately and is willing to pay down the $5,000 right away, then we might view debt reduction as the right feature to provide the subject in an explanation. Thus, the "right" explanation to give a decision subject may depend heavily on temporal aspects of their life.

Meanwhile, models are often retrained to react to changes in the overall environment or borrowers' behavioral patterns. Perhaps there is another recession. Perhaps lenders gain access to new data sources or otherwise figure out how to better model borrowers' behavior. Or borrowers might even change their behavior en masse based on the very explanations that banks offer, as the data distribution shifts over time (Liu et al., 2018). Any of these changes would necessitate model retraining by the lender.

At the same time, rejected applicants may be acting on recommendations that are frozen in time, despite ongoing changes made by lenders. Wachter et al. (2018) have argued that the law should treat a counterfactual explanation as a promise given to the rejected applicant rather than just an explanation. They argue that if a rejected applicant makes the recommended changes, the promise should be honored and credit granted, irrespective of the changes to the model that have occurred in the meantime. Whether this is the right approach or not, it is a recognition that without such a guarantee, counterfactual explanations might not serve their purpose when one considers the time it takes to make the suggested changes.

12.3 Unavoidable tensions

We have argued so far that the need to disclose a limited subset of features infuses feature-based explanations with subjective choices and creates a number of challenges that makes their promise harder to realize in practice than advocates of such techniques would have us believe. They also present a number of unavoidable normative tensions. Decision-makers start with a great deal of power over decision subjects, and the purpose of explanations—and the legal requirements for them—is to restore some degree of power to the decision subjects. Yet the fact that decision-makers must, by necessity, withhold information creates three unavoidable tensions. First, in order to generate genuinely helpful explanations, decision-makers must be both paternalistic and privacy-invasive. That is, they must interfere with decision subjects' autonomy to offer some back to them. Second, while designed to restore power to decision subjects, partial explanations grant a new kind of power to the decision maker, to use for good or to abuse as desired. Finally, the additional transparency that might help overcome some of the problems with feature-highlighting explanations render these techniques anathema to the decision makers that we would want to use them.

12.3.1 The autonomy paradox

Feature-highlighting explanations are motivated by either the desire to make recommendations to decision subjects or to justify the model's decision. Both these motivations are fundamentally justified by reference to the autonomy of a decision subject. Recommendations appeal to an instrumental vision of autonomy, where information enables action. Justification, however, is more focused

on a moral conception; the information is due because the subject deserves to know. Both of these motivations are complicated by the need to withhold some information.

It is worth recalling two predicates to our analysis. First, feature-highlighting explanations are employed as an alternative to direct oversight or audits of the model. We recognize that both can be deployed simultaneously, but the purpose of this chapter is to examine the feature-highlighting explanation as a tool separate from direct oversight; if employed in tandem, it is just as worthwhile to ask what they might add. Second, disclosure of the entire model is not practical. This is true both because of trade secret and gaming concerns, but also because models with even a moderate number of features would overwhelm the decision subject, defeating the very autonomy rationales that justify this approach.

Ironically, respecting a decision subject's autonomy requires making assumptions about which information will be valuable to a given decision subject. The decision maker will not know how features correspond to actions in the real world, and thus which features a decision subject could most readily change. The decision maker does not know how features in its model interact with features in the real world, and thus how features relate to each other. The decision maker does not know how a change in the measured feature, or the action required to make such a change, affects other aspects of a person's life positively or negatively. And assuming the decision maker could achieve a general sense of these facts, they further do not know how they vary from person to person. All this means that the choice of features to disclose can have unintended effects for decision subjects, which would have been avoided with

a different disclosure. Given the informational position of the decision maker, there is simply no way to fully realize its commitment to respecting a decision subject's autonomy.

One might suggest that these problems can be solved with even more data. If explaining a credit decision, the relevant factors might be affected by information about a person's health, family situation, or future educational plans. A person's choices about whether to look for a new job will be affected if they are sick, have a new baby or aging relative to care for, or are saving to go back to school. This information can directly or indirectly be mined from other sources in the world, such as social media data. If a decision maker understands other aspects of a person's life that may interact with the decision, then it might be able to offer explanations that are appropriate and tailored, or might be able to focus on features that are relevant to decisions in multiple contexts. So perhaps the answer is to collect it all.

Unfortunately, allowing a decision maker, such as a lender, to collect and connect every bit of information about a person's life is not really a solution. Rather, it is a privacy disaster. Under every major theory of privacy this would be impermissible (e.g. Solove, 2006; Cohen, 2012a; Nissenbaum, 2009), and because privacy is a fundamental aspect of autonomy (Cohen, 2012b; Reiman, 1976), this leads to an autonomy paradox. The problem is clearest through the lens of Helen Nissenbaum's theory of contextual integrity (Nissenbaum, 2009). Contextual integrity argues that a privacy violation occurs where information flows between actors in social context in ways that violate the informational norms relative to that context. So in one sense, this solution—allowing lenders in financial context to access social media information—is definitionally prob-

lematic. A primary concern of contextual integrity is for social contexts to keep operating as they should. If creditors have access to social media data, the worry is that they will make credit determinations based on public friendships, for example, and as a result people will have incentives to change or hide their social relationships in order to get better credit (Packin and Lev-Aretz, 2016). This will harm the social contexts in which friendships flourish for the sake of credit. Ironically, then, while in Section 12.2 we argued that explanations would not be useful unless the decision maker understood facts about people's lives beyond those considered in the model, contextual integrity would suggest that the fact of decision makers knowing this information is itself harmful to autonomy.

But imagine a decision subject who, facing an adverse credit decision, is shown the complete model and finds it overwhelming. Such a person may instead prefer the counterfactual explanation, even if it involves disclosing all the information necessary for the decision maker to offer an appropriately tailored set of instructions. This exact scenario motivates much of the work on counterfactual explanations, so we should not discount that a more informed explanation can still be autonomy-enhancing on balance. In a sense, this tension is reflective of a common concern in discussions of autonomy: When can giving up information and agency be autonomy-enhancing? For example, if a person hires an attorney, she outsources some important decisions, gives up very private information, and often gets answer back that she cannot understand, but it is her decision to do so for her own good. It is doubtful that anyone would consider hiring a lawyer to be a loss of autonomy. At the same time, we would immediately recognize that furnishing lenders with detailed information and relinquishing control over decision-making is not an obvious mechanism for enhancing one's autonomy. Notably, there are no requirements that they act in

your best interest, while such fiduciary obligations do apply to lawyers. (e.g. Balkin, 2015; Khan and Pozen, 2019). This is a difficult tension to resolve, and may depend on the relative power of and constraints upon the decision maker, rather than the quality of the explanation.

12.3.2 The burden and power to choose

One of the reasons model explanations are so appealing is that they appear to offer complete automation: whenever a decision is made, an explanation can be provided without any further human intervention. But this veneer of mechanization belies the fact that feature-highlighting explanations cannot be completely formulaic. They require decisions about what to disclose and assumptions about the real world.

The need for partial disclosure grants new power to the decision maker. Of course, a decision maker—by virtue of being one—has always had power over the decision subject. But by attempting to return power to the decision subject via an explanation that, for her own sake, cannot be a complete explanation, we grant a new form of largely unanticipated power to the decision maker. Furthermore, the requirement to make certain assumptions about the real world also grants power to the decision maker. Whenever there is ambiguity in the individual's preferences, the decision-maker has the power to resolve it however they see fit. This leaves the decision-maker with significant room to maneuver, the choice of when and where to further investigate, and more degrees of freedom to make choices that promote their own welfare than we might realize.

This new power can be used for good or ill. Consider, for example, a de-

cision maker providing a counterfactual explanation for why an individual did not qualify for a loan. As discussed in Section 12.2.4, this decision (and therefore explanation) is not simply binary—in its explanation, the decision maker must give the decision subject a counterfactual that would result in a *specific* interest rate. At best, it might allow the subject to choose their target interest rate. Alternatively, it might—somewhat paternalistically—choose the interest rate that it believes is "right" for this subject. More insidiously, it might choose the interest rate that is likely to maximize its profit. Ultimately, the point is that in the absence of standards or robust avenues for user input, the decision maker is left with the power to make this decision on its own.

This power is not simply limited to the choice of outcome. As we have argued, many aspects of explanations are unspecified by the law or by technical proposals, including what factors can be included in an explanation, what the relative costs of various features are, and how to account for real-world dependencies between them. The key point here is that left to their own devices, decision makers are afforded a remarkable degree of power to pursue their own welfare through these choices.

12.3.3 Too much transparency

Decision makers might seek out different ways to address the difficulty of taking decision subjects' real-world circumstances into account when generating counterfactual explanations. A number of recent papers propose presenting the user with a diverse set of counterfactual explanations (Mothilal et al., 2020; Russell, 2019; Wachter et al., 2018), allowing the decision subject to choose among

several possible ways to achieve a favorable outcome. This approach accepts that decision makers may lack the capacity to ever fully account for the unique constraints and preferences of decision subjects, instead providing a wide range of possible paths to success from which the decision subjects can choose. Doing so allows the decision subject to rely on knowledge of her own particular circumstances in selecting among these.

Others have advocated in favor of interactive tools that allow decision subjects to explore the effect of making changes to certain features (Citron and Pasquale, 2014; Hildebrandt, 2006). Industry has even implemented some such tools.[2] This approach gives decision subjects greater freedom to explore the space, using a deep understanding of their own constraints and preferences to investigate the effect of certain adjustments.

Still other work adopts an entirely different approach, focusing instead on finding ways for the decision maker to learn more about decision subjects. In particular, there have been recent proposals to devise mechanisms for soliciting input from decision subjects, allowing them to communicate whether they find certain counterfactuals helpful, whether changes to certain features are out of the question or less desirable, and what other preferences they might hold (Mothilal et al., 2020).

These approaches could also work in concert, seeding decisions subjects with an initial set of diverse explanations that could serve as starting points for interactive exploration. In theory, this would have the benefit of helping to ensure that decision subjects do not fail to explore the space sufficiently, concluding their investigation after only making a small number of adjustments

[2]See, e.g., Credit Karma's Credit Score Simulator: https://www.creditkarma.com/tools/credit-score-simulator/

from one initial starting point.

Unfortunately, each of these approaches runs the risk of revealing a sufficient amount of information about the underlying model to reconstruct it (Tramèr et al., 2016). As a result, while these approaches may be the most promising to overcome certain difficulties, they create difficulties of their own. Firms concerned with intellectual property and gaming are unlikely to afford decision subjects extensive freedom to explore.

12.4 Conclusion

Feature-highlighting explanations have been embraced as a way to help decision makers avoid a number of difficult trade-offs, granting firms the capacity to provide meaningful and useful explanations of machine-learned models without having to compromise on model performance, while also respecting concerns with trade secrecy, gaming, and legal compliance. Advocates have championed this style of explanation as an elegant way to honor and enhance decision subjects' autonomy even as machine learning models grow in complexity and ubiquity.

Yet as we have shown, these explanations lack a connection to the actions required to change features. They fail to consider the cost of these actions, decision subjects' preferences, and the effects of the necessary action on other parts of decision subjects' lives. Worse, attempts to correct these deficiencies undermine the very goals of explanation by violating decision subjects' autonomy in the name of enhancing it and granting more power to decision makers when trying to return it to decision subjects.

So what can be done? How can feature-highlighting explanations be useful, while protecting the autonomy of decision subjects? Much more work is needed to address the issues we have raised here, but we see three concrete avenues worth exploring.

First, at an absolute minimum, given the power that these explanations grant to decision makers, they should disclose the method by which they generate explanations. Additionally, legal requirements for explanation and adverse action notices should be amended to require this. Without understanding the method of explanation, decision subjects have no hope of understanding how to effectively realize their goals.

Second, we need to understand what actions people actually take when confronted with feature-highlighting explanations—and which disclosures help people act most effectively. Empirical research is essential to answer these questions. One obvious place to start such work is with longitudinal data documenting the successful paths that previous decision subjects have taken to receive positive outcomes when starting from various circumstances. Another is to engage directly with decision subjects to develop a richer account of their everyday strategies for responding to models and explanations of their decisions, as Malte Ziewitz and Ranjit Singh have done over the past few years.[3] This approach rests on the idea that explanations should focus on communicating what had worked well for other people under seemingly similar conditions.

Third, the concerns about power imbalance created by counterfactual explanations suggests that as they become more prominent, policy responses centered around the concept of "information fiduciaries" should be consid-

[3]https://zwtz.org/restoring-credit/

ered (Balkin, 2015). These proposals are not universally accepted (Khan and Pozen, 2019), and may not be an appropriate policy response in all contexts involving decision made with data, but the concern raised in Section 12.3.2—that someone has decision-making power over another person, yet that second person must rely on the decision maker to act in her best interest—is precisely the concern that motivates fiduciary duties in other spheres.

These proposals will only address some of the issues with feature-highlighting explanations raised here. There is still more work to do, in computer science, social science, and policy, if we want to understand when and where feature-highlighting explanations can be useful to decision makers and decision subjects alike.

Part V

Conclusion and Future Work

CHAPTER 13

FUTURE DIRECTIONS

The research presented in this thesis seeks to leverage the potential benefits of formal, algorithmic decision making while highlighting and mitigating the risks involved in so doing. Combining insights from a variety of fields including computer science, economics, sociology, and legal studies, we hope to inform ongoing work in theory, practice, and policy on the interface between algorithms and society.

We conclude with a number of directions for future research, roughly divided into the following categories: fairness in machine learning and mechanism design (Finocchiaro et al., 2021); algorithmic discrimination; and transparent and meaningful explanations. These topics, while not intended to be exhaustive, encompass a wide range of potential and ongoing research areas seeking to make algorithms work better for society.

13.1 Fairness in Machine Learning and Mechanism Design

As researchers, policy-makers, and practitioners increasingly turn their attention to issues of bias, fairness, and discrimination in algorithmic decision-making, there has been a growing realization that the complexity of modern decision-making systems makes it necessary to draw upon insights, tools, and methods from a variety of technical fields. In a number of application domains, like advertising, healthcare, and hiring, both machine learning and mechanism design are crucial to the design and analysis of decision-making tools. In re-

cent work, we survey the relationship between mechanism design and machine learning with a particular focus on questions of fairness (Finocchiaro et al., 2021). Building on this discussion, we highlight a few directions for future work.

Allocation and prediction. At a high level, much of the work on fairness in mechanism design concerns the *allocation* of scarce resources; in contrast more recent work on fair decision-making in machine learning settings focuses on *predictions* (Finocchiaro et al., 2021). Increasingly, modern decision-making systems (for example, online advertising platforms) incorporate both allocative and predictive elements. In such cases, what constitutes fair or equitable decision-making implicates notions of fairness from each of these fields. Future research is needed to fully integrate these ideas of fairness, with a particular eye towards implementation feasibility in large, complex systems.

Accounting for preferences. In motivating settings like hiring and lending in the literature on fair machine learning, we often assume that every decision subject has a known preference over outcomes (e.g., every applicant wants a job or a loan). In settings like content delivery, however, a decision-maker typically does not know the individual's preferences. In such cases, the decision-maker must learn to *personalize* decisions, and ideally, do so in a way that may be considered fair. This question of fair personalization incorporates ideas from both machine learning (What complexities are introduced by the process of learning?) and mechanism design (How might we model and elicit individual preferences?), and recent work seeks to develop frameworks to express ideas of fairness in such settings (Celis and Vishnoi, 2017; Kim et al., 2020; Çapan et al., 2020).

Designing around behavior. A natural question to consider in mechanism design is the strategic response of participants. There has been a recent resurgence of interest in strategic behavior in the machine learning setting (Dalvi et al., 2004; Brückner and Scheffer, 2011; Hardt et al., 2016a; Milli et al., 2019; Hu et al., 2019). In order to develop principles for the responsible deployment of algorithmic decision-making systems, we must carefully consider the incentives the produce. Not only will this make systems more robust to strategic behavior and the disparities it might produce (Hu et al., 2019; Milli et al., 2019), it can help ensure that decision-making systems do not produce perverse incentives for participants (Eubanks, 2018a; Kleinberg and Raghavan, 2020; Barocas et al., 2020). In particular, a growing line of work seeks to understand how to align decision subjects' incentives with actions that are in their own interests (Kleinberg and Raghavan, 2020; Tang et al., 2021; Alon et al., 2020; Miller et al., 2020; Haghtalab et al., 2020; Shavit et al., 2020).

13.2 Algorithmic Discrimination

Much of the recent technical work on bias in algorithmic decision-making centers around defining normative terms like "fairness" in quantitatively rigorous ways. In parallel, a growing body of work in the legal literature seeks to understand how exiting antidiscrimination protections apply to machine learning and algorithmic decision-making more broadly. There is a natural and emerging connection between these communities: while the law does not seek to provide prescriptive definitions of what fair decision-making should look like, it does offer some clear guidance on what constitutes discriminatory behavior. A central challenge for the future is the interpretation of existing laws in light of

new technical tools like machine learning (Barocas and Selbst, 2016). This will require innovation, both in terms of novel technical methods as well as new legislation or the re-interpretation or clarification of existing legislation.

As it pertains to discrimination law, the work in this thesis (Chapter 11) pertains primarily to algorithmic discrimination in the context of hiring, which has been the subject of recent work in the legal and computer science literatures (Ajunwa, 2021, 2020; Bornstein, 2018; Cofone, 2018; Kim, 2016, 2017, 2018; Harned and Wallach, 2019; Raghavan et al., 2020; Sanchez-Monedero et al., 2020). Beyond employment, concerns over algorithmic discrimination manifest in contexts like advertising (Lambrecht and Tucker, 2019; Blass, 2019; US Department of Housing and Urban Development, 2018), finance (Gillis and Spiess, 2019; Morse and Pence, 2020; Tantri, 2020), education (Porter, 2020; Kizilcec and Lee, Forthcoming), and housing (US Department of Housing and Urban Development, 2019; Selbst, 2019).

Kleinberg et al. (2019) point out that algorithms have the potential to make the detection of discrimination easier and more standardized; however, this is not inevitable. Further work at the intersection of computer science and law is needed to ensure that strong protections against discrimination apply to algorithmic decision-making. From a technical perspective, we need new algorithmic tools to diagnose and mitigate discrimination, particularly in complex, modern decision-making systems that go beyond simple classification tasks (Finocchiaro et al., 2021). Beyond our existing technical limitations, we need further inquiry from legal and policy perspectives into the potential to impose new regulatory frameworks to prevent discrimination. For example, algorithms can be audited before their deployment, opening the door to new antidiscrimi-

nation legislation (Kim, 2017; Wilson et al., 2021; New York City Council, 2020). In prior work, we have provided policy recommendations in the context of algorithmic discrimination in hiring (Raghavan and Barocas, 2019; Raghavan, 2020); more generally, efforts to prevent algorithmic discrimination will require a nuanced understanding of both technical and legal constraints.

13.3 Transparent and Meaningful Explanations

A key step in ensuring that algorithms have positive societal impact is to deepen our understanding of how to design transparent and explainable decision-making systems. Lipton (2018) enumerates a number of goals and challenges in the quest for *interpretable* models, and researchers continue to develop new techniques to explain models or the decisions they produce (Ribeiro et al., 2016; Caruana et al., 2020). We refer the reader to Gilpin et al. (2018); Carvalho et al. (2019) for comprehensive surveys on this topic.

Beyond the development of new technical methods, further work is needed to make models accessible to humans. One line of research seeks to standardize the reporting of information about models and the data on which they are trained (Gebru et al., 2018; Mitchell et al., 2019). As these efforts begin to see more practical usage, further research is needed to determine how best to implement or modify them in order to be useful to practitioners.

When algorithms are used to assist humans in making decisions, it is crucial to ensure that algorithmic outputs are constructed and presented in such a way that humans can readily use them. Do humans trust an algorithm's outputs? How can an algorithm express uncertainty? Can a human and algorithm op-

erating together perform better than either in isolation? Building on literature studying the effects like *automation bias* – the tendency for humans to defer to automated decisions (Goddard et al., 2012) – recent work seeks to address questions like these (Stevenson, 2018; Chouldechova et al., 2018; Skeem et al., 2020; De-Arteaga et al., 2020, 2021) from both theoretical and applied perspectives.

Finally, based on recent legal directives like the EU's General Data Protection Regulation (GDPR), scholars have begun to consider the "right to explanation" and how it might manifest in various algorithmic systems (Selbst and Powles, 2017; Wachter et al., 2017; Kaminski, 2019). What constitutes a meaningful explanation, and from a computational perspective, how do we provide individuals with such explanations? Questions like these will require continued collaboration at the interface between computer science and legal scholarship.

Part VI

Appendices

Here, we present supplementary information, omitted proofs, and additional results for the work in this thesis. The correspondence between appendices and chapters in this thesis is as follows:

App.	Chap.	Contents
A	3	A proof demonstrating that the integral risk assignment problem in Section 3.4.2 is NP-complete.
B	4	Additional theoretical results and details on experiments.
C	5	Supplementary lemmas and omitted proofs.
D	7	Supplementary lemmas and omitted proofs.
E	8	A characterization of strategic behavior in response to a linear mechanism.
F	9	Supplementary lemmas, omitted proofs, and counterexamples.
G	11	A table containing administrative information on vendors.

Table 1: Contents of appendices.

APPENDIX A

INHERENT TRADE-OFFS IN THE FAIR DETERMINATION OF RISK SCORES

A.1 NP-Completeness of Non-Trivial Integral Fair Risk Assignments

We can reduce to the integral assignment problem, parameterized by $a_{1\sigma}, a_{2\sigma}$, and p_σ, from subset sum as follows.

Suppose we have an instance of the subset sum problem specified by m numbers $w_1, \ldots, w_m$ and a target T; the goal is to determine whether a subset of the w_i add up to T. We create an instance of the integral assignment problem with $\sigma_1, \ldots, \sigma_{2m+2}$. $a_{1,\sigma_i} = 1/2$ if $i \in \{2m+1, 2m+2\}$ and 0 otherwise. $a_{2,\sigma_i} = 1/(2m)$ if $i \leq 2m$ and 0 otherwise. We make the following definitions:

$$\begin{aligned}
\hat{w}_i &= w_i/(Tm^4) \\
\varepsilon_i &= \sqrt{\hat{w}_i/2} \\
p_{\sigma_{2i-1}} &= i/(m+1) - \varepsilon_i && (1 \leq i \leq m) \\
p_{\sigma_{2i}} &= i/(m+1) + \varepsilon_i && (1 \leq i \leq m) \\
\gamma &= 1/m \sum_{i=1}^{2m} p_{\sigma_i}^2 - 1/m^5 \\
p_{\sigma_{2m+1}} &= (1 - \sqrt{2\gamma - 1})/2 \\
p_{\sigma_{2m+2}} &= (1 + \sqrt{2\gamma - 1})/2
\end{aligned}$$

With this definition, the subset sum instance has a solution if and only if the integral assignment instance given by $a_{1,\sigma}, a_{2,\sigma}, p_{\sigma_1}, \ldots, p_{\sigma_{2m+2}}$ has a solution.

Before we prove this, we need the following lemma.

Lemma A.1. *For any* $z_1, \ldots, z_k \in \mathbb{R}$,

$$\sum_{i=1}^{k} z_i^2 - \frac{1}{k}\left(\sum_{i=1}^{m} z_i\right)^2 = \frac{1}{k}\sum_{i<j}^{k}(z_i - z_j)^2$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{k} z_i^2 - \frac{1}{k}\left(\sum_{i=1}^{m} z_i\right)^2 &= \sum_{i=1}^{k} z_i^2 - \frac{1}{k}\left(\sum_{i=1}^{k} z_i^2 + 2\sum_{i<j}^{k} z_i z_j\right) \\
&= \frac{k-1}{k}\sum_{i=1}^{k} z_i^2 - \frac{2}{k}\sum_{i<j}^{k} z_i z_j \\
&= \frac{1}{k}\sum_{i<j}^{k}(z_i^2 + z_j^2) - \frac{2}{k}\sum_{i<j}^{k} z_i z_j \\
&= \frac{1}{k}\sum_{i<j}^{k} z_i^2 - 2z_i z_j + z_j^2 \\
&= \frac{1}{k}\sum_{i<j}^{k}(z_i - z_j)^2
\end{aligned}$$

□

Now, we can prove that the integral assignment problem is NP-hard.

Proof. First, we observe that for any nontrivial solution to the integral assignment instance, there must be two bins $b \neq b'$ such that $X_{\sigma_{2m+1},b} = 1$ and $X_{\sigma_{2m+2},b'} = 1$. In other words, the people with σ_{2m+1} and σ_{2m+2} must be split up. If not, then all the people of group 1 would be in the same bin, meaning that bin must be labeled with the base rate $\rho_1 = 1/2$. In order to maintain fairness, the same would have to be done for all the people of group 2, resulting in the trivial solution. Moreover, b and b' must be labeled $(1 \pm \sqrt{2\gamma - 1})/2$ respectively

because those are the fraction of people of group 1 in those bins who belong to the positive class.

This means that $\gamma_1 = 1/\rho \cdot (a_{1,\sigma_{2m+1}} p^2_{\sigma_{2m+1}} + a_{1,\sigma_{2m+2}} p^2_{\sigma_{2m+2}}) = p^2_{\sigma_{2m+1}} + p^2_{\sigma_{2m+2}} = \gamma$ as defined above. We know that a well-calibrated assignment is fair only if $\gamma_1 = \gamma_2$, so we know $\gamma_2 = \gamma$.

Next, we observe that $\rho_2 = \rho_1 = 1/2$ because all of the positive $a_{2,\sigma}$'s are $1/(2m)$, so ρ_2 is just the average of $\{p_{\sigma_1}, \ldots, p_{\sigma_{2m}}\}$, which is $1/2$ by symmetry.

Let Q be the partition of $[2m]$ corresponding to the assignment, meaning that for a given $q \in Q$, there is a bin b_q containing all people with σ_i such that $i \in q$. The label on that bin is

$$\begin{aligned} v_q &= \frac{\sum_{i \in q} a_{2,\sigma_i} p_{\sigma_i}}{\sum_{i \in q} a_{2,\sigma_i}} \\ &= \frac{1/(2m) \sum_{i \in q} p_{\sigma_i}}{|q|/(2m)} \\ &= \frac{1}{|q|} \sum_{i \in q} p_{\sigma_i} \end{aligned}$$

Furthermore, bin b_q contains $\sum_{i \in q} a_{2,\sigma_i} p_{\sigma_i} = 1/(2m) \sum_{i \in q} p_{\sigma_i}$ positive fraction. Using this, we can come up with an expression for γ_2.

$$\begin{aligned} \gamma_2 &= \frac{1}{\rho} \sum_{q \in Q} \left(v_b \cdot \frac{1}{2m} \sum_{i \in q} p_{\sigma_i} \right) \\ &= \frac{1}{m} \sum_{q \in Q} \frac{1}{|q|} \left(\sum_{i \in q} p_{\sigma_i} \right)^2 \end{aligned}$$

Setting this equal to γ, we have

$$\begin{aligned} \frac{1}{m} \sum_{q \in Q} \frac{1}{|q|} \left(\sum_{i \in q} p_{\sigma_i} \right)^2 &= \frac{1}{m} \sum_{i=1}^{2m} p^2_{\sigma_i} - \frac{1}{m^5} \\ \sum_{q \in Q} \frac{1}{|q|} \left(\sum_{i \in q} p_{\sigma_i} \right)^2 &= \sum_{i=1}^{2m} p^2_{\sigma_i} - \frac{1}{m^4} \end{aligned}$$

Subtracting both sides from $\sum_{i=1}^{2m} p_{\sigma_i}^2$ and using Lemma A.1, we have

$$\sum_{q \in Q} \frac{1}{|q|} \sum_{i<j \in q} (p_{\sigma_i} - p_{\sigma_j})^2 = \frac{1}{m^4} \tag{A.1}$$

Thus, Q is a fair nontrivial assignment if and only if (A.1) holds.

Next, we show that there exists Q that satisfies (A.1) if and only if there there exists some $S \subseteq [m]$ such that $\sum_{i \in S} \hat{w}_i = 1/m^4$.

Assume Q satisfies (A.1). Then, we first observe that any $q \in Q$ must either contain a single i, meaning it does not contribute to the left hand side of (A.1), or $q = \{2i-1, 2i\}$ for some i. To show this, observe that the closest two elements of $\{p_{\sigma_1}, \ldots, p_{\sigma_{2m}}\}$ not of the form $\{p_{\sigma_{2i-1}}, p_{\sigma_{2i}}\}$ must be some $\{p_{\sigma_{2i}}, p_{\sigma_{2i+1}}\}$. However,

we find that

$$\begin{aligned}
(p_{\sigma_{2i+1}} - p_{\sigma_{2i}})^2 &= \left(\frac{i+1}{m+1} - \varepsilon_{i+1} - \left(\frac{i}{m+1} + \varepsilon_i\right)\right)^2 \\
&= \left(\frac{1}{m+1} - \varepsilon_{i+1} - \varepsilon_i\right)^2 \\
&= \left(\frac{1}{m+1} - \sqrt{\frac{\hat{w}_{i+1}}{2}} - \sqrt{\frac{\hat{w}_i}{2}}\right)^2 \\
&\geq \left(\frac{1}{m+1} - \sqrt{\frac{2}{m^4}}\right)^2 && (\hat{w}_i \leq 1/m^4) \\
&= \left(\frac{1}{m+1} - \frac{\sqrt{2}}{m^2}\right)^2 \\
&\geq \left(\frac{1}{2m} - \frac{\sqrt{2}}{m^2}\right)^2 \\
&= \left(\frac{m - 2\sqrt{2}}{2m^2}\right)^2 \\
&\geq \left(\frac{m}{4m^2}\right)^2 \\
&= \left(\frac{1}{4m}\right)^2 \\
&= \frac{1}{16m^2}
\end{aligned}$$

If any q contains any j, k not of the form $2i-1, 2i$, then (A.1) will have a term on the left hand side at least $1/m \cdot 1/(16m^2) = 1/(16m^3) > 1/m^4$ for large enough m, and since there can be no negative terms on the left hand side, this immediately makes it impossible for Q to satisfy (A.1).

Consider every $2i-1, 2i \in [2m]$. Let $q_i = \{2i-1, 2i\}$. As shown above, either $q_i \in Q$ or $\{2i-1\} \in Q$ and $\{2i\} \in Q$. In the latter case, neither $p_{\sigma_{2i-1}}$ nor $p_{\sigma_{2i}}$ contributes to (A.1). If $q_i \in Q$, then q_i contributes $1/2(p_{\sigma_{2i}} - p_{\sigma_{2i-1}})^2 = 1/2(2\varepsilon_i)^2 = \hat{w}_i$ to the overall sum on the left hand side. Therefore, we can write the left hand

side of (A.1) as

$$\sum_{q \in Q} \frac{1}{|q|} \sum_{i<j \in q} (p_{\sigma_i} - p_{\sigma_j})^2 = \sum_{q_i \in Q} \frac{1}{2} (p_{\sigma_{2i}} - p_{\sigma_{2i-1}})^2 = \sum_{q_i \in Q} \hat{w}_i = \frac{1}{m^4}$$

Then, we can build a solution to the original subset sum instance as $S = \{i : q_i \in Q\}$, giving us $\sum_{i \in S} \hat{w}_i = \frac{1}{m^4}$. Multiplying both sides by Tm^4, we get $\sum_{i \in S} w_i = T$, meaning S is a solution for the subset sum instance.

To prove the other direction, assume we have a solution $S \subseteq [m]$ such that $\sum_{i \in S} w_i = T$. Dividing both sides by Tm^4, we get $\sum_{i \in S} \hat{w}_i = 1/m^4$. We build a partition Q of $2m$ by starting with the empty set and adding $q_i = \{2i - 1, 2i\}$ to Q if $i \in S$ and $\{2i - 1\}$ and $\{2i\}$ to Q otherwise. Clearly, each element of $[2m]$ appears in Q at most once, making this a valid partition. Moreover, when checking to see if (A.1) is satisfied (which is true if and only if Q is a fair assignment), we can ignore all $q \in Q$ such that $|q| = 1$ because they don't contribute to the left hand side. Since, we again have

$$\sum_{q \in Q} \frac{1}{|q|} \sum_{i<j \in q} (p_{\sigma_i} - p_{\sigma_j})^2 = \sum_{q_i \in Q} \frac{1}{2} (p_{\sigma_{2i}} - p_{\sigma_{2i-1}})^2 = \sum_{q_i \in Q} \hat{w}_i = \frac{1}{m^4}$$

meaning Q is a fair assignment. This completes the reduction. □

We have shown that the integral assignment problem is NP-hard, and it is clearly in NP because given an integral assignment, we can verify in polynomial time whether such an assignment satisfies the conditions (A), (B), and (C). Thus, the integral assignment problem is NP-complete.

APPENDIX B

ON FAIRNESS AND CALIBRATION

For simplicity, our focus thus far has been on classifiers that are perfectly calibrated. Here, we introduce an approximate notion of calibration, which we will use in subsequent proofs.

