ABSTRACT

Title of Dissertation: EXPLORING DIVERSITY AND FAIRNESS IN MACHINE LEARNING

Candice Schumann,
Doctor of Philosophy, 2020

Dissertation Directed by: Professor John P. Dickerson
Department of Computer Science

With algorithms, artificial intelligence, and machine learning becoming ubiquitous in our society, we need to start thinking about the implications and ethical concerns of new machine learning models. In fact, two types of biases that impact machine learning models are social injustice bias (bias created by society) and measurement bias (bias created by unbalanced sampling). Biases against groups of individuals found in machine learning models can be mitigated through the use of diversity and fairness constraints. This dissertation introduces models to help humans make decisions by enforcing diversity and fairness constraints.

This work starts with a call to action. Bias is rife in hiring, and since algorithms are being used in multiple companies to filter applicants, we need to pay special attention to this application. Inspired by this hiring application, I introduce new multi-armed bandit frameworks to help assign human resources in the hiring process while enforcing diversity through a submodular utility function. These frameworks increase diversity while using less resources compared to original admission decisions of the Computer Science graduate program at the University of Maryland. Moving outside of hiring I present a contextual multi-armed bandit algorithm that enforces group fairness by learning a societal bias term and correcting for it. This algorithm is tested on two real world datasets and shows marked improvement over other in-use algorithms. Additionally I take a look at fairness in traditional machine learning domain adaptation. I provide the first theoretical analysis of this setting and test the resulting model on two deal world datasets. Finally I explore extensions to my core work, delving into suicidality, comprehension of fairness definitions, and student evaluations.

EXPLORING DIVERSITY AND FAIRNESS IN MACHINE LEARNING

by

Candice Schumann

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor John P. Dickerson, Chair/Advisor
Professor Jeffrey S. Foster
Professor Hal Daumé III
Dr. Alex Beutel
Professor Stuart N. Vogel

Dedication

For Mow, Midi, Maggie, Max, and Mom.

Acknowledgements

First and foremost I want to thank my advisor John P. Dickerson for the unbelievable support during my PhD. Thank you to my intern hosts Alex Beutel, Jilin Chen, and Susanna Ricco. All of this work would not be possible without my collaborators Jeffrey S. Foster, Samsara N. Counts, Willy Lang, Debjani Saha, Duncan C. McElfresh, Nicholas Mattei, Michelle L. Mazurek, Michael Carl Tschantz, Xuezhi Wang, Hai Qian, and Ed H. Chi.

Table of Contents

Chapter 1: Introduction

Algorithms and machine learning models, including decisions made by these models, are becoming ubiquitous in our daily lives. These algorithms and models are not cold, hard, rational, decision-making machines. Instead, as pointed out by Cathy O'Neil in Weapons of Math Destruction, "the math-powered applications powering the data economy were based on choices made by fallible human beings" [175]. Indeed, as stated by psychiatrist M. Scott Peck, "human beings are poor examiners, subject to superstition, bias, prejudice, and a PROFOUND tendency to see what they want to see rather than what is really there." [179].

From the perspective of a computer scientist working on these "math-powered applications" we can view the world in three layers (Figure 1.1). The top layer is the world as it could be - where everyone is treated equally with no dependence on race, age, gender, or orientation. The second layer is the world as it is currently. In between these two layers we find social injustice bias. This social injustice bias is born from inequality, societal biases, superstitions, prejudices, unconscious biases, and just the general unfairness of our world. Social injustice bias can be found throughout history, starting as early as 6500 BCE in Mesopotamian Ubaid [52], continuing with the slave trade in Africa [157], the treatment of Jews in the Holocaust [96], through modern times with stop-and-frisk in New York [215] and discrimination in hiring [191]. Finally, the last layer is the world as it is measured. Machine learning algorithms learn to make decisions from data. In between the world as it is (middle

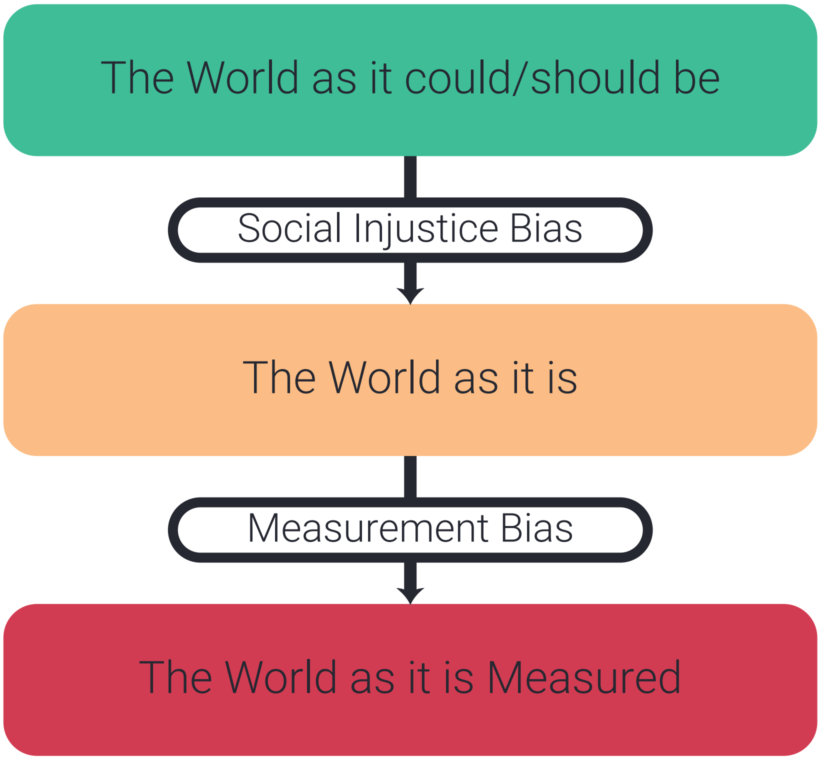


FIGURE 1.1: Three layered view of the world.

layer) and the world as it is measured, we have measurement bias. The term measurement bias here is used in the traditional sense of unbalanced data where one sensitive group is underrepresented. Every decision made when creating a dataset can further imbalance which datapoints are collected. Unbalanced datasets can be found in image datasets [212], natural language datasets [223], and even bug-fix datasets [34]. Of course, this layered view of the world is simplified. In reality measurement bias is influenced by societal bias, societal bias can be magnified given the world as it is measured, and the world as it could be is highly subjective. That being said, the three layered view of the world allows us to find solutions or at least mitigations to the problem of bias found in machine learning models.

Indeed, my thesis is as follows:

> Biases found in machine learning models, whether they are from social injustice or measurement bias, can be combated through the use of diversity objectives and fairness constraints.

Here, a diversity objective would favor a more diverse outcome while a fairness constraint would remove untenable solutions such that groups of individuals are not discriminated against. We do, however, need to keep human stakeholders in mind when making diversity objective and/or fairness constraint decisions. Technology should be built around human actors to help support them and push them in (hopefully) better directions. My dissertation work focuses on creating models to help humans make decisions that lead to more diversity in outcomes, arrived at in less biased ways. From helping to hire a diverse set of candidates, to actively learning societal bias and correcting for it, to transferring learned debiasing to new domains.

This dissertation does not necessarily follow my work linearly through time. Instead after setting up basic vocabulary and terminology in Chapter 2, Chapter 3 starts with a call to action for introducing fairness and diversity into automated hiring systems. The majority of my dissertation was inspired by the task of removing bias from hiring and admissions. This chapter lays out the groundwork for where we are in terms of hiring systems and supportive decision making technologies, and where we need to go. Chapter 4 sets up the foundation of a combinatorial multi-armed bandit algorithm that assigns interviewing and reviewing resources to help select a diverse cohort of applicants. I introduce different types of arm pulls that relate to reviewing and interviewing resources. A maximization oracle is then used to enforce diversity in the final cohort selection. Chapter 5 extends this idea of multi-armed bandits in hiring and admissions to a more realistic tiered interviewing setting where applicants move through a series of interview stages. When running simulations on a real world dataset of graduate admissions I find that algorithms provided in both Chapter 4 and Chapter 5 choose more diverse cohorts while using fewer resources than the original admissions decision. Chapter 6 steps away from hiring and combinatorial bandits to a contextual bandit approach. In this chapter I introduce a societal bias term that learns disparities between sensitive groups and corrects for the disparities when pulling arms. By including a societal bias term the algorithm outperforms general contextual mutli-armed bandit algorithms. Chapter 7 takes a theoretical look at fairness in machine learning in the application of domain adaptation. I take the first theoretical look at transferring fairness in domain adaptation and provide a modeling approach to transferring learned debiasing to a new domain. Finally Chapter 8 details extensions of my main research with applications in suicidality, comprehension of fairness, and student evaluations. This dissertation sets the groundwork for many more research directions, as discussed in Chapter 9. Each chapter deals with mitigating either social injustice bias or measurement bias.

Chapter 2: Preliminaries

2.1 Multi-Armed Bandits

The multi-armed bandit problem allows for modeling resource allocation during sequential decision making. Examples of practical applications of MAB algorithms include algorithms for selecting what advertisements to display to users on a webpage [165], systems for dynamic pricing [169], and content recommendation services [144]. Indeed, such ML-based decision-making systems continue to expand in scope, making ever more important decisions in our lives such as setting bail [68], making hiring decisions [39, 197], and policing [194]. Thus, the study of the properties of these algorithms is of tantamount importance [60].

2.1.1 Classical multi-armed bandits

The multi-armed bandit problem allows for modeling resource allocation during sequential decision making. Bubeck et al. [45] provide a general overview of historic research in this field. In a MAB setting there is a set of n arms A. Each arm has a true utility of $u(a) \in [0, 1]$, which is unknown [12, 137]. When an arm a is pulled, a reward is pulled from a distribution with mean $u(a)$ and a σ-sub-Gaussian tail and a cost of 1 is paid. These pulls give an empirical estimate $\hat{u}(a)$ of the underlying utility, and an uncertainty bound $rad(a)$ around the empirical estimate. With some probability δ we know that the true utility lies somewhere inside of the uncertainty bounds i.e., $\hat{u}(a) - rad(a) < u(a) < \hat{u}(a) + rad(a)$. Once arm a is pulled, $\hat{u}(a)$ and $rad(a)$ are updated. The goal of the agent is to maximize the collected reward over all timesteps, or to find the top arm. Therefore, the optimal strategy would be to pull the arm with the highest true utility $u(a)$ forever. In practice, since we do not know the true utilities we have to trade off exploring arms where we are uncertain of the true utility and exploiting arms that we know have high utilities.

2.1.2 Top-K multi-armed bandits

Classical multi-armed bandits are limited in that they only return the top arm. Recently, MAB formulations have been proposed that select an optimal subset of K arms.

$$M^* = \operatorname*{arg\,max}_{\{M \subset A \big| |M|=K\}} \sum_{a \in M} u(a) \tag{2.1}$$

Bubeck et al. [46] propose a budgeted algorithm (SAR) that successively accepts and rejects arms. Chen et al. [58] build on that work by generalizing SAR to a setting with a combinatorial objective. In the Chen et al. [58] formulation the overall goal is to choose an optimal cohort M^*, or subset of arms, from a decision class $\mathcal{M}$. They provide both a fixed confidence and a fixed budget algorithm. Cao et al. [51] tighten the bounds of Chen et al. [58] where the objective function is Top-K, defined as

$$w_{\text{TOP}}(\mathbf{u}, M) = \sum_{a \in M} u(a). \tag{2.2}$$

Locatelli et al. [153] address the thresholding bandit problem, finding the arms above and below threshold τ with precision ϵ. Jun et al. [124] look at the Top-K MAB problem with batch arm pulls and Singla et al. [204] look at the Top-K problem from a crowdsourcing point of view.

To select the best subset while satisfying a submodular function, Singla et al. [205] propose an algorithm maximizing an unknown function accessed through noisy evaluations. Radlinski et al. [186] learn a diverse ranking from the behavior patterns of different users and then greedily select the next document to rank. They treat each rank as a separate MAB instance, rather than our approach using a single MAB to model the whole system. Yue and Guestrin [231] introduce the *linear submodular bandits problem* to select diverse sets of content in an online learning setting for optimizing a class of feature-rich submodular utility models.

2.1.3 Sensitive groups

When dealing with fairness and/or diversity in a MAB setting we have a set of arms $a \in A$, such that each applicant is an arm a, and where A is partitioned into L groups $A = P_1 \cup P_2 \cup \cdots \cup P_L$ corresponding to specific sensitive attribute groups. These attributes could represent self-reported gender, race, and country of origin.

2.1.4 Variable cost multi-armed bandits

In many real-world settings, there are different ways to gather information, each of which vary in cost and effectiveness. Previous work uses stochastic costs in the MAB setting. However, our costs are fixed for specific types of arm pulls. Ding et al. [79] looked at a regret minimization MAB problem that has variable rewards and costs. When an arm is pulled a random reward is received and a random cost is taken from the budget. Xia et al. [228] extend this work to a batch arm pull setting.

Jain et al. [120] use MABs with variable rewards and costs to solve a crowdsourcing problem.

2.1.5 Contextual multi-armed bandits

A generalization of MAB is the contextual multi-armed bandit (CMAB) where the agent observes a d-dimensional *context* $x_{i,t} \in \mathcal{X}_i = \mathbb{R}^d$ for each arm $i \in A$, at each timestep t, to use along with the observed rewards of the arms played to choose a new arm [144]. In the CMAB problem the agent learns the relationship between contexts and rewards $f(x_i, t)$ and selects the best arm [3]. At a timestep t, let i^* denote the optimal arm that could be selected and i be the selected arm. Then, the regret for choosing arm i is

$$R(t) = f(x_{i^*,t}) - f(x_{a,t}). \tag{2.3}$$

2.2 Domain Adaptation

Both Pan et al. [177], and Weiss et al. [221] provide a survey on current work in transfer learning. One case of transfer learning is domain adaptation, where the task remains the same, but the distribution of features that the model is trained on (the source domain) does not match the distribution that the model is tested against (the target domain). Ben-David et al. [23] provide theoretical analysis of domain adaptation. Ben-David et al. [24] extend this analysis to provide a theoretical understanding of how much source and target data should be used to successfully transfer knowledge. Mansour et al. [164] provide theoretical bounds on domain adaptation using Rademacher Complexity analysis. In later research, Ganin et al. [92] build on this theory to use an adversarial training procedure over latent representations to improve domain adaptation.

2.3 Diversity

Quantifying the diversity of a set of elements is of interest to a variety of fields, including recommender systems, information retrieval, computer vision, and others [5, 11, 25, 151, 185, 186, 201]. A recent formalization from Lin and Bilmes [148] assumes that individuals can be split into L partitions where a partition is denoted as P_i and a cohort is defined as $M = P_1 \cup P_2 \cup \ldots \cup P_L$. At a high level, the diversity function w_{DIV} is defined as

$$w_{\text{DIV}}(M) = \sum_{i=1}^{L} \sqrt{\sum_{a \in P_i} u(a)}. \tag{2.4}$$

Lin and Bilmes [148] showed that w_{DIV} is submodular and monotone. Under $w_{\text{DIV}}(M)$ there is typically more benefit to selecting an arm from a class that is not already represented in the cohort, if the empirical utility of an arm is not substantially low. As soon as an arm is selected from a class, other arms from that class experience diminishing gain due to the square root function. Example 2.1 illustrates when w_{DIV} results in a different cohort selection than the top-K function $w_{\text{TOP}}(M) = \sum_{a \in M} u(a)$.

Example 2.1. *Assume we have three individuals*$\{a_1, a_2, a_3\}$ *with true utilities* $u(a_1) =$

0.6, $u(a_2) = 0.5$, and $u(a_3) = 0.3$. Assume there exist $L = 2$ classes, and let individuals a_1 and a_2 belong to class 1, and individual a_3 belong to class 2. Then, for a cohort of size $K = 2$, w_{TOP} will select cohort $M^*_{\text{TOP}} = \{a_1, a_2\}$, while w_{DIV} will select cohort $M^*_{\text{DIV}} = \{a_1, a_3\}$. Indeed, $w_{\text{TOP}}(M^*_{\text{TOP}}) = 1.1 > 0.9 = w_{\text{TOP}}(M^*_{\text{DIV}})$, while $w_{\text{DIV}}(M^*_{\text{TOP}}) = \sqrt{1.1} \approx 1.05 < 1.3 \approx \sqrt{0.6} + \sqrt{0.3} = w_{\text{DIV}}(M^*_{\text{DIV}})$.

Ashkan et al. [11] define two other diversity functions which look for coverage of a group. The first is the token membership function where only one member of a sensitive group needs to be selected. The second function is a quota diversity function where each group i has a quota requirement of N_i.

Each of these diversity functions look at different definitions of human concepts of diversity. w_{DIV} may work well in the hiring setting where you want a diverse set of individuals, while w_{token} may work well in document summarization where you only really need one example for each type of document. w_{quota} has the potential to work in both settings.

2.4 Fairness

Fairness in machine learning has become one of the most active topics in computer science [60]. The idea of using formal notions of fairness, i.e. axioms or properties, to design decision schemes has a long history in economics and political economy [188, 230]. Their work underscores that in many cases statistical parity is not sufficient to ensure individual fairness, as we may treat groups fairly but in doing so may be very unfair to some specific individual. Determining when, how, and if to define fairness is an ongoing discussion with roots well before the time of computer science [206]; indeed, it is known that many natural conditions for fairness cannot be achieved in tandem [91, 133]. Still, group fairness is found in many fielded systems [22, 222], and we focus on it in this dissertation.

When looking at group fairness we typically look at some sensitive attribute $\mathcal{A}$ such as gender, race, region of origin, sexual orientation, and others. A model is unfair in terms of group fairness when outcomes differ depending on the membership to the sensitive group. For example, a sentiment classifier may be considered unfair towards LGBTQ words if the model consistently assigns a negative sentiment label on a sentence that contains an LGBTQ word. In current research there are three major fairness metrics: demographic parity, equal opportunity, and equalized odds [103].

Demographic Parity A classifier is said to be fair under demographic parity if the probability of assigning a positive does not change across sensitive group membership. For instance if we have a binary sensitive attribute then,

$$\Pr(\hat{Y} = 1|A = 0) = \Pr(\hat{Y} = 1|A = 1). \tag{2.5}$$

Equal opportunity A classifier is said to be fair under equal opportunity if the false positive rates do not change across sensitive group membership. If we have a binary sensitive attribute then,

$$\Pr(\hat{Y} = 1|A = 0, Y = 0) = \Pr(\hat{Y} = 1|A = 1, Y = 0). \tag{2.6}$$

Equalized odds Equality of odds is similar to equal opportunity with the additional constraint of the false negative rates being the same across sensitive group membership. If we have a binary sensitive attribute then,

$$\Pr(\hat{Y} = -Y | A = 0, Y = y) = \Pr(\hat{Y} = -Y | A = 1, Y = y) \; \forall y \in \{-1, 1\}. \tag{2.7}$$

2.4.1 Fairness in Machine Learning

A large thread of recent research has studied how to optimize for fairness metrics during model training. Li et al. [146] empirically show that adversarial learning helps preserve privacy over sensitive attributes. Beutel et al. [30] focus on using adversarial learning to optimize different fairness metrics, and Madras et al. [162] provides a theoretical framework for understanding how adversarial learning optimizes these fairness goals. Zhang et al. [234] use adversarial training over logits rather than hidden representations. Other work has focused on constraint-based optimization of fairness objectives [2, 97]. Tsipras et al. [214] however, provide a theoretical bound on the accuracy of adversarial robust models. They show that even with infinite data there will still be a trade-off of accuracy for robustness. Kallus and Zhou [126] look at fairness in personalization when sensitive attributes are missing. Similarly, Chen et al. [57] look at measuring disparity when sensitive attributes are unknown.

2.4.2 Domain Adaptation & Fairness

Despite the prevalence of using one model across multiple domains, in practice little work has studied domain adaptation and transfer learning of fairness metrics. Coston et al. [72] look at domain adaptation for fairness where sensitive attribute labels are not available in both the source and target domains. Kallus and Zhou [125] use covariate shift correction when computing fairness metrics to address bias in label collection. More related, Madras et al. [162] show empirically that their method allows for fair transfer. The transfer learning here corresponds to preserving fairness for a single sensitive attribute but over different tasks. However, Lan and Huan [138] found empirically that fairness does not transfer well to a new domain. They found that as accuracy increased in the transfer process, fairness decreases in the new domain. It is concerning that these papers show opposing effects. Both of these papers offer empirical results on the UCI adult dataset, but neither provide a theoretical understanding of how and when fairness in one domain transfers to another.

2.4.3 Fairness in MAB

The study of fairness in bandits was initiated by Joseph et al. [123], who showed for both classical and contextual bandits that one can implement a fairness definition where within a given pool of applicants (say, for college admission or mortgages), a worse applicant is not favored over a better one, despite a learning algorithm's uncertainty over the true payoffs. However, Joseph et al. [123] only focus on individual fairness, and do not formally treat the idea of group fairness. Individual fairness is, in some sense, group fairness taken to an extreme, where every arm is its own singleton group; it offers strong guarantees, but under equally strong assumptions [33, 129].

Celis et al. [53] propose a bandit-based approach to personalization where arm pulls are constrained to fit some probability distribution defined by a fairness metric such as demographic parity. For example, when recommending news articles their algorithm provides personalized articles from both left and right sources. Their formulation is perhaps closest in the literature to our formulation in Chapter 6 as it deals with group fairness, however it does not explicitly assume biased feedback. Instead it enforces a fair probability distribution without learning about the bias present in the data.

There are a number of other recent studies of fairness in the MAB literature. Liu et al. [152] look at fairness between arms under the assumption that arm reward distributions are similar (another interpretation of equal treatment of equals). Patil et al. [178] define fairness such that each arm must be pulled for a predetermined required fraction over the total available rounds. Claure et al. [62] use the MAB framework to distribute resources amongst teammates in human-robot interaction settings; again, fairness is defined as a pre-configured minimum rate that each arm must be pulled.

There is also significant recent work in constrained reasoning in the MAB setting. Balakrishnan et al. [17] study the idea of learning constraints over pulling arms by observation in a pre-training phase. Wu et al. [226] study constraints in both number of pulls per arm, as well as number of rounds where arms are available to be pulled. Wu et al. [227] study a different flavor of constrained bandits where the learned policy cannot fall below a certain threshold; modeling the case where one wants to explore, but not suffer too much of a penalty over a status-quo policy. A related and perhaps interesting direction for future work is the work on bandits that are budget-constrained (without fairness considerations). Ding et al. [79] study budget-constrained bandits where each arm also has an unknown cost distribution and one must learn a policy that maximizes reward and minimizes cost.

Fairness in bandits is a particularly important area as the online, dynamic nature makes the task challenging and the use of bandits in a number of areas makes the problem particularly relevant. The motivating factor for group fairness is that one does not want to cause disparate impact, or the idea that groups should be treated differently based only on non-relevant aspects [87]. Indeed, discrimination in certain areas including housing, credit, and jobs is forbidden in the US by the Civil Rights Act of 1965. It is specifically in these areas where bandit algorithms are deployed: advertising (where discrimination has been found), [210] college admissions (Chapter 4), and interviewing (Chapter 5).

2.5 Machine Learning in Hiring and Admissions

Lux et al. [158] and Waters and Miikkulainen [220] use supervised learning to model admissions decisions. They develop accurate classifiers; none decide how to allocate interviewing resources or maximize a certain objective, unlike our aim to select a more diverse cohort via a principled semi-automated system.

The behavioral science literature shows that scoring candidates via the same rubric, asking the same questions, and spending the same amount of time are interviewing best practices [10, 105, 196, 224]. Such *structured interviews* reduce bias and provide better job success predictors [143, 184]. We incorporate these results into our model through our assumption that we can spend the same budget and get the same information gain across different arms.

Chapter 3: We Need Fairness and Explainability in Algorithmic Hiring

Algorithms and machine learning models, including the decisions made by these models, are becoming ubiquitous in our daily life, including hiring. We make no value judgment regarding this development; rather, we simply acknowledge that it is quickly becoming reality that automation plays a role in hiring. Increasingly, these technologies are used in all of the small decisions that make up the modern hiring pipeline: from which resumes get selected for a first screen to who gets an on site interview. Thus, these algorithms and models may potentially amplify bias and (un)fairness issues for many historically marginalized groups. While there is a rapidly expanding literature on algorithmic decision making and fairness, there has been limited work on fairness specifically for online, multi-stakeholder decision making processes such as those found in hiring. We outline broad challenges including formulating definitions for fair treatment and fair outcomes in hiring, and incorporating these definitions into the algorithms and processes that constitute the modern hiring pipeline.

3.1 Introduction

“Hiring is rarely a single decision point, but rather a cumulative series of small decisions.” So begins a recent report on *automated hiring processes* released by the non-profit group UpTurn [38], before recommending that digital sourcing firms begin explicitly addressing concerns of fairness and bias at every step of the hiring process. Indeed, at various decision points in the hiring process, algorithms already determine who sees which job advertisements; estimate the expected performance of an applicant; select which applicants to screen more heavily and with whom to match them; and forecast salary and other benefits necessary to ensure a successful offer. Thus, issues of bias or fairness at one stage of this procedure may lead to unexpected or amplified issues at a later stage of the process.

In addition to the difficulty of these decisions on their own, there are a number of regulatory and legal requirements that must be met at each stage of the hiring process. As a recent Facebook settlement[1] showed, the tools, platforms, and techniques developed to streamline hiring can be subtly—or blatantly—illegal. These requirements are complicated by the presence of multiple stakeholders: governmental regulators, hiring managers, employees, line managers, and myriad others involved in modern hiring and employment.

While one can argue that we may not need algorithmic hiring, the fact is that platforms and websites such as LinkedIn, ZipRecruiter, and Indeed are making these tools available to businesses of any size, and that large businesses are experimenting

[1] https://www.propublica.org/article/facebook-ads-discrimination-settlement\-housing-employment-credit

or have experimented with automated hiring techniques.[2]Thus, algorithmic processes are being deployed in the real-world, and it is incumbent on computer science researchers to ensure that the algorithms we create are aware of both fairness and legal compliance for these processes. There is already ample evidence from the areas of lending and pre-trial detention (bail) and policing that the algorithms that are deployed can have significant, and sometimes harmful, impacts on individuals lives [68]. There is a need for novel techniques from data science, artificial intelligence, and machine learning to ensure our algorithms act within the constraints set forth by business process, laws, social norms, and ethical guidelines [192].

One shortcoming of current research into algorithmic fairness is its focus on a *single* decision point [68]. As depicted in Figure 3.1, modern hiring is rarely a single step process [38]. It is the culmination of a series of steps, much like pre-trial detention and other decisions of consequence, and we currently lack the algorithmic tools and techniques to adequately address this challenge. Techniques developed to address these challenges can also be applied to many settings where we have a "prioritization funnel" setting, such as customer acquisition or government sourcing.

We argue for concentrated research around the thesis that:

> *Data-driven approaches to measuring and promoting fairness and explainability to each of the concerned stakeholders at a single stage of the hiring process can be extended—in a principled way—to the full, multi-stage hiring process.*

[2] https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against

It is important to note that the application of research in this area will not just be in the hiring scenario. The techniques developed here, along with a number of results in peer evaluation [13] and other areas of social choice including matching [42], will enable the creation of algorithmic tools that are both fair and efficient. These tools can and should be deployed in any situation where we are attempting to select a set of candidates (or items, or interventions) from a large pool or allocate other scarce resources, subject to various constraints over the selection and reviewing process [192]. These technologies could be applied to internal product ideation and review [219], academic proposal reviewing [107], advertisement/campaign selection [145], or indeed any setting where we need to collect recommendations over a large set from experts.

We detail the limits of current research into fairness and its shortcomings with respect to the challenge of algorithmic hiring. We detail both past and current work that demonstrates the research communities potential impact in the area. Finally, we close with additional ideas we see as research directions for the community.

3.2 Fairness in Online, Multi-Stage Decision-Making Algorithms

Within computer science, economics, and operations research circles many of the problems that are encountered in hiring are typically modeled in the *multi-armed bandit (MAB)* setting [209]. Indeed, bandit-based algorithms have received significant attention in the literature for their use in content recommendation [144], advertising, and hiring [39](See Chapter 4).

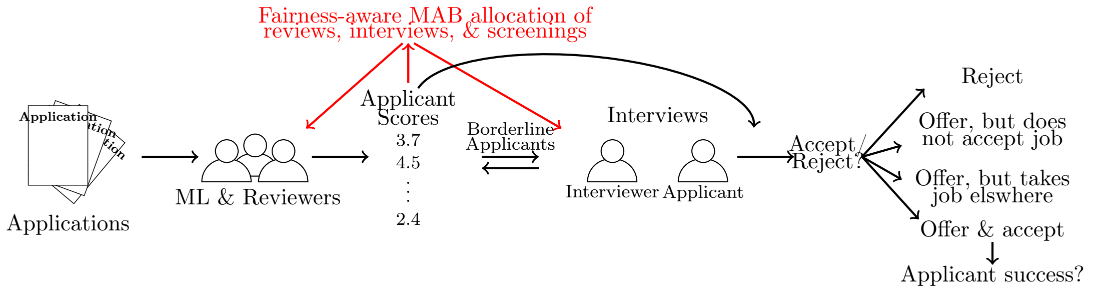


FIGURE 3.1: A sample current tiered hiring process (in black) and interventions proposed by this blue sky submission (in red).

There are many practical applications of MAB algorithms that are making ever more important decisions in our lives (See Section 2.1). Thus, the study of the properties of these algorithms is of tantamount importance [60].

Yet, the use of MAB-based systems often results in behavior that is societally repugnant. Sweeney [210] noted that queries for public records on Google resulted in different associated contextual advertisements based on whether the query target had a traditionally African American or Caucasian name; in the former case, advertisements were more likely to contain text relating to criminal incidents. In the years following, similar instances continue to be observed, both in the bandit setting and in the more general machine learning world [175]. In lockstep, the academic community has begun developing approaches to tackling issues of (un)fairness in a variety of learning settings.

Recently, a Computing Community Consortium (CCC) whitepaper on fairness research specifically identified that most studies of fairness are focused on classification problems [60]. Two fundamental issues identified by Chouldechova and Roth [60] that we believe are unaddressed by the current literature are extensions to notions of *group fairness* and looking at fairness in *online, dynamic systems*, e.g., the contextual bandit setting. We envision the research community addressing these gaps by formalizing and providing algorithms for myriad definitions of fairness and bias. We see the following research communities specifically as both sources of ideas and nexuses for collaboration around fairness in sequential decision making.

Markets and Game Theory. Mechanism and market design are both interested in fairness towards the agents that participate. We see the game theory community as being particularly helpful when it comes to analyzing the incentives at play among classes of stakeholders in the hiring process, e.g., competing firms, or a single firm and a single candidate, or hiring managers within a firm.

Learning and Adaptation. There have been numerous MAB papers recently that also deal with humans/crowdsourcing [187], fairness and diversity (See Chapter 4), and/or incorporating biased human feedback [211], to name just a few. Hence, we feel that the Learning and Adaptation community is able to help with this core topic.

Coordination, Organizations, and Norms. Many of the algorithmic hiring systems are both learning *from* and interacting *with* with mutliple stakeholders including hiring managers, line managers, and employees, in real time. The systems are are making decisions in environments with multiple competing interests. Much like Markets and Game Theory, researchers in this area will be key in advancing this overall agenda. Furthermore, we believe research into multi-stage fairness could more closely tie together all three of these research communities.

3.3 Fairness in the Hiring Process

The pipeline of a typical algorithmic hiring process is depicted in Figure 3.1. In this process, a set of applications is screened by either humans, algorithms, or a combination of both. After this initial screening and selection, applications are scored/ranked and many are discarded. After this an iterative process of allocating resources, e.g., requests for additional documentation; online or in-person interviews; and group discussion are committed to refine the initial ranking. After this, offers and/or rejections are sent to one or more candidates from the pool and the candidate provides a response.

We are proposing a focused research plan into a data-driven decision support process that draws inferences in part based on observed and estimated features of humans—and such tools are increasingly known to result in unexpected or adverse impact on dimensions such as fairness and bias [175]. We acknowledge that both our and others' initial work in this space, as well as our proposed extension to the more realistic multi-stage selection setting, may exacerbate issues of *fairness*. Thus, we also propose to incorporate recent definitions of fairness from the machine learning community into our tiered model. Such definitions do not fully capture the needs or wants of practitioners [110]; yet, we believe developing systems that are amenable to general definitions of fairness will be useful, because those definitions are evolving, and will continue to evolve, over time. In our exploratory work, we adopt a subset of the standard notions of fairness, and we perform analysis on real admissions data (As discussed in Chapters 4 and 5); still, much work remains to align systems to be fielded with the aggregate preferences of stakeholders.

It is important to ensure that the entire pipeline is capable of recognizing *fair* *treatment* and/or *fair outcome* (and possibly others) in the multi-armed bandit setting [118]. We have begun work in this direction, described in more detail in Chapter 6. We re-emphasize that, throughout, our models will be built to accept a host of fairness and parity measures; still, it is important to provide concrete plans for specific definitions of each.

We note that notions of "fairness," "bias," and "explainability" are (i) definable in many ways [68] and (ii) necessarily different based on application areas, societal norms, and policy-maker preferences. However, in hiring, credit, and housing there are a number of federally protected features that one must not use in the decision making process and also must not use for explanation. Simply removing these features from consideration by our algorithms is not enough, and we must actively ensure the fairness criteria is enforced across these features [49]. Thus, we endeavor to remain somewhat definition-agnostic in our modeling work, and then explicitly instantiate a definition when needed (e.g., we plan to use the well-known *equality of opportunity* [103] definition of fairness in our earliest experiments). However, our proposed approaches should generalize to a whole host of fairness or parity measures, so long as the measure of bias/fairness can be written as a linear constraint on conditional moments of predicted distributions over predictions, ground truth, and protected attributes [2].

A closely related area to our work is the research into fairness in rankings [203], multi-stakeholder recommender systems [1], and item allocation [25, 26]. When algorithms return rankings for an individual to select from one must pay attention to the ordering and the positioning of various groups [203]. One can see this as an application of the group fairness concept to the slates that are chosen for display. A particular aspect of recommendation systems that one needs to keep in mind is that often there are different stakeholders: the person receiving the recommendation, the company giving the recommendation, and the businesses that are the subjects of recommendation [1]. Finally, when goods are allocated, such as housing or subsidies one may need to observe both individual and group fairness [25]. Indeed, group fairness is specifically important in, e.g., Singapore, which has specifically enforced notions of group fairness when allocating public housing [26].

3.4 A First Step: An Initial Framework to Model "Fair" Tiered Hiring

In Chapter 5, we use a multidimensional approach to tackling issues in the efficient and fair *gathering* and *aggregation* of information by hiring managers, which jointly compose part of a decision support system for potential job offer *decisions*. We use the concept of *structured interviews* [43, 217], used widely in industry as well as in some academic programs (e.g., Fisk-Vanderbilt [208]); and develope a tiered-hiring MAB approach. Figure 3.1 gives an example tiered hiring process, and shows (in red text) where our proposed interventions fit into the present hiring system.

The presently-developed methods allow for the promotion of diversity in the final cohort of applicants (e.g., graduate students). Dovetailing with this, the **fairness of the review process** is also important. In the MAB setting, we propose that the research community build on work in incorporating constraints into the MAB framework [17] and extend this work with methods from the fairness in machine learning literature [19, 60] such as those developed within the silos of fairness of treatment and fairness of outcome. Of particular value would be merging these criteria into the single-level and multi-tiered settings, exploring theoretical metrics such as the impact on overall economic efficiency due to the use of a “fair” objective, and experimental validation on sensitive attributes such as self-reported gender, race, and country of origin that are available in our real data sets.

3.5 Blue Sky Research Challenges

As noted earlier, we are *not* making a value judgment regarding the use of automated systems in hiring; rather, we note that this is, increasingly, reality. We are also *not* making value judgments regarding particular definitions of fairness and/or bias in machine learning. Our goal here is to develop *general* and *principled* systems for tiered hiring that can incorprate *many* definitions of fairness.

We are working on extending our current research to incorporate different notions of fairness that could be deployed on a number of already-fielded MAB-based systems (See Chapter 6). We plan to extend these definitions to a tiered model (See Chapter 5) and investigate theoretically the “price of fairness” [28] in these systems.

This initial work may close the gap on a single point (the hiring), but there is still much work to be done. Some of our initial research has addressed questions of transparency, constraints, and fairness when working with multi-armed bandit algorithms (See Chapters 4 AND 6) [16, 17, 118]. Yet, these are small steps taken toward a larger research goal. We see the following issues as still omnipresent concerns, ripe for work by researchers from the research community.

1. How should we allocate effort—e.g., budget, interview slots—along the hiring

pipeline? While we have begun to address this gap there are still challenges that remain. Included in these challenges is maintaining notions of diversity at every stage of the pipeline, and not just at discrete points.

2. How can we explain the decisions made by the complete algorithmic process in a transparent and compliant way? With (inter-)national regulation like the newly-established GDPR [218] and the right to object and right to rectification, we need to build pipelines for decisions that are not only fair but capable of being audited.

3. How can we incorporate fairness into other automated screening tools that we are beginning to see? For instance, chatbots are starting to be used to gather pre-interview data with clients and the need to address concerns around usability and access are almost completely untouched.

4. How do we chose which features to select when building models for hiring? Which features are predictive, which are not, and which are protected? While the UpTurn study [39] states that employers should disclose all relevant features, the selection of these features is a ethically-laden decsion. While there has been recent work in this area [172] further exploration is necessary.

5. There has been extensive recent work in budget-limited and other constrained bandit models including limiting rounds [226], policy thresholds [227], and unknown, budget constrained cost distributions [79]. Exploring models with resource and budget constraints necessary for the hiring process is an important direction.

6. So far, we have assumed individuals have fixed group membership and that these group memberships do not overlap. Generalizing fairness definitions to work for intersectional fairness and settings where memberships in protected groups may change at every timestep t would fit more real world applications. One step forward might leverage results from work on bandits with non-stationary rewards [29]. Additionally, other group fairness definitions such as Equalized Opportunity should be converted to the MAB setting [103].

7. Algorithmic transparency to the end user is important, as discussed, but equally important is maintaining human involvement in the training, validation, and deployment process. We conjecture (and sincerely hope!) that no hiring process will become entirely automated—so we must ensure that the algorithms and systems we build are capable of working with, potentially biased, human input at every stage.

8. In our previous work (see Chapters 4 and 5) we explored an objective that balances both individual utility and the diversity of the set of arms returned. Research has shown that a more diverse workforce produces better products and increases productivity [78, 113]. Thus, such an objective is of interest to our application of hiring workers. Note that diversity, while related, is distinct from fairness. Trying to balance both diversity and fairness should be looked at more deeply since both diversity and fairness are important in the hiring process.