Definition B.1. *The* calibration gap $\epsilon(h_t)$ *of a classifier* h_t *with respect to a group* G_t *is*

$$\epsilon(h_t) = \int_0^1 \Big| \Pr_{(\mathbf{x},y)\sim G_t}\big[y{=}1 \mid h(\mathbf{x}){=}p\big] - p \Big| \Pr_{(\mathbf{x},y)\sim G_t}\big[h(\mathbf{x}) = p\big]\, dp. \tag{B.1}$$

Thus, a classifier h_t *is* perfectly calibrated *if* $\epsilon(h_t) = 0$.

A majority of this supplementary material is devoted to proving approximate versions of our major findings. In all cases, our results degrade smoothly as the calibration condition is relaxed. In addition, we also provide extended details on the experiments run in this chapter.

Note that we will use the notational abuse $\Pr_{G_t}$ and $\mathbb{E}_{G_t}$ in place of $\Pr_{(\mathbf{x},y)\sim G_t}$ and $\mathbb{E}_{(\mathbf{x},y)\sim G_t}$.

B.1 Linearity of Calibrated Classifiers

In Section 4.1, we claim that the set of all calibrated classifiers $\mathcal{H}_t^*$ for group G_t form a line in the generalized false-positive/false-negative plane. The following proof of this claim is adapted from Kleinberg et al. (2017).

Lemma B.2. *For a group G_t, if a classifier h_t has $\epsilon(h_t) \leq \delta_{cal}$, then*

$$\left|\mu_t c_{fn}(h_t) - (1-\mu_t)\, c_{fp}(h_t)\right| \leq 2\delta_{cal}.$$

where $c_{fp}(h_t)$ and $c_{fn}(h_t)$ are the generalized false-positive and false-negative and μ_t is the base rate of group G_t.

Proof. First, note that

$$\begin{aligned}
c_{fp}(h_t) &= \mathop{\mathbb{E}}_{G_t}\big[h_t(\mathbf{x}) \mid y=0\big] \\
&= \int_0^1 p \Pr_{G_t}\left[h_t(\mathbf{x})=p \mid y=0\right] dp \\
&= \int_0^1 p \,\frac{1-\Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right]}{1-\Pr_{G_t}\left[y=1\right]} \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&= \frac{1}{1-\mu_t}\int_0^1 p\,(1-\Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right]) \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \qquad \text{(B.2)}
\end{aligned}$$

Next, observe that

$$\begin{aligned}
&\int_0^1 p \cdot \Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right] \cdot \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&\quad= \int_0^1 p(p + \Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right] - p) \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&\quad\leq \int_0^1 \left(p^2 + |\Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right] - p|\right) \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&\quad\leq \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] + \delta_{cal}
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int_0^1 p \cdot \Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right] \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&\quad\geq \int_0^1 \left(p^2 - |\Pr_{G_t}\left[y=1 \mid h_t(\mathbf{x})=p\right] - p|\right) \Pr_{G_t}\left[h_t(\mathbf{x})=p\right] dp \\
&\quad\geq \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] - \delta_{cal}
\end{aligned}$$

Plugging these into (B.2), we have

$$\frac{1}{1-\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] - \delta_{cal}) \leq c_{fp}(h_t) \leq \frac{1}{1-\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] + \delta_{cal}) \tag{B.3}$$

We follow a similar procedure for $c_{fn}(h_t)$:

$$\begin{aligned} c_{fn}(h_t) &= \mathop{\mathbb{E}}_{G_t}\big[1 - h_t(\mathbf{x}) \mid y=0\big] \\ &= \int_0^1 (1-p)\ \Pr_{G_t}[h_t(\mathbf{x})=p \mid y=0]\, dp \\ &= \int_0^1 (1-p)\ \frac{\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p]}{\Pr_{G_t}[y=1]} \Pr_{G_t}[h_t(\mathbf{x})=p]\, dp. \\ &= \frac{1}{\mu_t}\int_0^1 (1-p)\,(\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p]) \Pr_{G_t}[h_t(\mathbf{x})=p]\, dp. \end{aligned}$$

We use the fact that

$$\begin{aligned} &(1-p)(\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p]) \\ &\quad = (1-p)(p + \Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p] - p) \\ &\quad \leq p(1-p) + |\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p] - p| \end{aligned}$$

and

$$\begin{aligned} &(1-p)(\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p]) \\ &\quad \geq p(1-p) - |\Pr_{G_t}[y=1 \mid h_t(\mathbf{x})=p] - p| \end{aligned}$$

to get

$$\frac{1}{\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] - \delta_{cal}) \leq c_{fn}(h_t) \leq \frac{1}{\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] + \delta_{cal}) \tag{B.4}$$

Combining (B.3) and (B.4), we have

$$\begin{aligned} c_{fn}(h_t) &\leq \frac{1}{\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] + \delta_{cal}) \\ &= \frac{1}{\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] - \delta_{cal} + 2\delta_{cal}) \\ &\leq \frac{1}{\mu_t}((1-\mu_t)\, c_{fp}(h_t) + 2\delta_{cal}) \\ &= \frac{1-\mu_t}{\mu_t} c_{fp}(h_t) + \frac{2\delta_{cal}}{\mu_t} \end{aligned}$$

We can get a similar lower bound for $c_{fn}(h_t)$ as

$$\begin{aligned} c_{fn}(h_t) &\geq \frac{1}{\mu_t}(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2] - \delta_{cal}) \\ &\geq \frac{1}{\mu_t}((1-\mu_t)\, c_{fp}(h_t) - 2\delta_{cal}) \\ &= \frac{1-\mu_t}{\mu_t} c_{fp}(h_t) - \frac{2\delta_{cal}}{\mu_t} \end{aligned}$$

Multiplying these inequalities by μ_t completes this proof. □

Corollary B.3. *Let $\mathcal{H}_t$ be the set of perfectly calibrated classifiers for group G_t — i.e. for any $h_t^* \in \mathcal{H}_T$, we have $\epsilon(h_t^*) = 0$. The generalized false-positive and false-negative rates of h_t^* are given by*

$$c_{fp}(h_t^*) = \frac{1}{1-\mu_t}\left(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2]\right) \tag{B.5}$$

$$c_{fn}(h_t^*) = \frac{1}{\mu_t}\left(\mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})] - \mathop{\mathbb{E}}_{G_t}[h_t(\mathbf{x})^2]\right) \tag{B.6}$$

Proof. This is a direct consequence of (B.3) and (B.4). □

Corollary B.4. *For a group G_t, any perfectly calibrated classifier h_t^* satisfies*

$$c_{fn}(h_t^*) = \frac{1-\mu_t}{\mu_t} c_{fp}(h_t). \tag{B.7}$$

In other words, all perfectly calibrated classifiers $h_t^* \in \mathcal{H}_t$ for group G_t lie on a line in the generalized false-positive/false-negative plane, where the slope of the line is uniquely determined by the group's base-rate μ_t.

B.2 Cost Functions

We will prove a few claims about cost functions g_t of the form given by (4.2) — i.e.

$$g_t(h_t) = a_t c_{fp}(h_t) + b_t c_{fn}(h_t)$$

for some non-negative constants a_t and b_t. First, we show that h^{μ_t} is the calibrated classifier that maximizes g_t.

Lemma B.5. *For any cost function g_t that follows the form of* (4.2)*, the trivial classifier h^{μ_t} is the calibrated classifier for G_t with maximum cost.*

Proof. Again, let g_t be a cost function:

$$g_t(h) = a_t c_{fp}(h_t) + b_t c_{fn}(h_t).$$

Using (B.5) and (B.6), we have that, for every classifier h_t that is perfectly calibrated for group G_t,

$$\begin{aligned} g_t(h_t) &= a_t c_{fp}(h_t) + b_t c_{fn}(h_t) \\ &= \left(\frac{a_t}{1-\mu_t} + \frac{b_t}{\mu_t}\right)\left(\mathop{\mathbb{E}}_{G_t}\left[h_t(x)\right] - \mathop{\mathbb{E}}_{G_t}\left[h_t(x)^2\right]\right) \\ &= \left(\frac{a_t}{1-\mu_t} + \frac{b_t}{\mu_t}\right)\left(\mu_t - \mathop{\mathbb{E}}_{G_t}\left[h_t(x)^2\right]\right). \end{aligned}$$

The last equation holds because $\mathbb{E}_{G_t}\left[h_t(\mathbf{x})\right] = \mu_t$ for any calibrated classifier – a fact which can easily be derived from Definition B.1.

We would like to find, $h_t^{\max} \in \mathcal{H}_t^*$, the calibrated classifier with the highest

weighted cost. Because $\left(\frac{a_t}{1-\mu_t} + \frac{b_t}{\mu_t}\right)$ and μ_t are non-negative constants, we have

$$\begin{aligned}
h_t^{\max} &= \arg\max_{h \in \mathcal{H}_t^*} \left[\left(\frac{a_t}{1-\mu_t} + \frac{b_t}{\mu_t} \right) \left(\mu_t - \mathop{\mathbb{E}}_{G_t} \left[h(x)^2 \right] \right) \right] \\
&= \arg\max_{h \in \mathcal{H}_t^*} \left[- \mathop{\mathbb{E}}_{G_t} \left[h(x)^2 \right] \right] \\
&= \arg\min_{h \in \mathcal{H}_t^*} \left[\mathop{\mathbb{E}}_{G_t} \left[h(x)^2 \right] \right] \\
&= \arg\min_{h \in \mathcal{H}_t^*} \left[\mathop{\mathbb{E}}_{G_t} \left[h(x)^2 \right] - \mu_t^2 \right]
\end{aligned}$$

Thus, the calibrated classifier with minimum variance will have the highest cost. This translates to a classifier that outputs the same probability for every sample. By the calibration constraint, this constant must be equal to μ_t, so this classifier must be the trivial classifier h^{μ_t} — i.e. for all $\mathbf{x}$

$$h_t^{\max}(\mathbf{x}) = h^{\mu_t}(\mathbf{x}) = \mu_t.$$

□

Next, we show that g_t is linear under randomized interpolations.

Lemma B.6. *Let $\tilde{h}_2$ be the classifier derived from* (4.3) *with interpolation parameter $\alpha \in [0, 1]$. The cost of $\tilde{h}_2$ is given by*

$$g_2(\tilde{h}_2) = (1-\alpha) g_2(h_2) + \alpha g_2(h^{\mu_2})$$

Proof. The cost of $\tilde{h}_2$ can be calculated using linearity of expectation. Let B be a

Bernoulli random variable with parameter α.

$$
\begin{aligned}
g_2(\tilde{h}_2) &= a_2 c_{fp}(\tilde{h}_2) + b_2 c_{fn}(\tilde{h}_2) \\
&= a_2 \mathop{\mathbb{E}}_{G_2}\left[1 - \tilde{h}_2(\mathbf{x}) \mid y{=}1\right] + b_2 \mathop{\mathbb{E}}_{G_2}\left[\tilde{h}_2(\mathbf{x}) \mid y{=}0\right] \\
&= a_2 \mathop{\mathbb{E}}_{B,G_2}\left[1 - [(1-B)h_2(\mathbf{x}) + Bh^{\mu_2}(\mathbf{x})] \mid y{=}1\right] \\
&+ b_2 \mathop{\mathbb{E}}_{B,G_2}\left[[(1-B)h_2(\mathbf{x}) + Bh^{\mu_2}(\mathbf{x})] \mid y{=}0\right] \\
&= a_2 \mathop{\mathbb{E}}_{B,G_2}\left[(1-B)\left(1 - h_2(\mathbf{x})\right) \mid y{=}1\right] + a_2 \mathop{\mathbb{E}}_{B,G_2}\left[B\left(1 - h^{\mu_2}(\mathbf{x})\right) \mid y{=}1\right] \\
&\quad + b_2 \mathop{\mathbb{E}}_{B,G_2}\left[(1-B)h_2(\mathbf{x}) \mid y{=}0\right] + b_2 \mathop{\mathbb{E}}_{B,G_2}\left[Bh^{\mu_2}(\mathbf{x}) \mid y{=}0\right] \\
&= a_2 \mathop{\mathbb{E}}_{B}[1-B] \mathop{\mathbb{E}}_{G_2}\left[1 - h_2(\mathbf{x}) \mid y{=}1\right] + a_2 \mathop{\mathbb{E}}_{B}[B] \mathop{\mathbb{E}}_{G_2}\left[1 - h^{\mu_2}(\mathbf{x}) \mid y{=}1\right] \\
&\quad + b_2 \mathop{\mathbb{E}}_{B}[1-B] \mathop{\mathbb{E}}_{G_2}\left[h_2(\mathbf{x}) \mid y{=}0\right] + b_2 \mathop{\mathbb{E}}_{B}[B] \mathop{\mathbb{E}}_{G_2}\left[h^{\mu_2}(\mathbf{x}) \mid y{=}0\right] \\
&= a_2(1-\alpha)c_{fp}(h_2) + b_2(1-\alpha)c_{fn}(h_2) + a_2(\alpha)c_{fp}(h^{\mu_2}) + b_2(\alpha)c_{fn}(h^{\mu_2}) \\
&= (1-\alpha)g_2(h_2) + \alpha g_2(h^{\mu_2}).
\end{aligned}
$$

□

B.3 Relationship Between Cost and Error

In Section 4.1, we claim that there is a tight connection between reducing any cost function $g_t(h_t)$ and reducing the generalized error rates $c_{fp}(h_t)$ and $c_{fn}(h_t)$ for approximately calibrated classifiers. In other words, assuming we are approximately calibrated, improving cost will approximately improve our error rates. We formalize this notion in this section:

Lemma B.7. *Let h_t be a classifier with $\epsilon(h_t) = \delta_{cal}$ and cost $g_t(h_t)$. For any other classifier h'_t, if $c_{fp}(h'_t) < c_{fp}(h_t) - \frac{4\delta_{cal}}{1-\mu_t}$ or $c_{fn}(h'_t) < c_{fn}(h_t) - \frac{4\delta_{cal}}{\mu_t}$, then $g_t(h'_t) < g_t(h_t)$ or $\epsilon(h_t) > \delta_{cal}$.*

Proof. First, assume that $c_{fp}(h'_t) < c_{fp}(h_t) - \frac{4\delta_{cal}}{1-\mu_t}$. Then, there are two cases: either $c_{fn}(h'_t) < c_{fn}(h_t)$ or $c_{fn}(h'_t) \geq c_{fn}(h_t)$. In the first case, $g_t(h'_t) < g_t(h_t)$ because $c_{fp}(h'_t) < c_{fp}(h_t)$ and $c_{fn}(h'_t) < c_{fn}(h_t)$. In the second case, if $\epsilon(h'_t) \leq \delta_{cal}$, we can use Lemma B.2 to get

$$\begin{aligned} c_{fp}(h'_t) &\geq \frac{\mu_t}{1-\mu_t} c_{fn}(h'_t) - \frac{2\delta_{cal}}{1-\mu_t} \\ &\geq \frac{\mu_t}{1-\mu_t} c_{fn}(h_t) - \frac{2\delta_{cal}}{1-\mu_t} \\ &\geq c_{fp}(h_t) - \frac{4\delta_{cal}}{1-\mu_t} \end{aligned}$$

Since this contradicts the initial assumption that $c_{fp}(h'_t) < c_{fp}(h_t) - \frac{4\delta_{cal}}{1-\mu_t}$, it cannot be the case that $\epsilon(h'_t) \leq \delta_{cal}$. This proves the lemma when $c_{fp}(h'_t) < c_{fp}(h_t) - \frac{4\delta_{cal}}{1-\mu_t}$.

To prove the second part, we now assume that $c_{fn}(h'_t) < c_{fn}(h_t) - \frac{4\delta_{cal}}{\mu_t}$. We again break this into two cases. If $c_{fp}(h'_t) < c_{fp}(h_t)$, then $g_t(h'_t) < g_t(h_t)$. If $c_{fp}(h'_t) \geq c_{fp}(h_t)$, then under the assumption that $\epsilon(h'_t) \leq \delta_{cal}$, we can rearrange Lemma B.2 to get

$$\begin{aligned} c_{fn}(h'_t) &\geq \frac{1-\mu_t}{\mu_t} c_{fp}(h'_t) - \frac{2\delta_{cal}}{\mu_t} \\ &\geq \frac{1-\mu_t}{\mu_t} c_{fp}(h_t) - \frac{2\delta_{cal}}{\mu_t} \\ &\geq c_{fn}(h_t) - \frac{4\delta_{cal}}{\mu_t} \end{aligned}$$

Again, this contradicts the assumption that $c_{fn}(h'_t) < c_{fn}(h_t) - \frac{4\delta_{cal}}{\mu_t}$, so it cannot be the case that $\epsilon(h'_t) \leq \delta_{cal}$. This completes the proof. □

From this result we can derive a stronger claim for perfectly calibrated classifiers.

Lemma B.8. *Let h_t and h'_t be perfectly calibrated classifiers with cost $g_t(h_t) \leq g_t(h'_t)$ for some cost function g_t. Then $c_{fp}(h_t) \leq c_{fp}(h'_t)$*

B.4 Proof of Algorithm 1 Optimality and Approximate Optimality

In Section 4.2, we claim that Algorithm 1 produces optimal non-discriminatory classifiers in exact calibration scenarios, and near-optimal classifiers in approximate calibration scenarios.

We begin with classifiers h_1 and h_2 be classifiers for groups G_1 and G_2, with calibrations $\epsilon(h_1) \leq \delta_{cal}$ and $\epsilon(h_2) \leq \delta_{cal}$. As before, assume that we cannot strictly improve the cost of either h_1 or h_2 without worsening calibration: i.e. h_1 and h_2 is g_t and calibration. We will now show that Algorithm 1 produces classifiers that are near-optimal with respect to both the false-positive and false-negative rates among calibrated classifiers satisfying the equal-cost constraint.

First, we show that interpolation preserves approximate calibration:

Theorem B.9 (Approximate Optimality of Algorithm 1)**.** *Given $\tilde{h}_2$, which is the classifier produced by Algorithm 1, we have that $\epsilon(\tilde{h}_2) \leq (1-\alpha)\,\epsilon(h_2)$, where $\alpha \in [0,1]$ is the interpolation parameter in* (4.3)*.*

Proof. We can calculate the calibration of $\tilde{h}_2(\mathbf{x})$ as follows:

$$\begin{aligned}\epsilon(\tilde{h}_2) &= \mathop{\mathbb{E}}_{B,G_2}\Big|\Pr_{G_2}\big[y{=}1 \mid \tilde{h}_2(\mathbf{x}){=}p\big] - p\Big| \\ &= \int_0^1 \Big|\Pr_{G_2}\big[y{=}1 \mid \tilde{h}_2(\mathbf{x}){=}p\big] - p\Big| \Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}){=}p\big]\,dp\end{aligned}$$

For $p \neq \mu_2$, $|\Pr_{G_2}\big[y=1 \mid \tilde{h}_2(\mathbf{x})=p\big] - p| = |\Pr_{G_2}\big[y=1 \mid h_2(\mathbf{x})=p\big] - p|$ and $\Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}){=}p\big] = (1-\alpha)\Pr_{G_2}\big[h_2(\mathbf{x}){=}p\big]$.

For $p = \mu_2$, let $\beta = \Pr_{B,G_2}\big[B=1 \mid \tilde{h}_2(\mathbf{x}) = p\big] / \Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}) = p\big]$. Note that

$\beta \geq \alpha$. Then,

$$\begin{aligned}\Pr_{G_2}\big[y{=}1 \mid \tilde{h}_2(\mathbf{x}){=}p\big] &= (1-\beta)\Pr_{G_2}\big[y{=}1 \mid h_2(\mathbf{x}){=}p\big] + \beta \Pr_{G_2}\big[y{=}1 \mid h^{\mu_2}(\mathbf{x}){=}p\big] \\ &= (1-\beta)\Pr_{G_2}\big[y{=}1 \mid h_2(\mathbf{x}){=}p\big] + \beta p\end{aligned}$$

because h^{μ_2} is perfectly calibrated. Moreover, note that $\Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}) = p\big] = \Pr_{G_2}\big[h_2(\mathbf{x}) = p\big]/(1-\beta)$ Using this, we have $|\Pr_{G_2}\big[y = 1 \mid \tilde{h}_2(\mathbf{x}) = p\big] - p|\Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}){=}p\big] = |\Pr_{G_2}\big[y{=}1 \mid h_2(\mathbf{x}){=}p\big] - p|\Pr_{G_2}\big[h_2(\mathbf{x}){=}p\big]$. Thus,

$$\begin{aligned}\epsilon(\tilde{h}_2) &= \int_0^1 \Big|\Pr_{G_2}\big[y{=}1 \mid \tilde{h}_2(\mathbf{x}){=}p\big] - p\Big| \Pr_{G_2}\big[\tilde{h}_2(\mathbf{x}){=}p\big]dp \\ &\leq \int_0^1 \Big|\Pr_{G_2}\big[y{=}1 \mid h_2(\mathbf{x}){=}p\big] - p\Big| \Pr_{G_2}\big[h_2(\mathbf{x}){=}p\big]dp \\ &= \epsilon(h_2)\end{aligned}$$

□

Next, we observe that by Lemma B.7, for any classifiers h_1' and h_2' with $\epsilon(h_1') \leq \delta_{cal}$ and $\epsilon(h_2') \leq \delta_{cal}$ satisfying the equal-cost constraint, it must be the case that $c_{fp}(h_t') \geq c_{fp}(\tilde{h}_t) - \frac{4\delta_{cal}}{1-\mu_t}$ and $c_{fn}(h_t') \geq c_{fp}(\tilde{h}_t) - \frac{4\delta_{cal}}{\mu_t}$ for $t = 1, 2$.

Thus, approximately calibrated classifiers will be approximately optimal. From this result, it is easy to derive the optimality result for perfectly-calibrated classifiers.

Theorem B.10 (Exact Optimality of Algorithm 1)**.** *Algorithm 1 produces the classifiers h_1 and $\tilde{h}_2$ that satisfy both perfect calibration and the equal-cost constraint with the lowest possible generalized false positive and false negative rates.*

B.5 Proof of Impossibility and Approximate Impossibility

In this section, we prove that it is impossible to satisfy multiple equal-cost constraints while simultaneously satisfying calibration. We will first prove this in an exact sense, and then show that the result holds approximately as well.

B.5.1 Exact Impossibility Theorem

Theorem 4.5 (Restated). Let h_1 and h_2 be calibrated classifiers for G_1 and G_2 with equal cost with respect to g_t. If $\mu_1 \neq \mu_2$, and if h_1 and h_2 have equal cost with respect to g'_t, then h_1 and h_2 must be perfect classifiers.

Proof. First, observe that the perfect classifier always satisfies any equal-cost constraint simply because if $c_{fp}(t) = c_{fn}(t) = 0$, $g_t(h_t) = 0$. Moreover, the perfect classifier is always calibrated.

For any classifier, as shown by Kleinberg et al. (2017), $c_{fp}(h_t)$ and $c_{fn}(h_t)$ are linearly related by (B.7). Furthermore, each equal-cost constraint is linear in $c_{fp}(h_t)$ and $c_{fn}(h_t)$. We define $g_t(h_t)$ and $g'_t(h_t)$ to be identical cost functions if the equal-cost constraints that they impose are identical, meaning one constraint is satisfied if and only if the other is satisfied. If this is not the case, then $g_t(h_t)$ and $g'_t(h_t)$ are *distinct*, meaning that the equal-cost constraints are linearly independent for $\mu_1 \neq \mu_2$. Moreover, these are also linearly independent from the calibration constraints because by assumption, they both have nonzero coefficients for at least one of $(c_{fp}(h_1), c_{fn}(h_1))$ and $(c_{fp}(h_2), c_{fn}(h_2))$. As a result, we have four linearly independent constraints (2 from calibration and at least 2 equal-cost constraints) on 4 variables ($c_{fp}(h_1)$, $c_{fn}(h_1)$, $c_{fp}(h_2)$, $c_{fn}(h_2)$), mean-

ing that these constraints yield a unique solution. From above, we know that all the constraints are simultaneously satisfied when $c_{fp}(h_t) = c_{fn}(h_t) = 0$ for $t = 1, 2$, meaning that the perfect classifier is the only classifier for which they are simultaneously satisfied. □

B.5.2 Approximate Impossibility Theorem

Now, we will show that this impossibility result holds in an approximate sense — i.e., approximately satisfying the calibration and equal-cost constraints is only possible if the classifiers approximately perfect.

Since the calibration and equal-cost constraints are all linear, let A be the matrix that encodes them. With two equal-cost constraints g_t and g'_t,

$$A = \begin{bmatrix} 1 & -\frac{\mu_1}{1-\mu_1} & 0 & 0 \\ 0 & 0 & 1 & -\frac{\mu_2}{1-\mu_2} \\ a_1 & b_1 & -a_2 & -b_2 \\ a'_1 & b'_1 & -a'_2 & -b'_2 \end{bmatrix}.$$

Note that the first two rows of A encode the calibration conditions — see (B.7). The bottom two rows encode two equal-cost constraints. Furthermore, let

$$\vec{q} = [c_{fp}(h_1)\ c_{fn}(h_1)\ c_{fp}(h_2)\ c_{fn}(h_2)]^\top.$$

If all constraints are required to hold exactly, then we have $A\vec{q} = 0$. Consider the case where the calibration and equal-cost constraints hold approximately.

Theorem B.11 (Generalized approximate impossibility result)**.** *Let h_1 and h_2 be classifiers with calibration δ_{cal} and cost difference at most δ_{cost} with respect to distinct cost functions g_t and g'_t. Furthermore, assume that every entry of A is rational with*

some common denominator D and is upper bounded by some maximum value M. Then, there is a constant L that depends on D and M such that

$$c_{fp}(h_t) \leq L \cdot \max\left\{\frac{2\delta_{cal}}{1-\mu_1}, \frac{2\delta_{cal}}{1-\mu_2}, \delta_{cost}\right\}$$

and

$$c_{fn}(h_t) \leq L \cdot \max\left\{\frac{2\delta_{cal}}{1-\mu_1}, \frac{2\delta_{cal}}{1-\mu_2}, \delta_{cost}\right\}$$

for $t = 1, 2$.

Proof. By Lemma B.2,

$$\left|\mu_t c_{fn}(h_t) - (1-\mu_t)\, c_{fp}(h_t)\right| \leq 2\delta_{cal}.$$

Since the first two rows in A correspond to the calibration constraints, and the second to correspond to the equal-cost constraints, it must be the case that

$$|A\vec{q}| \leq \begin{bmatrix} \frac{2\delta_{cal}}{1-\mu_1} \\ \frac{2\delta_{cal}}{1-\mu_2} \\ \delta_{cost} \\ \delta_{cost} \end{bmatrix},$$

i.e. the absolute value of each entry in $A\vec{q}$ is bounded by the vector on the right hand side. Let $\vec{\nu} = [2\delta_{cal}/(1-\mu_1)\ \ 2\delta_{cal}/(1-\mu_2)\ \ \delta_{cost}\ \ \delta_{cost}]^\top$. Let $\vec{s} = sign(A\vec{q})$, and multiply the ith row of A by the ith entry of $\vec{s}$ to produce $\widehat{A}$ This allows us to drop the absolute value, meaning we have

$$\widehat{A}\vec{q} \leq \vec{\nu}$$

Furthermore, since g_t and g_t' were assumed to be distinct, $\widehat{A}$ is invertible, so this is equivalent to

$$\vec{q} \leq \widehat{A}^{-1}\vec{\nu}.$$

Taking ℓ_∞ norms of both sides,

$$\|\vec{q}\|_\infty \leq \|\widehat{A}^{-1}\vec{\nu}\|_\infty \leq \|\widehat{A}^{-1}\|_\infty \|\vec{\nu}\|_\infty.$$

The (i,j) entry of $\widehat{A}^{-1}$ can be expressed as $\widehat{A}_{ji}/\det(\widehat{A})$, where $\widehat{A}_{ji}$ is the (j,i) cofactor. Note that $\widehat{A}_{ji}$ is a 3×3 determinant, so it is the sum of 6 cubic polynomials in entries of $\widehat{A}$. However, since every 3×3 submatrix of $\widehat{A}$ has at least one 0 entry, only 4 of those cubics can be nonnegative. By assumption, the maximum value of any entry of $\widehat{A}$ is M, so $|\widehat{A}_{ji}| \leq 4M^3$.

We can lower bound $\det(\widehat{A})$ by noting that since $\widehat{A}$ is not singular, its determinant is nonzero. However, because the determinant can be expressed as a 4×4 polynomial, and each term has common denominator D by assumption, $|\det(\widehat{A})| \geq 1/D^4$. As a result, $|\widehat{A}_{ji}/\det(\widehat{A})| \leq 4M^3D^4$.

Let d_{ij} be the (i,j) entry of $\widehat{A}$. We know that

$$\|\widehat{A}^{-1}\|_\infty \leq \max_j \sum_{i=1}^{4} |d_{ij}| \leq 16M^3D^4 = L$$

As a result,

$$\|\vec{q}\|_\infty \leq L\|\nu\|_\infty$$

which proves the claim. □

Note that Theorem B.11 is not intended to be a tight bound. It simply shows that impossibility result degrades smoothly for approximate constraints.

B.6 Details on Experiments

Post-processing for Equalized Odds To derive classifiers that satisfy the Equalized Odds notion of fairness, we use the method introduced by Hardt

et al. (2016b). Essentially, the false-positive and false-negative constraints are satisfied by randomly flipping some of the predictions of the original classifiers. Let $q^{(t)}_{\text{n2p}}$ be the probability for group G_t of "flipping" a negative prediction to positive, and $q^{(t)}_{\text{p2n}}$ be that of flipping a positive prediction to negative. The derived classifiers h^{eo}_1 and h^{eo}_2 essentially flip predictions according to these probabilities:

$$h^{eo}_t(\mathbf{x}) = \begin{cases} (1 - h_t(\mathbf{x}))\, B^{(t)}_{\text{p2n}} + h_t(\mathbf{x}) \left(1 - B^{(t)}_{\text{p2n}}\right) & h_t(\mathbf{x}) \geq 0.5 \\ (1 - h_t(\mathbf{x}))\, B^{(t)}_{\text{n2p}} + h_t(\mathbf{x}) \left(1 - B^{(t)}_{\text{n2p}}\right) & h_t(\mathbf{x}) < 0.5 \end{cases}$$

where $B^{(t)}_{\text{n2p}}$ and $B^{(t)}_{\text{p2n}}$ are Bernoulli random variables with expectations $q^{(t)}_{\text{n2p}}$ and $q^{(t)}_{\text{p2n}}$ respectively. Note that this is a probabilistic generalization of the derived classifiers presented in Hardt et al. (2016b). If all outputs of h_t were either 0 or 1 we would arrive at the original formulation.

We can find the best rates $q^{(1)}_{\text{n2p}}$, $q^{(1)}_{\text{p2n}}$, $q^{(2)}_{\text{n2p}}$, and $q^{(2)}_{\text{p2n}}$ through the following optimization problem:

$$\begin{aligned} \min_{q^{(1)}, q^{(1)}_{\text{p2n}}, q^{(2)}_{\text{n2p}}, q^{(2)}_{\text{p2n}}} \quad & \mathcal{L}(h^{eo}_1) + \mathcal{L}(h^{eo}_2) \\ \text{s.t.} \quad & c_{fp}(h^{eo}_1) = c_{fn}(h^{eo}_1), \\ & c_{fp}(h^{eo}_2) = c_{fn}(h^{eo}_2) \end{aligned}$$

where $\mathcal{L}$ represents the $0/1$ loss of the classifier:

$$\mathcal{L}(h_t) = \Pr_{G_t}\left[h_t(\mathbf{x}) \geq 0.5 \mid y{=}0\right] + \Pr_{G_t}\left[h_t(\mathbf{x}) < 0.5 \mid y{=}1\right].$$

The two constraints enforce the Equalized Odds constraints. Hardt et al. (2016b) show that this can be solved via a linear program.

Constrained-learning for Equalized Odds Zafar et al. (2017a) introduce a method to achieve Equalized Odds (under the name *Disparate Mistreatment*) at

training time using optimization constraints. The problem is set up at learning a logistic classifier under the Equalized Odds constraints. While these constraints make the problem non-convex, Zafar et al. (2017a) show how to formulate the problem as a disciplined convex-concave program. Though this is generally intractable, it can be solved in many instances. We refer the reader to Zafar et al. (2017a) for details.

Training Procedure for Income Prediction. We train three models: a random forest, a multi-layer perceptron, and a SVM with an RBF kernel. We convert the categorical features into one-hot encodings. 10% of the data is reserved for hyperparameter tuning and post-processing, and an additional 10% is saved for final evaluation. The random forest and MLP are naturally probabilistic and well calibrated. We use Platt scaling to calibrate the SVM. The hyperparameters were tuned by grid search based on 3-fold cross validation. Figure 4.3a displays the average false-positive and false-negative costs across all models.

Training Procedure for Health Prediction. We train a random forest and a linear SVM on this dataset. We use the same dataset split and hyperparameter selection as with Income Prediction. Figure 4.3b displays the average false-positive and false-negative costs across all models.

Training Procedure for Recidivism Prediction. For the trained Equalized Odds baseline we train a constrained logistic classifier using the method proposed by Zafar et al. (2017a). We derive the post-processed classifiers (both for Equalized Odds and its calibrated relaxation) from the original COMPAS classifier (Dieterich et al., 2016).

APPENDIX C

THE EXTERNALITIES OF EXPLORATION AND HOW DATA DIVERSITY HELPS EXPLOITATION

Throughout the chapter, we use a number of tools that are either known or easily follow from something that is known. Here, we provide the proofs for the sake of completeness.

C.1 (Sub)gaussians and Concentration

We rely on several known facts about Gaussian and subgaussian random variables. A random variable X is called σ-subgaussian, for some $\sigma > 0$, if $E[e^{\sigma X^2}] < \infty$. This includes variance-σ^2 Gaussian random variables as a special case.

Lemma C.1. *If $X \sim \mathcal{N}(0, \sigma^2)$, then for any $t \geq 0$,*

$$\mathbb{E}\left[X \mid X \geq t\right] \leq \begin{cases} 2\sigma & t \leq \sigma \\ t + \frac{\sigma^2}{t} & t > \sigma \end{cases}$$

Proof. We begin with

$$\mathbb{E}\left[X \mid X \geq t\right] = \frac{\frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty x \exp\left(x^2/(2\sigma^2)\right)\, dx}{\Pr\left[X \geq t\right]}. \tag{C.1}$$

X can be represented as $X = \sigma Y$, where Y is a standard normal random variable. Using a tail bound for the latter (from Cook (2009)),

$$\Pr\left[X \geq t\right] = \Pr\left[Y \geq \frac{t}{\sigma}\right] \geq \frac{1}{\sqrt{2\pi}} \frac{t/\sigma}{(t/\sigma)^2 + 1} \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

The numerator in (C.1) is

$$\frac{1}{\sigma\sqrt{2\pi}}\int_t^\infty x\exp\left(x^2/(2\sigma^2)\right)\,dx = -\frac{1}{\sigma\sqrt{2\pi}}\cdot\sigma^2 e^{-x^2/(2\sigma^2)}\Big|_t^\infty \cdot e^{-t^2/(2\sigma^2)}$$
$$= \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Combining, we have

$$\mathbb{E}\left[X \mid X \geq t\right] \leq \frac{\frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}}\frac{t/\sigma}{(t/\sigma)^2+1}\exp\left(-\frac{t^2}{2\sigma^2}\right)} = \frac{\sigma^2((t/\sigma)^2+1)}{t} = t + \frac{\sigma^2}{t}$$

For $t \leq \sigma$, $\mathbb{E}\left[X \mid X \geq t\right] \leq \mathbb{E}\left[X \mid X \geq \sigma\right] \leq 2\sigma$ by the above bound. □

Lemma C.2. *Suppose $X \sim \mathcal{N}(0, \Sigma)$ is a Gaussian random vector with covariance matrix Σ. Then*

$$\mathbb{E}\left[\,\|X\|_2 \mid \|X\|_2 > \alpha\,\right] \leq d\left(\alpha + \frac{\lambda_{\max}(\Sigma)}{\alpha}\right) \quad \textit{for any } \alpha \geq 0.$$

Proof. Assume without loss of generality that Σ is diagonal, since the norm is rotationally invariant. Observe that $\|X\|_2 \mid \forall i\ X_i > \alpha$ stochastically dominates $\|X\|_2 \mid \|X\|_2 > \alpha$. (Geometrically, the latter conditioning shifts the probability mass away from the origin.) Therefore,

$$\mathbb{E}\left[\,\|X\|_2 \mid \|X\|_2 > \alpha\,\right] \leq \mathbb{E}\left[\,\|X\|_2 \mid \forall i\ X_i > \alpha\,\right]$$
$$= \mathbb{E}\left[\sum_{i=1}^d X_i \mid \forall i\ X_i > \alpha\right] \leq \sum_{i=1}^d \left(t + \frac{\lambda_i(\Sigma)}{\alpha}\right)$$

by Lemma C.1, where $\lambda_i(\Sigma) \leq \lambda_{\max}(\Sigma)$ is the ith eigenvalue of Σ. □

Fact C.3. *If X is a σ-subgaussian random variable, then*

$$\Pr[|X - \mathbb{E}\left[X\right]| > t] \leq 2e^{-t^2/(2\sigma^2)}.$$

Lemma C.4. *If $X_1, \ldots, X_n$ are independent σ-subgaussian random variables, then*

$$\Pr\left[\max_i |X_i - \mathbb{E}[X_i]| > \sigma\sqrt{2\log\frac{2n}{\delta}}\right] \leq \delta.$$

Proof. For any X_i, we know from Fact C.3 that

$$\begin{aligned}\Pr\left[|X_i - \mathbb{E}[X_i]| > \sigma\sqrt{2\log\frac{2n}{\delta}}\right] &\leq 2\exp\left(-\frac{2\sigma^2\log\frac{2n}{\delta}}{2\sigma^2}\right) = 2\exp\left(-\log\frac{2n}{\delta}\right) \\ &= \frac{\delta}{n}.\end{aligned}$$

A union bound completes the proof. □

Lemma C.5. *If $X_1, \ldots, X_K$ are independent zero-mean σ-subgaussian random variables, then*

$$\mathbb{E}[\max_i X_i] \leq \sigma\sqrt{2\log K}.$$

Proof. Let $X = \max X_i$. Since each X_i is σ-subgaussian, it follows that

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Using Jensen's inequality, we have

$$\begin{aligned}\exp(\lambda\mathbb{E}[X]) \leq \mathbb{E}[\exp(\lambda X)] = \mathbb{E}[\max\exp(\lambda X_i)] &\leq \sum_i \mathbb{E}[\exp(\lambda X_i)] \\ &\leq K\exp\left(\frac{\lambda^2\sigma^2}{2}\right).\end{aligned}$$

Rearranging, we have

$$\mathbb{E}[X] \leq \frac{\log K}{\lambda} + \frac{\lambda\sigma^2}{2}.$$

Setting $\lambda = \frac{\sqrt{2\log K}}{\sigma}$, we have $\mathbb{E}[X] \leq \sigma\sqrt{2\log K}$ as needed □

Lemma C.6. *If $\theta \sim \mathcal{N}(\overline{\theta}, \Sigma)$ where $\overline{\theta} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$, then $\mathbb{E}\left[\ \|\theta - \overline{\theta}\|_2\ \right] \leq \sqrt{d\lambda_{\max}(\Sigma)}$.*

Proof. From Chandrasekaran et al. (2012), the expected norm of a standard normal d-dimensional Gaussian is at most $\sqrt{d}$. Using the fact that $\Sigma^{-1/2}(\theta - \overline{\theta}) \sim \mathcal{N}(0, I)$, we have

$$\begin{aligned}\mathbb{E}\left[\|\theta - \overline{\theta}\|_2\right] = \mathbb{E}\left[\|\Sigma^{1/2}\Sigma^{-1/2}(\theta - \overline{\theta})\|_2\right] &\leq \|\Sigma^{1/2}\|_2 \mathbb{E}\left[\|\Sigma^{-1/2}(\theta - \overline{\theta})\|_2\right] \\ &\leq \sqrt{d\lambda_{\max}(\Sigma)}\end{aligned}$$

□

Lemma C.7 (Lemma 2.2 in Dasgupta and Gupta (2003))**.** *If $X \sim \chi^2(d)$, i.e., $X = \sum_{i=1}^d X_i^2$, where $X_1, \ldots, X_d$ are independent standard Normal random variables, then*

$$\begin{aligned}&\Pr\left[X \leq \beta d\right] \leq (\beta e^{1-\beta})^{d/2} && \text{for any } \beta \in (0,1), \\ &\Pr\left[X \geq \beta d\right] \leq (\beta e^{1-\beta})^{d/2} && \text{for any } \beta > 1.\end{aligned}$$

Lemma C.8 (Hoeffding bound)**.** *If $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, where the X_i's are independent σ-subgaussian random variables with zero mean, then*

$$\begin{aligned}&\max\left(\Pr\left[\bar{X} \geq t\right],\ \Pr\left[\bar{X} \leq -t\right]\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right) && \text{for all } t > 0, \\ &\max\left(\Pr\left[\overline{X} \leq -\sigma\sqrt{\tfrac{2}{n}\log\tfrac{1}{\delta}}\right],\quad \Pr\left[\overline{X} \geq \sigma\sqrt{\tfrac{2}{n}\log\tfrac{1}{\delta}}\right]\right) \leq \delta && \text{for all } \delta > 0.\end{aligned}$$

C.2 KL-divergence

We use some basic facts about KL-divergence. Let us recap the definition: given two distributions P, Q on the same finite outcome space Ω, KL-divergence from

P to Q is

$$\mathrm{KL}(P \,||\, Q) := -\sum_{\omega \in \Omega} P(\omega) \log \tfrac{Q(\omega)}{P(\omega)}.$$

Lemma C.9 (High-probability Pinsker Inequality (Tsybakov, 2009))**.** *For any probability distributions P and Q over the same sample space and any arbitrary event E,*

$$P(E) + Q(\overline{E}) \geq \tfrac{1}{2}\, e^{-KL(P \,||\, Q)}.$$

Lemma C.10. *Let P and Q be Bernoulli distributions with means $p \in [1/2 - \varepsilon, 1/2 + \varepsilon]$ and $q \in [1/2 - \varepsilon, 1/2 + \varepsilon]$ respectively, with $\varepsilon \leq 1/4$. Then $KL(P \,||\, Q) \leq \frac{7}{3}\,\varepsilon^2$.*

Proof. For any $\varepsilon \leq 1/4$,

$$\log\left(\frac{p(1-p)}{q(1-q)}\right) \leq \log\left(\frac{1/4}{1/4 - \varepsilon^2}\right) \leq \log\left(\frac{1}{1 - 4\varepsilon^2}\right) \leq \frac{14\,\varepsilon^2}{3} \quad \text{(By Lemma C.16)}$$

$$\begin{aligned}\mathrm{KL}(P \,||\, Q) &= p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\ &\leq \left(\frac{1}{2} + \varepsilon\right) \log\left(\frac{p(1-p)}{q(1-q)}\right) = \left(\frac{1}{2} + \varepsilon\right) \frac{14\varepsilon^2}{3} \leq \frac{7\varepsilon^2}{2}.\end{aligned}$$

□

C.3 Linear Algebra

We use several facts from linear algebra. In what follows, recall that $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimal and the maximal eigenvalues of matrix M, resp. For two matrices A, B, let us write $B \succeq A$ to mean that $B - A$ is positive semidefinite.

Lemma C.11. $\lambda_{\max}(vv^\top) = \|v\|_2^2$ *for any* $v \in \mathbb{R}^d$.

Proof. $vv^\top$ has rank one, so it has one eigenvector with nonzero eigenvalue. v is an eigenvector since $(vv^\top)v = (v^\top v)v$, and it has eigenvalue $v^\top v = \|v\|_2^2$. This is the only nonzero eigenvalue, so $\lambda_{\max}(vv^\top) = \|v\|_2^2$. □

Lemma C.12. *For symmetric matrices A, B with B invertible,*

$$B \succeq A \iff I \succeq B^{-1/2} A B^{-1/2}$$

Proof.

$$\begin{aligned}
B \succeq A &\iff x^\top B x \geq x^\top A x && (\forall x) \\
&\iff x^\top (B - A) x \geq 0 && (\forall x) \\
&\iff x^\top B^{1/2} (I - B^{-1/2} A B^{-1/2}) B^{1/2} x \geq 0 && (\forall x) \\
&\iff x^\top (I - B^{-1/2} A B^{-1/2}) x \geq 0 && (\forall x) \\
&\iff I \succeq B^{-1/2} A B^{-1/2}.
\end{aligned}$$

□

Lemma C.13. *If $A \succeq 0$ and $B \succeq 0$, then $\lambda_{\min}(A + B) \geq \lambda_{\min}(A)$.*

Proof.

$$\begin{aligned}
\lambda_{\min}(A + B) &= \min_{\|x\|_2 = 1} x^\top (A + B) x \\
&= \min_{\|x\|_2 = 1} x^\top A x + x^\top B x \\
&\geq \min_{\|x\|_2 = 1} x^\top A x && \text{(because } x^\top B x \geq 0\text{)} \\
&= \lambda_{\min}(A)
\end{aligned}$$

□

C.4 Logarithms

We use several variants of standard inequalities about logarithms.

Lemma C.14. $x \geq \log(ex)$ *for all* $x > 0$.

Proof. This is true if and only if $x - \log(ex) \geq 0$ for $x > 0$. To show this, observe that

1. At $x = 1$, this holds with equality.
2. At $x = 1$, the derivative is
$$\frac{d}{dx}x - \log(ex)\bigg|_{x=1} = 1 - \frac{1}{x}\bigg|_{x=1} = 0.$$
3. The entire function is convex for $x > 0$, since
$$\frac{d^2}{dx^2}x - \log(ex) = \frac{d}{dx}1 - \frac{1}{x} = \frac{1}{x^2} > 0.$$

This proves the lemma. □

Corollary C.15. $x - \log x \geq \frac{e-1}{e}x$.

Proof. Using Lemma C.14 and letting $z = x/e$,

$$x - \log x = \frac{e-1}{e}x + \frac{1}{e}x - \log x = \frac{e-1}{e}x + z - \log(ez) \geq \frac{e-1}{e}x$$

□

Lemma C.16. $\log\left(\frac{1}{1-x}\right) \leq \frac{7x}{6}$ *for any* $x \in [0, 1/4]$.

Proof. First, we note that

$$\tfrac{d}{dx}\log\left(\tfrac{1}{1-x}\right) = 1 - x(-(1-x)^{-2})\cdot(-1) = \tfrac{1}{1-x} = \sum_{i=0}^{\infty} x^i.$$

Integrating both sides, we have

$$\log\left(\tfrac{1}{1-x}\right) = C + \sum_{i=0}^{\infty} \frac{x^i}{i},$$

for some constant C that does not depend on x. Taking $x = 0$ yields $C = 0$. Therefore,

$$\log\left(\frac{1}{1-x}\right) \leq x + \frac{x^2}{2}\sum_{i=0}^{\infty} x^i = x + \frac{x^2}{2(1-x)} = x\left(1 + \frac{x}{2(1-x)}\right) \leq \frac{7x}{6}.$$

□

APPENDIX D

SELECTION PROBLEMS IN THE PRESENCE OF IMPLICIT BIAS

D.1 Missing Proofs for Section 7.2

Proof of Theorems 7.6 and 7.9. We can expand the statement in Theorem D.1 to

$$\begin{aligned}
&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\
&\approxi \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right]\left[1-(1+c^{-1})^{-\delta/(1+\delta)}\sum_{j=0}^{k-1}\binom{j-\frac{1}{1+\delta}}{j}(1+c)^{-j}\right] \\
&\approxi (\alpha n)^{1/(1+\delta)}\Gamma\left(\frac{\delta}{1+\delta}\right)\left[1-(1+c^{-1})^{-\delta/(1+\delta)}\sum_{j=0}^{k-1}\binom{j-\frac{1}{1+\delta}}{j}(1+c)^{-j}\right]
\end{aligned}$$

(By Lemma D.22)

This gives us a ratio

$$
\begin{aligned}
& r_k(\alpha, \beta, \delta) \\
&= \frac{\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]}{\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right]} \\
&\approx\!\!\approx \frac{(\alpha n)^{1/(1+\delta)} \Gamma\left(\frac{\delta}{1+\delta}\right) \left[1 - (1+c^{-1})^{-\delta/(1+\delta)} \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right]}{(1+c)^{-(k-1/(1+\delta))} \frac{\Gamma\left(k - \frac{1}{1+\delta}\right)}{\Gamma(k)} n^{1/(1+\delta)}}
\end{aligned}
$$

(Using Theorem D.2)

$$
\begin{aligned}
&= \frac{\alpha^{1/(1+\delta)} \Gamma(k) \Gamma\left(\frac{\delta}{1+\delta}\right) (1+c)^{k-1/(1+\delta)}}{\Gamma\left(k - \frac{1}{1+\delta}\right)} \\
&\cdot \left[1 - (c^{-1}(1+c))^{-\delta/(1+\delta)} \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right] \\
&= \frac{\alpha^{1/(1+\delta)} c^{\delta/(1+\delta)} \Gamma(k) \Gamma\left(\frac{\delta}{1+\delta}\right) (1+c)^{k-1}}{\Gamma\left(k - \frac{1}{1+\delta}\right)} \\
&\cdot \left[(1+c^{-1})^{\delta/(1+\delta)} - \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right] \\
&= \frac{\alpha^{1/(1+\delta)} c^{\delta/(1+\delta)} (1+c)^{k-1}}{\binom{k-1-\frac{1}{1+\delta}}{k-1}} \left[(1+c^{-1})^{\delta/(1+\delta)} - \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right]
\end{aligned}
$$

□

Proof of Theorem 7.7. Since the only influence of β is through c and c is decreasing in β, it is sufficient to show that

$$
\frac{\alpha^{1/(1+\delta)} \left[1 - (1+c^{-1})^{-\delta/(1+\delta)} \left[1 + \frac{\delta}{1+\delta}(1+c)^{-1}\right]\right]}{\frac{\delta}{1+\delta}(1+c)^{-1-\delta/(1+\delta)}}
$$

is decreasing in c. Ignoring constants, this is

$$
\begin{aligned}
&\propto c^{\delta/(1+\delta)}(1+c)\left[(1+c^{-1})^{\delta/(1+\delta)} - 1 - \frac{\delta}{1+\delta}(1+c)^{-1}\right] \\
&= (1+c)^{1+\delta/(1+\delta)} - c^{\delta/(1+\delta)}(1+c) - \frac{\delta}{1+\delta} c^{\delta/(1+\delta)} \\
&= (1+c)^{1+\delta/(1+\delta)} - c^{1+\delta/(1+\delta)} - \left(1 + \frac{\delta}{1+\delta}\right) c^{\delta/(1+\delta)}
\end{aligned}
$$

This has derivative

$$\begin{aligned}
&\frac{\delta}{dc}(1+c)^{1+\delta/(1+\delta)} - c^{1+\delta/(1+\delta)} - \left(1+\frac{\delta}{1+\delta}\right)c^{\delta/(1+\delta)} \\
&\quad = \left(1+\frac{\delta}{1+\delta}\right)(1+c)^{\delta/(1+\delta)} - \left(1+\frac{\delta}{1+\delta}\right)c^{\delta/(1+\delta)} \\
&\quad + \left(\frac{\delta}{1+\delta}\right)\left(1+\frac{\delta}{1+\delta}\right)c^{-1/(1+\delta)},
\end{aligned}$$

which is negative if and only if

$$\begin{aligned}
&(1+c)^{\delta/(1+\delta)} < c^{\delta/(1+\delta)} + \frac{\delta}{1+\delta}c^{-1/(1+\delta)} \\
\iff &(1+c)^{\delta/(1+\delta)}c^{-\delta/(1+\delta)} < 1 + \frac{\delta}{1+\delta}c^{-1} \\
\iff &(1+c^{-1})^{\delta/(1+\delta)} < 1 + \frac{\delta}{1+\delta}c^{-1}.
\end{aligned}$$

This is true by Lemma D.21, which proves the theorem. □

Proof of Theorem 7.10. By Theorem 7.9,

$$\begin{aligned}
\phi_k(\alpha,\beta,\delta) = &\frac{\alpha^{1/(1+\delta)}c^{\delta/(1+\delta)}\Gamma(k)\Gamma\left(\frac{\delta}{1+\delta}\right)(1+c)^{k-1}}{\Gamma\left(k-\frac{1}{1+\delta}\right)} \\
&\cdot\left[(1+c^{-1})^{\delta/(1+\delta)} - \sum_{j=0}^{k-1}\binom{j-\frac{1}{1+\delta}}{j}(1+c)^{-j}\right]
\end{aligned}$$

We use the fact that for $a, b \in \mathbb{Z}$ and $s \in \mathbb{R}$

$$\frac{\Gamma(s-a+1)}{\Gamma(s-b+1)} = (-1)^{b-a}\frac{\Gamma(b-s)}{\Gamma(a-s)}.$$

If the summation went to ∞, it would be

$$\begin{aligned}
\sum_{j=0}^{\infty} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j} &= \sum_{j=0}^{\infty} (1+c)^{-j} \frac{\Gamma\left(j + \frac{\delta}{1+\delta}\right)}{\Gamma\left(\frac{\delta}{1+\delta}\right)\Gamma(j+1)} \\
&= \sum_{j=0}^{\infty} (1+c)^{-j} (-1)^j \frac{\Gamma\left(1 - \frac{\delta}{1+\delta}\right)}{\Gamma\left(-j+1+\frac{\delta}{1+\delta}\right)\Gamma(j+1)} \\
&= \sum_{j=0}^{\infty} \binom{-\frac{\delta}{1+\delta}}{j} (-(1+c)^{-1})^j \\
&= (1-(1+c)^{-1})^{-\delta/(1+\delta)} \\
&= (1+c^{-1})^{\delta/(1+\delta)}
\end{aligned}$$

Therefore,

$$\sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j} = (1+c^{-1})^{\delta/(1+\delta)} - \sum_{j=k}^{\infty} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}.$$

Plugging this in,

$$\begin{aligned}
\phi_k(\alpha, \beta, \delta) &= \frac{\alpha^{1/(1+\delta)} c^{\delta/(1+\delta)} \Gamma(k) \Gamma\left(\frac{\delta}{1+\delta}\right) (1+c)^{k-1}}{\Gamma\left(k - \frac{1}{1+\delta}\right)} \sum_{j=k}^{\infty} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j} \\
&= \frac{\alpha^{1/(1+\delta)} c^{\delta/(1+\delta)} \Gamma(k) \Gamma\left(\frac{\delta}{1+\delta}\right)}{\Gamma\left(k - \frac{1}{1+\delta}\right)(1+c)} \sum_{j=0}^{\infty} \binom{j + k - \frac{1}{1+\delta}}{j+k} (1+c)^{-j}
\end{aligned}$$

With this, we can take

$$
\begin{aligned}
&\phi_{k+1}(\alpha,\beta,\delta) - \phi_k(\alpha,\beta,\delta) \\
&= \frac{\alpha^{1/(1+\delta)}c^{\delta/(1+\delta)}\Gamma\left(\frac{\delta}{1+\delta}\right)}{(1+c)}\left[\frac{\Gamma(k+1)}{\Gamma\left(k+\frac{\delta}{1+\delta}\right)}\sum_{j=0}^{\infty}\binom{j+k+1-\frac{1}{1+\delta}}{j+k+1}(1+c)^{-j}\right. \\
&\left. - \frac{\Gamma(k)}{\Gamma\left(k-\frac{1}{1+\delta}\right)}\sum_{j=0}^{\infty}\binom{j+k-\frac{1}{1+\delta}}{j+k}(1+c)^{-j}\right] \\
&= \frac{\alpha^{1/(1+\delta)}c^{\delta/(1+\delta)}\Gamma(k)\Gamma\left(\frac{\delta}{1+\delta}\right)}{\Gamma\left(k-\frac{1}{1+\delta}\right)(1+c)}\left[\frac{k}{k-\frac{1}{1+\delta}}\sum_{j=0}^{\infty}\binom{j+k+1-\frac{1}{1+\delta}}{j+k+1}(1+c)^{-j}\right. \\
&\left. - \sum_{j=0}^{\infty}\binom{j+k-\frac{1}{1+\delta}}{j+k}(1+c)^{-j}\right] \\
&= \frac{\alpha^{1/(1+\delta)}c^{\delta/(1+\delta)}\Gamma(k)\Gamma\left(\frac{\delta}{1+\delta}\right)}{\Gamma\left(k-\frac{1}{1+\delta}\right)(1+c)}\sum_{j=0}^{\infty}(1+c)^{-j} \\
&\cdot\left[\frac{k}{k-\frac{1}{1+\delta}}\binom{j+k+1-\frac{1}{1+\delta}}{j+k+1} - \binom{j+k-\frac{1}{1+\delta}}{j+k}\right]
\end{aligned}
$$

Thus, to show that $\phi_{k+1} > \phi_k$, it is sufficient to show that for $j \geq 0$,

$$
\begin{aligned}
\frac{k}{k-\frac{1}{1+\delta}}\binom{j+k+1-\frac{1}{1+\delta}}{j+k+1} - \binom{j+k-\frac{1}{1+\delta}}{j+k} &> 0 \\
\frac{k}{k-\frac{1}{1+\delta}}\frac{\Gamma\left(j+k+1+\frac{\delta}{1+\delta}\right)}{\Gamma(j+k+2)\Gamma\left(\frac{\delta}{1+\delta}\right)} - \frac{\Gamma\left(j+k+\frac{\delta}{1+\delta}\right)}{\Gamma(j+k+1)\Gamma\left(\frac{\delta}{1+\delta}\right)} &> 0 \\
\frac{k}{k-\frac{1}{1+\delta}}\frac{j+k+\frac{\delta}{1+\delta}}{j+k+1} - 1 &> 0 \qquad (\Gamma(x+1) = x\Gamma(x)) \\
\frac{k-\frac{1}{1+\delta}+(j+1)}{k+(j+1)} &> \frac{k-\frac{1}{1+\delta}}{k}
\end{aligned}
$$

The last inequality holds by Lemma D.16. As a result a result, $\phi_{k+1} > \phi_k$, proving the theorem. □

Proof Theorem 7.11. We want to find

$$\Pr\left[X_{(\alpha n:\alpha n)} > Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right],$$

or equivalently,

$$\Pr\left[X_{(\alpha n:\alpha n)} < Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right].$$

This can be written as

$$\frac{\Pr\left[X_{(\alpha n:\alpha n)} < Y_{(n-k+1:n)} \cap X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]}{\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]} = \frac{\Pr\left[X_{(\alpha n:\alpha n)} < Y_{(n-k+1:n)}\right]}{\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right]}. \tag{D.1}$$

By Theorem D.3, the numerator can be approximated by $(1+\alpha)^{-k}$ while the denominator is approximately $(1+\alpha\beta^{-(1+\delta)})^{-k}$. Thus, we have

$$\begin{aligned}\Pr\left[X_{(\alpha n:\alpha n)} < Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] &\approxeq \frac{(1+\alpha\beta^{-(1+\delta)})^k}{(1+\alpha)^k} \\ &= \left(\frac{1+\alpha\beta^{-(1+\delta)}}{1+\alpha}\right)^k,\end{aligned}$$

and therefore

$$\Pr\left[X_{(\alpha n:\alpha n)} > Y_{(n-k+1:n)} \mid X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] \approxeq 1 - \left(\frac{1+\alpha\beta^{-(1+\delta)}}{1+\alpha}\right)^k.$$

□

D.2 Additional Theorems for Power Laws

Theorem D.1.

$$\begin{aligned}&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\ &\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right]\left[1 - (1+c^{-1})^{-\delta/(1+\delta)} \sum_{j=0}^{k-1} \binom{j - \frac{1}{1+\delta}}{j} (1+c)^{-j}\right]\end{aligned}$$

where $c = \alpha\beta^{-(1+\delta)}$.

Proof. First, observe that

$$\begin{aligned}&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\ &= \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] - \mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} \geq \beta Y_{(n-k+1:n)}\}}\right].\end{aligned}$$

Next, we use the fact that

$$\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} \geq \beta Y_{(n-k+1:n)}\}}\right] = \int_{\beta}^{\infty} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx.$$

We know that

$$\begin{aligned}\int_{\beta}^{\infty} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx = &\int_{\beta}^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx \\ &+ \int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx,\end{aligned} \tag{D.2}$$

and

$$\begin{aligned}&\int_{\beta}^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx \\ &\qquad \leq \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} F_{(\alpha n:\alpha n)}\left(\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}\right) \\ &\qquad \leq \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} \cdot \frac{1}{n}\end{aligned}$$

by Lemma D.18. The second term of (D.2) is

$$\begin{aligned}&(1+\delta)\alpha n \int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} (1 - x^{-(1+\delta)})^{\alpha n - 1} x^{-(1+\delta)} \\ &\sum_{j=0}^{k-1} \binom{n}{j} \left(1 - \left(\frac{x}{\beta}\right)^{-(1+\delta)}\right)^{n-j} \left(\frac{x}{\beta}\right)^{-j(1+\delta)}\, dx \\ &= (1+\delta)\alpha n \sum_{j=0}^{k-1} \binom{n}{j} \beta^{j(1+\delta)} \int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} (1 - x^{-(1+\delta)})^{\alpha n - 1} \left(x^{-(1+\delta)}\right)^{j+1} \\ &\cdot \left(1 - \left(\frac{x}{\beta}\right)^{-(1+\delta)}\right)^{n-j}\, dx\end{aligned}$$

Next, we show that for $x \geq \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}$,

$$\left(1 - \left(\frac{x}{\beta}\right)^{-(1+\delta)}\right)^{n-j} \approxeq (1 - x^{-(1+\delta)})^{\beta^{1+\delta} n - j}.$$

We begin with

$$\left(1-\left(\frac{x}{\beta}\right)^{-(1+\delta)}\right)^{n-j} \approxeq (1-x^{-(1+\delta)})^{\beta^{1+\delta}(n-j)}$$
$$= (1-x^{-(1+\delta)})^{\beta^{1+\delta}n-j}(1-x^{-(1+\delta)})^{-j(\beta^{1+\delta}-1)}.$$

Note that $(1-x^{-(1+\delta)})^{-j(\beta^{1+\delta}-1)} \geq 1$, and by Lemma D.17,

$$(1-x^{-(1+\delta)})^{-j(\beta^{1+\delta}-1)} = 1 + j(\beta^{1+\delta}-1)x^{-(1+\delta)} + O\left(\frac{1}{n}\right) \approxeq 1.$$

because $j \leq \ln n$. Thus, $(1-x^{-(1+\delta)})^{\beta^{1+\delta}n-j}(1-x^{-(1+\delta)})^{-j(\beta^{1+\delta}-1)} \approxeq (1-x^{-(1+\delta)})^{\beta^{1+\delta}n-j}$. Therefore, this becomes

$$(1+\delta)\alpha n \sum_{j=0}^{k-1} \binom{n}{j} \beta^{j(1+\delta)} \int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} \left(1-x^{-(1+\delta)}\right)^{\beta^{1+\delta}n(1+c)-j-1} \left(x^{-(1+\delta)}\right)^{j+1} dx.$$

We'll now try to relate the jth term in this summation to the order statistic $Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}$. We know that

$$\begin{aligned}
&\mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \\
&= \int_1^{\infty} z f_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}(z)\, dz \\
&= (1+\delta)(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1} \int_1^{\infty} \left(1-z^{-(1+\delta)}\right)^{\beta^{1+\delta}n(1+c)-j-1} \left(z^{-(1+\delta)}\right)^{j+1} dz.
\end{aligned}$$

Using this, we have

$$\begin{aligned}
&\int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right) dx \\
&\qquad \approxeq \sum_{j=0}^{k-1} \frac{\alpha n \beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \left[\mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right]\right. \\
&\qquad \left. - \int_1^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} z f_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}(z)\, dz\right]
\end{aligned}$$

We'll show that this last multiplicative term is approximately

$$\mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right].$$

Observe that

$$\begin{aligned}
&\int_1^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} z f_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}(z)\,dz \\
&\qquad \le \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} \int_1^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} f_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}(z)\,dz \\
&\qquad = \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} F_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\left(\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}\right) \\
&\qquad \le \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} \frac{\sqrt{k}}{n}
\end{aligned}$$

by Lemma D.15. This means

$$\begin{aligned}
&\sum_{j=0}^{k-1} \frac{\alpha n \beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \\
&\quad \ge \sum_{j=0}^{k-1} \frac{\alpha n \beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \left[\mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] - \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} \frac{\sqrt{k}}{n}\right] \\
&\quad \approxgeq \sum_{j=0}^{k-1} \frac{\alpha n \beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \qquad \text{(by Lemma D.19)}
\end{aligned}$$

Next, we deal with the $n\beta^{j(1+\delta)}\binom{n}{j}/((j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1})$ terms. These are

$$\begin{aligned}
&\frac{n\beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \\
&= \frac{n(n-1)\cdots(n-j+1)}{\beta^{1+\delta}n(1+c)(\beta^{1+\delta}n(1+c)-1)\cdots(\beta^{1+\delta}n(1+c)-j+1)} \cdot \frac{n\beta^{j(1+\delta)}}{\beta^{1+\delta}n(1+c)-j}.
\end{aligned} \tag{D.3}$$

Each term $(n-\ell)/(\beta^{1+\delta}n(1+c)-\ell)$ is between $1/(\beta^{1+\delta}(1+c))$ and $1/(\beta^{1+\delta}(1+c))\cdot(1-\ell/n)$. This means

$$\begin{aligned}
\frac{1}{(\beta^{1+\delta}(1+c))^j} &\ge \prod_{\ell=0}^{j} \frac{n-\ell}{\beta^{1+\delta}n(1+c)-\ell} \\
&\ge \prod_{\ell=0}^{j} \frac{1}{\beta^{1+\delta}(1+c)}\left(1-\frac{\ell}{n}\right) \\
&\ge \left(1-\frac{j^2}{n}\right) \\
&\approx \frac{1}{(\beta^{1+\delta}(1+c))^j}
\end{aligned}$$

since $j \le k \le ((1-c^2)/2)\ln n$. Multiplying by the second term in (D.3), which is

$$\frac{n\beta^{j(1+\delta)}}{\beta^{1+\delta}n(1+c)-j} \approxeq \frac{\beta^{(j-1)(1+\delta)}}{1+c},$$

we have

$$\frac{n\beta^{j(1+\delta)}\binom{n}{j}}{(j+1)\binom{\beta^{1+\delta}n(1+c)}{j+1}} \approxeq \frac{1}{\beta^{1+\delta}(1+c)^{j+1}}.$$

As a result,

$$\begin{aligned}
&\int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx \\
&\qquad \approxeq \sum_{j=0}^{k-1} \frac{\alpha}{\beta^{1+\delta}(1+c)^{j+1}} \mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \\
&\qquad = \sum_{j=0}^{k-1} \frac{c}{(1+c)^{j+1}} \mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \qquad \text{(D.4)}
\end{aligned}$$

Finally, note that

$$\begin{aligned}
&\mathbb{E}\left[Z_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\right] \\
&= \mathbb{E}\left[Z_{(n\beta^{1+\delta}(1+c):\beta^{1+\delta}n(1+c))}\right] \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)} \\
&\approxeq (\beta^{1+\delta}n(1+c))^{1/(1+\delta)} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(j+1)} \\
&= \beta(1+c)^{1/(1+\delta)} n^{1/(1+\delta)} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(j+1)} \\
&= \frac{\beta}{\alpha^{1/(1+\delta)}}(1+c)^{1/(1+\delta)}(\alpha n)^{1/(1+\delta)} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(j+1)} \\
&\approxeq c^{-1/(1+\delta)}(1+c)^{1/(1+\delta)} \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)}
\end{aligned}$$

Substituting back to (D.4),

$$\begin{aligned}
&\int_{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}}^{\infty} x f_{(\alpha n:\alpha n)}(x) \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] c^{\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-(j+\delta/(1+\delta))} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)}
\end{aligned}$$

Going back to (D.2),

$$\begin{aligned}
&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} > \beta Y_{(n-k+1:n)}\}}\right] \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] c^{\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-(j+\delta/(1+\delta))} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)} \\
&+ \int_{\beta}^{\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}} x f_{(\alpha n:\alpha n)}(x) F_{(n-k+1:n)}\left(\frac{x}{\beta}\right)\, dx \\
&\leq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] c^{\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-(j+\delta/(1+\delta))} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)} \\
&+ \left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)} \frac{(\ln n)^2}{n} \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] c^{\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-(j+\delta/(1+\delta))} \frac{\Gamma(j+\delta/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(j+1)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] \left[1 - c^{\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-(j+\delta/(1+\delta))} \frac{\Gamma\left(j+\frac{\delta}{1+\delta}\right)}{\Gamma\left(\frac{\delta}{1+\delta}\right)\Gamma(j+1)}\right].
\end{aligned}$$

We can simplify this to

$$\begin{aligned}
&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] \left[1 - (1+c^{-1})^{-\delta/(1+\delta)} \sum_{j=0}^{k-1} (1+c)^{-j} \frac{\Gamma\left(j+\frac{\delta}{1+\delta}\right)}{\Gamma\left(\frac{\delta}{1+\delta}\right)\Gamma(j+1)}\right].
\end{aligned}$$

Using the definition

$$\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)},$$

this is

$$\begin{aligned}
&\mathbb{E}\left[X_{(\alpha n:\alpha n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \\
&\approxeq \mathbb{E}\left[X_{(\alpha n:\alpha n)}\right] \left[1 - (1+c^{-1})^{-\delta/(1+\delta)} \sum_{j=0}^{k-1} \binom{j-\frac{1}{1+\delta}}{j} (1+c)^{-j}\right].
\end{aligned}$$

□

Theorem D.2.

$$\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] \approxeq (1+\alpha\beta^{-(1+\delta)})^{-(k-1/(1+\delta))}\mathbb{E}\left[Y_{(n-k+1:n)}\right]$$

Proof. We begin with

$$\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\}}\right] = \int_1^\infty y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy.$$

Let $c = \alpha\beta^{-(1+\delta)}$. Break this up into

$$\begin{aligned}\int_1^\infty y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy = &\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\ &+ \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy.\end{aligned} \tag{D.5}$$

The first term is

$$\begin{aligned}&\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\ &\qquad \le F_{(\alpha n:\alpha n)}\left(\beta\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right)\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n-k+1:n)}(y)\, dy \\ &\qquad \le F_{(\alpha n:\alpha n)}\left(\beta\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right)\mathbb{E}\left[Y_{(n-k+1:n)}\right] \\ &\qquad \le \frac{\mathbb{E}\left[Y_{(n-k+1:n)}\right]}{n}\end{aligned}$$

by Lemma D.18.