9. We need a new definition of *fair outcomes for the MAB setting*. Typically,

equality of opportunity fairness is used in classification tasks. We can formulate a strict definition of equal opportunity for bandits, but a hard constraint may be too strict a definition, or may not align with the expressed preferences of stakeholders. Instead, it may be necessary to define notions of fairness that straddle the line between individual and (sub-)community [130]. And, indeed, it may be necessary to balance notions of fairness and economic efficiency across both sides of the market, so as to promote truthful participation of both firms and workers in this ubiquitous and increasingly automated process.

3.6 Authors and Publication

This chapter was written by Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. It was published at the International Conference on Autonomous Agents and Multi-Agent Systems 2020 in the Blue Sky track.

Chapter 4: The Diverse Cohort Selection Problem

How should a firm allocate its limited interviewing resources to select the optimal cohort of new employees from a large set of job applicants? How should that firm allocate cheap but noisy resume screenings and expensive but in-depth in-person interviews? We view this problem through the lens of combinatorial pure exploration (CPE) in the multi-armed bandit setting, where a central learning agent performs costly exploration of a set of arms before selecting a final subset with some combinatorial structure. We generalize a recent CPE algorithm to the setting where arm pulls can have different costs and return different levels of information. We then prove theoretical upper bounds for a general class of arm-pulling strategies in this new setting. We apply our general algorithm to a real-world problem with combinatorial structure: incorporating diversity into university admissions. We take real data from admissions at one of the largest US-based computer science graduate programs and show that a simulation of our algorithm produces a cohort with hiring overall utility while spending comparable budget to the current admissions process at that university.

> *"It should come as no surprise that more diverse companies and institutions are achieving better performance." – McKinsey & Company,* Diversity Matters *(2015)*

4.1 Introduction

How should a firm, school, or fellowship committee allocate its limited interviewing resources to select the optimal cohort of new employees, students, or awardees from a large set of applicants? Here, the central decision maker must first form a belief about the true quality of an applicant via costly information gathering, and then select a subset of applicants that maximizes some objective function. Furthermore, various types of information gathering can be performed—reviewing a resume, scheduling a Skype interview, flying a candidate out for an all-day interview, and so on—to gather greater amounts of information, but also at greater cost.

In this work, we model the allocation of structured interviewing resources and subsequent selection of a cohort as a combinatorial pure exploration problem in the multi-armed bandit (MAB) setting. Here, each applicant is an arm, and a decision maker can *pull* the arm, at some cost, to receive a noisy signal about the underlying quality of that applicant. We further model two different levels of interviews as *strong* and *weak* pulls—the former costing more to perform than the latter, but also resulting in a less noisy signal. We introduce the strong-weak arm-pulls (SWAP) algorithm, generalizing an algorithm by Chen et al. [58], and provide theoretical upper bounds for a general class of our various arm-pull strategies. To complement these bounds, we provide simulation results comparing pulling strategies on a toy problem that mimics our theoretical assumptions.

We then validate our proposed method on a real-world scenario: admitting an optimal cohort of graduate students. We take recent data from one of the largest US-based Computer Science graduate programs—applications including recommendation letters, statements of purpose, transcripts, as well as the department's reviews of applications and final admissions decisions—and run experiments comparing our algorithm's performance under a variety of assumptions to reviews and decisions made in reality. We find that our simulation of SWAP produced a cohort with higher top-K utility using equivalent resources as in practice.

We also explore the empirical performance of our algorithm optimizing a nonlinear objective function, motivated by the real-world scenario of admitting a diverse cohort of graduate students. In experiments, our simulations of SWAP increased a diversity score (over gender and region of origin) with little loss in fit using roughly the same amount of resources as in practice. This gain suggests that SWAP can serve as a useful decision support tool to promote diversity in practice.

4.2 Problem Formulation

We now formally describe the stochastic multi-armed bandit setting in which we operate. For exposition's sake, we do so in the context of a decision-maker reviewing a set of job applicants. However, the formulation itself is fully general. Following the classical MAB formulation defined in Section 2.1, we represent a set of n applications A as arms $a_i \in A$ for $i \in [n]$. Each arm has a true utility, $u(a_i) \in [0, 1]$, which is unknown; an empirical estimate $\hat{u}(a_i) \in [0, 1]$ of that underlying true utility; and an uncertainty bound $rad(a_i)$. Once arm a_i is pulled (e.g., application reviewed or applicant interviewed), $\hat{u}(a_i)$ and $rad(a_i)$ are updated.

The set of potential *cohorts*, or subsets of arms, is defined by a decision class $\mathcal{M} \subseteq 2^{[n]}$. Note that $\mathcal{M}$ need not be the power set of arms, but can include cardinality and other constraints. The total utility for a cohort is given by some linear function $w : \mathbb{R}^n \times \mathcal{M} \to \mathbb{R}$ that takes as input the (unknown) true utilities $u(\cdot)$ of the arms and the selected cohort. Throughout the chapter, we assume a maximization oracle, defined as $\mathit{Oracle}(\mathcal{M}) = \arg\max_{M \in \mathcal{M}} w(M)$, where $\mathbf{v} \in \mathbb{R}^n$ is a vector of weights—in this case, estimated or true utilities for each arm. Our overall goal is to accurately estimate the true utilities of arms and then select the optimal subset of arms using the maximization oracle.

Problem hardness. Following the notation of Chen et al. [58], we define a *gap* score for each arm. For each arm a that is in the optimal cohort M^*, the gap is the difference in optimality between M^* and the best set without a. For each arm a that is not in the optimal set M^*, the gap is the sub-optimality of the best set that includes a. Formally, the gap is defined as

$$\Delta_a = \begin{cases} w(M^*) - \max_{M \in \mathcal{M}: a \in M} w(M), & \text{if } a \notin M^* \\ w(M^*) - \max_{M \in \mathcal{M}: a \notin M} w(M), & \text{if } a \in M^*. \end{cases} \tag{4.1}$$

This gap score serves as a useful signal for problem hardness, which we use in our theoretical analysis. Formally, the hardness of the problem can be defined as the sum of inverse squared gaps

$$\mathbf{H} = \sum_{a \in A} \Delta_a^{-2}. \tag{4.2}$$

Chen et al. [58] defined the concept of $\mathit{width}(\mathcal{M})$. When comparing all combinations of two sets $A, A' \in \mathcal{M}$, where $A \neq A'$, define $\mathit{dist}(A, A') = |A - A'| + |A' - A|$. Therefore, define $\mathit{width}(\mathcal{M}) = \min_{\{A, A' | A, A' \in \mathcal{M} \wedge A \neq A'\}} \mathit{dist}(A, A')$. In other words, the width is the smallest distance between any two sets in $\mathcal{M}$. See Chen et al. for an in-depth explanation of $width(\mathcal{M})$.

Strong and weak pulls. In reality, there is more than one way to gather information or receive rewards. Therefore, we introduce two kinds of arm pulls which vary in cost j and information gain s. Information gain s is defined as how sure one is the reward is close to the true utility. We model the information gain as s parallel arm pulls with the resulting rewards being averaged together. A *weak arm pull* has cost $j = 1$ but results in a small amount of information $s = 1$. In our domain of graduate admissions, weak arm pulls are standard application reviews, which involve reading submitted materials and then making a recommendation. A *strong arm pull*, in contrast, has cost $j > 1$, but results in $s > 1$ times the information as a weak arm pull. In our domain, strong arm pulls extend reading submitted materials with a structured Skype interview, followed by note-taking and a recommendation.

In our experience, the latter can reduce uncertainty considerably, which we quantify and discuss in Section 4.4. However, due to their high cost, such interviews are allocated relatively sparingly. We formally explore this problem in Section 4.3 and provide an algorithm for selecting which arms to pull, along with nonasymptotic upper bounds on total cost.

4.3 SWAP: An Algorithm for Allocating Interview Resources

In this section, we propose a new multi-armed bandit algorithm, strong-weak arm-pulls (SWAP), that is parameterized by s and j. SWAP uses a combination of strong and weak arm pulls to gain information about the true utility of arms and then selects the optimal cohort. Our setting and the algorithm we present generalize the CLUCB algorithm proposed by Chen et al. [58], which can be viewed as a special case with $s = j = 1$.

Algorithm 1 gives pseudocode for SWAP. It starts by weak pulling all arms once to initialize an empirical estimate of the true underlying utility of each arm. It then iteratively pulls arms, chooses to weak or strong pull based on a general strategy, updates empirical estimates of arms, and terminates with the optimal (i.e., objective-maximizing) subset of arms with probability $1-\delta$, for some user-supplied parameter δ.

During each iteration t, SWAP starts by finding the set of arms M_t that, according to current empirical estimates of their means, maximizes the objective function via an oracle. It then computes a confidence radius, $rad_t(a)$, for each arm a and estimates the worst-case utility of that arm with the corresponding bound. If an arm a is in the set M_t then the worst case is when the true utility of a is less than our estimate (a might not be in the true optimal set M^*). Alternatively, if an arm is not in the set M_t then the worst case is when the true utility of a is greater than our estimate (a might be in the true optimal set M^*). Using the worst-case estimates, SWAP computes an alternate subset of arms $\tilde{M}_t$. If the utility of the initial set M_t

and the worst-case set $\tilde{M}_t$ are equal, then SWAP terminates with output M_t, which is correct with probability $1 - \delta$ as we show in Theorems 4.1 and 4.2. If $w(M_t)$ and $w(\tilde{M}_t)$ differ, SWAP looks at a set of candidate arms in the symmetric difference of M_t and $\tilde{M}_t$ and chooses the arm p_t with the largest uncertainty bound $rad_t(p_t)$.

SWAP then chooses to either strong or weak pull the selected arm p_t using a *strong pull policy*, depending on parameters s and j. A strong pull policy is defined as $spp : \mathbb{R} \geq 1 \times (\mathbb{R} \geq 1) \rightarrow [0, 1]$. For example, in the experiments in Section 4.4, we use the following pull policy:

$$spp(s, j) = \frac{s - j}{s - 1}. \tag{4.3}$$

This policy tries to balance information gain and cost. When the strong pull gain is high relative to cost then many more strong pulls will be performed. When the weak pull gain is low relative to cost then fewer strong pulls will be performed, as discussed in Example 4.1.

Once an arm is pulled, the empirical mean $\hat{u}_{t+1}(p_t)$ and the information gain $T_{t+1}(p_t)$ is updated. A reward from a strong arm is counted s times more than a weak pull.

Example 4.1. *Suppose we wish to find a cohort of size $K = 2$ from three arms $A = \{a_1, a_2, a_3\}$. Run SWAP for t iterations. Figure 4.1 shows that SWAP maintains empirical utilities $\hat{u}_t(\cdot)$ and uncertainty bounds $rad_t(\cdot)$. In this case $M = \{a_1, a_2\}$ and $\tilde{M} = \{a_1, a_3\}$. Arm a_3, therefore, is the arm in the symmetric difference $\{a_2, a_3\}$ with the highest uncertainty, which therefore needs to be pulled. Further, assume that a_3 needs x information gain for SWAP to end. When $j = 1$ and $s = 1$, the best*

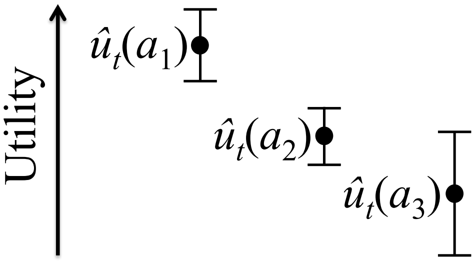


FIGURE 4.1: Example with $n = 3$ after running SWAP for t steps. Dots are the empirical utility $u_t(a)$ while flags represent the radius of confidence $rad_t(a)$. Here, $rad_t(a_2)$ and $rad_t(a_3)$ overlap; SWAP may pull a_3.

pulling strategy would be to weak pull a_3 for x times. When $j = 1$ and $s = y$ where $y > 1$, the best pulling strategy would be to strong pull a_3 for ceil$(\frac{x}{y})$ times. Finally when $j = z$ and $s = y$ where $y > z > 1$, the best pulling strategy would be to strong pull a_3 for floor$(\frac{x}{y}) + \mathbf{1}[z - (x \mod y)]$ times and weak pull a_3 for $\mathbf{1}[z - (x \mod y)] * (x \mod y)$ times, where $\mathbf{1}[a] = 1$ when $a \geq 0$ and 0 otherwise. In reality, we do not know how many times an arm needs to be pulled, which is why we introduce a probabilistic strong pull policy, like that in Equation 4.3.

Analysis. We now formally analyze SWAP. We define $\bar{X}_{Cost} = E[Cost]$ as the expected cost (or expected j value) and $\bar{X}_{Gain} = E[Gain]$ as the expected gain (or the expected s value). Assume that each arm $a \in [n]$ has mean $u(a)$ with an σ-sub-Gaussian tail. Following Chen et al. [58] set $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4nCost_t^3}{\delta}\right)/T_t(a)}$ for all $t > 0$.

Notice that if we use strong pull policy $spp(s, j) = 0$, then we only perform weak arm pulls, and SWAP reduces to Chen et al. [58]'s CLUCB. We call this reduction the *weak only pull problem.* Chen et al. proved that CLUCB returns the optimal set M^* and uses at most $\tilde{O}(width(\mathcal{M})^2\mathbf{H})$ samples. Similarly, if we set $spp(s, j) = 1$

then we only perform strong arm pulls—dubbed the *strong only pull problem.* We show that this version of SWAP returns the optimal set M^* and costs at most $\tilde{O}(width(\mathcal{M})^2\mathbf{H}/s)$.[1]

Theorem 4.1. *Given any $\delta \in (0,1)$, any decision class $\mathcal{M} \subseteq 2^{[n]}$, and any expected rewards $\mathbf{u} \in \mathbb{R}^n$, assume that the reward distribution φ_a for each arm $a \in [n]$ has mean $u(a)$ with an σ-sub-Gaussian tail. Let $M^* = \arg\max_{M\in\mathcal{M}} w(M)$ denote the optimal set. Set $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4nt^3j^3}{\delta}\right)/T_t(a)}$ for all $t > 0$ and $a \in [n]$. Then, with probability at least $1-\delta$, the SWAP algorithm with only strong pulls where $j \geq 1$ and $s > j$ returns the optimal set $\texttt{Out} = M^*$ and*

$$T \leq O\left(\frac{\sigma^2 width(\mathcal{M})^2\mathbf{H}\log(nj^3\sigma^2\mathbf{H}/\delta)}{s}\right) \tag{4.4}$$

where T denotes the total cost used by the SWAP algorithm and $\mathbf{H}$ is defined in Eq.4.2.

Although s and j are problem-specific, it is important to know when to use the strong only pull problem over the weak only pull problem. Corollary 4.1.1 provides weak bounds for s and j for the strong only pull problem. We also explore its ramifications experimentally in Figure 4.3a as discussed in Section 4.4.1.

Corollary 4.1.1. *SWAP with only strong pulls is equally or more efficient than SWAP with only weak pulls when $s > 0$ and $0 < j \leq C^{\frac{s}{3}-\frac{1}{3}}$ where $C = 4n\tilde{\mathbf{H}}/\delta$.*

We now address the general case of SWAP, for any probabilistic strong pull policy parameterized by s and j. In Theorem 4.2 we show that SWAP returns M^* in $\tilde{O}\left(width(\mathcal{M})^2\mathbf{H}/\bar{X}_{Gain}\right)$ samples.

[1] Note all proofs for this chapter can be found in Appendix A.3

Theorem 4.2. *Given any $\delta_1, \delta_2, \delta_3 \in (0,1)$, any decision class $\mathcal{M} \subseteq 2^{[n]}$, and any expected rewards $\mathbf{u} \in \mathbb{R}^n$, assume that the reward distribution φ_a for each arm $a \in [n]$ has mean $u(a)$ with an σ-sub-Gaussian tail. Let $M^* = \arg\max_{M \in \mathcal{M}} w(M)$ denote the optimal set. Set $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4n\,Cost_t^3}{\delta}\right)/T_t(a)}$ for all $t > 0$ and $a \in [n]$, set $\epsilon_1 = \sigma\sqrt{2\log\left(\frac{1}{2}\delta_2/T\right)}$, and set $\epsilon_2 = \sigma\sqrt{2\log\left(\frac{1}{2}\delta_3/n\right)}$. Then, with probability at least $(1-\delta_1)(1-\delta_2)(1-\delta_3)$, the SWAP algorithm (Algorithm 1) returns the optimal set* $\texttt{Out} = M^*$ *and*

$$T \leq O\left(\frac{\sigma^2 width(\mathcal{M})^2 \mathbf{H} \log\left(n\sigma^2\left(\bar{X}_{Cost} - \epsilon_1\right)^3 \mathbf{H}/\delta_1\right)}{\bar{X}_{Gain} - \epsilon_2}\right), \tag{4.5}$$

where T denotes the total cost used by Algorithm 1, and $\mathbf{H}$ is defined in Eq. 4.2.

It is nontrivial to determine where the general version of SWAP is better than both the SWAP algorithm with only strong pulls and the SWAP algorithm with only weak pulls, given the non-asymptotic nature of all three bounds (Chen et al. results and Theorems 4.1 and 4.2). Based on our experiments (§4.4), we conjecture that there is a of s and j pairs where SWAP is the optimal algorithm, even for relatively low numbers of arm pulls, though it is problem-specific. This is discussed more in Section 4.6.3.

4.4 Top-K Experiments

In this section, we experimentally validate the SWAP algorithm under a variety of arm pull strategies. We first explore (§4.4.1) the efficacy of our bounds in Theorem 4.2 and Corollary 4.1.1 in simulation. Then we deploy SWAP on real data

(§4.4.2) drawn from one of the largest computer science graduate programs in the United States. We show that SWAP provides a higher overall utility with equivalent cost to the actual admissions process.

4.4.1 Gaussian Arm Experiment

We begin by validating the tightness of our theoretical results in a simulation setting that mimics the assumptions made in Section 4.3. We pull from a Gaussian distribution around each arm. When arm a is weak pulled, a reward is pulled from a Gaussian distribution with mean u_a, the arm's true utility, and standard deviation σ. Similarly, when arm a is strong pulled, the algorithm is charged j cost, and a reward is pulled from a distribution with mean u_a and standard deviation $\sigma/\sqrt{s}$. This strong pull distribution is equivalent to pulling the arm s times and averaging the reward, thus ensuring an information gain of s.

We ran all three algorithms—SWAP with the strong pull policy defined in Equation 4.3, SWAP with only strong pulls, and SWAP with only weak pulls—while varying s and j. For each s and j pair we ran the algorithms at least $4,000$ times with a randomly generated set of arm values. Random seeds were maintained across policies. We then compared the cost of running each of the algorithms.[2]

To test Corollary 4.1.1, Figure 4.3a compares SWAP with only weak pulls to SWAP with only strong pulls. We found that Corollary 4.1.1 is a weak bound on the boundary value of j. The general version of SWAP should be used when it performs better—costs less—than both the strong only and weak only versions of SWAP. The zone where SWAP is effective varies with the problem (See §4.6.3 for

[2] All code to replicate this experiment can be found here: https://github.com/principledhiring/SWAP.

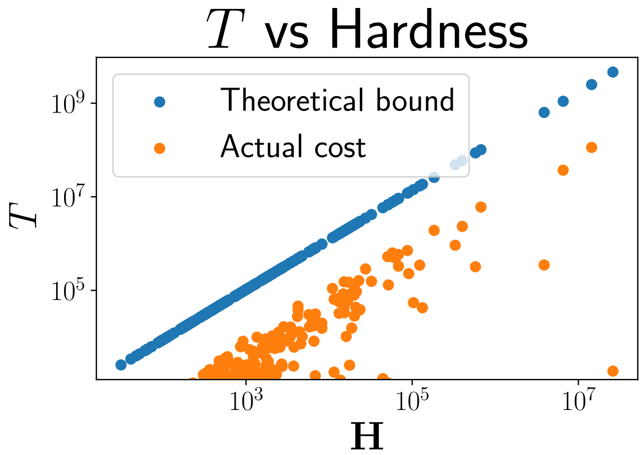
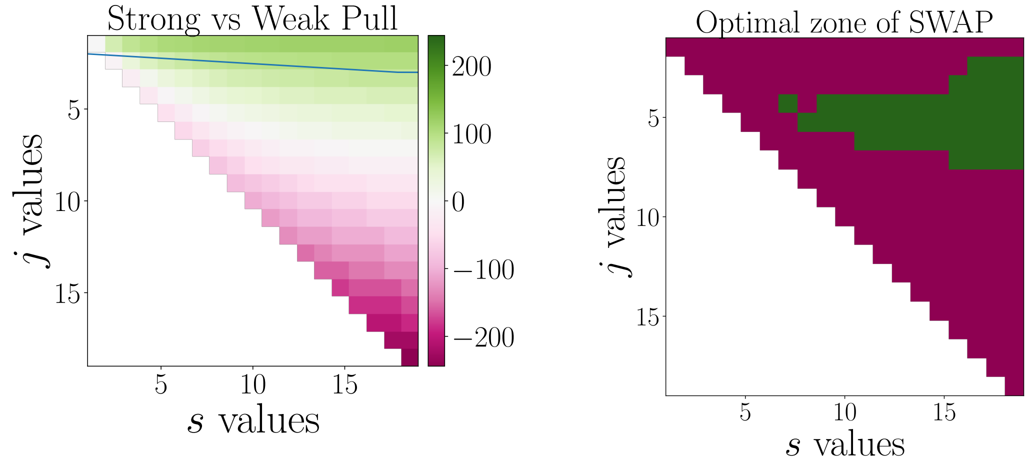


FIGURE 4.2: Exploration of bounds in practice vs. the theoretical bounds of Theorem 4.2 with respect to hardness (note that both axes are a log scale).

a deeper discussion). Figure 4.3b shows the optimal zone for the Gaussian Arm Experiment.

4.4.2 Graduate Admissions Experiment

Finally, we describe a preliminary exploration of SWAP on real graduate admissions data from one of the largest CS graduate programs in the United States. The experiment was approved by the university's Institutional Review Board. Our dataset consists of three years of graduate admissions applications, graduate committee application review text and ratings, and final admissions decisions. Information was gathered from the first two academic years (treated as a training set), while the data from last academic year was used to evaluate the performance of SWAP (treated as a test set).



(A) Weak vs Strong

(B) SWAP Optimal Zone

FIGURE 4.3: Cost comparisons. Figure 4.3a compares only strong to only weak pulls. Green indicates better performance by strong pulls, and intensity indicates magnitude. The blue line is the Corollary 4.1.1 bound on j. Figure 4.3b shows where the general version of SWAP outperformed (green) both SWAP with only strong pulls as well as SWAP with only weak pulls, and (maroon) where it outperformed at least one of the latter.

Dataset. During the admissions process, potential students from all over the world send in their applications. A single application consists of quantitative information such as GPA, GRE scores, TOEFL scores, nationality, gender, previous degrees and so on, as well as qualitative information in the form of recommendation letters and statements of purpose. In the 2016-17 academic year, the department received approximately 1,600 applications, with roughly 4,500 applications over all three years. The most recent 1,600 applications are roughly split into 1,000 Master's applications and 600 Ph.D. applications. The acceptance rate is 3% for Masters students and 20% for Ph.D. students.

Once all applications are submitted, they are sent to a review committee. Generally, applicants at the top (who far exceed expectations) and applicants at the bottom (who do not fulfill the program's strict requirements) only need one review. Applicants on the boundary, however, may go through multiple reviews with different committee members. Once all reviews have been made, the graduate chair chooses the final applicants to admit.

By administering an anonymous survey of past admissions committee members, we estimated that interviews are approximately six times longer than reviewing a written application. Therefore, we set our j value (the cost of a strong pull) to be 6. The gain of an interview is uncertain, so we ran tests over a wide range of s values (the information gain of a strong pull). The number of reviews and interviews ($\times 6$) were summed to get a cost T of the actual review process.

Experimental Setup. We simulate an arm pull by returning a real score that a reviewer gave during the admissions process (in the order of the original reviews) or a score from a probabilistic classifier (if all committee members' reviews have

	w	T
SWAP	80.1 (0.5)	1978 (53)
Actual	73.96	~2000

Table 4.1: Graduate Admissions Simulation of SWAP. Comparison of top-K utility w and cost T of SWAP with results of the actual admissions process. The values in parentheses are the standard deviations.

been used). An arm pull returns a score drawn from a distribution around the probabilistic result from the classifier to simulate some human error or bias.

We ran SWAP using the strong pull policy defined in Eq. 4.3, where we define the utility of each arm by the probabilistic result from the classifier. For our results, we compare SWAP's selections with the real decisions made during the admissions process.

Results. Running SWAP consistently resulted in a higher overall utility than the actual admissions process while using roughly equivalent cost (Table 4.1). We see that the overall top-K utility w is higher in SWAP than in practice. We also see that SWAP uses roughly equivalent resources T than what is used in practice. This suggests that SWAP is a viable option for admissions. There are, however, some limitations of only using a top-K policy, such as potentially overlooking the value diverse candidates bring to a cohort. For instance, when hiring a software engineering team, if the top candidates are all back-end developers, it may be worthwhile to hire a front-end developer with slightly lower utility.

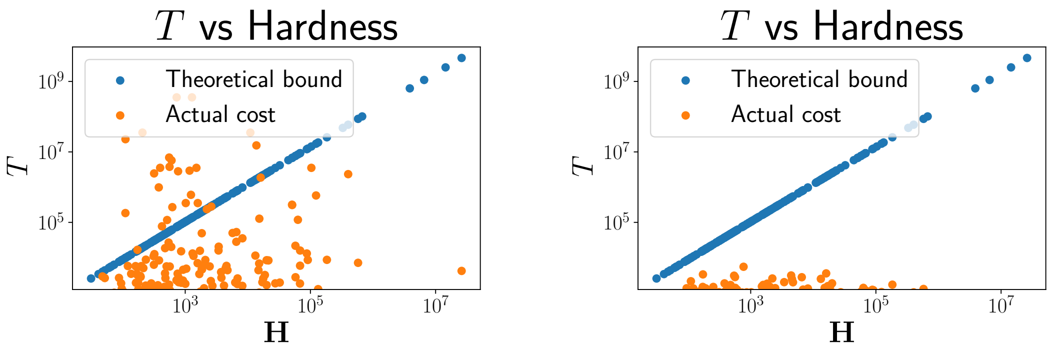
4.5 Promoting diversity through a submodular function

Motivated by recent evidence that diversity in the workforce can increase productivity [78, 113], we explore the effect of formally promoting diversity in the cohort selection problem. In this section we use the diversity function discussed in Section 2.3. Empirically, we show that SWAP performs well with a submodular objective function (Section 4.5.1). In experiments on real data, we show a significant increase in diversity with little loss in fit while using roughly the same resources as in practice (Section 4.5.2).

4.5.1 Diverse Gaussian Arm Experiments

To determine if SWAP works in this submodular setting, we ran simulations over a variety of hardness levels. We instantiated the problem similarly to that of Section 4.4.1 with the added complexity of dividing the arms into three partitions.

Figure 4.4a shows the cost of running SWAP compared to the theoretical bounds of the linear model over increasing hardness levels. The results show that SWAP performs well for the majority of cases. However, for some cases, the cost becomes very large. To deal with those situations, we can use a probably approximately correct (PAC) relaxation of Algorithm 1 where Line 13 becomes `If` $\left|w(\tilde{M}_t) - w(M_t)\right| \leq \epsilon$. The results from this PAC relaxation where $\epsilon = 0.01$ can be found in Figure 4.4b. Note that the definition of hardness found in Equation 4.2 does not quite fit this situation since the graphs in Figure 4.4 have higher costs for some lower hardness problems while having lower cost for some higher hardness problems. Given that the



(A) SWAP with w_{DIV}

(B) PAC relaxation with w_{DIV}

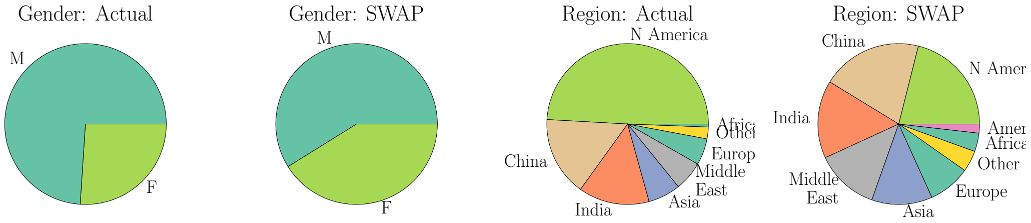
FIGURE 4.4: Exploration of bounds in practice for SWAP with w_{DIV} (4.4a) and the PAC relaxation of SWAP with w_{DIV} (4.4b) vs. the theoretical bounds of Theorem 4.2 with respect to hardness (Note that both axes are a log scale).

PAC relaxation performs well with low costs over all of the tested hardness problems, we propose that SWAP can be used with w_{DIV} and perhaps other submodular and monotone functions.

4.5.2 Diverse Graduate Admissions Experiment

Using the same setting as described in Section 4.4.2, we simulate a SWAP admissions process with the submodular function w_{DIV}. We partition groups by gender (which is binary in our dataset) and multi-class region of origin. We found that we did not have to resort to the PAC version of SWAP to tractably run the simulation over various partitions of the graduate admissions data.

Results. We compare two objective functions, w_{TOP} and w_{DIV}. w_{TOP} treats all applicants as members of one global class. This mimics a top-K objective, where applicants are valued based on individual merit alone. w_{DIV} promotes diversity using



(A) Actual (B) SWAP (C) Actual (D) SWAP

FIGURE 4.5: Comparison of true and SWAP-simulated admissions: gender (4.5a, 4.5b) & region (4.5c), 4.5d).

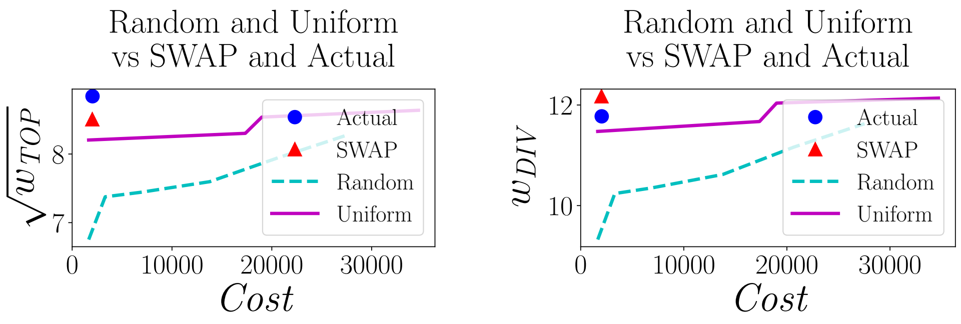
	Gender		Region of Origin	
	$\sqrt{w_{\text{TOP}}}$	w_{DIV}	$\sqrt{w_{\text{TOP}}}$	w_{DIV}
SWAP	8.5 (0.03)	12.1 (0.06)	8.0 (0.03)	22.1 (0.03)
Actual	8.6	11.8	8.6	20.47

TABLE 4.2: SWAP's average gain in diversity over different classes.

reported gender and region of origin for class memberships. We use those classes as our objective during separate runs of SWAP.

Table 4.2 and Figure 4.5 show experimental results on the test set (most recent year) of real admissions data. We report $\sqrt{w_{\text{TOP}}}$ instead of w_{TOP} to align units across objective functions. Because the square root function is monotonic, this conversion does not impact the maximum utility cohort. Since SWAP uses a diversity oracle, we notice a slight drop in top-K utility. However, there is a large gain in diversity.

SWAP, on average, used 1.17 pulls per arm, of which 5% were strong. During the last admissions decision process each applicant was reviewed on average 1.21 times. Interviews were not consistently documented. SWAP performed more strong pulls (interviews) of applicants than our estimation of interviews by the graduate admissions committee, but did fewer weak pulls. SWAP spent roughly the same amount of total resources as the committee did with strong pull cost $j = 6$ and



(A) $\sqrt{w_{\text{TOP}}}$

(B) w_{DIV}

FIGURE 4.6: Cost vs utility function comparisons of Actual, SWAP, Random, and Uniform.

weak pull cost of 1. Given the gains in diversity, this supports SWAP's potential use in practice.

We also compare SWAP to both uniform and random pulling strategies, shown in Figure 4.6. The uniform strategy weak pulls each arm once and strong pulls each arm once. This had a cost approximately 9 times that of SWAP and resulted in a general utility of 8.3 and a diversity value of 11.8. The random strategy weak or strong pulls arms randomly. Even when spending 10 times the cost of running SWAP, the random strategy has only a general utility of 7.9 and a diversity value of 11.16. SWAP significantly outperforms both of these strategies.

4.6 Discussion

Admissions and hiring are extremely important processes that affect individuals in very real ways. Lack of structure and systematic bias in these processes, present in application materials or in resource allocation, can negatively affect applicants from traditionally underrepresented minority groups. We suggest a formally structured process to help prevent disadvantaged people from falling through the cracks. We discuss benefits (Section 4.6.1) and limitations (Section 4.6.2) to this approach, as well as mechanism design suggestions for deploying SWAP in practice (Section 4.6.3).

4.6.1 Benefits

We established SWAP, a clear-cut way to model a sequential decision-making process where the aim is to select a subset using two kinds of information-gathering strategies as a multi-armed bandit algorithm. This process could have a number of benefits when used in practical hiring/admissions settings.

Over the course of designing and running our experiments, we noticed what seemed like bias in the application materials of candidates belonging to underrepresented minority groups. Our initial observations were similar to those of scholars such as Schmader et al. [195], who found that recommendation letters for female applicants to faculty jobs contained fewer work-specific terms than male applicants. After revisiting and coding application materials in our experiments, we found similar results for female and other minority candidates.

Our process hopes to mitigate this bias by providing a completely structured process, informed by the many studies showing that structured interviewing reduces bias (see Section 2.5). As we showed in our experiments, one can take additional steps to encourage diversity (by using w_{DIV}) to select a more diverse team, which can result in a less biased, more productive work environment [113].

Furthermore, by including a diversity measure in the objective function, candidates from disadvantaged groups are given a higher chance of being pulled through the cracks since we prioritize recommending diverse candidates for additional resource allocation.

A practical benefit to SWAP is that it avoids spending unnecessary resources on outlier candidates and quickly finds uncertain candidates. This give us more information about the applicant pool as whole, allowing us to make better decisions when choosing a cohort while using roughly equivalent resources.

Finally, in our simulations of running SWAP during the graduate admissions process, we also select a more diverse student cohort at low cost to cohort utility.

4.6.2 Limitations

One significant limitation of a large-scale system like SWAP is that it relies on having a utility score for each applicant. In our graduate admissions experiment, we assume the true utility of an applicant can be modeled by our classifier, which is not entirely accurate. In reality, the true utility of an applicant is nontrivial to estimate as it is subjective and depends on a wide range of factors. Finding an applicant's true utility would require following and evaluating the applicant through the end of the program, perhaps even after they have left the university. Even if that were possible, being able to quantify true utility is nontrivial due to the subjectivity of success and its qualitative properties. This problem is not limited to SWAP–it is present in any admissions, hiring, peer review, and other processes that attempt to quantify the value of qualitative properties. Therefore in these settings there is no choice but to rely on proxy values for the true utility, such as reviewer scores.

Similarly, even though the cost of a resource, j, may be inherently quantifiable, the information gain s, is harder to define in such a process. For example, how much more information one gains from an interview over a resume review is subjective and, by nature, more qualitative than quantitative. Also, the information gain from expending the same resource may vary over applicants, though this is slightly mitigated by using structured interviews.

Another limiting factor is that not every admitted applicant will matriculate into the program. We assume that all applicants will accept our offer, but in reality, that is not the case. Therefore, we potentially reject applicants that would matriculate, as opposed to accepting higher quality applicants that will ultimately not.

Finally, our graduate admissions experiment *simulated* strong arm pulls: reviewers did not give additional interviews of applicants during the experiment. Although our results are promising, SWAP should be run in conjunction with an actual admissions process to assess its true performance.

4.6.3 Design Choices

Our motivation in designing SWAP and exploring related extensions is to aid hiring and admissions processes that use structured interviewing practices and aim to hire a diverse cohort of workers. As with any algorithm deployed in practice, actually *running* SWAP alongside a hiring process requires adaptation to the specific environment in which it will be used (e.g., batch versus sequential review), as well as estimation of parameters involving correctness guarantees (e.g., δ and ϵ) or population estimates (e.g., σ).

In general, we recommend that the policymaker or mechanism designer tasked with setting parameters for SWAP, or a SWAP-style algorithm, should conduct a study on past admissions/hiring decisions. This study should include quantitative information (e.g., how many people applied, how many were accepted, how many were interviewed, how long did interviews take) and qualitative information (e.g., how confident was reviewer A after reviewing an applicant B). From this a mechanism designer could determine estimates of population parameters like σ, information gain parameters s, and interview cost parameter j.

To estimate σ, a policymaker could perform a study on past reviews and interviews to determine the range of scores for arms. However, this method could incorporate various biases that may already exist in prior review and scoring processes. That consideration should be taken into account, but exactly how is situation-specific. The introduction of and strict adherence to the structured interview paradigm is a general method to alleviate some of these concerns.

To estimate the value of s, the information gain of a strong pull, one could quantify the difference in confidence level for a particular applicant after performing weak and strong pulls; e.g., how confident was reviewer A after reviewing an applicant B, how much more confident was A after interviewing B, and so on. For j, policy makers could use the average relative difference in time (and possibly monetary) resources spent on different information gathering strategies.

The choice of δ and ϵ could be determined via a sensitivity-analysis-style study, where simulations are run using various settings of δ and ϵ. Policymakers can then judge the simulated risks and rewards to define the parameters.

Once the hyper-parameters have been found, simulations can be performed to find the optimal zone (as discussed in Section 4.4.1). This will allow the designer to determine the best strong pull policy.

Ideally, both studies should include a run focused on past decisions and one run every time the selection process occurs, to ensure SWAP's parameters align with the experiences and values of human decision-makers.

4.7 Conclusion

In this work, we modeled the allocation of interviewing resources and subsequent selection of a cohort of applicants as a combinatorial pure exploration (CPE) problem in the multi-armed bandit setting. We generalized a recent CPE algorithm to the setting where arm pulls can have different costs–where a decision maker can perform *strong* and *weak* pulls, with the former costing more than the latter, but also resulting in a less noisy signal. We presented the strong-weak arm-pulls (SWAP) algorithm and proved theoretical upper bounds for a general class of arm pulling strategies in that setting. We also provided simulation results to test the tightness of these bounds. We then applied SWAP to a real-world problem with combinatorial structure: incorporating diversity into university admissions. On real admissions data from one of the largest US-based computer science graduate programs, we showed that SWAP produces more diverse student cohorts at low cost to student quality while spending a budget comparable to that of the current admissions process.