For the second term in (D.5), we have

$$\begin{aligned}&\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\ &= (1+\delta)k\binom{n}{k}\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty \left(1-y^{-(1+\delta)}\right)^{n-k}\left(y^{-(1+\delta)}\right)^k\left(1-(\beta y)^{-(1+\delta)}\right)^{\alpha n} dy\end{aligned}$$

By Lemma D.14, for all $y \ge (cn/\ln n)^{1/(1+\delta)}$,

$$\left(1-(\beta y)^{-(1+\delta)}\right)^{\alpha n} \approxeq \left(1-y^{-(1+\delta)}\right)^{cn}.$$

Therefore,

$$\begin{aligned}
&\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\
&\qquad \approxeq (1+\delta)k\binom{n}{k}\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty}\left(1-y^{-(1+\delta)}\right)^{n-k+cn}\left(y^{-(1+\delta)}\right)^{k} dy \\
&\qquad = (1+\delta)k\binom{n}{k}\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty}\left(1-y^{-(1+\delta)}\right)^{n(1+c)-k}\left(y^{-(1+\delta)}\right)^{k} dy.
\end{aligned}$$

We'll now try to relate this to the order statistic $Z_{(n(1+c)-k+1:n(1+c))}$. We know that

$$\begin{aligned}
&\mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] \\
&= \int_1^{\infty} z f_{(n(1+c)-k+1:n(1+c))}(z)\, dz \\
&= (1+\delta)k\binom{n(1+c)}{k}\int_1^{\infty}\left(1-z^{-(1+\delta)}\right)^{n(1+c)-k}\left(z^{-(1+\delta)}\right)^{k} dz.
\end{aligned}$$

Using this, we have

$$\begin{aligned}
&\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\
&\approxeq \frac{\binom{n}{k}}{\binom{n(1+c)}{k}}\left[\mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] - \int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n(1+c)-k+1:n(1+c))}(y)\, dy\right].
\end{aligned} \tag{D.6}$$

From here, we'll show that the term being subtracted is only a $\frac{\sqrt{\ln n}}{n}$ fraction of $\mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right]$. To do so, note that

$$\begin{aligned}
&\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n(1+c)-k+1:n(1+c))}(y)\, dy \\
&\qquad \le \left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} f_{(n(1+c)-k+1:n(1+c))}(y)\, dy \\
&\qquad = \left(\frac{cn}{\ln n}\right)^{1/(1+\delta)} F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) \\
&\qquad \le \left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\left(\frac{\sqrt{k}}{n}\right)
\end{aligned}$$

By Lemma D.13. Lemma D.19 gives us

$$\begin{aligned}
&\mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] \\
&\geq \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] - \int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} y f_{(n(1+c)-k+1:n(1+c))}(y)\, dy \\
&\geq \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] - \left(\frac{cn}{\ln n}\right)^{1/(1+\delta)} \left(\frac{\sqrt{k}}{n}\right) \\
&\geq \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] \left(1 - \frac{\sqrt{k}}{n}\right) \\
&\approxgeq \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right]
\end{aligned}$$

Combining with (D.6), Lemma D.9 yields

$$\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \approxgeq \frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] \quad \text{(D.7)}$$

By Lemma D.20,

$$\frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \approxgeq \frac{1}{(1+c)^k}.$$

Putting this into (D.7),

$$\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} y f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \approxgeq \frac{1}{(1+c)^k} \mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right].$$

Finally, note that

$$\begin{aligned}
\mathbb{E}\left[Z_{(n(1+c)-k+1:n(1+c))}\right] &= \mathbb{E}\left[Z_{(n(1+c):n(1+c))}\right] \frac{\Gamma(k - 1/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(k)} \\
&\approxgeq (n(1+c))^{1/(1+\delta)} \frac{\Gamma(k - 1/(1+\delta))}{\Gamma(k)} \\
&= (1+c)^{1/(1+\delta)} n^{1/(1+\delta)} \frac{\Gamma(k - 1/(1+\delta))}{\Gamma(k)} \\
&\approxgeq (1+c)^{1/(1+\delta)} \mathbb{E}\left[Y_{(n:n)}\right] \frac{\Gamma(k - 1/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(k)} \\
&= (1+c)^{1/(1+\delta)} \mathbb{E}\left[Y_{(n-k+1:n)}\right]
\end{aligned}$$

Substituting into (D.5),

$$\mathbb{E}\left[Y_{(n-k+1:n)} \cdot \mathbf{1}_{\{X_{(\alpha n:\alpha n)} \le \beta Y_{(n-k+1:n)}\}}\right] \approxeq \mathbb{E}\left[Y_{(n-k+1:n)}\right]\left((1+c)^{-(k-1/(1+\delta))} + \frac{1}{n}\right)$$
$$\approxeq \mathbb{E}\left[Y_{(n-k+1:n)}\right](1+\alpha\beta^{-(1+\delta)})^{-(k-1/(1+\delta))}$$

since $c = a\beta^{-1(1+\delta)}$, proving the theorem. □

Theorem D.3.

$$\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] \approxeq (1+c)^{-k}.$$

Proof. Begin with

$$\begin{aligned}\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] &= \int_1^\infty f_{(n-k+1:n)}(y)F_{(\alpha n:\alpha n)}(\beta y)\,dy \\ &= \int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} f_{(n-k+1:n)}(y)F_{(\alpha n:\alpha n)}(\beta y)\,dy \\ &\quad + \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty f_{(n-k+1:n)}(y)F_{(\alpha n:\alpha n)}(\beta y)\,dy \qquad \text{(D.8)}\end{aligned}$$

Observe that

$$\begin{aligned}&\int_1^{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}} f_{(n-k+1:n)}(y)F_{(\alpha n:\alpha n)}(\beta y)\,dy \\ &\quad\le F_{(\alpha n:\alpha n)}\left(\beta\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right)F_{(n-k+1:n)}\left(\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) \\ &\quad\le F_{(\alpha n:\alpha n)}\left(\beta\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) \\ &\quad\le \left(1-\beta^{-(1+\delta)}\left(\frac{\ln n}{cn}\right)\right)^{\alpha n} \\ &\quad\le \exp\left(-\alpha\beta^{-(1+\delta)}\frac{\ln n}{cn}\right) \\ &\quad= \frac{1}{n}\end{aligned}$$

Next, we have

$$\begin{aligned}&\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty f_{(n-k+1:n)}(y)F_{(\alpha n:\alpha n)}(\beta y)\,dy \\ &\qquad= \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^\infty (1-(\beta y)^{-(1+\delta)})^{\alpha n} f_{(n-k+1:n)}(y)\,dy\end{aligned}$$

By Lemma D.14, for $y \geq (cn/\ln n)^{1/(1+\delta)}$,

$$\left(1 - (\beta y)^{-(1+\delta)}\right)^{\alpha n} \approxgeq \left(1 - y^{-(1+\delta)}\right)^{cn},$$

so

$$\begin{aligned}
&\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \\
&\qquad \approxgeq \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} (1 - y^{-(1+\delta)})^{cn} f_{(n-k+1:n)}(y)\, dy \\
&\qquad = (1+\delta)k \binom{n}{k} \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} (1 - y^{-(1+\delta)})^{n(1+c)-k} (y^{-(1+\delta)})^{k} y^{-1}\, dy \\
&\qquad = \frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} f_{(n(1+c)-k+1:n(1+c))}\, dy
\end{aligned}$$

From Lemma D.13, we have

$$F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}^{1/(1+\delta)}\right)\right) \leq \frac{\sqrt{k}}{n},$$

so

$$\begin{aligned}
\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} f_{(n(1+c)-k+1:n(1+c))}\, dy &= 1 - F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}^{1/(1+\delta)}\right)\right) \\
&\geq 1 - \frac{\sqrt{k}}{n} \\
&\approx 1.
\end{aligned}$$

Therefore,

$$\int_{\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}}^{\infty} f_{(n-k+1:n)}(y) F_{(\alpha n:\alpha n)}(\beta y)\, dy \approxgeq \frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \approxgeq \frac{1}{(1+c)^k}$$

by Lemma D.20. By (D.8), this means

$$\Pr\left[X_{(\alpha n:\alpha n)} < \beta Y_{(n-k+1:n)}\right] \approxgeq (1+c)^{-k}.$$

□

D.3 Lemmas for the Equivalence Definition

Lemma D.4 (Transitivity). *If* $f(n) \approxeq_{a_1;n_1} g(n)$ *and* $g(n) \approxeq_{a_2;n_2} h(n)$*, then* $f(n) \approxeq h(n)$.

Proof.

$$\begin{aligned}\frac{f(n)}{h(n)} &= \frac{f(n)}{g(n)} \cdot \frac{g(n)}{h(n)} \\ &\leq \left(1 + \frac{a_1(\ln n)^2}{n}\right)\left(1 + \frac{a_2(\ln n)^2}{n}\right) \\ &\leq 1 + \frac{(a_1 + a_2)(\ln n)^2}{n} + \frac{a_1 a_2 (\ln n)^4}{n^2} \\ &\leq 1 + \frac{(a_1 + a_2 + a_1 a_2)(\ln n)^2}{n}\end{aligned}$$

for all $n \geq \max(n_1, n_2)$, since $n \geq (\ln n)^2$. A symmetric argument holds for $h(n)/f(n)$. Thus, $f(n) \approxeq_{a_1+a_2+a_1a_2;\max(n_1,n_2)} h(n)$. □

Lemma D.5 (Linearity). *If* $f_1(n) \approxeq_{a_1;n_1} g_1(n)$ *and* $f_1(n) \approxeq_{a_2;n_2} g_2(n)$*, then* $bf_1(n) + cf_2(n) \approxeq bg_1(n) + cg_2(n)$.

Proof. By Lemma D.11,

$$\frac{bf_1(n) + cf_2(n)}{bg_1(n) + cg_2(n)} \leq \max\left(\frac{f_1(n)}{g_1(n)}, \frac{f_2(n)}{g_2(n)}\right) \leq \frac{\max(a_1, a_2)(\ln n)^2}{n}$$

for $n \geq \max(n_1, n_2)$. A symmetric argument holds for the reciprocal. Therefore,

$$bf_1(n) + cf_2(n) \approxeq_{\max(a_1,a_2);\max(n_1,n_2)} bg_1(n) + cg_2(n).$$

□

Lemma D.6 (Integrals). *If* $f(x, n) \approxeq_{a;n_0} g(x, n)$*, then*

$$\int f(x, n)\, dx \approxeq \int g(x, n)\, dx$$

Proof.

$$\frac{\int f(x,n)\,dx}{\int g(x,n)\,dx} = \frac{\int g(x,n)\frac{f(x,n)}{g(x,n)}\,dx}{\int g(x,n)\,dx} \leq \frac{\int g(x,n)\left(1+\frac{a(\ln n)^2}{n}\right)\,dx}{\int g(x,n)\,dx} \leq 1+\frac{a(\ln n)^2}{n}$$

for $n \geq n_0$. A symmetric argument holds for the reciprocal, proving the lemma.

□

Lemma D.7. *If $f_1(n) \approxeq_{a_1;n_1} g_1(n)$ and $f_2(n) \approxeq_{a_2;n_2} g_2(n)$, then*

$$f_1(n)f_2(n) \approxeq g_1(n)g_2(n).$$

Proof.

$$\begin{aligned}\frac{f_1(n)f_2(n)}{g_1(n)g_2(n)} &= \frac{f_1(n)}{g_1(n)} \cdot \frac{f_2(n)}{g_2(n)} \\ &\leq \left(1+\frac{a_1(\ln n)^2}{n}\right)\left(1+\frac{a_2(\ln n)^2}{n}\right) \\ &\leq 1+\frac{(a_1+a_2)(\ln n)^2}{n}+\frac{a_1a_2(\ln n)^4}{n^2} \\ &\leq 1+\frac{(a_1+a_2+a_1a_2)(\ln n)^2}{n}\end{aligned}$$

for all $n \geq \max(n_1, n_2)$, since $n \geq (\ln n)^2$. A symmetric argument holds for the reciprocal. Thus, $f_1(n)f_2(n) \approxeq_{a_1+a_2+a_1a_2;\max(n_1,n_2)} g_1(n)g_2(n)$. □

Lemma D.8. *If $f(n) \approxeq_{a;n_0} g(n)$, then $\frac{1}{f(n)} \approxeq \frac{1}{g(n)}$.*

Proof.

$$\frac{1/f(n)}{1/g(n)} = \frac{g(n)}{f(n)} \leq 1+\frac{a(\ln n)^2}{n}$$

for $n \geq n_0$. A symmetric argument holds for the reciprocal. □

Lemma D.9. *If $g_1(n) \leq f(n) \leq g_2(n)$, $g_1(n) \approxeq h(n)$, and $g_2(n) \approxeq h(n)$, then $f(n) \approxeq h(n)$.*

Proof.

$$\frac{f(n)}{h(n)} \leq \frac{g_2(n)}{h(n)}$$

and

$$\frac{h(n)}{f(n)} \leq \frac{h(n)}{g_1(n)},$$

proving the lemma by definition. □

Fact D.10. *For all $x \geq 1$, $\ln x \leq x$ and $(\ln x)^2 \leq x$.*

Lemma D.11. *For $a, b, c, d > 0$, if $\frac{a}{b} \leq \frac{c}{d}$, then*

$$\frac{a}{b} \leq \frac{a+c}{b+d} \leq \frac{c}{d}.$$

Proof. Since $\frac{a}{b} \leq \frac{c}{d}$, $\frac{d}{b} \leq \frac{c}{a}$. Therefore,

$$\frac{a+c}{b+d} = \frac{a}{b} \cdot \frac{1+c/a}{1+d/b} \geq \frac{a}{b} \cdot \frac{1+d/b}{1+d/b} = \frac{a}{b}.$$

Similarly,

$$\frac{a+c}{b+d} = \frac{c}{d} \cdot \frac{1+a/c}{1+b/d} \leq \frac{c}{d} \cdot \frac{1+b/d}{1+b/d} = \frac{c}{d}.$$

□

Lemma D.12.

$$\frac{a - (\ln n)^2/n}{b} \approxeq \frac{a}{b}$$

Proof.

$$\frac{\frac{a-(\ln n)^2/n}{b}}{\frac{a}{b}} = 1 - \frac{(\ln n)^2}{n} \leq 1$$

$$\frac{\frac{a}{b}}{\frac{a-(\ln n)^2/n}{b}} = \frac{1}{1 - \frac{(\ln n)^2}{an}} = 1 + \frac{\frac{(\ln n)^2}{an}}{1 - \frac{(\ln n)^2}{an}} \leq 1 + \frac{2(\ln n)^2}{an}$$

for $n \geq 16/a^4$. □

D.4 Lemmas for Appendix D.2

Lemma D.13. *For $k \leq (1-c)\ln n$,*

$$F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}^{1/(1+\delta)}\right)\right) \leq \frac{\sqrt{k}}{n}.$$

Proof. We can write

$$\begin{aligned}
&F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) \\
&= \sum_{j=0}^{k-1}\binom{n(1+c)}{j}\left(1-\frac{\ln n}{cn}\right)^{n(1+c)-j}\left(\frac{\ln n}{cn}\right)^{j} \\
&\leq \sum_{j=0}^{k-1}\frac{(n(1+c))^{j}}{j!}\exp\left(-\left(\frac{\ln n}{cn}\right)(n(1+c)-j)\right)\left(\frac{\ln n}{cn}\right)^{j} \\
&= \sum_{j=0}^{k-1}\frac{1}{j!}\left(\frac{(1+c)\ln n}{c}\right)^{j}\exp\left(-\ln n\left(1+c^{-1}\left(1-\frac{j}{n}\right)\right)\right) \\
&= \frac{1}{n}\sum_{j=0}^{k-1}\frac{1}{j!}\left((1+c^{-1})\ln n\right)^{j}\left(\frac{1}{n}\right)^{c^{-1}\left(1-\frac{j}{n}\right)} \\
&\leq \frac{1}{n^2}+\frac{1}{n}\sum_{j=1}^{k-1}\frac{1}{\sqrt{2\pi j}}\left(\frac{e(1+c^{-1})\ln n}{j}\right)^{j}\left(\frac{1}{n}\right)^{c^{-1}\left(1-\frac{j}{n}\right)} \qquad \text{(D.9)}
\end{aligned}$$

by Stirling's approximation. The term

$$\left(\frac{e(1+c^{-1})\ln n}{j}\right)^{j}$$

is increasing whenever it's natural log,

$$j\left(1+\ln(1+c^{-1})+\ln\ln n-\ln j\right),$$

is increasing. This has derivative

$$1+\ln(1+c^{-1})+\ln\ln n-\ln j-1=\ln(1+c^{-1})+\ln\ln n-\ln j\geq \ln\ln n-\ln j.$$

Thus, it is increasing for $j \le \ln n$. For $j \le (1-c)\ln n$, we have

$$\begin{aligned}
\left(\frac{e(1+c^{-1})\ln n}{j}\right)^j &\le \left(\frac{e(1+c^{-1})\ln n}{(1-c)\ln n}\right)^{(1-c)\ln n} \\
&= \left(\frac{e(1+c^{-1})}{1-c}\right)^{(1-c)\ln n} \\
&= \exp\left(1+\ln(1+c^{-1})-\ln(1-c)\right)^{(1-c)\ln n} \\
&= \exp\left(\ln n\right)^{(1+\ln(1+c^{-1})-\ln(1-c))(1-c)} \\
&= n^{(1+\ln(1+c^{-1})-\ln(1-c))(1-c)} \\
&\le n^{(1+c^{-1}+c+c^2)(1-c)} \\
&= n^{1+c^{-1}+c+c^2-c-1-c^2-c^3} \\
&= n^{c^{-1}-c^3} \\
&\le n^{c^{-1}(1-j/n)}
\end{aligned}$$

for sufficiently large n, since $j \le (1-c)\ln n$. Combining this with (D.9), we have

$$\begin{aligned}
F_{(n(1+c)-k+1:n(1+c))}\left(\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) &\le \frac{1}{n^2} + \frac{1}{n\sqrt{2\pi}}\sum_{j=1}^{k-1}\frac{1}{\sqrt{j}} \\
&\le \frac{1}{n^2} + \frac{1}{n\sqrt{2\pi}}\left(1+\int_1^k \frac{1}{\sqrt{j}}\,dj\right) \\
&\le \frac{1}{n^2} + \frac{\sqrt{k}}{n\sqrt{2\pi}} \\
&\le \frac{\sqrt{k}}{n}
\end{aligned}$$

□

Lemma D.14. *For $y \ge (cn/\ln n)^{1/(1+\delta)}$,*

$$\left(1-(\beta y)^{-(1+\delta)}\right)^{\alpha n} \approxeq \left(1-y^{-(1+\delta)}\right)^{cn},$$

Proof. We know that $1-(\beta y)^{-(1+\delta)} \geq (1-y^{-(1+\delta)})^{\beta^{-(1+\delta)}}$ from the Taylor expansion, giving us

$$\left(1-(\beta y)^{-(1+\delta)}\right)^{\alpha n} \geq \left((1-y^{-(1+\delta)})^{\beta^{-(1+\delta)}}\right)^{\alpha n} = (1-y^{-(1+\delta)})^{cn}.$$

On the other hand, for $y \geq (cn/\ln n)^{1/(1+\delta)}$,

$$\begin{aligned}
\left(1-(\beta y)^{-(1+\delta)}\right)^{\alpha n} &\leq \exp\left(-cy^{-(1+\delta)}n\right) \\
&\leq \frac{\left(1-y^{-(1+\delta)}\right)^{cn}}{1-cny^{-2(1+\delta)}} \\
&\leq \frac{\left(1-y^{-(1+\delta)}\right)^{cn}}{1-\frac{(\ln n)^2}{cn}} \\
&\approxeq \left(1-y^{-(1+\delta)}\right)^{cn}.
\end{aligned}$$

□

Lemma D.15. *For $k \leq ((1-c^2)\ln n)/2$,*

$$F_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\left(\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}\right) \leq \frac{\sqrt{k}}{n},$$

Proof. We begin with

$$\begin{aligned}
&F_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\left(\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}\right) \\
&= \sum_{\ell=0}^{j}\binom{\beta^{1+\delta}n(1+c)}{\ell}\left(1-\frac{\ln n}{\alpha n}\right)^{\beta^{1+\delta}n(1+c)-\ell}\left(\frac{\ln n}{\alpha n}\right)^{\ell} \\
&\leq \sum_{\ell=0}^{j}\frac{(\beta^{1+\delta}n(1+c))^{\ell}}{\ell!}\exp\left(-\left(\frac{\ln n}{\alpha n}\right)(\beta^{1+\delta}n(1+c)-\ell)\right)\left(\frac{\ln n}{\alpha n}\right)^{\ell} \\
&= \sum_{\ell=0}^{j}\frac{1}{\ell!}\left(\frac{\beta^{1+\delta}(1+c)\ln n}{c}\right)^{\ell}\exp\left(-\ln n\left(1+c^{-1}\left(1-\frac{\ell}{n}\right)\right)\right) \\
&= \frac{1}{n}\sum_{\ell=0}^{j}\frac{1}{\ell!}\left(\beta^{1+\delta}(1+c^{-1})\ln n\right)^{\ell}\left(\frac{1}{n}\right)^{c^{-1}\left(1-\frac{\ell}{n}\right)} \\
&\leq \frac{1}{n^2}+\frac{1}{n}\sum_{\ell=1}^{j}\frac{1}{\sqrt{2\pi\ell}}\left(\frac{e\beta^{1+\delta}(1+c^{-1})\ln n}{\ell}\right)^{\ell}\left(\frac{1}{n}\right)^{c^{-1}\left(1-\frac{\ell}{n}\right)} \qquad \text{(D.10)}
\end{aligned}$$

We apply a similar argument as in Lemma D.13, showing that for $\ell \leq ((1 - c^2)/2) \ln n$,

$$\begin{aligned}
\left(\frac{e\beta^{1+\delta}(1+c^{-1})\ln n}{\ell}\right)^{\ell} &\leq \left(\frac{ec^{-1}(1+c^{-1})\ln n}{(1-c)\ln n}\right)^{((1-c^2)/2)\ln n} \\
&= \left(\frac{ec^{-1}(1+c^{-1})}{1-c}\right)^{(1-c)\ln n} \\
&= \exp\left(1 + \ln c^{-1} + \ln(1+c^{-1}) - \ln(1-c)\right)^{((1-c^2)/2)\ln n} \\
&= \exp\left(\ln n\right)^{(1+\ln c^{-1}+\ln(1+c^{-1})-\ln(1-c))((1-c^2)/2)} \\
&= n^{(1+\ln c^{-1}+\ln(1+c^{-1})-\ln(1-c))((1-c^2)/2)} \\
&\leq n^{(1+(c^{-1}-1)+c^{-1}+c+c^2)((1-c^2)/2)} \\
&= n^{(2c^{-1}+c+c^2)((1-c^2)/2)} \\
&\leq n^{(2c^{-1}+2c)((1-c^2)/2)} && (c \leq 1) \\
&= n^{c^{-1}(1+c^2)(1-c^2)} \\
&= n^{c^{-1}(1-c^4)} \\
&\leq n^{c^{-1}(1-\ell/n)}
\end{aligned}$$

for sufficiently large n. This gives us

$$F_{(\beta^{1+\delta}n(1+c)-j:\beta^{1+\delta}n(1+c))}\left(\left(\frac{\alpha n}{\ln n}\right)^{1/(1+\delta)}\right) \leq \frac{1}{n^2} + \frac{1}{n}\sum_{\ell=1}^{j}\frac{1}{\sqrt{2\pi\ell}} \leq \frac{\sqrt{k}}{n},$$

□

Lemma D.16. *For $0 < a < b$ and $c > 0$,*

$$\frac{a+c}{b+c} > \frac{a}{b}$$

Proof.

$$\frac{a+c}{b+c} = \frac{a(1+c/a)}{b(1+c/b)} > \frac{a(1+c/b)}{b(1+c/b)} = \frac{a}{b}$$

□

Lemma D.17. *For $0 \le y \le a_1 \cdot \frac{\ln n}{n}$ and $|z| \le a_2 \ln n$,*

$$|(1-y)^z - (1-yz)| = O\left(\frac{1}{n}\right)$$

Proof. By Taylor's theorem,

$$f(y) = (1-y)^z = 1 - yz \pm \frac{f''(\varepsilon)}{2} y^2$$

for some $0 \le \varepsilon \le y$. Note that

$$f''(\varepsilon) = z(z-1)(1-\varepsilon)^{z-2} \le |z(z-1)| \exp(-\varepsilon(z-2)) \le |z(z-1)| \exp(\varepsilon|z-2|).$$

Since $\varepsilon \le y \le a_1 \cdot \frac{\ln n}{n}$ and $|z| \le a_2 \ln n$,

$$|z(z-1)| \exp(\varepsilon|z-2|) \le a_2^2 (\ln n)^2 n^{-2} n^{a_1|z-2|/n}.$$

This gives us

$$\frac{f''(\varepsilon)}{2} y^2 \le \frac{a_2^2 (\ln n)^2 n^{-2} n^{a_1|z-2|/n}}{2} a_1^2 (\ln n)^2 n^{-2} \le a_1^2 a_2^2 (\ln n)^4 n^{-(2-a_1(a_2 \ln n+2)/n)}.$$

Using $\ln n = n^{\ln \ln n / \ln n}$, this is

$$\frac{a_1^2 a_2^2}{n} n^{-(1-a_1(a_2 \ln n+2)/n - 4 \ln \ln n/\ln n)}.$$

For sufficiently large n, $a_1(a_2 \ln n + 2)/n + 4 \ln \ln n / \ln n \le 1$, so

$$\frac{a_1^2 a_2^2}{n} n^{-(1-a_1(\ln n+2)/n - 4 \ln \ln n/\ln n)} \le \frac{a_1^2 a_2^2}{n} = O(1/n),$$

which proves the lemma. □

Lemma D.18.

$$F_{(an:an)}\left(b\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) \le n^{-ab^{-(1+\delta)}/c}$$

Proof.

$$\begin{aligned} F_{(an:an)}\left(b\left(\frac{cn}{\ln n}\right)^{1/(1+\delta)}\right) &= \left(1 - b^{-(1+\delta)}\frac{\ln n}{cn}\right)^{an} \\ &\leq \exp\left(-\frac{ab^{-(1+\delta)}}{c}\ln n\right) \\ &= n^{-ab^{-(1+\delta)}/c} \end{aligned}$$

□

Lemma D.19.

$$\mathbb{E}\left[Z_{(Cn-\ln n+1:Cn)}\right] \geq \left(\frac{Cn}{\ln n}\right)^{1/(1+\delta)}$$

for $C \geq 1$ and sufficiently large n.