This work lies in the social injustice bias level in the three tiered view of the world found in Chapter 1. By using the diversity function we address the disparities found between sensitive groups and ensure that those with artificially low scores are pushed higher.

It would be of both practical and theoretical interest to tighten the upper bounds on convergence for SWAP, either for a reduced or general set of arm pulling strategies. We would also like to extend SWAP to include more than two types of pulls or information gathering strategies. We aim to incorporate a more realistic version of diversity and achieve a provably *fair* multi-armed bandit algorithm, as formulated by Joseph et al. [123] and Liu et al. [152]. Additionally, we aim to create a version of SWAP that incorporates applicant matriculation into the candidate-recommending and selection process.

An interesting direction that may be worth pursuing is drawing connections between our work—the selection of a diverse subset of arms—to recent work in *multi-winner voting* [86], a setting in social choice where a subset of alternatives are selected instead of a single winner. Recent work in that space looks at selecting a "diverse but good" committee of alternatives via social choice methods [14, 44]. Similarly, drawing connections to diversity in allocation and matching problems [4, 25, 147] is also potentially of interest.

4.8 Authors and Publication

This chapter was written by Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. It was published at the International Conference on Autonomous Agents and Multiagent Systems 2019 [197]. Initial versions of this paper was published at the Women in Machine Learning Workshop at NeurIPS 2017, the Aligned AI Workshop at NeurIPS 2017, and the Ph.D. in Research Conference at Google 2018.

Algorithm 1 Strong Weak Arm Pulls (SWAP)

```
Require: Confidence $\delta \in (0,1)$; Maximization oracle: $Oracle(\cdot) : \mathbb{R}^n \to \mathcal{M}$
Weak pull each arm $a \in [n]$ once to initialize empirical means $\hat{\mathbf{u}}_n$
$\forall i \in [n]$ set $T_n(a_i) \leftarrow 1$,
$Cost_n \leftarrow n$, total resources spent
for $t = n, n+1, \ldots$ do
    $M_t \leftarrow Oracle(\hat{\mathbf{u}}_t)$
    for $a_i = 1, \ldots, n$ do
        $rad_t(a_i) = \sigma\sqrt{2\log\left(\frac{4nCost_t^3}{\delta}/T_t(a_i)\right)}$
        if $a_i \in M_t$ then
            $\tilde{u_t}(a_i) \leftarrow \hat{u}_t(a_i) - rad_t(a_i)$
        else
            $\tilde{u_t}(a_i) \leftarrow \hat{u}_t(a_i) + rad_t(a_i)$
    $\tilde{M_t} \leftarrow Oracle(\tilde{\mathbf{u}}_t)$
    if $w(\tilde{M_t}) = w(M_t)$ then
        Out $\leftarrow M_t$
        return Out
    $p_t \leftarrow \arg\max_{a \in (\tilde{M}_t \setminus M_t) \cup (M_t \setminus \tilde{M}_t)} rad_t(a)$
    $\alpha \leftarrow spp(s, j)$
    with probability $\alpha$ do
        Strong pull $p_t$
        $T_{t+1}(p_t) \leftarrow T_t(p_t) + s$
        $Cost_{t+1} \leftarrow Cost_t + j$
    else
        Weak pull $p_t$
        $T_{t+1}(p_t) \leftarrow T_t(p_t) + 1$
        $Cost_{t+1} \leftarrow Cost_t + 1$
    Update empirical mean $\hat{\mathbf{u}}_{t+1}$ using observed reward
    $T_{t+1}(a) \leftarrow T_t(a)\ \forall a \neq p_t$
```

Chapter 5: Making the Cut: A Bandit-based Approach to Tiered Interviewing

Given a huge set of applicants, how should a firm allocate *sequential* resume screenings, phone interviews, and in-person site visits? In a tiered interview process, later stages (e.g., in-person visits) are more informative, but also more expensive than earlier stages (e.g., resume screenings). Using accepted hiring models and the concept of structured interviews, a best practice in human resources, we cast tiered hiring as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting. The goal is to select a subset of arms (in our case, applicants) with some combinatorial structure. We present new algorithms in both the probably approximately correct (PAC) and fixed-budget settings that select a near-optimal cohort with provable guarantees. We show via simulations on real data from one of the largest US-based computer science graduate programs that our algorithms make better hiring decisions or use less budget than the status quo.

> *‘... nothing we do is more important than hiring and developing people. At the end of the day, you bet on people, not on strategies.” – Lawrence Bossidy,* The CEO as Coach *(1995)*

5.1 Introduction

Hiring workers is expensive and lengthy. The average cost-per-hire in the United States is $4,129 [207], and with over five million hires per month on average, total annual hiring cost in the United States tops hundreds of billions of dollars [216]. In the past decade, the average length of the hiring process has doubled to nearly one month [54]. At every stage, firms expend resources to learn more about each applicant's true quality, and choose to either cut that applicant or continue interviewing with the intention of offering employment.

In this Chapter, we address the problem of a firm hiring a *cohort* of multiple workers, each with unknown true utility, over multiple stages of structured interviews. We operate under an assumption that a firm is willing to spend an increasing amount of resources—e.g., money or time—on applicants as they advance to later stages of interviews. Thus, the firm is motivated to aggressively "pare down" the applicant pool at every stage, culling low-quality workers so that resources are better spent in more costly later stages. This concept of tiered hiring can be extended to crowdsourcing or finding a cohort of trusted workers. At each successive stage, crowdsourced workers are given harder tasks.

Using techniques from the multi-armed bandit (MAB) and submodular optimization literature, we present two new algorithms—in the probably approximately correct (PAC) (§5.3) and fixed-budget settings (§5.4)—and prove upper bounds that select a near-optimal cohort in this restricted setting. We explore those bounds in simulation and show that the restricted setting is not necessary in practice (§5.5). Then, using real data from admissions to a large US-based computer science Ph.D. program, we show that our algorithms yield *better* hiring decisions at equivalent cost to the status quo—or *comparable* hiring decisions at lower cost (§5.5).

5.2 A Formal Model of Tiered Interviewing

In this section, we formally define our general multi-stage combinatorial MAB problem. For an overview of related work and background information see Chapter 2. Each of our n applicants is an arm a in the full set of arms A. Our goal is to select $K < n$ arms that maximize some objective w using a maximization oracle. We split up the review/interview process into m stages, such that each stage $i \in [m]$ has per-interview information gain s_i, cost j_i, and number of required arms K_i (representing the size of the "short list" of applicants who proceed to the next round). We want to solve this problem using either a confidence constraint (δ, ϵ), or a budget constraint over each stage (T_i). We rigorously define each of these inputs below.

Multi-armed bandits. In this chapter we follow the classical MAB approach described in Section 2.1.1 where each arm $a \in A$ has a true utility $u(a) \in [0, 1]$, which is unknown. When an arm $a \in A$ is pulled, a reward is pulled from a distribution with mean $u(a)$ and a σ-sub-Gaussian tail.

Top-K and subsets. In the previous chapter we relied on an oracle to chose the optimal cohort. In this Chapter we do the same with the modification of choosing an optimal cohort for the shortlist at each stage. As such, we use decision class $\mathcal{M}_K(A) = \{M \subseteq A \mid |M| = K\}$. A cohort is optimal if it maximizes a linear objective function $w : \mathbb{R}^n \times \mathcal{M}_K(A) \rightarrow \mathbb{R}$. Therefore, the maximization oracle can be defined as

$$Oracle_K(\hat{\mathbf{u}}, A) = \underset{M \in \mathcal{M}_K(A)}{\arg\max}\ w(\hat{\mathbf{u}}, M). \tag{5.1}$$

Additionally, for any arm $a \in A$, the gap score Δ_a is now defined as

$$\Delta_a = \begin{cases} w(M^*) - \max_{\{M \mid M \in \mathcal{M}_K \wedge a \in M\}} w(M), & \text{if } a \notin M^* \\ w(M^*) - \max_{\{M \mid M \in \mathcal{M}_K \wedge a \notin M\}} w(M), & \text{if } a \in M^*. \end{cases} \tag{5.2}$$

Using this gap score we estimate the hardness of a problem as the sum of inverse squared gaps:

$$\mathbf{H} = \sum_{a \in A} \Delta_a^{-2} \tag{5.3}$$

This helps determine how easy it is to differentiate between arms at the border of accept/reject.

Objectives. As in the previous setting we apply both a Top-K maximization oracle as well as a diversity oracle. For more information on the Diversity oracle see Section 2.3.

Variable costs. Interviews allow firms to compare applicants. *Structured interviews* treat each applicant the same by following the same questions and scoring strategy, allowing for meaningful cross-applicant comparison. A substantial body of research shows that structured interviews serve as better predictors of job success and reduce bias across applicants when compared to traditional methods [105, 184]. As decision-making becomes more data-driven, firms look to demonstrate a link between hiring criteria and applicant success—and increasingly adopt structured interview processes [132, 143].

In the previous Chapter we introduced a concept of “weak” and “strong” pulls in the Strong Weak Arm Pull (SWAP) algorithm. In this Chapter, however, we transform the concept of “weak” and “strong” pulls to multiple stages. As stages get more expensive, the estimates of utility become more precise - the estimate comes with a distribution with a lower variance. In practice, a resume review may make a candidate seem much stronger than they are, or a badly written resume could severely underestimate their abilities. However, in-person interviews give better estimates. In Section 5.5, we extend (as best we can) the SWAP model to our setting and compare as part of our experimental testbed.

Generalizing to multiple stages. This Chapter, to our knowledge, gives the first computational formalization of tiered structured interviewing. We build on hiring models from the behavioral science literature [43, 217] in which the hiring process starts at recruitment and follows several stages, concluding with successful hiring. We model these m successive stages as having an increased cost—in-person interviews cost more than phone interviews, which in turn cost more than simple résumé screenings—but return additional information via the score given to an applicant. For each stage $i \in [m]$ the user defines a cost j_i and an information gain s_i for the type of pull (type of interview) being used in that stage. During each stage, K_i arms move on to the next stage (we cut off $K_{i-1} - K_i$ arms), where $n = K_0 > K_1 > \cdots > K_{m-1} > K_m = K$). The user must therefore define K_i for each $i \in [m-1]$. The arms chosen to move on to the next stage are denoted as $A_m \subset A_{m-1} \subset \cdots \subset A_1 \subset A_0 = A$.

Tiered MAB and interviewing stages. Our formulation was initially motivated by the graduate admissions system run at our university. Here, at every stage, it is possible for *multiple* independent reviewers to look at an applicant. Indeed, our admissions committee strives to hit at least two written reviews per application package, before potentially considering one or more Skype/Hangouts calls with a potential applicant. (In our data, for instance, some applicants received up to 6 independent reviews per stage.)

While motivated by academic admissions, we believe our model is of broad interest to industry as well. For example, in the tech industry, it is common to allocate more (or fewer) 30-minute one-on-one interviews on a visit day, and/or multiple pre-visit programming screening teleconference calls. Similarly, in management consulting [113], it is common to repeatedly give independent "case study" interviews to borderline candidates.

5.3 Probably Approximately Correct Hiring

In this section, we present Cutting Arms using a Combinatorial Oracle (CACO), the first of two multi-stage algorithms for selecting a cohort of arms with provable guarantees. CACO is a probably approximately correct (PAC) [106] algorithm that performs interviews over m stages, for a user-supplied parameter m, before returning a final subset of K arms.

Algorithm 2 provides pseudocode for CACO. The algorithm requires several user-supplied parameters in addition to the standard PAC-style confidence parameters (δ - confidence probability, ϵ - error), including the total number of stages m;

pairs (s_i, j_i) for each stage $i \in [m]$ representing the information gain s_i and cost j_i associated with each arm pull; the number K_i of arms to remain at the end of each stage $i \in [m]$; and a maximization oracle. After each stage i is complete, CACO removes all but K_i arms. The algorithm tracks these "active" arms, denoted by A_{i-1} for each stage i, the total cost $Cost$ that accumulates over time when pulling arms, and per-arm a information such as empirical utility $\hat{u}(a)$ and total information gain $T(a)$. For example, if arm a has been pulled once in stage 1 and twice in stage 2, then $T(a) = s_1 + 2s_2$.

Algorithm 2 Cutting Arms using a Combinatorial Oracle (CACO)

```
Require: Confidence $\delta \in (0,1)$; Error $\epsilon \in (0,1)$; $Oracle$; number of stages $m$;
    $(s_i, j_i, K_i)$ for each stage $i$
 1: $A_0 \leftarrow A$
 2: for stage $i = 1, \ldots, m$ do
 3:     Pull each $a \in A_{i-1}$ once using the given $s_i, j_i$ pair
 4:     Update empirical means $\hat{\mathbf{u}}$
 5:     $Cost \leftarrow Cost + K_{i-1} \cdot j_i$
 6:     for $t = 1, 2, \ldots$ do
 7:         $A_i \leftarrow Oracle_{K_i}(\hat{\mathbf{u}})$
 8:         for $a \in A_{i-1}$ do
 9:             $rad(a) \leftarrow \sigma\sqrt{\frac{2\log(4|A|Cost^3)/\delta}{T(a)}}$
10:             if $a \in A_i$ then $\tilde{u}(a) \leftarrow \hat{u}(a) - rad_t(a)$
11:             else $\tilde{u}(a) \leftarrow \hat{u}(a) + rad(a)$
12:         $\tilde{A}_i \leftarrow Oracle_{K_i}(\tilde{\mathbf{u}})$
13:         if $|w(\tilde{A}_i) - w(A_i)| < \epsilon$ then break
14:         $p \leftarrow \arg\max_{a \in (\tilde{A}_i \setminus A_i) \cup (A_i \setminus \tilde{A}_i)} rad(a)$
15:         Pull arm $p$ using the given $s_i, j_i$ pair
16:         Update $\hat{u}(p)$ with the observed reward
17:         $T(p) \leftarrow T(p) + s_i$
18:         $Cost \leftarrow Cost + j_i$
19: Out $\leftarrow A_m$; return Out
```

CACO begins with all arms active (line 1). Each stage i starts by pulling each active arm once using the given (s_i, j_i) pair to initialize or update empirical utilities (line 3). It then pulls arms until a confidence level is triggered, removes all but K_i arms, and continues to the next stage (line 13).

In a stage i, CACO proceeds in rounds indexed by t. In each round, the algorithm first finds a set A_i of size K_i using the maximization oracle and the current empirical means $\hat{u}$ (line 7). Then, given a confidence radius (line 9), it computes pessimistic estimates $\tilde{u}(a)$ of the true utilities of each arm a and uses the oracle to find a set of arms $\tilde{A}_i$ under these pessimistic assumptions (lines 10-12). If those two sets are "close enough" (ϵ away), CACO proceeds to the next stage (line 13). Otherwise, across all arms a in the symmetric difference between A_i and $\tilde{A}_i$, the arm p with the most uncertainty over its true utility—determined via $rad(a)$—is pulled (line 14). At the end of the last stage m, CACO returns a final set of K active arms that approximately maximizes an objective function (line 19).

We prove a bound on CACO in Theorem 5.1. As a special case of this theorem, when only a single stage of interviewing is desired, and as $\epsilon \to 0$, then Algorithm 2 reduces to Chen et al. [58]'s CLUCB, and our bound then reduces to their upper bound for CLUCB. This bound provides insights into the trade-offs of *Cost*, information gain s, problem hardness $\mathbf{H}$ (Equation 5.3), and shortlist size K_i. Given the *Cost* and information gain s parameters Theorem 5.1 provides a tighter bound than those for CLUCB.[1]

Theorem 5.1. *Given any $\delta \in (0, 1)$, any $\epsilon \in (0, 1)$, any decision classes $\mathcal{M}_i \subseteq 2^{[n]}$ for each stage $i \in [m]$, any linear function w, and any expected rewards $u \in \mathbb{R}^n$,*

[1] All proofs for this Chapter can be found in Appendix B.2.

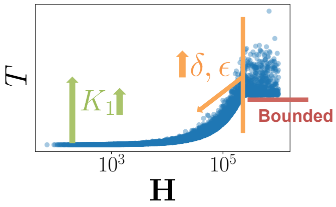


FIGURE 5.1: Hardness (**H**) vs theoretical cost (T) as user-specified parameters to the CACO algorithm change.

assume that the reward distribution φ_a *for each arm* $a \in [n]$ *has mean* $u(a)$ *with a* σ*-sub-Gaussian tail. Let* $M_i^* = \arg\max_{M \in \mathcal{M}_i}$ *denote the optimal set in stage* $i \in [m]$. *Set* $rad_t(a) = \sigma\sqrt{2\log(\frac{4K_{i-1}Cost_{i,t}^3}{\delta})/T_{i,t}(a)}$ *for all* $t > 0$ *and* $a \in [n]$. *Then, with probability at least* $1-\delta$*, the* CACO *algorithm (Algorithm 2) returns the set* `Out` *where* $w($`Out`$) - w(M_m^*) < \epsilon$ *and*

$$T \leq O\left(\sigma^2 \sum_{i\in[m]} \left(\frac{j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{1}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \log\left(\frac{\sigma^2 j_i^4}{s_i\delta} \sum_{a\in A_{i-1}} \min\left\{\frac{1}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\right)\right).$$

Theorem 5.1 gives a bound relative to problem-specific parameters such as the gap scores Δ_a (Equation 5.2), inter-stage cohort sizes K_i, and so on. Figure 5.1[2] lends intuition as to how CACO changes with respect to these inputs, in terms of problem hardness (defined in Eq. 5.3). When a problem is easy (gap scores Δ_a are large and hardness **H** becomes small), the min parts of the bound are dominated by gap scores Δ_a, and there is a smooth increase in total cost. When the problem gets harder (gap scores Δ_a are small and hardness **H** becomes large), the mins are dominated by K_i^2/ϵ^2 and the cost is noisy but bounded below. When ϵ or δ increases, the lower bounds of the noisy section decrease—with the impact of ϵ dominating that of δ. A policymaker can use these high-level trade-offs to determine hiring

[2] For detailed figures see Appendix B.5.

mechanism parameters. For example, assume there are two interview stages. As the number K_1 of applicants who pass the first interview stage increases, so too does total cost T. However, if K_1 is too small (here, very close to the final cohort size K), then the cost also increases.

5.4 Hiring on a Fixed Budget with BRUTaS

In many hiring situations, a firm or committee has a fixed budget for hiring (number of phone interviews, total dollars to spend on hosting, and so on). With that in mind, in this section, we present Budgeted Rounds Updated Targets Successively (BRUTaS), a tiered-interviewing algorithm in the fixed-budget setting.

Algorithm 3 provides pseudocode for BRUTaS, which takes as input fixed budgets $\bar{T}_i$ for each stage $i \in [m]$, where $\sum_{i\in[m]} \bar{T}_i = \bar{T}$, the total budget. In this version of the tiered-interview problem, we also know how many decisions—whether to accept or reject an arm—we need to make in each stage. This is slightly different than in the CACO setting (§5.3), where we need to remove all but K_i arms at the conclusion of each stage i. We make this change to align with the CSAR setting of Chen et al. [58], which BRUTaS generalizes. In this setting, let $\tilde{K}_i$ represent how many decisions we need to make at stage $i \in [m]$; thus, $\sum_{i\in[m]} \tilde{K}_i = n$. The $\tilde{K}_i$s are independent of K, the final number of arms we want to accept, except that the total number of accept decisions across all $\tilde{K}$ must sum to K.