Proof.

$$\begin{aligned} \mathbb{E}\left[Z_{(Cn-\ln n+1:Cn)}\right] &= \mathbb{E}\left[Z_{(Cn:Cn)}\right] \prod_{j=1}^{\ln n-1}\left(1 - \frac{1}{(1+\delta)j}\right) \\ &= \mathbb{E}\left[Z_{(Cn:Cn)}\right] \prod_{j=1}^{\ln n-1}\left(\frac{(1+\delta)j-1}{(1+\delta)j}\right) \\ &= \mathbb{E}\left[Z_{(Cn:Cn)}\right] \prod_{j=1}^{\ln n-1}\left(\frac{j-1/(1+\delta)}{j}\right) \\ &= \mathbb{E}\left[Z_{(Cn:Cn)}\right] \frac{\Gamma(\ln n - 1/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(\ln n)} \\ &\geq \Gamma\left(\frac{\delta}{1+\delta}\right)(Cn)^{1/(1+\delta)}\frac{\Gamma(\ln n - 1/(1+\delta))}{\Gamma(\delta/(1+\delta))\Gamma(\ln n)} \\ &\qquad \text{(by Lemma D.22)} \\ &= (Cn)^{1/(1+\delta)}\frac{\Gamma(\ln n - 1/(1+\delta))}{\Gamma(\ln n)} \\ &\geq \left(\frac{Cn}{\ln n}\right)^{1/(1+\delta)}\left(1 + \frac{\frac{1}{1+\delta}\cdot\frac{1+2\delta}{1+\delta}}{\ln n} - O\left(\frac{1}{(\ln n)^2}\right)\right) \\ &\qquad \text{(by Tricomi and Erdélyi (1951))} \\ &\geq \left(\frac{Cn}{\ln n}\right)^{1/(1+\delta)} \qquad \text{(for sufficiently large } n\text{)} \end{aligned}$$

□

Lemma D.20. *For $k = O(\ln n)$,*

$$\frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \approxeq \frac{1}{(1+c)^k}.$$

Proof.

$$\frac{\binom{n}{k}}{\binom{n(1+c)}{k}} = \frac{n(n-1)\cdots(n-k+1)}{n(1+c)(n(1+c)-1)\cdots(n(1+c)-k+1)}.$$

Each term $(n-j)/(n(1+c)-j)$ is between $1/(1+c)$ and $(1-j/n)/(1+c)$. Therefore, the entire product is at least

$$\prod_{j=0}^{k-1} \frac{1}{1+c}\left(1-\frac{j}{n}\right) = \frac{1}{(1+c)^k}\prod_{j=0}^{k-1}\left(1-\frac{j}{n}\right) \geq \frac{1}{(1+c)^k}\left(1-\frac{k^2}{n}\right)$$

and at most $1/(1+c)^k$. This means that

$$\frac{1}{(1+c)^k} \geq \frac{\binom{n}{k}}{\binom{n(1+c)}{k}} \geq \frac{1}{(1+c)^k}\left(1-\frac{(\ln n)^2}{n}\right) \approxeq \frac{1}{(1+c)^k}$$

□

Lemma D.21. *For $0 < z < 1$, and $y \geq 0$,*

$$(1+y)^z < 1 + yz.$$

Proof. Let $w = z^{-1}$. Then, the lemma is true if and only if for $w > 1$,

$$1 + y < \left(1 + \frac{y}{w}\right)^w.$$

Note that for $w = 1$, we have equality. We will show that the function

$$f(w) = \left(1 + \frac{y}{w}\right)^w$$

has nonnegative derivative for $w \geq 1$. This is equivalent to showing the same for its log, which is

$$\begin{aligned}\frac{d}{dw}\log f(w) &= \frac{d}{dw} w \log\left(1+\frac{y}{w}\right) \\ &= \log\left(1+\frac{y}{w}\right) + \frac{w}{1+\frac{y}{w}} \cdot \left(-\frac{y}{w^2}\right) \\ &= \log\left(1+\frac{y}{w}\right) - \frac{\frac{y}{w}}{1+\frac{y}{w}}\end{aligned}$$

Let $x = 1 + \frac{y}{w}$. Then, the lemma is true if for $x > 1$,

$$\log(x) - \frac{x-1}{x} > 0$$

$$x\log(x) > x - 1$$

Both are 0 at $x = 1$, but the left hand side has derivative $1 + \log(x)$ while the right hand side has derivative 1, so left hand side will be strictly larger than the right hand side for $x > 1$. □

Lemma D.22.

$$\mathbb{E}\left[Z_{(m:m)}\right] \approxeq \Gamma\left(\frac{\delta}{1+\delta}\right) m^{1/(1+\delta)}.$$

Also,

$$\mathbb{E}\left[Z_{(m:m)}\right] \geq \Gamma\left(\frac{\delta}{1+\delta}\right) m^{1/(1+\delta)}.$$

Proof. From Malik (1966), we have

$$\mathbb{E}\left[Z_{(m:m)}\right] = \frac{\Gamma(m+1)\Gamma\left(1-\frac{1}{1+\delta}\right)}{\Gamma\left(m+\frac{\delta}{1+\delta}\right)}.$$

By Tricomi and Erdélyi (1951),

$$\frac{\Gamma(m+1)}{\Gamma\left(m+\frac{\delta}{1+\delta}\right)} = m^{1/(1+\delta)}\left(1 + \frac{\left(\frac{1}{1+\delta}\right)\left(\frac{\delta}{1+\delta}\right)}{2m} + O\left(\frac{1}{m^2}\right)\right) \geq m^{1/(1+\delta)}.$$

This means

$$\Gamma\left(\frac{\delta}{1+\delta}\right) m^{1/(1+\delta)} \leq \mathbb{E}\left[Z_{(m:m)}\right] \leq \Gamma\left(\frac{\delta}{1+\delta}\right) m^{1/(1+\delta)} \left(1 + O\left(\frac{1}{n}\right)\right),$$

so

$$\Gamma\left(\frac{\delta}{1+\delta}\right) m^{1/(1+\delta)} \approxeq \mathbb{E}\left[Z_{(m:m)}\right].$$

□

Lemma D.23 ((Malik, 1966), Formula 1)**.**

$$\mathbb{E}\left[Z_{(m-k:m)}\right] = \left(1 - \frac{1}{k(1+\delta)}\right) \mathbb{E}\left[Z_{(m-k+1:m)}\right]$$

D.5 Lemmas and Proofs for Section 7.3

Proof of Theorem 7.12. To proceed, we need some notation. Let L be the event that $X_{(n-1:n)} \geq T \cap Y_{(n-1:n)} \geq T$ (the samples are "large"). Let G be the event that $b(X_{(n:n)}) < Y_{(n-1:n)}$, meaning G is the event that the policy has an effect. Let D be the random variable $X_{(n:n)} - Y_{(n-1:n)}$. We want to show that $\mathbb{E}[D \mid G] > 0$. To do so, we observe that by Lemma D.24, is sufficient to show that $\mathbb{E}[D \mid L] > \frac{\Pr[\overline{L}]}{\Pr[L]}$. By Lemma D.25, we know that $\Pr\left[\overline{L}\right] \leq 2nF(T)^{n-1}$. To complete the proof, we need to show that $\mathbb{E}[D \mid L]$ is large, which we do via Lemma D.26.

Since $\Pr[L] \geq 1 - 2nF(T)^{n-1}$, there exists N_1 such that for all $n \geq N_1$, $\Pr[L] \geq$

$1/2$. Using Lemma D.26, if $n \geq N_1$, it is sufficient to have

$$\begin{aligned}
\mathbb{E}\left[D \mid L\right] &> \frac{\Pr\left[\overline{L}\right]}{\frac{1}{2}} \\
K(F(T)+\eta)^{n-1} &> 4nF(T)^{n-1} \\
\left(1+\frac{\eta}{F(T)}\right)^{n} &> \frac{4n}{K} \\
n\log\left(1+\frac{\eta}{F(T)}\right) &> \log n + \log\left(\frac{4}{K}\right) \\
\sqrt{n}\log\left(1+\frac{\eta}{F(T)}\right) &> 2 \qquad (n \geq 4/K, \text{ using } \sqrt{n} > \log n) \\
n &> 4\left(\log\left(1+\frac{\eta}{F(T)}\right)\right)^{-2} = N_2
\end{aligned}$$

Thus, for $n > \max\{N_1, N_2, 4/K\}$, $\mathbb{E}\left[D \mid L\right] > \frac{\Pr\left[\overline{L}\right]}{\Pr[L]}$, which by Lemma D.26 implies that $\mathbb{E}\left[D \mid G\right] > 0$. This completes the proof of Theorem 7.12. □

Lemma D.24. *If $L \Rightarrow G$ and $D \geq -1$, then $\mathbb{E}\left[D \mid L\right] > \frac{\Pr\left[\overline{L}\right]}{\Pr[L]}$ implies $\mathbb{E}\left[D \mid G\right] > 0$.*

Proof.

$$\begin{aligned}
\mathbb{E}\left[D \mid G\right] &= \mathbb{E}\left[D \cdot \mathbf{1}_{\{L\}} \mid G\right] + \mathbb{E}\left[D \cdot \mathbf{1}_{\{\overline{L}\}} \mid G\right] \\
&= \frac{\mathbb{E}\left[D \cdot \mathbf{1}_{\{L\}} \cdot \mathbf{1}_{\{G\}}\right] + \mathbb{E}\left[D \cdot \mathbf{1}_{\{\overline{L}\}} \cdot \mathbf{1}_{\{G\}}\right]}{\Pr\left[G\right]} \\
&= \frac{\mathbb{E}\left[D \cdot \mathbf{1}_{\{L\}}\right] + \mathbb{E}\left[D \cdot \mathbf{1}_{\{\overline{L}\}} \cdot \mathbf{1}_{\{G\}}\right]}{\Pr\left[G\right]} && (L \Rightarrow G) \\
&\geq \frac{\mathbb{E}\left[D \cdot \mathbf{1}_{\{L\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{\overline{L}\}} \cdot \mathbf{1}_{\{G\}}\right]}{\Pr\left[G\right]} && (D \geq -1) \\
&\geq \frac{\mathbb{E}\left[D \cdot \mathbf{1}_{\{L\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{\overline{L}\}}\right]}{\Pr\left[G\right]} && (\mathbf{1}_{\{G\}} \leq 1) \\
&= \frac{\mathbb{E}\left[D \mid L\right]\Pr\left[L\right] - \Pr\left[\overline{L}\right]}{\Pr\left[G\right]}
\end{aligned}$$

$$\frac{\mathbb{E}\left[D \mid L\right]\Pr\left[L\right] - \Pr\left[\overline{L}\right]}{\Pr\left[G\right]} > 0$$
$$\iff \mathbb{E}\left[D \mid L\right]\Pr\left[L\right] - \Pr\left[\overline{L}\right] > 0$$
$$\iff \mathbb{E}\left[D \mid L\right] > \frac{\Pr\left[\overline{L}\right]}{\Pr\left[L\right]}$$

□

Lemma D.25. *For $X_{(n-1:n)}$, $Y_{(n-1:n)}$ order statistics from a distribution with support on $[0,1]$,*

$$\Pr\left[X_{(n-1:n)} \geq T \cap Y_{(n-1:n)} \geq T\right] \leq 2nF(T)^{n-1}.$$

Proof.

$$\begin{aligned}
\Pr\left[X_{(n-1:n)} \geq T \cap Y_{(n-1:n)} \geq T\right] &= \Pr\left[X_{(n-1:n)} \geq T\right]\Pr\left[Y_{(n-1:n)} \geq T\right] \\
&= (1 - F_{(n-1)}(T))^2 \\
&= (1 - nF(T)^{n-1}(1 - F(T)) - F(T)^n)^2 \\
&= (1 - nF(T)^{n-1} + (n-1)F(T)^n)^2 \\
&\geq (1 - nF(T)^{n-1})^2 \\
&\geq 1 - 2nF(T)^{n-1}
\end{aligned}$$

$$\Pr\left[\overline{L}\right] = 1 - \Pr\left[L\right] \leq 2nF(T)^{n-1}$$

□

Lemma D.26. *There exist constants $\eta > 0$ and $K > 0$ such that $\mathbb{E}\left[D \mid L\right] \geq K(F(T) + \eta)^{n-1}$*

Proof. First, let f_Z and F_Z be the pdf and cdf respectively of $Y \mid Y \geq T$, i.e.

$F_Z(x) = \frac{F(x)-F(T)}{1-F(T)}$ and $f_Z = F'_Z$. Note that

$$\begin{aligned}
\mathbb{E}\left[D \mid L\right] &= \mathbb{E}\left[X_{(n:n)} - Y_{(n-1:n)} \mid L\right] \\
&= \mathbb{E}\left[X_{(n:n)} \mid X_{(n-1:n)} \geq T\right] - \mathbb{E}\left[Y_{(n-1:n)} \mid Y_{(n-1:n)} \geq T\right] \\
&= \mathbb{E}\left[Y_{(n:n)} \mid Y_{(n-1:n)} \geq T\right] - \mathbb{E}\left[Y_{(n-1:n)} \mid Y_{(n-1:n)} \geq T\right] \\
&= \mathbb{E}\left[Y_{(n:n)} - Y_{(n-1:n)} \mid Y_{(n-1:n)} \geq T\right]
\end{aligned}$$

Let M be a random variable corresponding to the number of samples from $Y_1, \ldots, Y_n$ that are larger than T. We can rewrite this as

$$\begin{aligned}
&\mathbb{E}\left[D \mid L\right] \\
&= \sum_{m=2}^{M} \mathbb{E}\left[Y_{(n:n)} - Y_{(n-1:n)} \mid Y_{(n-1:n)} \geq T, M = m\right] \Pr\left[M = m \mid Y_{(n-1:n)} \geq T\right] \\
&= \sum_{m=2}^{M} \mathbb{E}\left[Y_{(n:n)} - Y_{(n-1:n)} \mid M = m\right] \Pr\left[M = m \mid Y_{(n-1:n)} \geq T\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (M \geq 2 \Longrightarrow Y_{(n-1:n)} \geq T)
\end{aligned}$$

Conditioning on $M = m$, $Y_{(n:n)}$ and $Y_{(n-1:n)}$ have the same distributions as $Z_{(m:m)}$ and $Z_{(m-1:m)}$ respectively, where $Z_{(k:m)}$ is the kth order statistic of random variables $Z_1, Z_2, \ldots, Z_m$ drawn from the distribution with cdf F_Z. We will use $F_{Z,(k:m)}$ to denote the cdf of $Z_{(k:m)}$. Thus, $\mathbb{E}\left[Y_{(n:n)} - Y_{(n-1:n)} \mid M = m\right] = \mathbb{E}\left[Z_{(m:m)} - Z_{(m-1:m)}\right]$. Using an analysis similar to that of Lopez and Marengo (2011),

$$\begin{aligned}
\mathbb{E}\left[Z_{(m:m)} - Z_{(m-1:m)}\right] &= \int_T^1 (1 - F_{Z,(m:m)}(x)) - (1 - F_{Z,(m-1:m)}(x))\, dx \\
&= \int_T^1 F_{Z,(m-1:m)} - F_{Z,(m:m)}(x)\, dx \\
&= \int_T^1 \binom{m}{m-1} F_Z(x)^{m-1}(1 - F_Z(x))\, dx \\
&\geq \int_T^1 F_Z(x)^{m-1}(1 - F_Z(x))\, dx
\end{aligned}$$

Choose $\eta \in (0, 1 - F(T))$ and $\eta' \in (\eta, 1 - F(T))$. Let $r = F_Z^{-1}(F(T) + \eta)$ and $r' = F_Z^{-1}(F(T) + \eta')$. Note that $T < r < r' < 1$ because otherwise F_Z would have infinite slope at r or r', which is impossible because f_Z is continuous over a compact set and therefore has a finite maximum. Moreover, it must be the case that $F(T) < 1$ because by assumption, $\sup_{x:f(x)>0} = 1$. If $F(T)$ were 1, this would imply that $\sup_{x:f(x)>0} = T < 1$, which is a contradiction.

$$
\begin{aligned}
&\int_T^1 F_Z(x)^{m-1}(1 - F_Z(x))\,dx \\
&\quad \geq \int_r^1 F_Z(x)^{m-1}(1 - F_Z(x)) \\
&\quad \geq \int_r^1 F_Z(r)^{m-1}(1 - F_Z(x)) \\
&\quad = (F(T) + \eta)^{m-1} \int_r^1 1 - F_Z(x)\,dx \\
&\quad \geq (F(T) + \eta)^{n-1} \int_r^1 1 - F_Z(x)\,dx \\
&\quad \geq (F(T) + \eta)^{n-1} \int_r^{r'} 1 - F_Z(x)\,dx \\
&\quad \geq (F(T) + \eta)^{n-1} \int_r^{r'} 1 - F_Z(r')\,dx \qquad (F_Z(x) \leq F_Z(r') \text{ for } x \leq r') \\
&\quad = (F(T) + \eta)^{n-1}(r' - r)(1 - (F(T) + \eta')) \\
&\quad = (F(T) + \eta)^{n-1}[F_Z^{-1}(F(T) + \eta') - F_Z^{-1}(F(T) + \eta)](1 - F(T) - \eta') \\
&\quad = K(F(T) + \eta)^{n-1}
\end{aligned}
$$

where $K = [F_Z^{-1}(F(T) + \eta') - F_Z^{-1}(F(T) + \eta)](1 - F(T) - \eta')$. Since this is independent of m, we have

$$
\begin{aligned}
\mathbb{E}[D \mid L] &= \sum_{m=2}^{n} \mathbb{E}\left[Y_{(n:n)} - Y_{(n-1:n)} \mid M = m\right] \Pr\left[M = m \mid Y_{(n-1:n)} \geq T\right] \\
&\geq \sum_{m=2}^{n} K(F(T) + \eta)^{n-1} \Pr\left[M = m \mid Y_{(n-1:n)} \geq T\right] \\
&= K(F(T) + \eta)^{n-1}
\end{aligned}
$$

□

APPENDIX E

HOW DO CLASSIFIERS INDUCE AGENTS TO BEHAVE STRATEGICALLY?

E.1 Characterizing the Agent's Response to a Linear Mechanism.

In this section, we'll characterize how a rational agent best-responds to a linear mechanism. Its utility is $H = \beta^\top F$, and therefore we can rewrite the optimization problem (8.2) with $M(F) = \beta^\top F$, which yields

$$\begin{aligned} \max_{x \in \mathbb{R}^m} \ & \sum_{i=1}^{n} \beta_i f_i([\alpha^\top x]_i) \\ \text{s.t. } & x \geq \mathbf{0} \\ & \sum_{j=1}^{m} x_j \leq B \end{aligned} \tag{E.1}$$

Note that this is a concave maximization since each f_i is weakly concave and $[\alpha^\top x]_i$ is linear in x. The Lagrangian is then

$$\mathcal{L}(x, \lambda) = \sum_{i=1}^{n} \beta_i f_i\left([\alpha^\top x]_i\right) + \lambda_0 \left(B - \sum_{j=1}^{m} x_j \right) + \sum_{j=1}^{m} \lambda_j x_j.$$

By the Karush-Kuhn-Tucker conditions, since (E.1) is convex, a solution x^* is optimal if and only if $\nabla_x \mathcal{L}(x^*, \lambda^*) = \mathbf{0}$, so for each $j \in [m]$,

$$\sum_{i=1}^{n} \alpha_{ji} \beta_i f_i'\left([\alpha^\top x^*]_i\right) - \lambda_0^* + \lambda_j^* = 0.$$

Note that we can write this as

$$\lambda_0^* = \left. \frac{\partial H}{\partial x_j} \right|_{x^*} + \lambda_j^*.$$

By complementary slackness, $\lambda_j^* > 0 \implies x_j^* = 0$. Therefore, it follows that at optimality, the gradients with respect to all nonzero components of the effort profile are λ_0^*. Furthermore, the gradients with respect to all effort components are at most λ_0^* since $\lambda_j^* \geq 0$ by definition. This proves the following lemma.

Lemma E.1. *For any $x \in \mathbb{R}^m$ such that $x \geq \mathbf{0}$, x is an optimal solution to* (E.1) *if and only if the following conditions hold*

1. $\sum_{j=1}^m x_j = B$
2. *For all j, j' such that $x_j > 0$ and $x_{j'} > 0$,*
$$\left.\frac{\partial H}{\partial x_j}\right|_x = \left.\frac{\partial H}{\partial x_{j'}}\right|_x$$
3. *For all j such that $x_j > 0$ and for all j',*
$$\left.\frac{\partial H}{\partial x_j}\right|_x \geq \left.\frac{\partial H}{\partial x_{j'}}\right|_x$$

Proof. Choose $\lambda_0^* = \left.\frac{\partial H}{\partial x_j}\right|_x$ for any j such that $x_j > 0$. Choose $\lambda_j^* = \lambda_0^* - \left.\frac{\partial H}{\partial x_j}\right|_x$ for all j. Then, (x, λ^*) satisfies stationarity (since $\nabla_x \mathcal{L}(x, \lambda^*) = \mathbf{0}$), primal and dual feasibility by definition, and complementary slackness (since $B - \sum_{j=1}^m x_j = 0$). Therefore, x is an optimal solution to (E.1).

To show the other direction, note that $\max_j \frac{\partial H}{\partial x_j} > 0$ because each $f_i(\cdot)$ is strictly increasing and there is some nonzero β_i. Therefore, $\lambda_0 > 0$, and by complementary slackness, every optimal solution must satisfy $\sum_{j=1}^m x_j = B$. $\square$

APPENDIX F

ALGORITHMIC MONOCULTURE AND SOCIAL WELFARE

F.1 Random Utility Models satisfying Definition 9.1

Theorem F.1. *Let f be the pdf of $\mathcal{E}$. The family of RUMs $\mathcal{F}_\theta$ given by ranking $x_i + \frac{\varepsilon_i}{\theta}$ with $\varepsilon_i \sim \mathcal{E}$ satisfies the conditions of Definition 9.1 if:*

- *f is differentiable*
- *f has positive support on $(-\infty, \infty)$*

Proof. We need to show that $\mathcal{F}_\theta$ satisfies the differentiability, asymptotic optimality, and monotonicity conditions in Definition 9.1.

Differentiability: The probability density of any realization of the n noise samples ε_i/θ is $\prod_{i=1}^n f(\varepsilon_i/\theta)$. Let $\varepsilon = [\varepsilon_1/\theta, \ldots, \varepsilon_n/\theta]$ be the vector of noise values and let $M(\pi) \subseteq \mathbb{R}^n$ be the region such that any $\varepsilon \in M(\pi)$ will produces the ranking π. The probability of any permutation π is

$$\Pr_\theta[\pi] = \int_{M(\pi)} \prod_{i=1}^{n} f\left(\frac{\varepsilon_i}{\theta}\right)\, d^n\mathbf{z}.$$

Because f is differentiable,

$$\frac{d}{d\theta} f\left(\frac{x}{\theta}\right) = f'\left(\frac{x}{\theta}\right) \cdot \left(-\frac{x}{\theta^2}\right)$$

Because $\Pr\theta(\pi)$ is an integral of the product of differentiable functions over a fixed region, it is differentiable.

Asymptotic optimality: We will show that for any pair of elements and any $\delta > 0$, there exists sufficiently large θ such that the probability that they incorrectly ranked is at most δ. We will conclude with a union bound over the $n - 1$

pairs of adjacent candidates that there exists sufficiently large θ such that the probability of outputting the correct ranking must be at least $1 - (n-1)\delta$.

Consider two candidates $x_i > x_{i+1}$. Let ν be the difference $x_i - x_{i+1}$. Then, they will be correctly ranked if

$$\frac{\varepsilon_i}{\theta} > -\frac{\nu}{2}$$
$$\frac{\varepsilon_{i+1}}{\theta} < \frac{\nu}{2}$$

Let $\overline{q}$ and $\underline{q}$ be the $1-\frac{\delta}{2}$ and $\frac{\delta}{2}$ quantiles of $\mathcal{E}$ respectively, and let $q = \max(|\overline{q}|, |\underline{q}|)$. For $\theta > \frac{2q}{\nu}$,

$$\begin{aligned}
\Pr\left[\frac{\varepsilon_i}{\theta} < -\frac{\nu}{2}\right] &= \Pr\left[\varepsilon_i < -\frac{\nu\theta}{2}\right] \\
&< \Pr\left[\varepsilon_i < -q\right] \\
&\leq \Pr\left[\varepsilon_i < \underline{q}\right] \\
&= \frac{\delta}{2} \\
\Pr\left[\frac{\varepsilon_{i+1}}{\theta} > \frac{\nu}{2}\right] &= \Pr\left[\varepsilon_{i+1} > \frac{\nu\theta}{2}\right] \\
&< \Pr\left[\varepsilon_{i+1} > q\right] \\
&\leq \Pr\left[\varepsilon_{i+1} > \overline{q}\right] \\
&= \frac{\delta}{2}
\end{aligned}$$

Thus, for sufficiently large θ, the probability that x_i and x_{i+1} are incorrectly ordered is at most δ.

Repeating this analysis for all $n-1$ pairs of adjacent elements, taking the maximum of all the θ's, and taking a union bound yields that the probability of incorrectly ordering any pair of elements is at most $(n-1)\delta$, meaning the probability of outputting the correct ranking is at least $1-(n-1)\delta$. Since δ

is arbitrary, this probability can be made arbitrarily close to 1, satisfying the asymptotic optimality condition.

Monotonicity: The removal of any elements does not alter the distribution of the remaining elements, meaning that the distribution of $\pi^{(-S)}$ is equivalent to a RUM with $n - |S|$ elements. Thus, it suffices to show that for a RUM with positive support on $(-\infty, \infty)$, the probability of ranking the best candidate first strictly increases with θ.

Recall that by definition, the candidates are ranked according to $x_i + \frac{\varepsilon_i}{\theta}$. The probability that x_1 is ranked first is

$$\begin{aligned}
\Pr\left[x_1 + \frac{\varepsilon_1}{\theta} > \max_{2\le i\le n} x_i + \frac{\varepsilon_i}{\theta}\right] &= \Pr\left[\frac{\varepsilon_1}{\theta} > \max_{2\le i\le n} x_i - x_1 + \frac{\varepsilon_i}{\theta}\right] \\
&= \Pr\left[\varepsilon_1 > \max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i\right] \\
&= \mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n} \Pr\left[\varepsilon_1 > \max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i \mid \varepsilon_2, \dots, \varepsilon_n\right]
\end{aligned} \tag{F.1}$$

We want to show that (F.1) is increasing in θ. Intuitively, this is because as θ increases, the right hand side of the inequality inside the probability decreases. To prove this formally, it suffices to show that the subderivative of (F.1) with respect to θ only includes strictly positive numbers. First, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n} \Pr\left[\varepsilon_1 > \max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i \mid \varepsilon_2, \dots, \varepsilon_n\right] \subset \mathbb{R}_{>0} \\
&\quad\iff \frac{\partial}{\partial\theta} \Pr\left[\varepsilon_1 > \max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i \mid \varepsilon_2, \dots, \varepsilon_n\right] \subset \mathbb{R}_{>0}
\end{aligned}$$

Let F and f be the cumulative density function and probability density function of $\mathcal{E}$ respectively. Then,

$$\Pr\left[\varepsilon_1 > \max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i \mid \varepsilon_2, \dots, \varepsilon_n\right] = 1 - F\left(\max_{2\le i\le n} \theta(x_i - x_1) + \varepsilon_i\right)$$

Note that $F(\cdot)$ is strictly increasing (since f is assumed to have positive support on $(-\infty, \infty)$), so it suffices to show that

$$\frac{\partial}{\partial\theta} \max_{2\leq i\leq n} \theta(x_i - x_1) + \varepsilon_i \subset \mathbb{R}_{<0}$$

For any i,

$$\frac{d}{d\theta}\theta(x_i - x_1) + \varepsilon_i = x_i - x_1 < 0.$$

Thus, the subderivative of the max of such functions includes only strictly negative numbers, which completes the proof. □

F.2 3-candidate RUM Counterexamples

F.2.1 Violating Definition 9.2

Here, we provide a noise mode $\mathcal{E}$, accuracy parameter θ, and candidate distribution $\mathcal{D}$ such that $U_{AH} < U_{AA}$.

Choose the noise distribution $\mathcal{E}$ and accuracy parameter θ such that

$$\frac{\varepsilon}{\theta} = \begin{cases} 1 & w.p. \frac{\delta}{2} \\ 0 & w.p. 1-\delta \\ -1 & w.p. \frac{\delta}{2} \end{cases}$$

Note that this distribution does not satisfy Definition 9.1 because it is neither differentiable nor supported on $(-\infty, \infty)$; however, we can provide a "smooth" approximation to this distribution by expressing it as the sum of arbitrarily tightly concentrated Gaussians with the same results.

We choose the candidate distribution $\mathcal{D}$ such that $x_1 - 1 > x_2 > x_3 > x_1 - 2$. For example,

$$x_1 = \frac{7}{4}$$
$$x_2 = \frac{1}{2}$$
$$x_3 = 0$$

Under this condition, assuming $x_3 = 0$ without loss of generality,

$$\begin{aligned} &U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta) \\ &\qquad = \frac{\delta^2}{32}\left(\delta^3 x_1 - 4\delta^2 x_1 + 4\delta x_1 + 2\delta^3 x_2 - 14\delta^2 x_2 + 20\delta x_2 - 8x_2\right) \end{aligned}$$

Notice that the lowest-power δ term is $-\frac{\delta^2 x_2}{4}$. Therefore, for sufficiently small δ, this is negative. For example, plugging in the values given above with $\delta = .1$, $U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta) \approx -0.00076$.

F.2.2 Violating Definition 9.3

Next, we'll give a 3-candidate RUM for which $U_{AH} < U_{HH}$ does not hold in general. Consider the following 3-candidate example.

$$x_1 = 3$$
$$x_2 = 2$$
$$x_3 = 0$$

Choose $\mathcal{E}$ and θ such that

$$\frac{\varepsilon}{\theta} = \begin{cases} 1 & \text{w.p. } \frac{1-\delta}{2} \\ -1 & \text{w.p. } \frac{1-\delta}{2} \\ 10 & \text{w.p. } \frac{\delta}{2} \\ -10 & \text{w.p. } \frac{\delta}{2} \end{cases}$$

Again, while this noise model doesn't satisfy Definition 9.1, we can approximate it arbitrarily closely with the sum of tightly concentrated Gaussians. Let the $\theta_A = 1.1\theta$ and $\theta_H = 0.9\theta$.

We will show that for these parameters, $U_{AH}(\theta_A, \theta_H) > U_{HH}(\theta_A, \theta_H)$, i.e., it is somehow better to choose after a better opponent than after a worse opponent. At a high level, the reasoning for this is as follows:

1. When choosing first, the only difference between the algorithm and the human evaluator is that the algorithm is more likely to choose x_2 than x_3. Both strategies have identical probabilities of selecting x_1.
2. When choosing second, the human evaluator's utility is higher when x_2 has already been chosen than when x_3 has already been chosen. This is because when x_2 is unavailable, the human evaluator is almost guaranteed to get x_1; when x_3 is unavailable, the human evaluator will choose x_2 with probability $\approx 1/4$.

Let τ and π be rankings generated by the algorithm and human evaluator respectively. First, we will show that

$$\Pr[\tau_1 = x_1] = \Pr[\pi_1 = x_1] \tag{F.2}$$

$$\Pr[\tau_1 = x_2] > \Pr[\pi_1 = x_2] \tag{F.3}$$

To do so, consider the realizations of $\varepsilon_1, \varepsilon_2, \varepsilon_3$ that result in different rankings under θ_A and θ_H. In fact, the only set of realizations that result in different rankings are when $\varepsilon_2/\theta = -1$ and $\varepsilon_3/\theta = 1$. Thus, the algorithm and human evaluator always rank x_1 in the same position, conditioned on a realization, which proves (F.2); the only difference is that the algorithm sometimes ranks x_2 above x_3 when the human evaluator does not. Moreover, whenever $\varepsilon_1/\theta = -10$, x_2 is more strictly more likely to be ranked first under the algorithm than the human evaluator, which proves (F.3).

Next, we must show that when choosing second, the human evaluator is better off when x_2 is unavailable than when x_3 is unavailable. This is clearly true because for the human evaluator,

$$\Pr\left[x_1 + \frac{\varepsilon_1\theta_H}{\theta} > x_3 + \frac{\varepsilon_3\theta_H}{\theta}\right] \approx 1 - O(\delta)$$
$$\Pr\left[x_1 + \frac{\varepsilon_1\theta_H}{\theta} > x_2 + \frac{\varepsilon_3\theta_H}{\theta}\right] \approx \frac{3}{4}$$

Thus, conditioned on x_2 being unavailable, the human evaluator gets utility ≈ 3, whereas when x_3 is unavailable, the human evaluator gets utility ≈ 2.75. Let u_{-i} be the expected utility for the human evaluator when x_i is unavailable. Putting

this together, we get

$$
\begin{aligned}
&U_{AH}(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) \\
&\quad = \sum_{i=1}^{3} (\Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i]) u_{-i} \\
&\quad = (\Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1]) u_{-1} + (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2]) u_{-2} \\
&\quad + (\Pr[\tau_1 = x_3] - \Pr[\pi_1 = x_3]) u_{-3} \\
&\quad = (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2]) u_{-2} + (\Pr[\tau_1 = x_3] - \Pr[\pi_1 = x_3]) u_{-3} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\Pr[\tau_1 = x_1 = \Pr[\pi_1 = x_1]) \\
&\quad = (\Pr[\tau_1 = x_2] - \Pr[\pi_1 = x_2])(u_{-2} - u_{-3}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\textstyle\sum_{i=1}^{3} \Pr[\tau_1 = x_i] = \sum_{i=1}^{3} \Pr[\pi_1 = x_i]) \\
&\quad > 0
\end{aligned}
$$

The last step follows from (F.3) and because $u_{-2} > u_{-3}$.