The budgeted setting uses a constrained oracle $COracle : \mathbb{R}^n \times 2^{[n]} \times 2^{[n]} \rightarrow \mathcal{M} \cup \{\perp\}$ defined as

$$COracle(\hat{\mathbf{u}}, A, B) = \mathop{\arg\max}_{\{M\in\mathcal{M}_K \mid A\subseteq M \,\wedge\, B\cap M=\emptyset\}} w(\hat{\mathbf{u}}, M),$$

where A is the set of arms that have been accepted and B is the set of arms that have been rejected.

In each stage $i \in [m]$, BRUTaS starts by collecting the accept and reject sets from the previous stage. It then proceeds through $\tilde{K}_i$ rounds, indexed by t, and selects a single arm to place in the accept set A or the reject set B. In a round t, it first pulls each *active* arm—arms not in A or B—a total of $\tilde{T}_{i,t} - \tilde{T}_{i,t-1}$ times using the appropriate s_i and j_i values. $\tilde{T}_{i,t}$ is set according to Line 6; note that $\tilde{T}_{i,0} = 0$. Once all the empirical means for each active arm have been updated, the constrained oracle is run to find the empirical best set $M_{i,t}$ (Line 9). For each active arm a, a new pessimistic set $\tilde{M}_{i,t,a}$ is found (Lines 11-15). a is placed in the accept set A if a is not in $M_{i,t}$, or in the reject set B if a is in $M_{i,t}$. This is done to calculate the gap that arm a creates (Equation 5.2). The arm $p_{i,t}$ with the largest gap is selected and placed in the accept set A if $p_{i,t}$ was included in $M_{i,t}$, or placed in the reject set B otherwise (Lines 16-20). Once all rounds are complete, the final accept set A is returned.

Theorem 5.2, provides an lower bound on the confidence that BRUTaS returns the optimal set. Note that if there is only a single stage, then Algorithm 3 reduces to Chen et al. [58]'s CSAR algorithm, and our Theorem 5.2 reduces to their upper bound for CSAR. Again Theorem 5.2 provides tighter bounds than those for CSAR given the parameters for information gain s_b and arm pull cost j_b.

Theorem 5.2. *Given any $\bar{T}_i$s such that $\sum_{i\in[m]} \bar{T}_i = \bar{T} > n$, any decision class $\mathcal{M}_K \subseteq 2^{[n]}$, any linear function w, and any true expected rewards $\mathbf{u} \in \mathbb{R}^n$, assume that reward distribution φ_a for each arm $a \in [n]$ has mean $u(a)$ with a σ-sub-Gaussian tail. Let $\Delta_{(1)}, \ldots, \Delta_{(n)}$ be a permutation of $\Delta_1, \ldots, \Delta_n$ (defined in Eq.*

5.2) such that $\Delta_{(1)} \leq \ldots \leq \Delta_{(n)}$. Define $\tilde{\mathbf{H}} \triangleq \max_{i\in[n]} i\Delta_{(i)}^{-2}$. Then, Algorithm 3 uses at most $\bar{T}_i$ samples per stage $i \in [m]$ and outputs a solution $\texttt{Out} \in \mathcal{M}_K \cup \{\perp\}$ such that

$$\Pr[\texttt{Out} \neq M_*] \leq n^2 \exp\left(-\frac{\sum_{b=1}^{m} s_b(\bar{T}_b - \tilde{K}_b)/(j_b \widetilde{\log}(\tilde{K}_b))}{72\sigma^2\tilde{\mathbf{H}}}\right) \tag{5.4}$$

where $\widetilde{\log}(n) \triangleq \sum_{i=1}^{n} i^{-1}$, and $M_* = \arg\max_{M\in\mathcal{M}_K} w(M)$.

Algorithm 3 Budgeted Rounds Updated Targets Successively (BRUTaS)

```
Require: Budgets $\bar{T}_i$ $\forall i \in [m]$; $(s_i, j_i, \tilde{K}_i)$ for each stage $i$; constrained oracle $\mathit{COracle}$
Define $\widetilde{\log}(n) \triangleq \sum_{i=1}^{n} \frac{1}{i}$
$A_{0,1} \leftarrow \varnothing$; $B_{0,1} \leftarrow \varnothing$
for stage $i = 1, \ldots, m$ do
    $A_{i,1} \leftarrow A_{i-1,\tilde{K}_{i-1}+1}$; $B_{i,1} \leftarrow B_{i-1,\tilde{K}_{i-1}+1}$; $\tilde{T}_{i,0} \leftarrow 0$
    for $t = 1, \ldots, \tilde{K}_i$ do
        $\tilde{T}_{i,t} \leftarrow \left\lceil \frac{\bar{T}_i - \left(n - \sum_{a=0}^{i-1} \tilde{K}_i\right)}{\widetilde{\log}\left(n - \sum_{a=0}^{i-1} \tilde{K}_i\right) j_i(\tilde{K}_i - t + 1)} \right\rceil$
        for all $a \in [n] \setminus (A_{i,t} \cup B_{i,t})$ do
            Pull $a$ $(\tilde{T}_{i,t} - \tilde{T}_{i,t-1})$ times; update $\hat{\mathbf{u}}_{i,t}(a)$
        $M_{i,t} \leftarrow \mathit{COracle}(\hat{\mathbf{u}}_{i,t}, A_{i,t}, B_{i,t})$
        if $M_{i,t} = \perp$ then return $\perp$
        for all $a \in [n] \setminus (A_{i,t} \cup B_{i,t})$ do
            if $a \in M_{i,t}$ then
                $\tilde{M}_{i,t,a} \leftarrow \mathit{COracle}(\hat{\mathbf{w}}_{i,t}, A_{i,t}, B_{i,t} \cup \{a\})$
            else
                $\tilde{M}_{i,t,a} \leftarrow \mathit{COracle}(\hat{\mathbf{w}}_{i,t}, A_{i,t} \cup \{a\}, B_{i,t})$
        $p_{i,t} \leftarrow \underset{a \in [n] \setminus (A_{i,t} \cup B_{i,t})}{\arg\max} \; w(M_{i,t}) - w(\tilde{M}_{i,t,a})$
        if $p_{i,t} \in M_t$ then
            $A_{i,t+1} \leftarrow A_{i,t} \cup \{p_{i,t}\}$; $B_{i,t+1} \leftarrow B_{i,t}$
        else
            $A_{i,t+1} \leftarrow A_{i,t}$; $B_{i,t+1} \leftarrow B_{i,t} \cup \{p_{i,t}\}$
Out $\leftarrow A_{m,\tilde{K}_m+1}$; return Out
```

When setting the budget for each stage, a policymaker should ensure there is sufficient budget for the number of arms in each stage i, and for the given exogenous cost values j_i associated with interviewing at that stage. There is also a balance between the number of decisions that must be made in a given stage i and the ratio $\frac{s_i}{j_i}$ of interview information gain and cost. Intuitively, giving higher budget to stages with a higher $\frac{s_i}{j_i}$ ratio makes sense—but one also would not want to make *all* accept/reject decisions in those stages, since more decisions corresponds to lower confidence. Generally, arms with high gap scores Δ_a are accepted/rejected in the earlier stages, while arms with low gap scores Δ_a are accepted/rejected in the later stages. The policy maker should look at past decisions to estimate gap scores Δ_a (Equation 5.2) and hardness $\mathbf{H}$ (Equation 5.3). There is a clear trade-off between information gain and cost. If the policy maker assumes (based on past data) that the gap scores will be high (it is easy to differentiate between applicants) then the lower stages should have a high K_i, and a budget to match the relevant cost j_i. If the gap scores are all low (it is hard to differentiate between applicants) then more decisions should be made in the higher, more expensive stages. By looking at the ratio of small gap scores to high gap scores, or by bucketing gap scores, a policy maker will be able to set each K_i.

5.5 Experiments

In this section, we experimentally evaluate BRUTaS and CACO in two different settings. The first setting uses data from a toy problem of Gaussian distributed

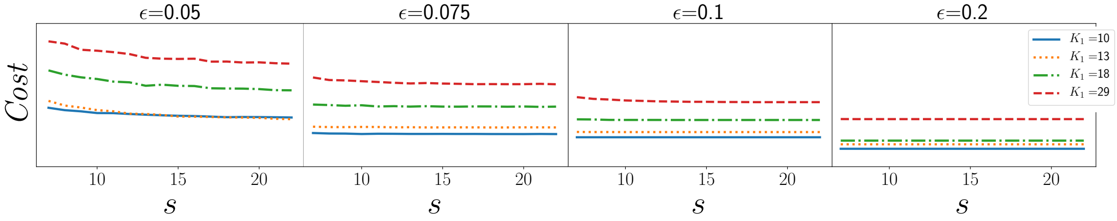


FIGURE 5.2: Comparison of $Cost$ vs information gain (s) as ϵ increases for CACO. Here, $\delta = 0.05$ and $\sigma = 0.2$. As ϵ increases, the cost of the algorithm also decreases. If the overall cost of the algorithm is low, then increasing s (while keeping j constant) provides diminishing returns.

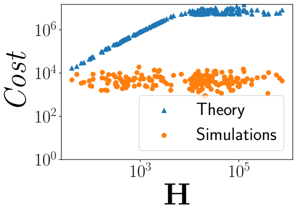


FIGURE 5.3: Hardness (**H**) vs Cost, comparing against Theorem 5.1.

arms. The second setting uses real admissions data from one of the largest US-based graduate computer science programs.

5.5.1 Gaussian Arm Experiments

We begin by using simulated data to test the tightness of our theoretical bounds. To do so, we instantiate a cohort of $n = 50$ arms whose true utilities, u_a, are sampled from a normal distribution. We aim to select a final cohort of size $K = 7$. When an arm is pulled during a stage with cost j and information gain s, the algorithm is charged a cost of j and a reward is pulled from a distribution with mean u_a and standard deviation of $\sigma/\sqrt{s}$. For simplicity, we present results in the setting of $m = 2$ stages.

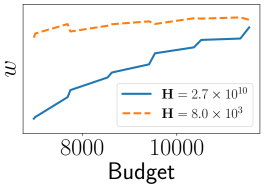


FIGURE 5.4: Effect of an increasing budget on the overall utility of a cohort. As hardness ($\mathbf{H}$) increases, more budget is needed to produce a high quality cohort.

CACO. To evaluate CACO, we vary δ, ϵ, σ, K_1, and s_2. We find that as δ increases, both cost and utility decrease, as expected. Similarly, Figure 5.2 shows that as ϵ increases, both cost and utility decrease. Higher values of σ increase the total cost, but do not affect utility. We also find diminishing returns from high information gain s values (x-axis of Figure 5.2). This makes sense—as s tends to infinity, the true utility is returned from a single arm pull. We also notice that if many "easy" arms (arms with very large gap scores) are allowed in higher stages, total cost rises substantially.

Although the bound defined in Theorem 5.1 assumes a linear function w, we empirically tested CACO using a submodular function w_{DIV}. We find that the cost of running CACO using this submodular function is significantly lower than the theoretical bound. This suggests that (i) the bound for CACO can be tightened and (ii) CACO could be run with submodular functions w.

BRUTaS. To evaluate BRUTaS, we varied σ and $(\tilde{K}_i, T_i)$ pairs for two stages. Utility varies as expected from Theorem 5.2: when σ increases, utility decreases. There is also a trade-off between $\tilde{K}_i$ and T_i values. If the problem is easy, a low budget and a high $\tilde{K}_1$ value is sufficient to get high utility. If the problem is hard (high **H** value), a higher overall budget is needed, with more budget spent in the second stage. Figure 5.4 shows this escalating relationship between budget and utility based on problem hardness. Again we found that BRUTaS performed well when using a submodular function w_{DIV}.

Finally, we compare CACO and BRUTaS to two baseline algorithms: Uniform and Random, which uniformly and randomly respectively, pulls arms in each stage. In both algorithms, the maximization oracle is run after each stage to determine which arms should move on to the next stage. When given a budget of 2,750, BRUTaS achieves a utility of 244.0, which outperforms both the Uniform and Random baseline utilities of 178.4 and 138.9, respectively. When CACO is run on the same problem, it finds a solution (utility of 231.0) that beats both Uniform and Random at a roughly equivalent cost of 2,609. This qualitative behavior exists for other budgets.

5.5.2 Graduate Admissions Experiment

We evaluate how CACO and BRUTaS might perform in the real world by applying them to a graduate admissions dataset from one of the largest US-based graduate computer science programs. These experiments were approved by the university's Institutional Review Board and did not affect any admissions decisions for the university. Our dataset consists of three years (2014–16) worth of graduate applications. For each application we also have graduate committee review scores (normalized to between 0 and 1) and admission decisions.

Experimental setup. Using information from 2014 and 2015, we used a random forest classifier [180], trained in the standard way on features extracted from the applications, to predict probability of acceptance. Features included numerical information such as GPA and GRE scores, topics from running Latent Dirichlet Allocation (LDA) on faculty recommendation letters [195], and categorical information such as region of origin and undergraduate school. In the testing phase, the classifier was run on the set of applicants A from 2016 to produce a probability of acceptance $P(a)$ for every applicant $a \in A$.

We mimic the university's application process of two stages: a first review stage where admissions committee members review the application packet, and a second interview stage where committee members perform a Skype interview for a select subset of applicants. The committee members follow a structured interview approach. We determined that the time taken for a Skype interview is roughly 6 times as long as a packet review, and therefore we set the cost multiplier for the second stage $j_2 = 6$. We ran over a variety of s_2 values, and we determined σ by looking at the distribution of review scores from past years. When an arm $a \in A$ is pulled with information gain s and cost j, a reward is randomly pulled from the arm's review scores (when $s_1 = 1$ and $j_1 = 1$, as in the first stage), or a reward is pulled from a Gaussian distribution with mean $P(a)$ and a standard deviation of $\frac{\sigma}{\sqrt{s}}$.

We ran simulations for BRUTaS, CACO, Uniform, and Random. In addition we compare to an adjusted version of SWAP(defined in Chapter 4). SWAP uses a strong pull policy to probabilistically weak or strong pull arms. In this adjusted version we use a strong pull policy of always weak pulling arms until some threshold time t and strong pulling for the remainder of the algorithm. Note that this

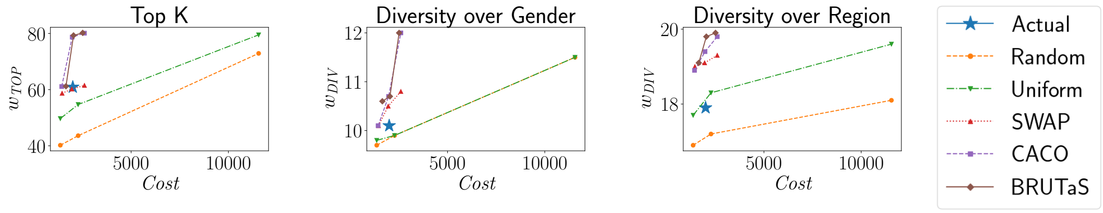


Figure 5.5: Utility vs Cost over four different algorithms (Random, Uniform, SWAP, CACO, BRUTaS) and the actual admissions decisions made at the university. Both CACO and BRUTaS produce equivalent cohorts to the actual admissions process with lower cost, or produce high quality cohorts than the actual admissions process with equivalent cost.

adjustment moves SWAP away from fixed confidence but not all the way to a budgeted algorithm like BRUTaS but fits into the tiered structure. For the budgeted algorithms BRUTaS, Uniform, and Random, (as well as the pseudo-budgeted SWAP) if there are K_i arms in round i, the budget is $K_i \cdot x_i$ where $x_i \in \mathbb{N}$. We vary δ and ϵ to control CACO's cost.

We compare the utility of the cohort selected by each of the algorithms to the utility from the cohort that was actually selected by the university. We maximize either objective w_{TOP} or w_{DIV} for each of the algorithms. We instantiate w_{DIV}, defined in Equation 2.4, in two ways: first, with self-reported gender, and second, with region of origin. Note that since the graduate admissions process is run entirely by humans, the committee does not explicitly maximize a particular function. Instead, the committee tries to find a good overall cohort while balancing areas of interest and general diversity.

Results. Figure 5.5 compares each algorithm to the actual admissions decision process performed by the real-world committee. In terms of utility, for both w_{TOP} and

w_{DIV}, BRUTaS and CACO achieve similar gains to the actual admissions process (higher for w_{DIV} over region of origin) when using less cost/budget. When roughly the same amount of budget is used, BRUTaS and CACO are able to provide higher predicted utility than the true accepted cohort, for both w_{TOP} and w_{DIV}. As expected, BRUTaS and CACO outperform the baseline algorithms Random, Uniform. The adjusted SWAP algorithm performs poorly in this restricted setting of tiered hiring. By limiting the strong pull policy of SWAP, only small incremental improvements can be made as *Cost* is increased.

5.6 Conclusions & Discussion of Future Research

We provided a formalization of tiered structured interviewing and presented two algorithms, CACO in the PAC setting and BRUTaS in the fixed-budget setting, which select a near-optimal cohort of applicants with provable bounds. We used simulations to quantitatively explore the impact of various parameters on CACO and BRUTaS and found that behavior aligns with theory. We showed empirically that both CACO and BRUTaS work well with a submodular function that promotes diversity. Finally, on a real-world dataset from a large US-based Ph.D. program, we showed that CACO and BRUTaS identify higher quality cohorts using equivalent budgets, or comparable cohorts using lower budgets, than the status quo admissions process. Moving forward, we plan to incorporate multi-dimensional feedback (e.g., with respect to an applicant's technical, presentation, and analytical qualities) into our model; recent work due to Katz-Samuels and Scott [127, 128] introduces that feedback (in a single-tiered setting) as a marriage of MAB and constrained optimization, and we see this as a fruitful model to explore combining with our novel tiered system.

Discussion. The results support the use of BRUTaS and CACO in a practical hiring scenario. Once policymakers have determined an objective, BRUTaS and CACO could help reduce costs and produce better cohorts of employees. Yet, we note that although this experiment uses real data, it is still a simulation. The classifier is not a true predictor of utility of an applicant. Indeed, finding an estimate of utility for an applicant is a nontrivial task. Additionally, the data that we are using incorporates human bias in admission decisions, and reviewer scores [9, 195]. Finally, defining an objective function on which to run CACO and BRUTaS is a difficult task. Recent advances in human value judgment aggregation [89, 172] could find use in this decision-making framework.

Indeed, this work again lies in the social injustice bias level found in the three level world view from Chapter 1. Similar to the previous chapter, our tiered MAB formulation makes use of a diversity function to deal with artificially low scores from reviewers.

5.7 Authors and Publication

This chapter was written by Candice Schumann, Zhi Lang, Jeffrey S. Foster, and John P. Dickerson. It was published at the Conference on Neural Information Processing Systems (NeurIPS) 2019 [198].

Chapter 6: Group Fairness in Bandit Arm Selection

Moving away from hiring, we propose a novel formulation of group fairness with biased feedback in the contextual multi-armed bandit (CMAB) setting. In the CMAB setting a sequential decision maker must at each time step choose an arm to pull from a finite set of arms after observing some context for each of the potential arm pulls. In our model arms are partitioned into two or more sensitive groups based on some protected feature (e.g., age, race, or socio-economic status). Initial rewards received from pulling an arm may be biased due to some unknown societal or measurement bias. We assume that in reality these groups are equal dispite this biased feedback. To alleviate this we learn a societal bias term which can be used to find the source of the bias to potentially fix the problem outside of the algorithm. Note that this societal bias term attempts to measure the societal bias mentioned in Chapter 1, however some form of measurement bias will be found in this term. We provide a novel algorithm that can accommodate this notion of fairness for an arbitrary number of groups, and provide a theoretical bound on the regret for our algorithm. We validate our algorithm using synthetic data and two real-world datasets for intervention settings wherein we want to allocate resources fairly across groups.

6.1 Introduction

In many online settings a computational or human agent must sequentially select an item from a slate, receive feedback on that selection, and then use that feedback to learn how to select the best items in the following rounds. Within computer science, economics, and operations research circles this is typically modeled as a *multi-armed bandit (MAB)* problem [209]. Examples include algorithms for selecting what advertisements to display to users on a webpage [165], systems for dynamic pricing [169], and content recommendation services [144]. Indeed, such decision-making systems continue to expand in scope, making ever more important decisions in our lives such as setting bail [68], making hiring decisions [39] (See Chapter 4), and policing [194]. Thus the study of the properties of these algorithms is of tantamount importance as highlighted by the recent work of Chouldechova and Roth [60] on priorities for fairness research in machine learning.

In the previous two chapters we focused on the classical MAB setting where at each time step $t \in T$, an agent pulls an arm and receives a reward that is independent of any previous action and follows the selected arm's probability distribution. In this chapter we instead focus on the generalization of MAB to the contextual multi-armed bandit (CMAB) where the agent observes a d-dimensional *context* of features to use along with the observed rewards of the arms played to choose a new arm. For more details on the CMAB formulation see Section 2.1.5.

However, the use of MAB- and CMAB-based systems often results in behavior that is societally repugnant. Sweeney [210] noted that queries for public records on Google resulted in different associated contextual advertisements based on whether the query target had a traditionally African American or Caucasian name; in the former case advertisements were more likely to contain text relating to criminal incidents. Following that initial report similar instances continue to be observed, both in the bandit setting and in the general machine learning world [175]. In lockstep, the academic community has begun developing approaches to tackling issues of (un)fairness in learning settings. We have an opportunity here to identify and understand why the data we have may be *causing* the bias. See Chapter 3 for a more in depth discussion on this behavior.

Recently, a Computing Community Consortium (CCC) whitepaper on fairness in machine learning specifically identified that most studies of fairness are focused on classification problems [60]. Two fundamental issues identified by Chouldechova and Roth [60] that we address in this Chapter are extensions to notions of *group fairness* and looking at fairness in *online dynamic systems*, e.g., the contextual bandit setting. We address these gaps by formalizing and providing an algorithm for fairness with biased feedback when the arms of the bandit can be partitioned into groups.

Running Example.

As a running example throughout the chapter, imagine the position of an agent at a bank or a lender on a micro-lending site. Here, the agent must sequentially pick loans to fund where the agent regrets picking a loan that fails repayment. In many cases, such as the micro-lending site Kiva,[1] a user is presented with a slate of potential loans to fund when they log in. Each of these loans, i.e. arms, has a context which includes attributes of the applicant including a personal statement,

[1] https://www.kiva.org/

repayment history, business plan, and other data related to the loans. The loans can also be partitioned into sets of m sensitive attributes, e.g. location, race, or gender. In the simplest case, assume we have two female applicants and two male applicants on the slate at a given time. We also assume that when pulling an arm from, for example, a female applicant, there is some societal bias introduced into the reward. We want to balance the number of times the agent selects women versus men given this societal bias built into the feedback.

Observe that while we use loans as our running example, our notion of regret could be extended to a number of other areas including recent work in MAB problems on hiring situations (See Chapter 5. One could imagine a situation where hiring decisions are made with respect to a short-term reward signal that is biased,[2] versus a longer-term reward of performance which is less biased, e.g., via an end-of-year review that is based on a more quantitative metric such as on-the-job performance. A similar argument can be made about school admissions or matching workers to online tasks in a crowdwork setting.

For a detailed discussion on related work refer back to Chapter 2.

6.2 Preliminaries

We follow the standard CMAB setting and assume that we are attempting to maximize a measure over a series of time steps $t \in T$. We assume that there is a d-dimensional domain for the context space, $\mathcal{X} = \mathbb{R}^d$. The agent is presented with a set A of arms from which to select, and we have $|A| = n$ total arms. Each of these

[2] Recent research shows that class-based bias presents itself within seconds of an in-person interview; see https://news.yale.edu/2019/10/21/yale-study-shows-class-bias-hiring-based-few-seconds-speech.

arms is associated with a possibly disjoint context space $\mathcal{X}_i \subseteq \mathcal{X}$. Additionally, we assume that we have m sensitive groups and that the arms are partitioned into these sensitive groups such that $P_1 \cap \cdots \cap P_m = \emptyset$ and $P_1 \cup \cdots \cup P_m = A$. For exposition's sake, we assume a binary sensitive attribute with $m = 2$ for the remaining of the chapter. However, we show the generality of our results to any number of groups in Section 6.3.

Each arm i has a true linear reward function $f_i : \mathcal{X} \to \mathbb{R}$ such that $f_i(x) = \beta_i \cdot x$ where β_i is a vector of coefficients that is unknown to the agent. During each round $t \in T$, a context $x_{t,i} \in \mathcal{X}_i$ is given for each arm i. One arm is pulled per round. When arm i is pulled during round t, a reward is returned: $r_{t,i} = f_i(x_{t,i}) + e_{t,i}$ where $e_{t,i} \sim \mathcal{N}(0, 1)$. The goal of the agent is to minimize the regret over all timesteps in T. Formally, the regret of the agent at timestep t is the difference between the arm selected and the best arm that could have been selected. Let i^* denote the optimal arm that could be selected and a be the selected arm. Then, the regret at t is

$$R(t) = f(x_{i^*,t}) - f(x_{a,t}). \tag{6.1}$$

In this chapter we compare our proposed algorithm against three other algorithms: TopInterval, a variation of LinUCB from Li et al. [144], NaiveFair which randomly picks a sensitive group and then applies TopInterval to that group[3], and IntervalChaining, an individually fair algorithm from Joseph et al. [122]. All algorithms use OLS estimators of the arm coefficients $\hat{\beta}_i$ with a confidence variable $w_{i,t}$ such that the true utility lies within $[\hat{\beta}_i \cdot x_{i,t} - w_{i,t}, \hat{\beta}_i \cdot x_{i,t} + w_{i,t}]$ with probability $1 - \delta$. NaiveFair implements a naive version of demographic parity

[3] See Appendix C.2 for more information

without explicitly looking at societal bias. TopInterval either explores by pulling an arm uniformly at random or exploits by pulling the arm with the highest upper confidence $\hat{\beta}_i \cdot x_{i,t} + w_{i,t}$. To ensure individual fairness, IntervalChaining either explores by choosing an arm uniformly at random or exploits by pulling arms that have overlapping confidence intervals with the arm with the highest upper confidence. For example if the top arm has a confidence interval of (1,4) and another arm has an interval of (0.5,2), these intervals overlapped and one arm is picked uniformly at random.

6.2.1 Regret with Societal Bias

As mentioned before, ground truth rewards for sensitive groups can be noisy due to societal or measurement bias. We now formalize this bias in terms of multi-armed bandits. Again, we assume that n arms can be partitioned into two sets P_1 and P_2 such that $P_1 \cap P_2 = \emptyset$ and $P_1 \cup P_2 = [n]$. We consider P_1 as the sensitive set or the set with some societal bias. Each arm i has a true utility function $f(x_{i,t}) = \beta_i \cdot x_{i,t}$ where β_i is a vector of coefficients, however, if arm i is pulled at timestep t the following reward is returned:

$$r_{i,t} = \beta_i \cdot x_{i,t} + \mathbb{1}[i \in P_1]\psi_{P_1} \cdot x_{i,t} + \mathcal{N}(0,1), \tag{6.2}$$

where $\mathbb{1}[i \in P_1] = 1$ when $i \in P_1$ and 0 otherwise, and ψ_{P_1} is a societal or systematic bias against group P_1. Note that ψ_{P_2} is a zero vector for the non-sensitive group.

Using our running example, let's assume that the down payment reward received has some bias against the male applicants compared to the female applicants, while the final repayment does not. Note that the final repayment is not measured after accepting a loan and is only measured much later. The loan agency should then take the bias into account while learning what good applications look like. Or, in a hiring setting, an applicant may have a biased interview (initial reward) while their true performance is measured only after working for a year (later true reward).

We therefore define regret for pulling an arm a at time t as

$$R(t) = f(x_{i^*,t}) - f(x_{a,t}) \tag{6.3}$$

where i^* is the optimal arm to pull at timestep t and $f(x_{i,t})$ is the true reward with no bias terms $\psi_{P_1} \cdot x_{i,t}$. We also assume that the average true reward (with no bias) for group P_1 should be the same as the average reward for group P_2. In the loan agency example, this real regret would measure the regret of the final repayments instead of the biased down payment regret.

One can view the societal bias term ψ_i that we learn for some group i as our algorithm learning how to automatically identify and adjust for anti-discrimination for group i compared to all other groups. Anti-discrimination is the practice of identifying a relevant feature in data and adjusting it to provide fairness under that measure [68]. One example of this, discussed by Dwork et al. [81], Joseph et al. [123], and in the official White House algorithmic decision making statement [174], comes up in college admissions. Given other factors, specifically income level, some colleges weight SAT scores *less* in wealthy populations due to the presence of tutors while increasing the weight of working-class populations [21]. While in these admissions settings the adjustments may be ad-hoc, we learn our bias term from data. Past work has compared the vector β learned for each arm as akin to adjusting for these biases [81]. While this is true at an *individual* level, our explicit modeling of bias allows us to discover these adjustments at a *group* level.

6.3 Group Fair Contextual Bandits

In this section, given our new definition of reward (Equation 6.2) and corresponding new definition of regret (Equation 6.3), we present the algorithm GroupFairTopInterval (Algorithm 4) which takes societal bias into account. We also give a bound on its regret in this new reward and regret setting. Subsequently, we briefly describe the algorithm.

In GroupFairTopInterval, each round t is randomly chosen with probability $\frac{1}{t^{1/3}}$ to be an exploration round. The exploration round randomly chooses an arm to learn more about.

The remaining rounds become exploitation rounds, where linear estimates are used to pull arms. GroupFairTopInterval learns two different types of standard OLS linear estimators [135]. The first is a coefficient vector $\hat{B}_{i,t}$ for each arm i (line 7). Additionally, GroupFairTopInterval learns a group coefficient vector $\hat{\psi}_{P_j,t}$ for each group P_j (lines 4 and 5). As mentioned previously, we treat P_1 as the sensitive group of arms. An arm i in the non-sensitive group P_2 has a reward estimation of $\hat{\beta}_{i,t} \cdot x_{i,t}$, while an arm i in the sensitive group P_1 has a bias corrected reward estimation of $\hat{\beta}_{i,t} \cdot x_{i,t} - \hat{\psi}_{P_1,t} + \hat{\psi}_{P_2,t}$.

For each arm i, the algorithm calculates confidence intervals $w_{i,t}$ around the linear estimates $\hat{B}_{i,t} \cdot x_{i,t}$ using a Quantile function Q (line 9). This means that the true utility (including some bias) falls within $[\hat{B}_{i,t} \cdot x_{i,t} - w_i, \hat{B}_{i,t} \cdot x_{i,t} + w_i]$ with probability $1 - \delta$ at every arm i and every timestep t. Similarly, for each group P_j and context

$w_{i,t}$ for a given arm i at timestep t, the algorithm calculates a confidence interval $b_{P_j,i,t}$ using a Quantile function Q (lines 4 and 5). This means that the true *group* utility (or true average group utility) falls within $[\hat{\psi}_{P_j,i,t} \cdot x_{i,t} - b_{P_j,i,t}, \hat{\psi}_{P_j,i,t} \cdot x_{i,t} + b_{P_j,i,t}]$ with probability $[1-\delta]$. Using the confidence intervals $w_{i,t}$ and $b_{P_j,i,t}$, and the linear estimates $\hat{B}_{i,t} \cdot x_{i,t}$ and $\hat{\psi}_{P_j,i,t} \cdot x_{i,t}$ we can calculate the upper bound of the estimated reward for each arm i (lines 15 and 17). The algorithm then pulls the arm with the highest upper bound (line 18).

Algorithm 4 GroupFairTopInterval

```
Require: $\delta$, $P_1$, $P_2$
 1: for $t = 1 \ldots T$ do
 2:     With probability $\frac{1}{t^{1/3}}$, play $i_t \in_R \{1, \ldots, n\}$
 3:     otherwise:
 4:         $\hat{\psi}_{P_1,t} \leftarrow \left(\mathcal{X}_{P_1,t}^T \mathcal{X}_{P_1,t}\right)^{-1} \mathcal{X}_{P_1,t}^T \mathcal{Y}_{P_1,t}$
 5:         $\hat{\psi}_{P_2,t} \leftarrow \left(\mathcal{X}_{P_2,t}^T \mathcal{X}_{P_2,t}\right)^{-1} \mathcal{X}_{P_2,t}^T \mathcal{Y}_{P_2,t}$
 6:         for $i = 1 \ldots n$ do
 7:             $\hat{\beta}_{i,t} \leftarrow \left(X_{i,t}^T X_{i,t}\right)^{-1} X_{i,t}^T Y_{i,t}^T$
 8:             $F_{i,t} \leftarrow \mathcal{N}\left(0, \sigma^2 x_{i,t} \left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t}^T\right)$
 9:             $w_{i,t} \leftarrow Q_{F_{i,t}}\left(\frac{\delta}{2nt}\right)$
10:             if $i \in P_1$ then
11:                 $\mathcal{F}_{P_1,i,t} \leftarrow \mathcal{N}\left(0, \sigma^2 x_{i,t} \left(\mathcal{X}_{P_1,t}^T \mathcal{X}_{P_1,t}\right) x_{i,t}^T\right)$
12:                 $\mathcal{F}_{P_2,i,t} \leftarrow \mathcal{N}\left(0, \sigma^2 x_{i,t} \left(\mathcal{X}_{P_2,t}^T \mathcal{X}_{P_2,t}\right) x_{i,t}^T\right)$
13:                 $b_{P_1,i,t} \leftarrow Q_{\mathcal{F}_{P_1,i,t}}\left(\frac{\delta}{2\frac{n}{|P_1|}T}\right)$
14:                 $b_{P_2,i,t} \leftarrow Q_{\mathcal{F}_{P_2,i,t}}\left(\frac{\delta}{2\frac{n}{|P_2|}T}\right)$
15:                 $\hat{u}_{i,t} \leftarrow \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t} - \hat{\psi}_{P_1,t} \cdot x_{i,t} + b_{P_1,i,t} + \hat{\psi}_{P_2,t} \cdot x_{i,t} + b_{P_2,i,t}$
16:             else
17:                 $\hat{u}_{i,t} \leftarrow \hat{\beta}_{i,t} \cdot x_{i,t}$
18:         Play $\arg\max_i \hat{u}_{i,t}$ and observe reward $y_{i,t}$
```

Returning to our running example, using GroupFairTopInterval, the loan agency would learn a down payment reward function for each of the arms, i.e., a coefficient vector β_i where $i \in$ [young female arm, young male arm, older female arm, older male arm], as well as the group average coefficients for the gender-grouped arms, ψ_{P_j}, for male and female. Using the gender-grouped coefficients, expected rewards for male arms are reweighted to account for the bias in down payment.

Standard algorithms like TopInterval[4] would choose an arm $i = \arg\max(\hat{\beta} \cdot x_{i,t} + w_{i,t})$, ignoring societal bias (Equation 6.2, leading to a larger true regret (Equation 6.3)). Note that GroupFairTopInterval can be extended to multiple groups by defining an overall average reward.

GroupFairTopInterval is fair—in the context of the group fairness definitions used throughout this chapter—and satisfies the following theorem. A proof sketch follows the theorem and a full proof can be found in Appendix C.3.

Theorem 6.1. *For two groups P_1 and P_2, where P_1 has a bias offset in rewards,* GroupFairTopInterval *has regret*

$$R(T) \;=\; O\left(\sqrt{\frac{dn\ln\frac{2nT}{\delta}}{l}}T^{2/3} + \left(\frac{dnL}{l}\left(\ln^2\frac{2nT}{\delta} + \ln d\right)\right)^{2/3}\right). \tag{6.4}$$

Proof Sketch. We start by proving two lemmas. The first of which states that with probability at least $1-\delta$:

$$\left|\hat{\beta}_{i,t} \cdot x_{i,t} - (\beta_i \cdot x_{i,t} + \mathbb{1}[i \in P_1]\psi_{P_1} \cdot x_{i,t})\right| \leq w_{i,t} \tag{6.5}$$

[4] A variant of the contextual bandit LinUCB by Li et al. [144]

holds for any i at time t. Similarly, the second states that with probability at last $1-\delta$:

$$\left|\hat{\beta}_{i,t} \cdot x_{i,t} - \beta_i \cdot x_{i,t}\right| \leq w_{i,t} \tag{6.6}$$

holds for any group P_j, any arm i, and at any timestep t.

The regret for GroupFairTopInterval can be broken down into three terms:

$$\begin{aligned} R(T) = & \sum_{t\colon\, t \textit{ is an explore round}} \textit{regret}(t) \\ & + \sum_{t\colon\, t \textit{ is an exploit round and } t<T_1} \textit{regret}(t) \\ & + \sum_{t\colon\, t \textit{ is an exploit round and } t\geq T_1} \textit{regret}(t). \end{aligned} \tag{6.7}$$

First, for any t we have:

$$\sum_{t'<t} \frac{1}{t^{1/3}} = \Theta(t^{2/3}). \tag{6.8}$$

We then show that the number of rounds T_1 after which we have sufficient samples such that the estimators are well concentrated is:

$$T_1 = \Theta\left(\min_a \left(\frac{dnL}{\lambda_{min_{a,d}}}\left(\ln^2 \frac{2}{\delta} + \ln d\right)\right)^{3/2}\right). \tag{6.9}$$

Finally, we bound the third term in Equation 6.7 as follows:

$$\sum_{t\colon\, t \textit{ is an exploit round and } t\geq T_1} \textit{regret}(t) \leq O\left(\sqrt{dn\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}} T^{2/3} + \delta' T\right). \tag{6.10}$$

Combining Equations 6.7, 6.8, 6.9, and 6.10, we have Theorem 6.1. □

Note that we can extend Algorithm 4 to m groups. In this setting we make the strong assumption that true rewards are centered about ρ defined by the user.[5] In this adaption of the algorithm, we set the upper bound radius for arm i as:

$$\hat{u}_{i,t} = \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t} + \rho - \hat{\psi}_{P_j,t} \cdot x_{i,t} + b_{P_j,i,t}$$

where $i \in P_j$. We then have the following theorem for multiple groups:

Theorem 6.2. *For m groups $P_1, \ldots, P_m$, where ρ is the expected average reward,* GroupFairTopInterval (Multiple Groups) *has regret*

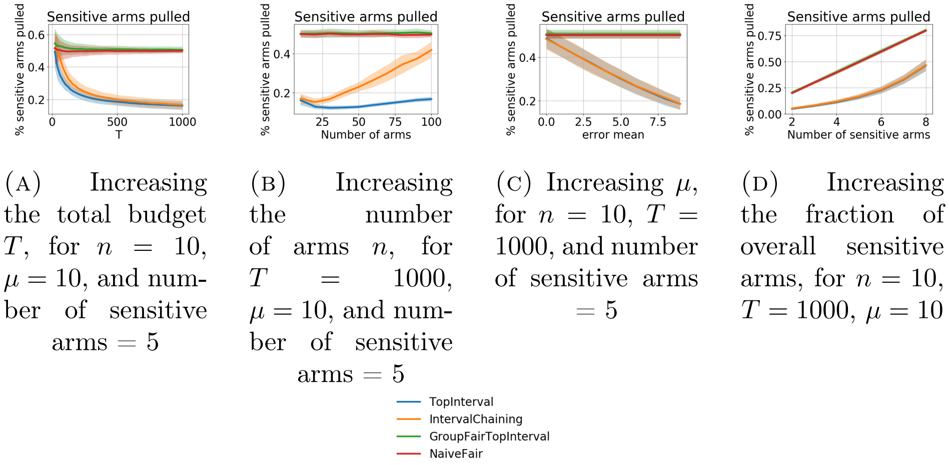
$$R(T) = O\left(\sqrt{\frac{dn \ln \frac{2nT}{\delta}}{l}} T^{2/3} + \left(\frac{dnmL}{l}\left(\ln^2 \frac{2nT}{\delta} + \ln d\right)\right)^{2/3}\right). \tag{6.11}$$

where $l = \min_i \lambda_{min_{i,d}}$ and $L > \max_t \lambda_{\max}(x_{i,t}^T x_{i,t})$.