F.3 Proof of Theorem 9.5

F.3.1 Verifying Definition 9.2

By (9.2), we can equivalently show that for any θ, $U_{AH}(\theta, \theta) > U_{AA}(\theta, \theta)$. Let τ and π be the algorithmic and human-generated rankings respectively. Note that they're identically distributed because $\theta_A = \theta_H$. Define

$$
Y \triangleq \begin{cases} \pi_1 & \pi_1 \neq \tau_1 \\ \pi_2 & \text{otherwise} \end{cases}
$$

Note that $U_{AH}(\theta, \theta) = \mathbb{E}[Y]$ and $U_{AA}(\theta, \theta) = \mathbb{E}[\tau_2]$. We want to show that $U_{AH}(\theta, \theta) - U_{AA}(\theta, \theta) = \mathbb{E}[x_Y - x_{\tau_2}] > 0$. It is sufficient to show that for any

k, $\mathbb{E}[Y - \tau_2 \mid \tau_1 = x_k] > 0$. Let $X_i = x_i + \varepsilon_i/\theta$. Note that for distinct i, j, k and $x_i > x_j$,

$$\begin{aligned}
&\mathbb{E}[Y - \tau_2 \mid \tau_1 = x_k] > 0 \\
&\Longleftarrow \frac{\Pr[Y = x_i \mid \tau_1 = x_k]}{\Pr[Y = x_j \mid \tau_1 = x_k]} > \frac{\Pr[\tau_2 = x_i \mid \tau_1 = x_k]}{\Pr[\tau_2 = x_j \mid \tau_1 = x_k]} \\
&\Longleftrightarrow \Pr[Y = x_i \mid \tau_1 = x_k] > \Pr[\tau_2 = x_i \mid \tau_1 = x_k] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(numerator and denominator sum to 1)} \\
&\Longleftrightarrow \Pr[X_i > X_j] > \Pr[X_i > X_j \mid X_k > X_i \cap X_k > X_j] \\
&\Longleftrightarrow \Pr[X_i > X_j] > \mathbb{E}_{X_k}[\Pr[X_i > X_j \mid X_k = a, X_i < a, X_j < a]].
\end{aligned}$$

Thus, it suffices to show that for any a,

$$\Pr[X_i > X_j] > \Pr[X_i > X_j \mid X_i < a, X_j < a]. \tag{F.4}$$

Since $\Pr[X_i > X_j] = \lim_{a \to \infty} \Pr[X_i > X_j \mid X_i < a, X_j < a]$, it suffices to show that for all a,

$$\frac{d}{da} \Pr[X_i > X_j \mid X_i < a, X_j < a] \geq 0, \tag{F.5}$$

and that it is strictly positive for some a. In other words, the higher a is, the more likely i and j are to be correctly ordered. In Theorems F.7 and F.8, we show that (F.5) holds for both Laplacian and Gaussian noise respectively, which proves that RUMs based on both distributions satisfy Definition 9.2.

F.3.2 Verifying Definition 9.3

Next, we show that for both Laplacian and Gaussian distributions, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ for all $\theta_A > \theta_H$. In fact, for 3-candidate RUM families, we will show that this is always true for any *well-ordered* distribution, defined as follows.

Definition F.2. *A noise model with density $f(\cdot)$ is **well-ordered** if for any $a > b$ and $c > d$,*

$$f(a-c)f(b-d) > f(a-d)f(b-c).$$

In other words, for a well-ordered noise model, given two numbers, two candidates are more likely to be correctly ordered than inverted conditioned on realizing those two numbers in some order. Lemma F.4 shows that both Gaussian and Laplacian distributions are well-ordered.

Thus, it suffices to show that for any 3-candidate RUM with a well-ordered noise model, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ when $\theta_A > \theta_H$.

Theorem F.3. *For 3 candidates with unique values $x_1 > x_2 > x_3$ and well-ordered i.i.d. noise with support $(-\infty, \infty)$, if $\theta_A > \theta_H$, then $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$.*

Proof. Define u_{-i} to be the expected utility of the maximum element of the human-generated ranking when i is not available. Because we're in the 3-candidate setting, we have

$$u_{-1} = \lambda_1 x_2 + (1-\lambda_1)x_3$$
$$u_{-2} = \lambda_2 x_1 + (1-\lambda_2)x_3$$
$$u_{-3} = \lambda_3 x_1 + (1-\lambda_3)x_2$$

where $1/2 < \lambda_i < 1$. This is because the noise has support everywhere, so it is impossible to correctly rank any two candidates with probability 1, and any two candidates are more likely than not to be correctly ordered:

$$\Pr\left[\frac{\varepsilon_i}{\theta} - \frac{\varepsilon_j}{\theta} > -\delta\right] = \Pr[\varepsilon_i - \varepsilon_j \geq 0] + \Pr\left[0 > \frac{\varepsilon_i - \varepsilon_j}{\theta} > -\delta\right] > \frac{1}{2}$$

Note that $\lambda_2 > \lambda_1$ and $\lambda_2 > \lambda_3$, since

$$\begin{aligned}\lambda_2 &= \Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_1 - x_3)] \\ &> \max\left\{\Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_2 - x_3)], \Pr[\varepsilon_1 - \varepsilon_3 > -\theta(x_1 - x_2)]\right\} \\ &= \max\{\lambda_1, \lambda_3\}.\end{aligned}$$

Let $\tau \sim \mathcal{F}_{\theta_A}$ and $\pi \sim \mathcal{F}_{\theta_H}$. With this, we can write

$$U_{AH}(\theta_A, \theta_H) = \sum_{i=1}^{3} \Pr[\tau_1 = i] u_{-i}$$
$$U_{HH}(\theta_A, \theta_H) = \sum_{i=1}^{3} \Pr[\pi_1 = i] u_{-i}$$

Define

$$\Delta p_i = \Pr[\tau_1 = i] - \Pr[\pi_1 = i]$$

Using Lemmas F.5, and F.6, we have

$\Delta p_1 > 0$ (By monotonicity of RUM families, see Appendix F.1)

$\Delta p_1 \geq \Delta p_2$

$\Delta p_3 \leq 0$

Also, $\Delta p_1 + \Delta p_2 + \Delta p_3 = 0$. We must show that

$$U_{AH}(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) = \sum_{i=1}^{3} \Delta p_i u_{-i} < 0.$$

We consider 2 cases.

Case 1: $\Delta p_2 \leq 0$.

Then, $\Delta p_1 = -(\Delta p_2 + \Delta p_3)$. This yields

$$\begin{aligned}
&\sum_{i=1}^{3} \Delta p_i u_{-i} \\
&\quad = \Delta p_1 u_{-1} + \Delta p_2 u_{-2} + \Delta p_3 u_{-3} \\
&\quad \leq \Delta p_1 u_{-1} - \Delta p_1 \min(u_{-2}, u_{-3}) \\
&\quad = \Delta p_1 \left(\lambda_1 x_2 + (1-\lambda_1) x_3 - \min\left\{\lambda_2 x_1 + (1-\lambda_2) x_3, \lambda_3 x_1 + (1-\lambda_3) x_2\right\}\right) \\
&\quad \leq \Delta p_1 \left(\lambda_1 x_2 + (1-\lambda_1) x_3 - \min\left\{\lambda_2 x_1 + (1-\lambda_2) x_3, x_2\right\}\right)
\end{aligned}$$

We can show that this is at most 0 regardless of which term attains the minimum. Because $\lambda_2 > \lambda_1$,

$$\begin{aligned}
\lambda_1 x_2 + (1-\lambda_1) x_3 - \lambda_2 x_1 - (1-\lambda_2) x_3 &= \lambda_1 x_2 + x_3 - \lambda_1 x_3 - \lambda_2 x_1 - x_3 + \lambda_2 x_3 \\
&= \lambda_1 x_2 - \lambda_1 x_3 - \lambda_2 x_1 + \lambda_2 x_3 \\
&= \lambda_1 (x_2 - x_3) + \lambda_2 (x_3 - x_1) \\
&< \lambda_1 (x_2 - x_3) + \lambda_1 (x_3 - x_1) \\
&= \lambda_1 (x_2 - x_1) \\
&< 0
\end{aligned}$$

For the second term, we have

$$\lambda_1 x_2 + (1-\lambda_1) x_3 - x_2 = (1-\lambda_1)(x_3 - x_2) < 0.$$

Thus,

$$\sum_{i=1}^{3} \Delta p_i u_{-i} < 0.$$

Case 2: $\Delta p_2 > 0$. Note that $u_{-1} < x_2 < u_{-3}$. Then, using $\Delta p_3 = -(\Delta p_1 + \Delta p_2)$,

$$\begin{aligned}
\sum_{i=1}^{3} \Delta p_i u_{-i} &= \Delta p_1 u_{-1} + \Delta p_2 u_{-2} + \Delta p_3 u_{-3} \\
&= \Delta p_1 (u_{-1} - u_{-3}) + \Delta p_2 (u_{-2} - u_{-3}) \\
&\leq \Delta p_2 (u_{-1} - u_{-3}) + \Delta p_2 (u_{-2} - u_{-3}) \quad (\Delta p_1 \geq \Delta p_2 \text{ and } u_{-1} < u_{-3}) \\
&= \Delta p_2 (u_{-1} + u_{-2} - 2u_{-3}) \\
&\leq \Delta p_2 (x_2 + x_1 - 2(\lambda_3 x_1 + (1 - \lambda_3) x_2)) \\
&< \Delta p_2 \left(x_2 + x_1 - 2 \left(\frac{1}{2} x_1 + \frac{1}{2} x_2 \right) \right) \qquad (\lambda_3 > \tfrac{1}{2}) \\
&= 0
\end{aligned}$$

Thus, $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$. □

F.3.3 Supplementary Lemmas for Random Utility Models

Lemma F.4. *Both Gaussian and Laplacian distributions are well-ordered.*

Proof. The Gaussian noise model is well-ordered:

$$\begin{aligned}
f(a-c)f(b-d) &= \frac{1}{2\sigma^2\pi} \exp(-(a-c)^2 - (b-d)^2) \\
&= \frac{1}{2\sigma^2\pi} \exp(-(a-d)^2 - (b-c)^2 - 2(ac + bd - ad - bc)) \\
&= f(a-d)f(b-c) \exp(-2((a-b)(c-d))) \\
&< f(a-d)f(b-c)
\end{aligned}$$

Laplacian noise is as well:

$$f(a-c)f(b-d) = \frac{1}{4}\exp(-|a-c| - |b-d|)$$
$$f(a-d)f(b-c) = \frac{1}{4}\exp(-|a-d| - |b-c|)$$

It suffices to show that for $a > b$ and $c > d$, $|a-c| + |b-d| < |a-d| + |b-c|$. To show this, plot (a, b) and (c, d) in the (x, y) plane. Note that they're both below the $y = x$ line, and that the ℓ_1 distance between them is $|a-c|+|b-d|$. Moreover, the ℓ_1 distance between any two points must be realized by some Manhattan path, which is a combination of horizontal and vertical line segments. Consider the point (b, a), which is above the $y = x$ line. Any Manhattan path from (b, a) to (c, d) must cross the $y = x$ line at some point (w, w). Since (b, a) and (a, b) are equidistant from (w, w), for any Manhattan path from (b, a) to (c, d), there exists a Manhattan path from (a, b) to (c, d) passing through (w, w) of the same length, meaning the ℓ_1 distance from (a, b) to (c, d) is smaller than the ℓ_1 distance from (b, a) to (c, d). As a result, $|a - c| + |b - d| < |a - d| + |b - c|$. □

Next, we show a few basic facts. Let $f_A(r)$ be the density function of the joint realization $R = [X_1, \ldots, X_n] = [x_1 + \varepsilon_1/\theta_A, \ldots, x_n + \varepsilon_n/\theta_A]$ under the algorithmic ranking and $f_H(r)$ be the similarly defined density function under the human-generated ranking. Consider the "contraction" operation $r' = \mathrm{cont}(r)$ such that $r'_i = x_i + (r_i - x_i) \cdot \frac{\theta_H}{\theta_A}$. Essentially, the contraction defines a coupling between $f_A(\cdot)$ and $f_H(\cdot)$, since for $r' = \mathrm{cont}(r)$, $f_A(r')\ dr' = f_H(r)\ dr$. Let $\pi(r)$ be the ranking induced by r. Note that contraction cannot introduce any new inversions in $\pi(r)$—that is, if i is ranked above j in $\pi(r)$ for $i < j$, then i is ranked above j in $\pi(\mathrm{cont}(r))$. Intuitively, this is because contraction pulls values closer to their

means, and can therefore only correct existing inversions, not introduce new ones. This fact will allow us to prove some useful lemmas.

Lemma F.5. *If F_θ is a RUM family satisfying Definition 9.1, then for $\tau \sim \mathcal{F}_{\theta_A}$ and $\pi \sim \mathcal{F}_{\theta_H}$,*

$$\Pr[\tau_1 = x_n] \leq \Pr[\pi_1 = x_n]$$

Proof. Consider any realization r. Because inversions can only be corrected, not generated, by contraction, if $\pi_1(r') = n$, then $\pi_1(r) = n$ where $r' = \mathrm{cont}(r)$. Since r' and r have equal measure under f_A and f_H respectively, we have

$$\begin{aligned}
\Pr[\pi_1 = x_n] &= \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{\pi_1(r)=x_n\}}\, dr \\
&= \int_{\mathbb{R}^n} f_A(\mathrm{cont}(r))\mathbf{1}_{\{\pi_1(r)=x_n\}}\, d\,\mathrm{cont}(r) \\
&\geq \int_{\mathbb{R}^n} f_A(\mathrm{cont}(r))\mathbf{1}_{\{\pi_1(\mathrm{cont}(r))=x_n\}}\, d\,\mathrm{cont}(r) \\
&= \int_{\mathbb{R}^n} f_A(r)\mathbf{1}_{\{\pi_1(r)=x_n\}}\, dr \\
&= \Pr[\tau_1 = x_n]
\end{aligned}$$

□

Next, we prove the following result for well-ordered noise models.

Lemma F.6. *For any $i > 1$, if the noise model $\mathcal{E}$ is well-ordered, for $\theta_A \geq \theta_H$, $\tau \sim \mathcal{F}_{\theta_A}$, and $\pi \sim \mathcal{F}_{\theta_H}$,*

$$\Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1] \geq \Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i]$$

Proof. For $j \neq i$, let $S_{j\to i} \subseteq \mathbb{R}^n$ be the set of realizations r such that $\pi_1(r) = x_j$ and $\pi_1(\text{cont}(r)) = x_i$. Note that $S_{j\to i} = \emptyset$ for $j < i$ because contraction cannot create inversions. Then, we have that

$$\begin{aligned}\Pr[\tau_1 = x_i] - \Pr[\pi_1 = x_i] &= \sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{r\in S_{j\to i}\}}\, dr - \sum_{j<i} \int_{R^n} f_H(r)\mathbf{1}_{\{r\in S_{i\to j}\}}\, dr \\ &\leq \sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{r\in S_{j\to i}\}}\, dr\end{aligned}$$

Define

$$\text{swap}_i(r) = r',$$

where

$$r'_j = \begin{cases} r_j & j \notin \{1, i\} \\ r_1 & j = i \\ r_i & j = 1 \end{cases}$$

Intuitively, the swap_i operation simply swaps the realizations in positions 1 and i. Note that this is a bijection. Also, if $r \in S_{j\to i}$, then $\text{swap}_i(r) \in S_{j\to 1}$, since

$$\text{cont}(\text{swap}_i(r))_1 \geq \text{cont}(r)_i \geq \max_j \text{cont}(r)_j \geq \max_{j\notin\{1,i\}} \text{cont}(\text{swap}_i(r))_j$$

$$\text{cont}(\text{swap}_i(r))_1 \geq \text{cont}(r)_i \geq \text{cont}(r)_1 \geq \text{cont}(\text{swap}_i(r))_i$$

Furthermore for $r \in S_{j\to i}$, $f_H(r) \leq f_H(\text{swap}_i(r))$ since

$$\frac{f_H(\text{swap}_i(r))}{f_H(r)} = \frac{f(r_i - x_1)f(r_1 - x_i)}{f(r_1 - x_1)f(r_i - x_i)} \geq 1$$

because the noise is well-ordered and $r \in S_{j\to i}$ implies $r_i > r_1$. Thus,

$$\begin{aligned}
\sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{r\in S_{j\to i}\}}\, dr &\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(\mathrm{swap}_i(r))\mathbf{1}_{\{r\in S_{j\to i}\}}\, dr \\
&\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(\mathrm{swap}_i(r))\mathbf{1}_{\{\mathrm{swap}_i(r)\in S_{j\to 1}\}}\, dr \\
&\le \sum_{j>i} \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{r\in S_{j\to 1}\}}\, dr \\
&\le \sum_{j>1} \int_{\mathbb{R}^n} f_H(r)\mathbf{1}_{\{r\in S_{j\to 1}\}}\, dr \\
&= \Pr[\tau_1 = x_1] - \Pr[\pi_1 = x_1]
\end{aligned}$$

□

Finally, we show that (F.5) holds for both Laplacian and Gaussian noise.

Theorem F.7. *For any $a \in \mathbb{R}$ and $X_i = x_i + \sigma\varepsilon_i$ where ε_i is Laplacian with unit variance,*

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] \ge 0.$$

Moreover, it is strictly positive for some a.

Proof. First, we must derive an expression for $\Pr[X_i > X_j \mid X_i < a, X_j < a]$. Recall that the Laplace distribution parameterized by μ and λ has pdf

$$f(x;\mu,\lambda) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|)$$

and cdf

$$F(x;\mu,\lambda) = \begin{cases} \frac{1}{2}\exp\left(-\lambda(\mu - x)\right) & x < \mu \\ 1 - \frac{1}{2}\exp\left(-\lambda(x-\mu)\right) & x \ge \mu \end{cases}$$

Note that x_i and x_j be the respective means of X_i and X_j, with $x_i > x_j$. Because the Laplace distribution is piecewise defined, we must consider 3 cases and

show that in all 3 cases, (F.5) holds. Note that

$$\Pr[X_i > X_j \mid X_i < a, X_j < a] = \frac{\int_{-\infty}^{a} f(x; x_i, \lambda) F(x; x_j, \lambda)\, dx}{F(a; x_i, \lambda) F(a; x_j, \lambda)} \tag{F.6}$$

Case 1: $a \leq x_j$.

Then, the numerator of (F.6) is

$$\begin{aligned}
&\int_{-\infty}^{a} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \cdot \frac{1}{2} \exp(-\lambda(x_j - x))\, dx \\
&\quad = \frac{\lambda}{4} \int_{-\infty}^{a} \exp(-\lambda(x_i + x_j - 2x))\, dx \\
&\quad = \frac{\lambda \exp(-\lambda(x_i + x_j))}{4} \int_{-\infty}^{a} \exp(2\lambda x)\, dx \\
&\quad = \frac{\lambda \exp(-\lambda(x_i + x_j))}{4} \frac{1}{2\lambda} \exp(2\lambda a) \\
&\quad = \frac{\exp(-\lambda(x_i + x_j - 2a))}{8}
\end{aligned}$$

The denominator is

$$\frac{1}{2} \exp(-\lambda(x_i - a)) \cdot \frac{1}{2} \exp(-\lambda(x_j - a)) = \frac{1}{4} \exp(-\lambda(x_i + x_j - 2a)).$$

Thus,

$$\Pr[X_i > X_j \mid X_i < a, X_j < a] = \frac{1}{2},$$

so its derivative is trivially nonnegative.

Case 2: $x_j < a \leq x_i$.

Then, the numerator of (F.6) is

$$\begin{aligned}
&\int_{-\infty}^{x_j} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \cdot \frac{1}{2} \exp(-\lambda(x_j - x))\, dx \\
&\quad + \int_{x_j}^{a} \frac{\lambda}{2} \exp\left(-\lambda(x_i - x)\right) \left(1 - \frac{1}{2} \exp(-\lambda(x - x_j))\right)\, dx \\
&\quad = \frac{\exp(-\lambda(x_i - x_j))}{8} + \frac{\lambda}{2} \int_{x_j}^{a} \exp(-\lambda(x_i - x))\, dx \\
&\quad - \frac{\lambda}{4} \int_{x_j}^{a} \exp(-\lambda(x_i - x_j))\, dx \\
&\quad = \frac{\exp(-\lambda(x_i - x_j))}{8} + \frac{\lambda}{2} \frac{1}{\lambda} (\exp(-\lambda(x_i - a)) \\
&\quad - \exp(-\lambda(x_i - x_j))) - \frac{\lambda}{4} (a - x_j) \exp(-\lambda(x_i - x_j)) \\
&\quad = \frac{1}{2} \exp(-\lambda(x_i - a)) - \left(\frac{3}{8} + \frac{\lambda}{4}(a - x_j)\right) \exp(-\lambda(x_i - x_j))
\end{aligned}$$

The denominator is

$$\begin{aligned}
&\left(1 - \frac{1}{2} \exp(-\lambda(a - x_j))\right) \cdot \frac{1}{2} \exp(\lambda(x_j - a)) \\
&= \frac{1}{2} \exp(-\lambda(x_i - a)) - \frac{1}{4} \exp(-\lambda(x_i - x_j))
\end{aligned}$$

We can factor out $\frac{1}{4} \exp(-\lambda(x_i - x_j))$ from both, so

$$\begin{aligned}
\Pr[X_i > X_j \mid X_i < a, X_j < a] &= \frac{2 \exp(\lambda(a - x_j)) - \left(\frac{3}{2} + \lambda(a - x_j)\right)}{2 \exp(\lambda(a - x_j)) - 1} \\
&= \frac{2 \exp(\lambda(a - x_j)) - 1 - \left(\frac{1}{2} + \lambda(a - x_j)\right)}{2 \exp(\lambda(a - x_j)) - 1} \\
&= 1 - \frac{\frac{1}{2} + \lambda(a - x_j)}{2 \exp(\lambda(a - x_j)) - 1}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] > 0 \\
\iff &\frac{d}{da}\frac{\frac{1}{2} + \lambda(a - x_j)}{2\exp(\lambda(a - x_j)) - 1} < 0 \\
\iff &(2\exp(\lambda(a - x_j)) - 1)\lambda < \left(\frac{1}{2} + \lambda(a - x_j)\right) 2\lambda\exp(\lambda(a - x_j)) \\
\iff &2 - \exp(-\lambda(a - x_j)) < 2\left(\frac{1}{2} + \lambda(a - x_j)\right) \\
\iff &1 - \exp(-\lambda(a - x_j)) < 2\lambda(a - x_j) \\
\iff &\exp(-\lambda(a - x_j)) > 1 - 2\lambda(a - x_j)
\end{aligned}$$

This is true because $\lambda(a - x_j) > 0$, and for $z > 0$,

$$\exp(-z) > 1 - z > 1 - 2z.$$

Case 3: $a > x_i$.

Then, the numerator of (F.6) is

$$\begin{aligned}
&\int_{-\infty}^{x_j} \frac{\lambda}{2}\exp\left(-\lambda(x_i - x)\right) \cdot \frac{1}{2}\exp(-\lambda(x_j - x))\, dx \\
&\quad + \int_{x_j}^{x_i} \frac{\lambda}{2}\exp\left(-\lambda(x_i - x)\right)\left(1 - \frac{1}{2}\exp(-\lambda(x - x_j))\right)\, dx \\
&\quad + \int_{x_i}^{a} \frac{\lambda}{2}\exp(-\lambda(x - x_i))\left(1 - \frac{1}{2}\exp(-\lambda(x - x_j))\right)\, dx \\
&\quad = \frac{1}{2} - \left(\frac{3}{8} + \frac{\lambda}{4}(x_i - x_j)\right)\exp(-\lambda(x_i - x_j)) + \frac{1}{2}(1 - \exp(-\lambda(a - x_i))) \\
&\quad - \frac{\lambda}{4}\int_{x_i}^{a} \exp(-\lambda(2x - x_i - x_j))\, dx \\
&\quad = 1 - \left(\frac{3}{8} + \frac{\lambda}{4}(x_i - x_j)\right)\exp(-\lambda(x_i - x_j)) - \frac{1}{2}\exp(-\lambda(a - x_i)) \\
&\quad + \frac{1}{8}\exp(\lambda(x_i + x_j))(\exp(-2\lambda a) - \exp(-2\lambda x_i)) \\
&\quad = 1 - \left(\frac{1}{2} + \frac{\lambda}{4}(x_i - x_j)\right)\exp(-\lambda(x_i - x_j)) - \frac{1}{2}\exp(-\lambda(a - x_i)) \\
&\quad + \frac{1}{8}\exp(-\lambda(2a - x_i - x_j))
\end{aligned}$$

The denominator is

$$\left(1 - \frac{1}{2}\exp(-\lambda(t - x_i))\right)\left(1 - \frac{1}{2}\exp(-\lambda(t - x_j))\right)$$
$$= 1 - \frac{1}{2}\exp(-\lambda(a - x_i)) - \frac{1}{2}\exp(-\lambda(a - x_j)) + \frac{1}{4}\exp(-\lambda(2a - x_i - x_j))$$

Thus,

$$\Pr[X_i > X_j \mid X_i < a, X_j < a]$$
$$= \frac{1 - \left(\frac{1}{2} + \frac{\lambda}{4}(x_i - x_j)\right)\exp(-\lambda(x_i - x_j)) - \frac{1}{2}\exp(-\lambda(a - x_i)) + \frac{1}{8}\exp(-\lambda(2a - x_i - x_j))}{1 - \frac{1}{2}\exp(-\lambda(a - x_i)) - \frac{1}{2}\exp(-\lambda(a - x_j)) + \frac{1}{4}\exp(-\lambda(2a - x_i - x_j))}$$
$$\propto \frac{8 - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(x_i - x_j)) - 4\exp(-\lambda(a - x_i)) + \exp(-\lambda(2a - x_i - x_j))}{4 - 2\exp(-\lambda(a - x_i)) - 2\exp(-\lambda(a - x_j)) + \exp(-\lambda(2a - x_i - x_j))}$$

We're interested in

$$
\begin{aligned}
&\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] > 0 \\
\iff &(4 - 2\exp(-\lambda(a - x_i)) - 2\exp(-\lambda(a - x_j)) + \exp(-\lambda(2a - x_i - x_j) \\
&\cdot (4\lambda\exp(-\lambda(a - x_i)) - 2\lambda\exp(-\lambda(2a - x_i - x_j))) \\
&> (8 - 4\exp(-\lambda(a - x_i)) + \exp(-\lambda(2a - x_i - x_j)) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(x_i - x_j))) \\
&\cdot (2\lambda\exp(-\lambda(a - x_i)) + 2\lambda\exp(-\lambda(a - x_j)) - 2\lambda\exp(-\lambda(2a - x_i - x_j))) \\
\iff &16\exp(-\lambda(a - x_i)) - 8\exp(-\lambda(2a - x_i - x_j)) - 8\exp(-2\lambda(a - x_i)) \\
&+ 4\exp(-\lambda(3a - 2x_i - x_j)) - 8\exp(-\lambda(2a - x_i - x_j)) \\
&+ 4\exp(-\lambda(3a - x_i - 2x_j)) + 4\exp(-\lambda(3a - 2x_i - x_j)) \\
&- 2\exp(-2\lambda(2a - x_i - x_j)) \\
&> 16\exp(-\lambda(a - x_i)) + 16\exp(-\lambda(a - x_j)) - 16\exp(-\lambda(2a - x_i - x_j)) \\
&- 8\exp(-2\lambda(a - x_i)) - 8\exp(-\lambda(2a - x_i - x_j)) + 8\exp(-\lambda(3a - 2x_i - x_j)) \\
&+ 2\exp(-\lambda(3a - 2x_i - x_j)) + 2\exp(-\lambda(3a - x_i - 2x_j)) \\
&- 2\exp(-2\lambda(2a - x_i - x_j)) - 2(4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) \\
&- 2(4 + 2\lambda(x_i - x_j))\exp(-\lambda(a + x_i - 2x_j)) \\
&+ 2(4 + 2\lambda(x_i - x_j))\exp(-2\lambda(a - x_j)) \\
\iff &\exp(-\lambda(3a - x_i - 2x_j)) \\
&> 8\exp(-\lambda(a - x_j)) - 4\exp(-\lambda(2a - x_i - x_j)) + \exp(-\lambda(3a - 2x_i - x_j)) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a + x_i - 2x_j)) \\
&+ (4 + 2\lambda(x_i - x_j))\exp(-2\lambda(a - x_j))
\end{aligned}
$$

$$
\begin{aligned}
&\iff \exp(-\lambda(2a - x_i - x_j)) \\
&\quad > 8 - 4\exp(-\lambda(a - x_i)) + \exp(-\lambda(2a - 2x_i)) \\
&\quad - (4 + 2\lambda(x_i - x_j)) - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(x_i - x_j)) \\
&\quad + (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) \\
&\iff \exp(-\lambda(2a - x_i - x_j)) - 8 + 4\exp(-\lambda(a - x_i)) - \exp(-2\lambda(a - x_i)) \\
&\quad + (4 + 2\lambda(x_i - x_j))(1 + \exp(-\lambda(x_i - x_j))) \\
&\quad - (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) \\
&\quad > 0
\end{aligned} \tag{F.7}
$$

Note that for any $z \geq 0$, we have

$$
\begin{aligned}
(4 + 2z)(1 + e^{-z}) - 8 \geq 0 &\iff (2 + z)(1 + e^{-z}) \geq 4 \\
&\iff z + 2e^{-z} + ze^{-z} \geq 2
\end{aligned}
$$

For $z = 0$, this holds with equality, and the left hand side is increasing since

$$
\begin{aligned}
\frac{d}{dx} z + 2e^{-z} + ze^{-z} \geq 0 &\iff 1 - 2e^{-z} + e^{-z} - ze^{-z} \geq 0 \\
&\iff 1 \geq e^{-z} + ze^{-z} \\
&\iff \frac{1}{1 + z} \geq e^{-z} \\
&\iff 1 + z \leq e^{z}
\end{aligned}
$$

Therefore, choosing $z = \lambda(x_i - x_j)$ and plugging back to (F.7), we have

$$\begin{aligned}
&\exp(-\lambda(2a - x_i - x_j)) - 8 + 4\exp(-\lambda(a - x_i)) - \exp(-2\lambda(a - x_i)) \\
&+ (4 + 2\lambda(x_i - x_j))(1 + \exp(-\lambda(x_i - x_j))) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) > 0 \\
\Longleftarrow\ &\exp(-\lambda(2a - x_i - x_j)) + 4\exp(-\lambda(a - x_i)) - \exp(-2\lambda(a - x_i)) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(a - x_j)) > 0 \\
\Longleftrightarrow\ &\exp(-\lambda(a - x_j)) + 4 - \exp(-\lambda(a - x_i)) \\
&- (4 + 2\lambda(x_i - x_j))\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftrightarrow\ &4(1 - \exp(-\lambda(x_i - x_j))) + \exp(-\lambda(a - x_i))(\exp(-\lambda(x_i - x_j)) - 1) \\
&- 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftrightarrow\ &(4 - \exp(-\lambda(a - x_i)))(1 - \exp(-\lambda(x_i - x_j))) \\
&- 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \\
\Longleftarrow\ &3(1 - \exp(-\lambda(x_i - x_j))) - 2\lambda(x_i - x_j)\exp(-\lambda(x_i - x_j)) > 0 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\exp(-\lambda(a - x_i)) < 1)
\end{aligned}$$

Again letting $z = \lambda(x_i - x_j)$, this is true if and only if

$$\begin{aligned}
3(1 - e^{-z}) > 2ze^{-z} &\Longleftrightarrow 3(e^z - 1) > 2z \\
&\Longleftrightarrow 3e^z > 3 + 2z
\end{aligned}$$

which is true because $e^z > 1 + z$ for $z > 0$. This completes the proof for Case 3.

As a result, we have that

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] \geq 0$$

for all a, with strict inequality for some a, which proves the theorem. □

Theorem F.8. *For any $a \in \mathbb{R}$ and $X_i = x_i + \sigma\varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, 1)$,*

$$\frac{d}{da}\Pr[X_i > X_j \mid X_i < a, X_j < a] > 0.$$

Proof. Assume $\sigma = 1/\sqrt{2}$. This is without loss of generality because for any instance with arbitrary σ', there is an instance with $\sigma = 1/\sqrt{2}$ that yields the same distribution over rankings (simply by scaling all item values by σ/σ'). First, we have

$$\begin{aligned}\Pr[X_i > X_j \mid X_i < a, X_j < a] &= \frac{\int_{-\infty}^{a} \Pr[X_i = x] \Pr[X_j < x]\ dx}{Pr[X_i < a] \Pr[X_j < a]} \\ &= \frac{\int_{-\infty}^{a} \exp(-(x - x_i)^2)/\sqrt{\pi} \cdot (1 + \operatorname{erf}(x - x_j))/2\ dx}{(1 + \operatorname{erf}(a - x_i))/2 \cdot (1 + \operatorname{erf}(a - x_j))/2} \\ &= \frac{2}{\sqrt{\pi}} \frac{\int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\ dx}{(1 + \operatorname{erf}(a - x_i)) \cdot (1 + \operatorname{erf}(a - x_j))}\end{aligned}$$

The derivative with respect to a is positive if and only if

$$\begin{aligned}&(1 + \operatorname{erf}(a - x_i))(1 + \operatorname{erf}(a - x_j)) \exp(-(a - x_i)^2)(1 + \operatorname{erf}(a - x_j)) \\ &> \int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\ dx \\ &\cdot \frac{2}{\sqrt{\pi}} \left((1 + \operatorname{erf}(a - x_i)) \exp(-(a - x_j)^2) + (1 + \operatorname{erf}(a - x_j)) \exp(-(a - x_i)^2)\right)\end{aligned} \tag{F.8}$$

Let $t = a - x_i$ and $\delta = x_i - x_j$. Then, using the fact that

$$\begin{aligned}&\int_{-\infty}^{a} \exp(-(x - x_i)^2)(1 + \operatorname{erf}(x - x_j))\ dx \\ &\quad = \int_{-\infty}^{a-x_i} \exp(-x^2)\ dx + \int_{-\infty}^{a-x_i} \exp(-x^2) \operatorname{erf}(x + \delta)\ dx \\ &\quad = \frac{\sqrt{\pi}}{2}(1 + \operatorname{erf}(a - x_i)) + \int_{-\infty}^{a-x_i} \exp(-x^2) \operatorname{erf}(x + \delta)\ dx,\end{aligned}$$

(F.8) becomes

$$\frac{\sqrt{\pi}}{2} \cdot \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t)) \exp(-(t+\delta)^2) + (1+\operatorname{erf}(t+\delta)) \exp(-t^2)}$$
$$> \frac{\sqrt{\pi}}{2}(1+\operatorname{erf}(t)) + \int_{-\infty}^{t} \exp(-x^2) \operatorname{erf}(x+\delta)\, dx$$
$$\iff \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t)) \exp(-(t+\delta)^2) + (1+\operatorname{erf}(t+\delta)) \exp(-t^2)} - (1+\operatorname{erf}(t))$$
$$- \frac{2}{\sqrt{\pi}} \int_{-\infty}^{t} \exp(-x^2) \operatorname{erf}(x+\delta)\, dx > 0 \tag{F.9}$$

To show that this is true, we will use the fact that $f(t) > 0$ whenever the following conditions are met:

1. $f(t)$ is continuous and differentiable everywhere
2. $\lim_{t \to -\infty} f(t) = 0$
3. $\frac{d}{dt} f(t) > 0$

We'll show that these conditions hold for the LHS of (F.9).

$$\lim_{t \to -\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t)) \exp(-(t+\delta)^2) + (1+\operatorname{erf}(t+\delta)) \exp(-t^2)} - (1+\operatorname{erf}(t))$$
$$- \frac{2}{\sqrt{\pi}} \int_{-\infty}^{t} \exp(-x^2) \operatorname{erf}(x+\delta)\, dx$$
$$= \lim_{t \to -\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t)) \exp(-(t+\delta)^2) + (1+\operatorname{erf}(t+\delta)) \exp(-t^2)} \tag{F.10}$$

Observe that both the numerator and denominator of (F.10) are positive, so this limit must be at least 0. We can upper bound it by

$$\lim_{t \to -\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t)) \exp(-(t+\delta)^2) + (1+\operatorname{erf}(t+\delta)) \exp(-t^2)}$$
$$\leq \lim_{t \to -\infty} \frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2 \exp(-t^2)}{(1+\operatorname{erf}(t+\delta)) \exp(-t^2)}$$
$$= \lim_{t \to -\infty} (1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))$$
$$= 0$$

Thus, the limit is 0. Now, we must show that the derivative is positive. The derivative is

$$\begin{aligned}
&\frac{d}{dt}\left[\frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)}-(1+\operatorname{erf}(t))\right.\\
&\left.-\frac{2}{\sqrt{\pi}}\int_{-\infty}^{t}\exp(-x^2)\operatorname{erf}(x+\delta)\,dx\right]\\
&=\frac{d}{dt}\left[\frac{(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))^2\exp(-t^2)}{(1+\operatorname{erf}(t))\exp(-(t+\delta)^2)+(1+\operatorname{erf}(t+\delta))\exp(-t^2)}\right]-\frac{2}{\sqrt{\pi}}\exp(-t^2)\\
&-\frac{2}{\sqrt{\pi}}\exp(-t^2)\operatorname{erf}(t+\delta) \qquad \text{(F.11)}
\end{aligned}$$

Taking this derivative and factoring out

$$\frac{2(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))\exp(4t^2)}{\sqrt{\pi}\left((\operatorname{erf}(t)+1)\,e^{t^2}+(\operatorname{erf}(\delta+t)+1)\,e^{(\delta+t)^2}\right)^2},$$

we get that (F.11) is positive if and only if

$$\begin{aligned}
&\delta\sqrt{\pi}\exp((t+\delta)^2)(1+\operatorname{erf}(t))(1+\operatorname{erf}(t+\delta))-\exp(2\delta t+t^2)(1+\operatorname{erf}(t+\delta))\\
&+(1+\operatorname{erf}(t))>0\\
\Longleftrightarrow&\delta\sqrt{\pi}\exp((t+\delta)^2)(1+\operatorname{erf}(t))+\frac{1+\operatorname{erf}(t)}{1+\operatorname{erf}(t+\delta)}-\exp(2\delta t+t^2)>0\\
\Longleftrightarrow&\delta\sqrt{\pi}\exp(t^2)(1+\operatorname{erf}(t))+\exp(-2\delta t-t^2)\frac{1+\operatorname{erf}(t)}{1+\operatorname{erf}(t+\delta)}-1>0\\
\Longleftrightarrow&(1+\operatorname{erf}(t))\left[\delta\sqrt{\pi}\exp(t^2)+\frac{\exp(-2\delta t-\delta^2)}{1+\operatorname{erf}(t+\delta)}\right]-1>0\\
\Longleftrightarrow&\frac{1+\operatorname{erf}(t)}{\exp(-t^2)}\left[\delta\sqrt{\pi}+\frac{\exp(-(t+\delta)^2)}{1+\operatorname{erf}(t+\delta)}\right]-1>0 \qquad \text{(F.12)}
\end{aligned}$$

Define

$$g(t)\triangleq\frac{1+\operatorname{erf}(t)}{\exp(-t^2)}.$$

Then, (F.12) is

$$\begin{aligned}
&g(t)\left[\delta\sqrt{\pi}+\frac{1}{g(t+\delta)}\right]-1>0\\
\Longleftrightarrow&\frac{1}{g(t)}-\frac{1}{g(t+\delta)}<\delta\sqrt{\pi}
\end{aligned}$$

By the Mean Value Theorem,

$$\frac{1}{g(t)} - \frac{1}{g(t+\delta)} = -\delta \left.\frac{d}{dt}\frac{1}{g(t)}\right|_{t=t^*}$$

for some $t \le t^* \le t + \delta$. Thus, it suffices to show that

$$\frac{d}{dt}\frac{1}{g(t)} > -\sqrt{\pi} \tag{F.13}$$

for all t. To do this, consider Mills Ratio (Mills, 1926)

$$R(t) \triangleq \exp(t^2/2) \int_t^\infty \exp(-x^2/2)\, dx.$$

Note that this is quite similar in functional form to $g(t)$, and with some manipulation, we can relate the two:

$$\begin{aligned}
R(t) &= \exp(t^2/2) \int_t^\infty \exp(-x^2/2)\, dx \\
R(\sqrt{2}t) &= \exp(t^2) \int_{\sqrt{2}t}^\infty \exp(-x^2/2)\, dx \\
&= \sqrt{2}\exp(t^2) \int_t^\infty \exp(-x^2)\, dx \\
&= \sqrt{2}\exp(t^2) \int_{-\infty}^{-t} \exp(-x^2)\, dx \qquad (\exp(-x^2) \text{ is symmetric}) \\
R(-\sqrt{2}t) &= \sqrt{2}\exp(t^2) \int_{-\infty}^{t} \exp(-x^2)\, dx \\
&= \sqrt{2}\exp(t^2) \cdot \frac{\sqrt{\pi}}{2}(1 + \operatorname{erf}(t)) \\
&= \sqrt{\frac{\pi}{2}} \left(\frac{1 + \operatorname{erf}(t)}{\exp(-t^2)}\right) \\
R(-\sqrt{2}t) &= \sqrt{\frac{\pi}{2}} g(t)
\end{aligned}$$

Sampford (1953, Eq. (3)) proved that $\frac{d}{dt}\frac{1}{R(t)} < 1$ for any t. Thus,

$$\frac{d}{dt}\frac{1}{g(t)} = \frac{d}{dt}\frac{1}{\sqrt{\frac{2}{\pi}}R(-\sqrt{2}t)} = \sqrt{\frac{\pi}{2}}\frac{d}{dt}\frac{1}{R(-\sqrt{2}t)} > \sqrt{\frac{\pi}{2}} \cdot 1 \cdot -\sqrt{2} = -\sqrt{\pi},$$

which proves (F.13) and completes the proof. □

F.4 Verifying that the Mallows Model Satisfies Definition 9.1

Theorem F.9. *The family of distributions $\mathcal{F}_\theta$ produced by the Mallows Model with Kendall tau distance with $\theta = \phi - 1$ satisfies the conditions of Definition 9.1.*

Proof. We must show that $\mathcal{F}_\theta$ satisfies the differentiability, asymptotic optimality, and monotonicity conditions of Definition 9.1.

Differentiability: Let Π be the set of all permutations on n candidates. The probability of a realizing a particular permutation π under the Mallows model is

$$\Pr_\theta[\pi] = \frac{\phi^{-d(\pi,\pi^*)}}{\sum_{\pi' \in \Pi} \phi^{-d(\pi',\pi^*)}}$$

Both the numerator and denominator are differentiable with respect to $\theta = \phi - 1$, so $\Pr_\theta[\pi]$ is differentiable with respect to θ.

Asymptotic optimality: For the correct ranking π^*,

$$\Pr_\theta[\pi^*] = \frac{1}{Z},$$

where the normalizing constant Z is

$$Z = \sum_{\pi \in \Pi} \phi^{-d(\pi,\pi^*)}$$

In the limit,

$$\begin{aligned}
\lim_{\theta\to\infty} Z &= \lim_{\phi\to\infty} Z \\
&= \lim_{\phi\to\infty} \sum_{\pi\in\Pi} \phi^{-d(\pi,\pi^*)} \\
&= \lim_{\phi\to\infty} 1 + \sum_{\pi\neq\pi^*\in\Pi} \phi^{-d(\pi,\pi^*)} \\
&= 1 + \sum_{\pi\neq\pi^*\in\Pi} \lim_{\phi\to\infty} \phi^{-d(\pi,\pi^*)} \\
&= 1
\end{aligned}$$

because for any $\pi \neq \pi^*$, $d(\pi, \pi^*) \geq 1$. Therefore,

$$\lim_{\theta\to\infty} \Pr_{\theta}[\pi^*] = \lim_{\theta\to\infty} \frac{1}{Z} = 1$$

Monotonicity: We must show that for any $S \subset \mathbf{x}$, if $\pi_1^{(-S)}$ denotes the value of the top-ranked candidate according to π excluding candidates in S,

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] \geq \mathbb{E}_{\mathcal{F}_{\theta}}\left[\pi_1^{(-S)}\right].$$

For any $i \notin S$, let j be the smallest index such that $j > i$ and $j \notin S$. Consider any π such that $\pi_1^{(-S)} = x_j$. Then, swapping i and j yields a permutation $\hat{\pi}$ such that $\hat{\pi}_1^{(-S)} = x_i$. Moreover,

$$\Pr[\hat{\pi}] = \Pr[\pi] \cdot \phi^{\mathrm{inv}(\pi)-\mathrm{inv}(\hat{\pi})}.$$

Since $i < j$, $\mathrm{inv}(\pi) - \mathrm{inv}(\hat{\pi}) \geq 1$. Finally, note that swapping i and j is a bijection between $\{\pi : \pi_1^{(-S)} = x_j\}$ and $\{\pi : \pi_1^{(-S)} = x_i\}$. Thus,

$$\frac{\Pr[\pi_1^{(-S)} = x_i]}{\Pr[\pi_1^{(-S)} = x_j]} = \sum_{\pi:\pi_1^{(-S)}=x_j} \frac{\Pr[\pi]}{\Pr[\pi_1^{(-S)} = x_j]} \cdot \phi^{\mathrm{inv}(\pi)-\mathrm{inv}(\hat{\pi})}$$

Note that the terms $\frac{\Pr[\pi]}{\Pr[\pi_1^{(-S)}=x_j]}$ sum to 1, so this is sum is some polynomial in ϕ with nonnegative weights and integer powers of ϕ. As a result, it must have a

positive derivative with respect to ϕ, i.e., for $i < j$,

$$\frac{d}{d\phi} \frac{\Pr[\pi_1^{(-S)} = x_i]}{\Pr[\pi_1^{(-S)} = x_j]} > 0$$

Let $\phi' > \phi$. Then,

$$\frac{\Pr_\phi[\pi_1^{(-S)} = x_i]}{\Pr_\phi[\pi_1^{(-S)} = x_j]} < \frac{\Pr_{\phi'}[\pi_1^{(-S)} = x_i]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_j]}$$

Rearranging,

$$\frac{\Pr_\phi[\pi_1^{(-S)} = x_i]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_i]} < \frac{\Pr_\phi[\pi_1^{(-S)} = x_j]}{\Pr_{\phi'}[\pi_1^{(-S)} = x_j]} \tag{F.14}$$

For $\theta' - \phi' - 1$ and $\theta = \phi - 1$,

$$\mathbb{E}_{\mathcal{F}_\theta}\left[\pi_1^{(-S)}\right] = \sum_{i \notin S} \Pr_\phi\left[\pi_1^{-(S)} = x_i\right] x_i$$

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] = \sum_{i \notin S} \Pr_{\phi'}\left[\pi_1^{-(S)} = x_i\right] x_i$$

By Lemma F.10,

$$\mathbb{E}_{\mathcal{F}_{\theta'}}\left[\pi_1^{(-S)}\right] > \mathbb{E}_{\mathcal{F}_\theta}\left[\pi_1^{(-S)}\right],$$

which completes the proof. Note that we apply Lemma F.10 indexing backwards from n to 1, ignoring elements in S, with $p_i = \Pr_\phi\left[\pi_1^{-(S)} = x_i\right]$ and $q_i = \Pr_{\phi'}\left[\pi_1^{-(S)} = x_i\right]$. (F.14) provides the condition that p_i/q_i is decreasing (as i decreases, since we are indexing backwards). □

F.5 Proof of Theorem 9.6

F.5.1 Verifying Definition 9.2

We must show that when $\pi, \tau \sim \mathcal{F}_\theta$,

$$\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1\right] > 0. \tag{F.15}$$

We begin by expanding:

$$\begin{aligned}
&\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] \\
&= \sum_{i=1}^{n-1}\sum_{j>i}(x_i - x_j) \\
&\cdot (\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1])
\end{aligned}$$

Since $x_i > x_j$ for $i < j$, it suffices to show that for all $i < j$,

$$\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] \geq \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1], \tag{F.16}$$

and that this holds strictly for some $i < j$. We simplify (F.16) as follows:

$$\begin{aligned}
&\Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_1 \neq \tau_1] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_1 \neq \tau_1] \\
&\iff \frac{\Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \pi_1 \neq \tau_1]}{\Pr[\pi_1 \neq \tau_1]} > \frac{\Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \pi_1 \neq \tau_1]}{\Pr[\pi_1 \neq \tau_1]} \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \pi_1 \neq \tau_1] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \pi_1 \neq \tau_1] \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j \cap \tau_1 \neq x_i] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i \cap \tau_1 \neq x_j] \\
&\iff \Pr[\pi_1 = x_i \cap \pi_2 = x_j]\Pr[\tau_1 \neq x_i] > \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\Pr[\tau_1 \neq x_j] \qquad \text{(F.17)}
\end{aligned}$$

We can simplify (F.17) using Lemmas F.11 and F.12. Let $|i - j|$ denote the difference in rank between x_i and x_j.

$$\begin{aligned}
&\Pr[\pi_1 = x_i \cap \pi_2 = x_j]\Pr[\tau_1 \neq x_i] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\Pr[\tau_1 \neq x_j] \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i]) \\
&- \phi^{-1}\Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_j]) \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i]) \\
&- \phi^{-1}\Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \phi^{-|i-j|}\Pr[\tau_1 = x_i]) \\
&= \Pr[\pi_1 = x_i \cap \pi_2 = x_j](1 - \Pr[\tau_1 = x_i] - \phi^{-1} - \phi^{-|i-j|-1}\Pr[\tau_1 = x_i]))
\end{aligned}$$

This is positive if and only if

$$
\begin{aligned}
&1 - \Pr[\tau_1 = x_i] - \phi^{-1} - \phi^{-|i-j|-1} \Pr[\tau_1 = x_i] > 0 \\
&\iff \Pr[\tau_1 = x_i] \left(1 - \phi^{-|i-j|-1}\right) < 1 - \phi^{-1} \\
&\iff \Pr[\tau_1 = x_i] < \frac{1 - \phi^{-1}}{1 - \phi^{-|i-j|-1}} \\
&\iff \frac{1 - \phi^{-1}}{\phi^{i-1}(1 - \phi^{-n})} < \frac{1 - \phi^{-1}}{1 - \phi^{-|i-j|-1}} \\
&\iff \phi^{i-1}(1 - \phi^{-n}) > 1 - \phi^{-|i-j|-1}
\end{aligned}
$$

This is weakly true for any $i < j$ because $\phi^{i-1} \geq 1$ and $|i - j| + 1 \leq n$, and it is strictly true for any i, j other than 1 and n. Thus, $\mathbb{E}\left[\pi_1 - \pi_2 \mid \pi_1 \neq \tau_1\right] > 0$.

F.5.2 Verifying Definition 9.3

Recall that Definition 9.3 is equivalent to $U_{AH}(\theta_A, \theta_H) < U_{HH}(\theta_A, \theta_H)$ for $\theta_A > \theta_H$. Let τ be the algorithmic ranking, and let π be a ranking from a human evaluator. Recall that $U_H(\theta_A, \theta_H) = \mathbb{E}\left[\pi_1\right]$. Throughout this proof, we will drop

the (θ_A, θ_H) notation and simply write U_H, U_{AH}, and U_{HH}.

$$\begin{aligned}
U_{AH} &= \sum_{i=1}^{n} (\Pr[\pi_1 = x_i \cap \tau_1 \neq x_i] + \Pr[\pi_2 = x_i \cap \pi_1 = \tau_1]) x_i \\
&= \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 \neq x_i] x_i + \sum_{i=1}^{n} \Pr[\pi_2 = x_i \cap \pi_1 = \tau_1] x_i \\
&= \sum_{i=1}^{n} \left(Pr[\pi_1 = x_i] - \Pr[\pi_1 = x_i \cap \tau_1 = x_i] \right) x_i \\
&+ \sum_{i=1}^{n} \sum_{j \neq i} \Pr[\pi_1 = x_j \cap \tau_1 = x_j \cap \pi_2 = x_i] x_i \\
&= U_H - \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 = x_i] x_i \\
&+ \sum_{i=1}^{n} \Pr[\pi_1 = x_i \cap \tau_1 = x_i] \mathbb{E}\left[\pi_2 \mid \pi_1 = x_i \cap \tau_1 = x_i\right] \\
&= U_H + \sum_{i=1}^{n} \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i] \left(\mathbb{E}\left[\pi_2 \mid \pi_1 = x_i\right] - x_i \right)
\end{aligned}$$

Similarly, because two human evaluators are independent,

$$U_{HH} = U_H + \sum_{i=1}^{n} \Pr[\pi_1 = x_i]^2 \left(\mathbb{E}\left[\pi_2 \mid \pi_1 = x_i\right] - x_i \right).$$

Let $V_{-i} = \mathbb{E}\left[\pi_2 \mid \pi_1 = x_i\right]$. Note that conditioned on $\pi_1 = x_i$, the remaining elements of π_1 follow a Mallows model distribution over $n-1$ candidates. Because the Mallows model is value-independent, increasing any item value increases the expected value of the top-ranked item (and in fact, the item ranked at any position). Thus, V_{-i} increases as i increases (since x_i, the value of the unavailable candidate, decreases). Moreover, x_i is strictly decreasing in i, so $V_{-i} - x_i$ is strictly increasing in i. With this, we have

$$\begin{aligned}
U_{AH} - U_H &= \sum_{i=1}^{n} \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i] \left(V_{-i} - x_i \right) \\
U_{HH} - U_H &= \sum_{i=1}^{n} \Pr[\pi_1 = x_i]^2 \left(V_{-i} - x_i \right)
\end{aligned}$$

Let $C_A = \Pr[\pi_1 = \tau_1] = \sum_{i=1}^n \Pr[\pi_1 = x_i] \Pr[\tau_1 = x_i]$, and similarly let $C_H = \sum_{i=1}^n \Pr[\pi_1 = x_i]^2$. $C_A > C_H$ by Lemma F.10 with $y'_i = \Pr[\pi_1 = x_{n-i+1}]$, $p'_i = \Pr[\pi_1 = x_{n-i+1}]$ and $q'_i = \Pr[\tau_1 = x_{n-i+1}]$.

Let $p_i = \Pr[\pi'_1 = i] \Pr[\pi_1 = i]/C_A$, $q_i = \Pr[\pi'_1 = i]^2/C_H$, and $y_i = V_{-i} - x_i$. Then, we have

$$\frac{U_{AH} - U_H}{C_A} = \sum_{i=1}^n p_i y_i$$
$$\frac{U_{HH} - U_H}{C_H} = \sum_{i=1}^n q_i y_i$$

With $\phi_A = \theta_A + 1$ and $\phi_H = \theta_H + 1$,

$$\frac{p_i}{q_i} = \frac{C_H}{C_A} \cdot \frac{\frac{1-\phi_A^{-1}}{\phi_A^{i-1}(1-\phi_A^{-n})}}{\frac{1-\phi_H^{-1}}{\phi_H^{i-1}(1-\phi_H^{-n})}} \propto \frac{\phi_H^{i-1}}{\phi_A^{i-1}},$$

which is decreasing in i since $\phi_H < \phi_A$. By Lemma F.10, $\sum_{i=1}^n p_i y_i < \sum_{i=1}^n q_i y_i$. Finally, note that $U_{HH} - U_H < 0$ by Lemma F.13, so

$$\begin{aligned}
\sum_{i=1}^n p_i y_i &< \sum_{i=1}^n q_i y_i \\
\frac{U_{AH} - U_H}{C_A} &< \frac{U_{HH} - U_H}{C_H} \\
\frac{C_H(U_{AH} - U_H)}{C_A} &< U_{HH} - U_H \\
U_{AH} - U_H &< U_{HH} - U_H && (C_A > C_H, \text{ and } U_{HH} - U_H < 0) \\
U_{AH} &< U_{HH}
\end{aligned}$$

F.6 Supplementary Lemmas for the Mallows Model

Lemma F.10. *Let $\{y_i\}_{i=1}^n$, $\{p_i\}_{i=1}^n$, and $\{q_i\}_{i=1}^n$ be sequences such that*

- *y_i is strictly increasing.*

- $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$.
- $\frac{p_i}{q_i}$ *is decreasing.*

Then, $\sum_{i=1}^{n} p_i y_i < \sum_{i=1}^{n} q_i y_i$.

Proof. First, note that there exists j such that $p_i > q_i$ for $i < j$ and $p_i \leq q_i$ for $i \geq j$. To see this, let j be the smallest index such that $p_j \leq q_j$. Such a j must exist because p_i and q_i both sum to 1, so it cannot be the case that $p_i > q_i$ for all i. This implies $p_i/q_i \leq 1$, and since p_i/q_i is decreasing, $p_i \leq q_i$ for $i \geq j$.

Next, note that

$$\begin{aligned} 0 &= \sum_{i=1}^{n} (p_i - q_i) \\ &= \sum_{i=1}^{j-1} (p_i - q_i) + \sum_{i=j}^{n} (p_i - q_i), \end{aligned}$$

meaning

$$\sum_{i=1}^{j-1} (p_i - q_i) = \sum_{i=j}^{n} (q_i - p_i).$$

Using this choice of j, we can write

$$\begin{aligned}
\sum_{i=1}^{n} p_i y_i - \sum_{i=1}^{n} q_i y_i &= \sum_{i=1}^{n} (p_i - q_i) y_i \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_i - \sum_{i=j}^{n} (q_i - p_i) y_i \\
&\leq \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=j}^{n} (q_i - p_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=j}^{n} (q_i - p_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i) y_{j-1} - \sum_{i=1}^{j-1} (p_i - q_i) y_j \\
&= \sum_{i=1}^{j-1} (p_i - q_i)(y_{j-1} - y_j) \\
&< 0
\end{aligned}$$

□

Lemma F.11. *For $x_i > x_j$,*

$$\Pr[\pi_1 = x_i \cap \pi_2 = x_j] = \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i]. \tag{F.18}$$

Proof. Let π_{-ij} be a permutation of all of the candidates except x_i and x_j. Then, we have

$$\begin{aligned}
\Pr[\pi_1 = x_i \cap \pi_2 = x_j] &= \sum_{\pi_{-ij}} \Pr[\pi_1 = x_i \cap \pi_2 = x_j \mid \pi_{-ij}] \Pr[\pi_{-ij}] \\
&= \sum_{\pi_{-ij}} \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i \mid \pi_{-ij}] \Pr[\pi_{-ij}] \\
&= \phi \Pr[\pi_1 = x_j \cap \pi_2 = x_i]
\end{aligned}$$

Intuitively, given that x_i and x_j are in the top 2 positions, x_i followed by x_j is ϕ times more likely than x_j followed by x_i regardless of the remainder of the permutation, and therefore, x_i followed by x_j is ϕ times more likely overall. □

Lemma F.12. *For* $1 \leq i \leq n$,

$$\Pr[\pi_1 = x_i] = \frac{1 - \phi^{-1}}{\phi^{i-1}(1 - \phi^{-n})}. \tag{F.19}$$

Proof. Let π_{-i} be a permutation over all items except i. Then,

$$\begin{aligned}
\Pr[\pi_1 = x_i] &= \sum_{\pi_{-i}} \Pr[\pi_1 = x_i \mid \pi_{-i}] \Pr[\pi_{-i}] \\
&= \sum_{\pi_{-i}} \phi^{-(i-1)} \Pr[\pi_{-i}] \\
&= \phi^{-(i-1)} \sum_{\pi_{-i}} \Pr[\pi_{-i}]
\end{aligned}$$

Note that $\Pr[\pi_{-i}]$ doesn't depend on *which* $n-1$ items are being ranked, so this term appears for any i. Moreover, $\sum_{i=1}^{n} \Pr[\pi_1 = x_i] = 1$. Therefore, we have

$$\Pr[\pi_1 = x_i] \propto \phi^{-(i-1)}.$$

Normalizing, we get

$$\begin{aligned}
\Pr[\pi_1 = x_i] &= \frac{\phi^{-(i-1)}}{\sum_{j=1}^{n} \phi^{-(j-1)}} \\
&= \frac{\phi^{-(i-1)}}{\frac{1-\phi^{-n}}{1-\phi^{-1}}} \\
&= \frac{1 - \phi^{-1}}{\phi^{i-1}(1 - \phi^{-n})}
\end{aligned}$$

Intuitively, any permutation over $n-1$ items is equally likely regardless of what those items are, and inserting any item at the front of this permutation yields

a likelihood proportional to the number of additional inversions this causes, which is equal to the item's position on the list.[1] □

Lemma F.13. *For the Mallows Model,* $U_H(\theta_A, \theta_H) > U_{HH}(\theta_A, \theta_H)$.

Proof. Intuitively, this is because selecting first is better than selecting second. To prove this, let π and τ be ranking generated by independent human evaluators under the Mallows Model, i.e., $\pi, \tau \sim \mathcal{F}_{\theta_H}$.

$$\begin{aligned} U_H(\theta_A, \theta_H) - U_{HH}(\theta_A, \theta_H) &= \mathbb{E}\left[\pi_1\right] - \mathbb{E}\left[\tau_1 \cdot \mathbf{1}_{\{\pi_1 \neq \tau_1\}} + \tau_2 \cdot \mathbf{1}_{\{\pi_1 = \tau_1\}}\right] \\ &= \mathbb{E}\left[(\pi_1 - \tau_2) \cdot \mathbf{1}_{\{\pi_1 = \tau_1\}}\right] \\ &= \mathbb{E}\left[(\pi_1 - \pi_2) \cdot \mathbf{1}_{\{\pi_1 = \tau_1\}}\right] \end{aligned}$$

For any $i < j$, conditioned on $\pi_1 = \tau_1$, they are more likely to be correctly

[1] Alternatively, we could prove this by showing that for any permutation with i in front, the permutation in which i and $i-1$ are swapped is ϕ times more likely, and thus, $i-1$ is ϕ times more likely to be in front than i.

ordered than not:

$$
\begin{aligned}
&\mathbb{E}\left[(\pi_1 - \pi_2) \cdot \mathbf{1}_{\{\pi_1 = \tau_1\}}\right] \\
&= \sum_{i<j} (\Pr[\pi_1 = x_i \cap \tau_1 = x_i \cap \pi_2 = x_j] \\
&- \Pr[\pi_1 = x_j \cap \tau_1 = x_j \cap \pi_2 = x_i])(x_i - x_j) \\
&= \sum_{i<j} (\Pr[\pi_1 = x_i \cap \pi_2 = x_j] \Pr[\tau_1 = x_i] \\
&- \Pr[\pi_1 = x_j \cap \pi_2 = x_i] \Pr[\tau_1 = x_j])(x_i - x_j) \\
&> \sum_{i<j} (\Pr[\pi_1 = x_i \cap \pi_2 = x_j] \Pr[\tau_1 = x_j] \\
&- \Pr[\pi_1 = x_j \cap \pi_2 = x_i] \Pr[\tau_1 = x_j])(x_i - x_j) \\
&= \sum_{i<j} \left(\Pr[\pi_1 = x_i \cap \pi_2 = x_j] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\right)(x_i - x_j) \\
&\geq \sum_{i<j} \left(\phi_H \Pr[\pi_1 = x_j \cap \pi_2 = x_i] - \Pr[\pi_1 = x_j \cap \pi_2 = x_i]\right)(x_i - x_j) \\
&> 0
\end{aligned}
$$

□

F.6.1 Proof of Theorem 9.7

To prove this theorem, we make use of the following lemma.

Lemma F.14. *Under the Mallows model, the probability that any two items $i < j$ are correctly ranked increases monotonically with the accuracy parameter ϕ.*

Proof. Let $\mathrm{inv}(\pi)$ be the number of inversions in a permutation π. Under the Mallows model, the probability of observing π is proportional to $\phi^{-\mathrm{inv}(\pi)}$. Let

$S_{i \succ j}$ (resp. $S_{j \succ i}$) be the set of permutations where i is ranked before j (resp. j is ranked before i). Then, the probability i is ranked before j is

$$\Pr[i \succ j] = \frac{\sum_{\pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi)}}{\sum_{\pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi)} + \sum_{\pi \in S_{j \succ i}} \phi^{-\operatorname{inv}(\pi)}}.$$

We will show that $\frac{d}{d\phi} \Pr[i \succ j] > 0$. Note that this is equivalent to showing $\frac{d}{d\phi} \frac{\Pr[i \succ j]}{\Pr[j \succ i]} > 0$. Note that

$$\frac{\Pr[i \succ j]}{\Pr[j \succ i]} = \frac{\sum_{\pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi)}}{\sum_{\pi \in S_{j \succ i}} \phi^{-\operatorname{inv}(\pi)}}.$$

Let $\pi_{i:j}$ be the subsequence of π containing elements i through j. Then, we have

$$\begin{aligned}
\frac{\Pr[i \succ j]}{\Pr[j \succ i]} &= \frac{\sum_{\pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi)}}{\sum_{\pi \in S_{j \succ i}} \phi^{-\operatorname{inv}(\pi)}} \\
&= \frac{\sum_{\pi_{i:j} : \pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi_{i:j})} \sum_{\pi' : \pi'_{i:j} = \pi_{i:j}} \phi^{\operatorname{inv}(\pi_{i:j}) - \operatorname{inv}(\pi')}}{\sum_{\pi_{i:j} : \pi \in S_{j \succ i}} \phi^{-\operatorname{inv}(\pi_{i:j})} \sum_{\pi' : \pi'_{i:j} = \pi_{i:j}} \phi^{\operatorname{inv}(\pi_{i:j}) - \operatorname{inv}(\pi')}} \\
&= \frac{\sum_{\pi_{i:j} : \pi \in S_{i \succ j}} \phi^{-\operatorname{inv}(\pi_{i:j})}}{\sum_{\pi_{i:j} : \pi \in S_{j \succ i}} \phi^{-\operatorname{inv}(\pi_{i:j})}}
\end{aligned}$$

Intuitively, the term $\sum_{\pi' : \pi'_{i:j} = \pi_{i:j}} \phi^{\operatorname{inv}(\pi_{i:j}) - \operatorname{inv}(\pi')}$ does not depend on $\pi_{i:j}$ because for any $\pi_{i:j}$, if we fix the order and positions of the remaining elements, the number of inversions involving at least one element outside of $i : j$ (i.e., $\operatorname{inv}(\pi') - \operatorname{inv}(\pi_{i:j})$) is a constant. For fixed $\pi_{i:j}$, there is a bijection between permutations $\pi' : \pi'_{i:j} = \pi_{i:j}$ and a fixed order and position of the remaining elements (excluding $i : j$), meaning this sum does not depend on $\pi_{i:j}$. Thus, for the remainder of this proof, we can assume without loss of generality that $i = 1$ and $j = n$. The quantity of interest becomes

$$\begin{aligned}
\frac{\Pr[1 \succ n]}{\Pr[n \succ 1]} &= \frac{\sum_{\pi_{1:n} : \pi \in S_{1 \succ n}} \phi^{-\operatorname{inv}(\pi_{1:n})}}{\sum_{\pi_{i:j} : \pi \in S_{n \succ 1}} \phi^{-\operatorname{inv}(\pi_{1:n})}} \\
&= \frac{\sum_{\pi \in S_{1 \succ n}} \phi^{-\operatorname{inv}(\pi)}}{\sum_{\pi \in S_{n \succ 1}} \phi^{-\operatorname{inv}(\pi)}}
\end{aligned}$$

Next, we observe that we can similarly ignore inversions between two elements that are neither 1 nor n. To see this, let $\mathrm{inv}_{1,n}(\pi)$ be the number of inversions involving at least one of 1 and n. Then, if we fix the order and positions of 1 and n, all possible permutations of the remaining elements 2 through $n-1$ produce the same number of inversions $\mathrm{inv}_{1,n}(\pi)$. More formally, let $\pi_{(1)}$ and $\pi_{(n)}$ be the respective positions of elements 1 and n. Then, this we have

$$\begin{aligned}
\sum_{\pi \in S_{1 \succ n}} \phi^{-\operatorname{inv}(\pi)} &= \sum_{k<\ell} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{-\operatorname{inv}(\pi)} \\
&= \sum_{k<\ell} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{-\operatorname{inv}_{1,n}(\phi)} \cdot \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)} \\
&= \sum_{k<\ell} \phi^{-(k-1)-(n-\ell)} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)}
\end{aligned}$$

As noted above, $\sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)}$ does not depend on k or ℓ, since every permutation of the remaining elements yields the same number of inversions among them regardless of k and ℓ. A similar argument yields

$$\sum_{\pi \in S_{n \succ 1}} \phi^{-\operatorname{inv}(\pi)} = \sum_{k>\ell} \phi^{-(k-1)-(n-\ell)+1} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)}$$

Putting these together, we have

$$\begin{aligned}
\frac{\Pr[1 \succ n]}{\Pr[n \succ 1]} &= \frac{\sum_{k<\ell} \phi^{-(k-1)-(n-\ell)} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)}}{\sum_{k>\ell} \phi^{-(k-1)-(n-\ell)+1} \sum_{\pi:\pi_{(1)}=k,\pi_{(n)}=\ell} \phi^{\operatorname{inv}_{1,n}(\phi)-\operatorname{inv}(\pi)}} \\
&= \frac{\sum_{k<\ell} \phi^{-(k-1)-(n-\ell)}}{\sum_{k>\ell} \phi^{-(k-1)-(n-\ell)+1}} \\
&= \frac{\sum_{k<\ell} \phi^{-(k-1)-(n-\ell)}}{\sum_{k>\ell} \phi^{-(k-1)-(n-\ell)+1}} \cdot \frac{\phi^{n-1}}{\phi^{n-1}} \\
&= \frac{\sum_{k<\ell} \phi^{\ell-k}}{\sum_{k>\ell} \phi^{\ell-k+1}}
\end{aligned}$$

Note that each term in the numerator is strictly increasing in ϕ, while each term in the denominator is weakly decreasing in ϕ. As a result, $\frac{d}{d\phi} \frac{\Pr[1 \succ n]}{\Pr[n \succ 1]} > 0$, meaning for any $i < j$, $\frac{d}{d\phi} \Pr[i \succ j] > 0$. $\square$

With this, we proceed inductively, showing that when $\phi_H \geq \phi_A$, each firm rationally chooses to use H. For the first firm, by Lemma F.14, H is more likely than A to correctly order any pair of candidates, meaning it produces higher expected utility. Similarly, for any subsequent firm, conditioned on the remaining candidates, H is still more likely to correctly order any pair of remaining candidates, meaning it leads to higher expected utility. A similar argument shows that all firms strictly prefer H when $\phi_H > \phi_A$.

APPENDIX G

MITIGATING BIAS IN ALGORITHMIC DECISION-MAKING: EVALUATING CLAIMS AND PRACTICES

G.1 Administrative Information on Vendors

Vendor name	Funding	# of employees	Location
8 and Above	–	1-10	WA, USA
ActiView	$6.5M	11-50	Israel
Assessment Innovation	$1.3M	1-10	NY, USA
Good&Co	$10.3M	51-100	CA, USA
Harver	$14M	51-100	NY, USA
HireVue	$93M	251-500	UT, USA
impress.ai	$1.4M	11-50	Singapore
Knockri	–	11-50	Canada
Koru	$15.6M	11-50	WA, USA
LaunchPad Recruits	£2M	11-50	UK
myInterview	$1.4M	1-10	Australia
Plum.io	$1.9M	11-50	Canada
PredictiveHire	A$4.3M	11-50	Australia
pymetrics	$56.6M	51-100	NY, USA
Scoutible	$6.5M	1-10	CA, USA
Teamscope	€800K	1-10	Estonia
ThriveMap	£781K	1-10	UK
Yobs	$1M	11-50	CA, USA

Table G.1: Administrative information on vendors of algorithmic pre-employment assessments.

BIBLIOGRAPHY

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2014.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Miro Dudik, John Langford, Lihong Li, Luong Hoang, Dan Melamed, Siddhartha Sen, Robert Schapire, and Alex Slivkins. Multiworld testing: A system for experimentation, learning, and decision-making. A white paper, available at https://github.com/Microsoft/mwt-ds/raw/master/images/MWT-WhitePaper.pdf, 2016.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, Siddhartha Sen, and Alex Slivkins. Making contextual decisions with low technical debt. *CoRR arXiv:1606.03966*, 2017.

Ifeoma Ajunwa. The paradox of automation as anti-bias intervention. *Cardozo Law Review*, 41:1671, 2020.

Ifeoma Ajunwa. The auditing imperative for automated hiring. *Harvard Journal of Law & Technology*, 34, 2021.

Tal Alon, Magdalen Dobson, Ariel D Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. In *AAAI*, pages 1774–1781, 2020.

Ricardo Alonso and Niko Matouschek. Optimal delegation. *The Review of Economic Studies*, 75(1):259–293, 2008.

Julia Angwin and Jeff Larson. Bias in criminal risk scores is mathematically inevitable, researchers say. *Propublica*, December 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016.

Kenneth J Arrow. Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5):941–73, 1963.

Kenneth J Arrow. The economics of moral hazard: further comment. *American Economic Review*, 58(3):537–539, 1968.

Article 29 Data Protection Working Party. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, 2017.

Illinois General Assembly. Artificial intelligence video interview act, 2019.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

Hossein Azari Soufiani, Hansheng Diao, Zhenyu Lai, and David C Parkes. Generalized random utility models with multiple types. In *Advances in Neural Information Processing Systems*, pages 73–81, 2013.

Lewis Baker, David Weisberger, Daniel Diamond, Mark Ward, and Joe Naso. audit-AI. https://github.com/pymetrics/audit-ai, 2018.

Jack M Balkin. Information fiduciaries and the first amendment. *UC Davis Law Review*, 49:1183–1234, 2015.

Ian Ball. Scoring strategic agents. Working paper, 2020.

Jane Bambauer and Tal Zarsky. The algorithm game. *Notre Dame L. Rev.*, 94(1): 1–48, 2018.

Loren Baritz. *The servants of power: A history of the use of social science in American industry.* Wesleyan University Press, 1960.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring

emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2020.

Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. Metric-free individual fairness in online learning. In *Advances in Neural Information Processing Systems*, 2020.

Jöran Beel, Bela Gipp, and Erik Wilde. Academic search engine optimization (ASEO) optimizing scholarly literature for Google Scholar & co. *Journal of scholarly publishing*, 41(2):176–190, 2009.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Marc Bendick, Jr. and Ana P Nunes. Developing the research basis for controlling bias in hiring. *Journal of Social Issues*, 68(2):238–262, 2012.

Marc Bendick, Jr., Charles W Jackson, and J Horacio Romero. Employment discrimination against older workers: An experimental study of hiring practices. *Journal of Aging & Social Policy*, 8(4):25–46, 1997.

Jason R Bent. Is algorithmic affirmative action legal? *Georgetown Law Journal*, 108(4):803, 2020.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Richard Berk. A primer on fairness in criminal justice risk assessments. *Criminology*, 41(6):6–9, 2016.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2018.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Daniel A Biddle. Are the uniform guidelines outdated? Federal guidelines, professional standards, and validity generalization (VG). *The Industrial-Organizational Psychologist*, 45(4):17–23, 2008.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *CoRR arXiv:1802.04064*, 2018.

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. Available at SSRN: https://ssrn.com/abstract=2846909, also appeared at the Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2016.

Kenneth P Birman and Fred B Schneider. The monoculture risk put into context. *IEEE Security & Privacy*, 7(1):14–17, 2009.

Joseph Blass. Algorithmic advertising discrimination. *Nw. UL Rev.*, 114(2):415, 2019.

HD Block and J Marschak. Random orderings and stochastic theories of responses. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B.

Mann, editors, *Contributions to probability and statistics*, pages 97–132. Stanford University Press, 1960.

Miranda Bogen and Aaron Rieke. Help wanted: An exploration of hiring algorithms, equity, and bias. Technical report, Upturn, 2018. URL https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf.

Iris Bohnet, Alexandra van Geen, and Max Bazerman. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62 (5):1225–1234, 2016.

Raphael Boleslavsky and Kyungmin Kim. Bayesian persuasion and moral hazard. *Available at SSRN 2913669*, 2018.

Stephanie Bornstein. Antidiscriminatory algorithms. *Ala. L. Rev.*, 70:519, 2018.

Dietrich Braess. Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung*, 12(1):258–268, 1968.

Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.

Maja Brkan. Do algorithms rule the world? algorithmic decision-making and data protection in the framework of the gdpr and beyond. *International journal of law and information technology*, 27(2):91–121, 2019.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555. ACM, 2011.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

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.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *KDD*, 2012.

Gökhan Çapan, Özge Bozal, İlker Gündoğdu, and Ali Taylan Cemgil. Towards fair personalization by avoiding feedback loops. *arXiv preprint arXiv:2012.12862*, 2020.

Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105 (2):536–63, 2015.

Rich Caruana, Scott Lundberg, Marco Tulio Ribeiro, Harsha Nori, and Samuel Jenkins. Intelligible and explainable machine learning: Best practices and practical challenges. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3511–3512, 2020.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Bryan Casey, Ashkon Farhangi, and Roland Vogl. Rethinking explainable machines: The GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34:143–188, 2019.

Marilyn Cavicchia. How to fight implicit bias? With conscious thought, diversity expert tells NABE. *American Bar Association: Bar Leader*, 40(1), 2015.

L. Elisa Celis and Nisheeth K Vishnoi. Fair personalization. *CoRR arXiv:1707.02260, also appeared at the Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, pages 28:1–28:15, 2018.

Tomas Chamorro-Premuzic, Dave Winsborough, Ryne A Sherman, and Robert Hogan. New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology*, 9(3):621–640, 2016.

Tomas Chamorro-Prezumic and Reece Akhtar. Should companies use AI to assess job candidates. *Harvard Business Review*, 2019.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2018.

Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26. ACM, 2018.

Steven NS Cheung. *The theory of share tenancy*. Arcadia Press Ltd., 1969.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

Aaron Clauset, Cosma R. Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *American Economic Review*, 83(5):1220–1240, 1993.

Ignacio N Cofone. Algorithmic discrimination is an information problem. *Hastings LJ*, 70:1389, 2018.

Julie E Cohen. *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press, 2012a.

Julie E Cohen. What privacy is for. *Harvard Law Review*, 126:1904–1933, 2012b.

Richard M Cohn. Statistical laws and the use of statistics in law: A rejoinder to Professor Shoben. *Ind. LJ*, 55:537, 1979a.

Richard M Cohn. On the use of statistics in employment discrimination cases. *Ind. LJ*, 55:493, 1979b.

Brian W. Collins. Tackling unconscious bias in hiring practices: The plight of the Rooney Rule. *NYU Law Review*, 82(3), 2007.

John D Cook. Upper and lower bounds for the normal distribution function, 2009.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Covington and Burling. Recommendations to Uber, 13 June 2017.

Bo Cowgill. Bias and productivity in humans and machines. *Columbia Business School, Columbia University*, 29, 2018.

Hamilton Cravens. *The triumph of evolution: The heredity–environment controversy, 1900–1941.* Johns Hopkins University Press, 1978.

Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25 (4):1692–1706, 2016.

Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 448–483, 2015.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108, 2004.

HE Daniels. Rank correlation and population models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(2):171–191, 1950.

Sanmay Das and Zhuoshu Li. The role of common and private signals in two-sided matching with interviews. In *International Conference on Web and Internet Economics*, pages 492–497. Springer, 2014.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

H. A. David and H. N. Nagaraja. *Basic Distribution Theory*, pages 9–32. John Wiley & Sons, Inc., 2005. ISBN 9780471722168.

Harold Davis. *Search engine optimization*. O'Reilly Media, Inc., 2006.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*, 2021.

Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe, July 2016. URL http://www.northpointeinc.com/northpointe-analysis.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Cynthia DuBois. What the NFL can teach congress about hiring more diverse staffs. *FiveThirtyEight*, 2017.

Cynthia DuBois and Diane Whitmore Schanzenbach. The effect of court-ordered hiring guidelines on teacher composition and student achievement. Technical report, National Bureau of Economic Research, 2017.

Philip Hunter DuBois. *A history of psychological testing*. Allyn and Bacon, 1970.

Marvin D Dunnette and Walter C Borman. Personnel selection and classification systems. *Annual review of psychology*, 30(1):477–525, 1979.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, 2012.

Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Lilian Edwards and Michael Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16:18–74, 2017.

Equal Credit Opportunities Act, Public Law 93-495. Codified at 15 u.s.c. § 1691, et seq., 1974.

Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures. *Federal Register*, 43(166):38290–38315, 1978.

Virginia Eubanks. Automating Bias. *Scientific American*, 319(5):68–71, 2018a.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018b.

Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report, May 2016.

Fair Credit Reporting Act, Public Law 91-508. Codified at 15 u.s.c. § 1681, et seq., 1970.

Federal Register. 50 fed. reg. 10915, 1985.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.

Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, 2021.

Anthony Flores, Christopher Lowenkamp, and Kristin Bechtel. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". Technical report, Crime & Justice Institute, September 2016. URL http://www.crj.org/cji/entry/false-positives-false-negatives-and-false-analyses-a-rejoinder.

Yuk-fai Fong and Jin Li. Information revelation in relational contracts. *The Review of Economic Studies*, 84(1):277–299, 2016.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2): 379–390, 1998.

Dean Foust and Aaron Pressman. Credit scores: Not-so-magic numbers. *Business Week*, 7, 2008.

Robert H Frank. Why is cost-benefit analysis so controversial? *The Journal of Legal Studies*, 29(S2):913–930, 2000.

Peter Frazier, David Kempe, Jon M. Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *ACM Conference on Economics and Computation (ACM EC)*, 2014.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001.

Jim Fruchterman and Joan Melllea. Expanding employment success for people with disabilities. Technical report, benetech, 2018.

Roland G Fryer Jr and Glenn C Loury. Valuing diversity. *Journal of Political Economy*, 121(4):747–774, 2013.

Qiang Fu and Jingfeng Lu. Micro foundations of multi-prize lottery contests: a perspective of noisy performance ranking. *Social Choice and Welfare*, 38(3): 497–517, 2012.

Howard N. Garb. Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice*, 4(2):99–120, 1997.

Stacia Sherman Garr and Carole Jackson. Diversity & inclusion technology: The rise of a transformative market. Technical report, RedThread Research, 2019. URL https://info.mercer.com/rs/521-DEV-513/images/Mercer_DI_Report_Digital.pdf.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

PW Gerhardt. Scientific selection of employees. *Electric Railway Journal*, 47, 1916.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2019.

Talia B Gillis and Jann L Spiess. Big data and discrimination. *The University of Chicago Law Review*, 86(2):459–488, 2019.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

Abe Gong. Ethics for powerful algorithms (1 of 4). *Medium*, July 2016. URL https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84#.dhsd2ut3i.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

Alexander R. Green, Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. Implicit bias among physi-

cians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, 22(9):1231–1238, 2007.

Ben Green. Data science as political action: Grounding data science in a politics of justice. *arXiv preprint arXiv:1811.03435*, 2018.

Anthony G. Greenwald and Mahzarin R. Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995.

Anthony G. Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *California Law Review*, 94:945–967, 2006.

Jeff Grimmett. Veterinary practitioners - personal characteristics and professional longevity. *VetScript*, 2017.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45, 1983.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hannah Wallach, and Meredith Ringel Morris. Toward fairness in ai for people with disabilities: A research roadmap. *ACM SIGACCESS*, 125, October 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Richard A Guzzo, Alexis A Fink, Eden King, Scott Tonidandel, and Ronald S Landis. Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology*, 8(4):491–508, 2015.

Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Wang. Maximiz-

ing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*, 2020.

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2013.

Patrick Hall, Wen Phan, and SriSatish Ambati. Ideas on interpreting machine learning. *O'Reilly*, 2017.

Craig Haney. Employment tests and employment discrimination: A dissenting psychological opinion. *Indus. Rel. LJ*, 5:1, 1982.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016a.

Moritz Hardt, Eric Price, and Srebro Nathan. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016b.

Zach Harned and Hanna Wallach. Stretching human laws to apply to machines: The dangers of a "colorblind" computer. *Fla. St. UL Rev.*, 47:617, 2019.

Kamala D. Harris, Patty Murray, and Elizabeth Warren. Letter to U.S. Equal Employment Opportunity Commission, 2018. URL https://www.scribd.com/embeds/388920670/content#from_embed.

Deborah Hellman. Measuring algorithmic fairness. *Va. L. Rev.*, 106(4):811, 2020.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

Benjamin E Hermalin and Michael L Katz. Moral hazard and verifiability: The effects of renegotiation in agency. *Econometrica*, 59(6):1735–1753, 1991.

Mireille Hildebrandt. Profiling: From data to knowledge. *Datenschutz und Datensicherheit-DuD*, 30(9):548–552, 2006.

Bengt Holmström. Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182, 1999.

Bengt Holmström and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, 55(2):303–328, 1987.

Bengt Holmström and Paul Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, 7:24, 1991.

Johannes Hörner and Nicolas S Lambert. Motivational ratings. *The Review of Economic Studies*, 2020.

Kimberly Houser. Can AI solve the diversity problem in the tech industry? mitigating noise and bias in employment decision-making. *Stanford Technology Law Review*, 22:290, 2019.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*, pages 1389–1398, 2018.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*, pages 259–268, 2019.

Amy E. Hurley-Hanson and Cristina M. Giannantonio. Journal of business management. In *Autism in the Workplace*, volume 22, 2016.

URL https://www.chapman.edu/business/_files/journals-and-essays/jbm-editions/jbm-vol-22-no-1-autism-in-the-workplace.pdf.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58. ACM, 2019.

Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing, FORC 2020*, volume 156, pages 2:1–2:11, 2020.

Svante Janson. Tail bounds for sums of geometric and exponential variables. *Statistics Probability Letters*, 135:1–6, 2018.

Josh Jarrett and Sarah Croft. The science behind the Koru model of predictive hiring for fit. Technical report, Koru, 2018.

Michael C Jensen and William H Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of financial economics*, 3(4):305–360, 1976.

Harry Joe. Inequalities for random utility models, with applications to ranking and subset choice data. *Methodology and computing in Applied Probability*, 2(4): 359–372, 2000.

Stefanie K Johnson, David R Hekman, and Elsa T Chan. If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, April 2016.

Christine Jolls and Cass R. Sunstein. The law of implicit bias. *California Law Review*, 94:969–996, 2006.

Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, 2016.

Daniel Kahneman and Amos Tversky. Intuitive prediction: Biases and corrective procedures. Technical report, Decisions and Designs Inc Mclean Va, 1977.

Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Margot E Kaminski. The right to explanation, explained. *Berkeley Tech. LJ*, 34: 189–218, 2019.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *International Conference on Computer Control and Communication*, 2009.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*, pages 2231–2241, 2018.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *The 23rd*

International Conference on Artificial Intelligence and Statistics, AISTATS 2020, volume 108 of *Proceedings of Machine Learning Research*, pages 895–905. PMLR, 2020.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning*, pages 1828–1836, 2017.

William F Kemble. Testing the fitness of your employees. *Industrial Management*, 1916.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93, 1938.

Steven Kerr. On the folly of rewarding A, while hoping for B. *Academy of Management journal*, 18(4):769–783, 1975.

Lina M Khan and David E Pozen. A skeptical view of information fiduciaries. *Harv. L. Rev.*, 133:497, 2019.

Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. Preference-Informed Fairness. In *11th Innovations in Theoretical Computer Science Conference*, volume 151, pages 16:1–16:23, 2020.

Pauline T Kim. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58: 857, 2016.

Pauline T Kim. Auditing algorithms for discrimination. *U. Pa. L. Rev. Online*, 166:189, 2017.

Pauline T Kim. Big data and artificial intelligence: New challenges for workplace equality. *U. Louisville L. Rev.*, 57:313, 2018.

Pauline T Kim. Manipulating opportunity. *Virginia Law Review*, 106(4):867–935, 2020.

Barbara Kiviat. *Prediction and the Moral Order: Contesting Fairness in Consumer Data Capitalism*. PhD thesis, Harvard University, 2019.

René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *Ethics in Artificial Intelligence in Education*. Taylor & Francis, Forthcoming.

Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. In *Innovations in Theoretical Computer Science*, volume 94, pages 33:1–33:17, 2018.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, volume 67, pages 43:1–43:23, 2017.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2019.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80:245–272, 2010.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Daniel Koretz, Robert Linn, Stephen Dunbar, and Lorrie Shepard. The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. In *American Educational Research Association and the National Council on Measurement in Education*, 1991.

Daniel M Koretz. *Measuring up*. Harvard University Press, 2008.

Robin SS Kramer and Robert Ward. Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology*, 63(11): 2273–2287, 2010.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122:988–1012, 2014.

Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 2017.

Jean-Jacques Laffont and David Martimort. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, 2009.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981, 2019.

John Langford and Tong Zhang. The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

California State Legislature. Fair employment and housing act, 1959.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International World Wide Web Conference (WWW)*, 2010.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57, 2018.

Zachary C Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. Calibrated fairness in bandits. *CoRR arXiv:1707.01875, also appeared at the Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

Manuel Lopez and James Marengo. An upper bound for the expected difference between order statistics. *Mathematics Magazine*, 84(5):365–369, 2011.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2012.

Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences. *International Conference on Machine Learning*, 2011.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959.

George F Madaus and Marguerite Clarke. The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data. Technical report, ERIC, 2001.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3384–3393, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Rahul Makhijani and Johan Ugander. Parametric models for intransitivity in pairwise rankings. In *The World Wide Web Conference*, pages 3056–3062, 2019.

Gianclaudio Malgieri and Giovanni Comandé. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*, 7:243–265, 2017.

Henrick John Malik. Exact moments of order statistics from the pareto distribution. *Scandinavian Actuarial Journal*, 1966(3-4):144–157, 1966.

Colin L Mallows. Non-null ranking models. I. *Biometrika*, 44(1/2):114–130, 1957.

Farhad Manjoo. This summer stinks. But at least we've got 'Old Town Road.'. *New York Times Opinion*, 2019.

Charles F Manski. The structure of random utility models. *Theory and decision*, 8(3):229, 1977.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *ACM Conference on Economics and Computation (ACM EC)*, 2015.

Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning under competition. In *Innovations in Theoretical Computer Science (ITCS)*, 2018.

Andrew Mariotti. Talent acquisition benchmarking report. Technical report, Society for Human Resource Management, 2017. URL https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2017-Talent-Acquisition-Benchmarking.pdf.

David Martens and Foster Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–100, 2014.

R Preston McAfee and John McMillan. Bidding for contracts: a principal-agent analysis. *The RAND Journal of Economics*, 17(3):326–338, 1986.

Michael A Mcdaniel, Sven Kepes, and George C Banks. The uniform guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, 4(4):494–514, 2011.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Isak Mendoza and Lee A Bygrave. The right not to be subject to automated decisions based on profiling. In *EU Internet Law*, pages 77–98. Springer, 2017.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*, pages 230–239, 2019.

John P Mills. Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika*, pages 395–400, 1926.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

Adair Morse and Karen Pence. Technological innovation and discrimination in

household finance. Technical report, National Bureau of Economic Research, 2020.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM, 2020.

Hugo Munsterberg. *Psychology and industrial efficiency*, volume 49. A&C Black, 1998.

Isabel Briggs Myers. *The Myers-Briggs type indicator*. Consulting Psychologists Press, 1962.

B K Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

National Research Council. *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. The National Academies Press, 1989.

National Research Council. *New directions in assessing performance potential of individuals and groups: Workshop summary*. The National Academies Press, 2013.

David Neumark, Roy J Bank, and Kyle D Van Nort. Sex discrimination in restaurant hiring: An audit study. *The Quarterly journal of economics*, 111(3):915–941, 1996.

New York City Council. A local law to amend the administrative code of the city of New York, in relation to the sale of automated employment decision tools, 2020. URL https://legistar.council.nyc.gov/

LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 2005.

Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

Warren T Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

Ziad Obermeyer and Sendhil Mullainathan. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 89–89. ACM, 2019.

Nizan Geslevich Packin and Yafit Lev-Aretz. On social credit and the right to be unnetworked. *Columbia Business Law Review*, pages 339–425, 2016.

Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva. Crowdsourcing exploration. *Management Science*, 64(4):1727–1746, 2018.

Christina Passariello. Tech firms borrow football play to increase hiring of women. *Wall Street Journal*, 27 September 2016.

Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48. ACM, 2019.

Mark V Pauly. The economics of moral hazard: comment. *American Economic Review*, 58(3):531–537, 1968.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *KDD*, 2008.

Eduardo Perez-Richet and Vasiliki Skreta. Test design under falsification. Working paper, 2018.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3): 61–74, 1999.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

Jon Porter. UK ditches exam results generated by biased algorithm after student protests. *The Verge*, 2020.

JF Power and RF Follett. Monoculture. *Scientific American*, 256(3):78–87, 1987.

Julia Powles and Helen Nissenbaum. The seductive diversion of 'solving' bias in artificial intelligence. *Medium.com*, 2018.

Ruchir Puri. Mitigating bias in AI models. *IBM Research Blog*, 2018.

Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. Meta-analysis of field experiments shows no change in racial discrimination in hir-

ing over time. *Proceedings of the National Academy of Sciences*, 114(41):10870–10875, 2017.

Stephen Ragain and Johan Ugander. Pairwise choice markov chains. In *Advances in Neural Information Processing Systems*, pages 3198–3206, 2016.

Manish Raghavan. Testimony on New York City int. no. 1894, 2020.

Manish Raghavan and Solon Barocas. Challenges for mitigating bias in algorithmic hiring. *Brookings*, 2019.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Vaughan Wortman, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory*, pages 1724–1738. PMLR, 2018.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *AAAI/ACM Conf. on AI Ethics and Society*, 2019.

McKenzie Raub. Bots, bias and big data: Artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Ark. L. Rev.*, 71:529, 2018.

Regulation B. 12 c.f.r. § 1002 et seq.

Jeffrey H Reiman. Privacy, intimacy, and personhood. *Philosophy & Public Affairs*, pages 26–44, 1976.

Lauren Rhue. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*, 2018.

Peter A Riach and Judith Rich. Field experiments of discrimination in the market place. *The economic journal*, 112(483):F480–F518, 2002.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

Philippe Rigollet and Assaf Zeevi. Nonparametric Bandits with Covariates. In *Conference on Learning Theory (COLT)*, 2010.

John Roach. Microsoft improves facial recognition technology to perform well across all skin tones, genders. *The AI Blog*, 2018.

David Rodina and John Farragut. Inducing effort through grades. Working paper, 2016.

Michael C Rodriguez and Yukiko Maeda. Meta-analysis of coefficient alpha. *Psychological methods*, 11(3):306, 2006.

Stephen A Ross. The economic theory of agency: The principal's problem. *American Economic Review*, 63(2):134–139, 1973.

Edward Ruda and Lewis E Albright. Racial differences on selection instruments related to subsequent job performance. *Personnel Psychology*, 1968.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

Chris Russell, Matt J Kusner, Joshua R Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems 30. Pre-proceedings*, 30, 2017.

Eduardo Salas. Reply to request for public comment on plan for retrospective analysis of significant regulations pursuant to executive order 13563, 2011.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

Michael R Sampford. Some inequalities on Mill's ratio and related functions. *The Annals of Mathematical Statistics*, 24(1):130–132, 1953.

Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to solve the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2020.

Heinz Schuler, James L Farr, and Mike Smith. *Personnel selection and assessment: Individual and organizational perspectives*. Psychology Press, 1993.

Andrew D Selbst. A new HUD rule would effectively encourage discrimination by algorithm. *Slate*, 2019.

Andrew D. Selbst and Solon Barocas. The intuitive appeal of explainable machines. *Fordham Law Review*, 87:1085, 2018.

Andrew D. Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

Andrew D. Selbst, danah boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68. ACM, 2019.

Senate Report No. 94-589, 1976.

Hamza Shaban. What is the "Rooney Rule" that Uber just adopted? *Washington Post*, 13 June 2017.

Yonadav Shavit, B Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Elaine W Shoben. Differential pass-fail rates in employment testing: Statistical proof under Title VII. *Harvard Law Review*, pages 793–813, 1978.

Elaine W Shoben. In defense of disparate impact analysis under Title VII: A reply to Dr. Cohn. *Ind. LJ*, 55:515, 1979.

Jim Sidanius and Marie Crane. Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, 19(2):174–197, 1989.

Jennifer Skeem, Nicholas Scurich, and John Monahan. Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior*, 44:51–59, 2020.

Society for Industrial and Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. American Psychological Association, 2018.

Daniel J Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154:477–560, 2006.

Michael Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374, 1973.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

Megan Stevenson. Assessing risk assessment in action. *Minn. L. Rev.*, 103:303, 2018.

Joseph E Stiglitz. Incentives and risk sharing in sharecropping. *The Review of Economic Studies*, 41(2):219–255, 1974.

David Strauss. Some results on random utility models. *Journal of Mathematical Psychology*, 20(1):35–52, 1979.

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42, 2019.

Ina Taneva. Information design. *American Economic Journal: Microeconomics*, 11 (4):151–85, 2019.

Wei Tang, Chien-Ju Ho, and Yang Liu. Linear models are robust optimal under strategic behavior. In *Proceedings of The 24th International Conference on Arti-*

ficial Intelligence and Statistics, volume 130 of *Proceedings of Machine Learning Research*, pages 2584–2592. PMLR, 13–15 Apr 2021.

Prasanna Tantri. Fintech for the poor: Financial intermediation without discrimination. *Review of Finance*, 2020.

Lewis Madison Terman. *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Houghton Mifflin, 1916.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4): 273, 1927.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

Francesco Giacomo Tricomi and Arthur Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math*, 1(1):133–142, 1951.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Leona E Tyler. *The psychology of human differences.* D Appleton-Century Company, 1947.

Tom R. Tyler. *Why people obey the law*. Princeton University Press, 2006.

Eric Luis Uhlmann and Geoffrey L. Cohen. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16(6):474–480, 2005.

U.S. Congress. Civil rights act, 1964.

U.S. Congress. Americans with disabilities act, 1990.

U.S. Congress. Civil rights act, 1991.

US Department of Housing and Urban Development. Charge of discrimination, HUD v. Facebook, 2018. URL https://archives.hud.gov/news/2019/HUD_v_Facebook.pdf.

US Department of Housing and Urban Development. HUD's implementation of the fair housing act's disparate impact standard, 2019. URL https://www.federalregister.gov/documents/2019/08/19/2019-17542/huds-implementation-of-the-fair-housing-acts-disparate-impact-standard.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM, 2019.

Linda van den Bergh, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W. Holland. The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Education Research Journal*, 47(2):497–527, 2010.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GPDR. *Harv. Journal of Law & Technology*, 31(2):841–887, 2018.

Christine Wenneras and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387:341–343, 1997.

David R. Williams and Selina A. Mohammed. Discrimination and racial disparities in health: Evidence and needed research. *J. Med. Behav.*, 32(1), 2009.

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

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.

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

John W Young. Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis. Research report no. 2001-6. *College Entrance Examination Board*, 2001.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, pages 609–616, 2001.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *World Wide Web Conference*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970, Fort Lauderdale, FL, USA, 20–22 Apr 2017b. PMLR.

Andriy Zapechelnyuk. Optimal quality certification. *American Economic Review: Insights*, 2(2):161–76, 2020.

Tal Z Zarsky. Law and online social networks: Mapping the challenges and promises of user-generated information flows. *Fordham Intell. Prop. Media & Ent. LJ*, 18(3):741–783, 2008.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578. ACM, 2017.

Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Zhibing Zhao, Tristan Villamil, and Lirong Xia. Learning mixtures of random utility models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Dawei Zhou, Jiebo Luo, Vincent MB Silenzio, Yun Zhou, Jile Hu, Glenn Currier, and Henry Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Malte Ziewitz. Rethinking gaming: The ethical work of optimization in web search engines. *Social studies of science*, 49(5):707–731, 2019.