6.4 Experiments

In this section, we empirically evaluate GroupFairTopInterval. We perform experiments on synthetic data to demonstrate the effects of various parameters, and on real datasets to demonstrate how GroupFairTopInterval performs in the wild. In each of these sections we compare to TopInterval, due to Li et al. [144], NaiveFair discussed in Appendix C.2, and IntervalChaining, due to Joseph et al. [123].

[5] See Appendix C.3.2 for further details



(A) Increasing the total budget T, for $n = 10$, $\mu = 10$, and number of sensitive arms $= 5$

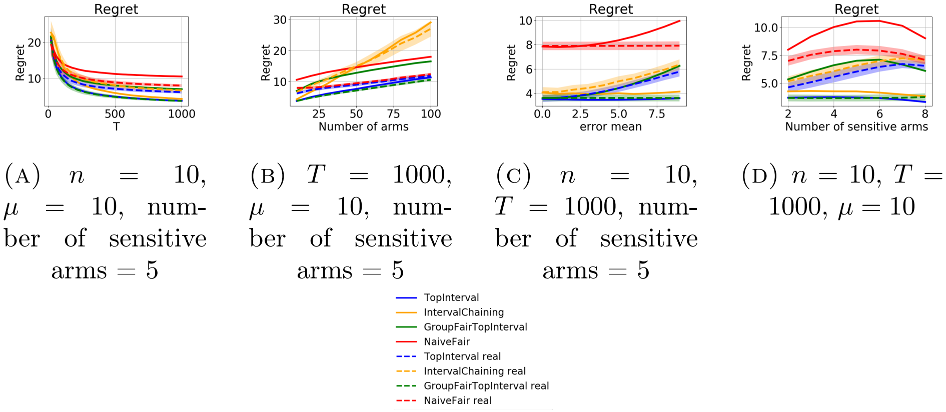
(B) Increasing the number of arms n, for $T = 1000$, $\mu = 10$, and number of sensitive arms $= 5$

(C) Increasing μ, for $n = 10$, $T = 1000$, and number of sensitive arms $= 5$

(D) Increasing the fraction of overall sensitive arms, for $n = 10$, $T = 1000$, $\mu = 10$

(E) Legend

FIGURE 6.1: Percentage of total arm pulls that were pulled using sensitive arms.



(A) $n = 10$, $\mu = 10$, number of sensitive arms $= 5$

(B) $T = 1000$, $\mu = 10$, number of sensitive arms $= 5$

(C) $n = 10$, $T = 1000$, number of sensitive arms $= 5$

(D) $n = 10$, $T = 1000$, $\mu = 10$

(E) Legend

FIGURE 6.2: Regret for synthetic experiments. The solid lines are regret given the rewards received from pulling the arms (including the group bias). The dashed lines is the true regret (without the group bias).

6.4.1 Synthetic Experiments

In each synthetic experiment, we generate true coefficient vectors β_i by choosing coefficients uniformly at random for each arm i. Contexts at each timestep t are chosen randomly for each arm i. Seeds are set at the beginning of each experiment to keep arms consistent between algorithms for a fair comparison. Additionally, bias coefficients ψ_1 are set uniformly at random with a given mean $\mu = 10$.

We run four different types of experiments:[6]

(a) Varying the total budget for pulling arms (T) while setting the number of arms $n = 10$, the error mean $\mu = 10$, the number of sensitive arms equal to 5, and the context dimension $d = 2$ (Figures 6.2a and 6.1a).

(b) Varying the total number of arms n while setting the total budget $T = 1000$, the error mean $\mu = 10$, the number of sensitive arms to 5, and the context dimension $d = 2$ (Figures 6.2b and 6.1b).

(c) Varying error mean μ while setting the total budget $T = 1000$, the number of arms $n = 10$, the number of sensitive arms equal to 5, and the context dimension $d = 5$ (Figures 6.2c and 6.1c).

(d) Varying the number of sensitive arms while setting the total budget $T = 1000$, the number of arms $n = 10$, the error mean $\mu = 10$, and the context dimension $d = 2$ (Figures 6.2d and 6.1d).

The plots in Figure 6.1 show the percentage of times an algorithm pulled a sensitive arm over the full budget T. In order to be fair, the percentage of sensitive

[6] Additional experiments can be found in Appendix C.4.

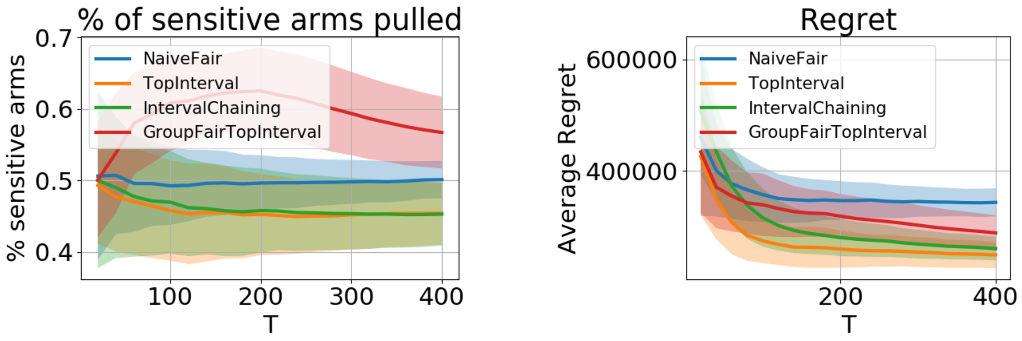
arms pulled should be proportional to the number of sensitive arms, i.e., when there are 2 sensitive arms out of the 10 total arms, the percentage of sensitive arms pulled is roughly 20%. The plots in Figure 6.2 show the perceived regret that includes bias ψ as solid lines, and real regret that corrects bias (See Equations 6.2 and 6.3) as dashed lines. Algorithms with low real regret are considered 'good'.

Figure 6.1a shows that once exploration is over, GroupFairTopInterval pulls sensitive arms roughly 50% of the time, matching the 50% of sensitive arms. Figure 6.2a shows that GroupFairTopInterval performs comparably on real regret as TopInterval performs on biased regret. This means GroupFairTopInterval should be used over TopInterval in contexts where bias is anticipated. Additionally NaiveFair performs poorly in the context of societal bias.

Figure 6.1b illustrates that IntervalChaining becomes more group fair as the number of arms increase. This is because many arms are chained together and therefore, arms are chosen uniformly at random. Figure 6.2b illustrates this random picking of arms as real regret and biased regret increases dramatically for IntervalChaining.

As expected, Figure 6.1c illustrates that when the error mean μ is large, both IntervalChaining and TopInterval choose fewer sensitive arms. This leads to a high real regret as shown in Figure 6.2c. Following Kleinberg et al. [133], Figure 6.2c also suggests that one cannot have both individual *and* group fairness in a scenario with high mean error. The randomness in NaiveFair leads to a very high regret for both perceived regret and real regret.

Figure 6.1d demonstrates the fairness property of proportionality. The percentage of sensitive arms pulled by GroupFairTopInterval matches the number of



(A) Sensitive arm pulls (%)

(B) Regret

FIGURE 6.3: Results of running contextual bandit algorithms on the family income and expenditure dataset. Figure 6.3a shows the percentage of pulls that were of sensitive arms. Figure 6.3b shows the biased regret for each of the algorithms. Note that the "real" regret like that shown in the synthetic experiments cannot be calculated.

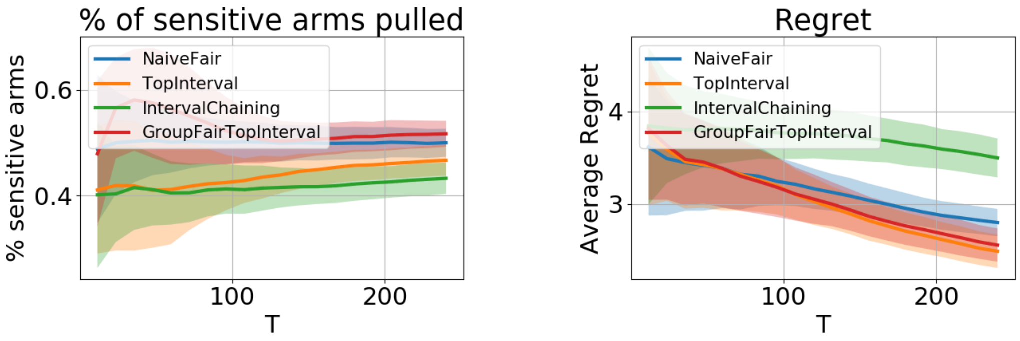
sensitive arms. As shown in Figure 6.2d, the number of sensitive arms does not affect the real regret of GROUPFAIRTOPINTERVAL.

6.4.2 Experiments on Real-World Data

After exploring GROUPFAIRTOPINTERVAL on synthetic data, we move on to using both the Philippines family income and expenditure dataset on Kaggle[7] and the ProPublica COMPAS dataset.[8] The family income dataset is from the Philippines and when one looks at the gender and age breakdown in the family income dataset, one can see that quite often female heads of households make more money than males in the Philippines." or some variation. This is most likely due to the large number of Filipino women who work out of the country. It is estimated that up to

[7] https://www.kaggle.com/grosvenpaul/family-income-and-expenditure

[8] https://www.kaggle.com/danofer/compass



(A) Sensitive arm pulls (%)

(B) Regret

FIGURE 6.4: Results on the COMPAS dataset. Figure 6.4a shows the percentage of pulls that were of sensitive arms. Figure 6.4b shows the biased regret for each of the algorithms. Note that the "real" regret like that shown in the synthetic experiments cannot be calculated.

20% of the GDP of the Philippines is actually remittances from these overseas—primarily female—workers.[9] In fact, almost 60% of overseas workers are women and 75% of these women are between the ages of 25 and 44.[10] ProPublica found that recidivism risk scores for African-Americans were generally higher than other races.[11]

Experimental setup.

Given the skew of high income coming from female head of households in the family income dataset, we treat the binary 'Household Head Sex' feature as the sensitive attribute. To create arms, we then split up households based on 'Household Head Age' bucketed into the following five groups: (8, 27], (27, 45], (45, 63], (63, 81],

[9] https://www.nationalgeographic.com/magazine/2018/12/filipino-workers-return-from-overseas-philippines-celebrates/

[10] https://psa.gov.ph/content/2017-survey-overseas-filipinos-results-2017-survey-overseas-filipinos

[11] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

and (81, 99]. We therefore have 10 different arms (for example, two arms would be Female head of household between 8 and 27, and Male head of household between 8 and 27).

Similarly, we treat African-American individuals from the COMPAS dataset as the sensitive attribute. We create arms by splitting up households based on the three age categories found in the data. We therefore have 6 different arms.

At each timestep t, we randomly select an individual from each arm. The context vector is the remaining features where any nominal features are transformed into integers. After an arm is pulled, a reward of the household income (for the family income dataset) or violent decile score (for the COMPAS dataset) is returned. Note that we use these datasets for illustrative purposes only.

Results.

We see the same behavior of arm pulls in the real world data. Figures 6.3a and 6.4a show that after a period of exploration, the percentage of sensitive arms (male-grouped arms) pulled gets very close to 50%, matching the proportion of sensitive-grouped arms.

Figures 6.3b and 6.4b are perhaps more interesting. Since we cannot measure the "real" regret without the bias we assumed from the sensitive-grouped arms, we consider the gap between GroupFairTopInterval and TopInterval as the price of fairness. The gap in regret is small compared to the increase in percentage of sensitive arms pulled. However, the gap in regret for NaiveFair is large in comparison. This suggests that explicitly learning a societal bias term will help in biased settings with low price to perceived regret.

6.5 Discussion and Conclusions

Our new definitions of reward (Equation 6.2) and regret (Equation 6.3) for the MAB setting provide an opportunity to look at biased data in a new light. In many cases, ground truths provided during learning are noisy with respect to sensitive groups. Additionally, debiased ground truths may be very expensive to receive or may take a long time to acquire. For instance, if looking at loans, true rewards of repayment may take years to receive. Or, for example, in hiring—the true reward of hiring an individual may take over a year to estimate, while the initial estimate may be influenced by a hiring team's unconscious bias over features such as ethnicity, gender, or orientation. Our proposed algorithm, GroupFairTopInterval, learns societal bias in the data while still being able to differentiate between individual arms. Previous solutions relied on setting ad-hoc thresholds, requiring some form of quota, or choosing groups uniformly at random. These solutions either lead to high regret, or require a large amount of domain knowledge for the chosen application. Indeed, our solution gets even closer to mitigating the social injustice bias found in the three tiered view of the world from Chapter 1 than the previous two chapters. It explicitly assumes disparate treatment of sensitive groups where there should be equal treatment.

Our main contributions are:

- We provide a new definition of reward and regret which captures societal bias.
- We provide an algorithm that learns and corrects for that definition of societal bias.

- We empirically explore the effects different CMAB algorithms have in the setting of societal bias.

This chapter provides an initial look at group fairness in the contextual multi-armed bandit (MAB) setting. Future work could expand GroupFairTopInterval to enforce individual fairness within groups. Intersectional group fairness is also important to look at in the MAB setting where more than one type of sensitive attribute needs to be protected. Additionally, other group fairness definitions such as Equalized Opportunity should be converted to the MAB setting [103]. Another interesting direction for future work is to mix ideas from the study of budget constrained bandits [79, 226] with our fairness definitions. We have also assumed individual arms have fixed group membership; generalizing to a setting where memberships in protected groups may change at every timestep t would fit more real world applications.

6.6 Authors and Publication

This chapter was written by Candice Schumann, Zhi Lang, Nicholas Mattei, and John P. Dickerson. It is under submission to the International Conference on Machine Learning 2020. It was presented at the Do the right thing: machine learning and causal inference for improved decision making Workshop at NeurIPS (2019).

Chapter 7: Transfer of Machine Learning Fairness Across Domains

If our models are used in new or unexpected cases, do we know if they will make fair predictions? Previously, researchers developed ways to debias a model for a single problem domain. However, this is often not how models are trained and used in practice. For example, labels and demographics (sensitive attributes) are often hard to observe, resulting in auxiliary or synthetic data to be used for training, and proxies of the sensitive attribute to be used for evaluation of fairness. A model trained for one setting may be picked up and used in many others, particularly as is common with pre-training and cloud APIs. Despite the pervasiveness of these complexities, remarkably little work in the fairness literature has theoretically examined these issues. We frame all of these settings as domain adaptation problems: how can we use what we have learned in a source domain to debias in a new target domain, without directly debiasing on the target domain as if it is a completely new problem? We offer new theoretical guarantees of improving fairness across domains, and offer a modeling approach to transfer to data-sparse target domains. We give empirical results validating the theory and showing that these modeling approaches can improve fairness metrics with less data.

7.1 Introduction

Much of machine learning research, and especially machine learning fairness, focuses on optimizing a model for a single use case [2, 30]. However, the reality of machine learning applications is far more chaotic. It is common for models to be used on multiple tasks, frequently different in a myriad of ways from the dataset that they were trained on, often coming at significant cost [200]. This is especially concerning for machine learning fairness – we want our models to obey strict fairness properties, but we may have far less data on how the models will actually be used. How do we understand our fairness metrics in these more complex environments?

In traditional machine learning, domain adaptation techniques are used when the distribution of training and validation data does not match the target distribution that the model will ultimately be tested against. Therefore, in this work we ask: if the model is trained to be "fair" on one dataset, will it be "fair" over a different distribution of data? Instead of starting again with this new dataset, can we use the knowledge gained during the original debiasing to more effectively debias in the new space?

It turns out that this framing covers many important cases for machine learning fairness. We will use, as a running example, the task of income prediction, where some decisions will be made based on the person's predicted income and we want the model to perform "fairly" over a sensitive attribute such as gender. We primarily follow the *equality of opportunity* [103] perspective where we are concerned with one group (broken down by gender or race) having worse accuracy than another. In this setting, there are a myriad of fairness issues that arise that we find domain adaptation can shed light on:

Lacking sensitive features for training: There may be few examples where we know the sensitive attribute. In these cases, a proxy of the sensitive attribute have been used [102], or researchers need very sample-efficient techniques [2, 30]. For distant proxies, researchers have asked how well fairness transfers across attributes [138]. Here the sensitive attribute differs in the source and target domains.

Data is not representative of application: Dataset augmentation, models offered as an API, or models used in multiple unanticipated settings, are all increasingly common design patterns. Even for machine learning fairness, researchers often believe limited training data is a primary source of fairness issues [56] and will employ dataset augmentation techniques to try to improve fairness [80]. How can we best make use of auxiliary data during training and evaluation when it differs in distribution from the real application?

Multiple tasks: In some cases having accurate labels for model training is difficult and instead proxy tasks with more labeled data are used to train the model, e.g., using pre-trained image or text models or using income brackets as a proxy for defaulting on a loan. Again we ask: when does satisfying a fairness property on the original task help satisfy that same property on the new task?

Each of these cases are common throughout machine learning but present challenges for fairness. In this work, we explore mapping domain adaptation principles to machine learning fairness. In particular, we offer the following contributions:

1. **Theoretical Bounds:** We provide theoretical bounds on transferring equality of opportunity and equality of odds metrics across domains. Perhaps more importantly, we discuss insights gained from these bounds.

2. **Modeling for Fairness Transfer:** We offer a general, theoretically-backed modeling objective that enables transferring fairness across domains.

3. **Empirical validation:** We demonstrate when transferring machine learning fairness works successfully, and when it does not, through both synthetic and realistic experiments.

See Chapter 2 for related works related to this Chapter.

7.2 Problem Formulation

We begin with some notation to make precise the problem formulation. Building on our running example we have two domains: a source domain $Z \sim \mathcal{D}_S$, which is a feature distribution influenced by sensitive attribute $A_S \in \mathcal{A}_S$ (e.g., $\Pr_{Z \sim \mathcal{D}_S}[Z|A_S = \textit{male}] \neq \Pr_{Z \sim \mathcal{D}_S}[Z|A_S = \textit{female}]$), as well as a target domain $\mathcal{D}_T$ influenced by sensitive attribute $A_T \in \mathcal{A}_T$ (e.g., $\Pr_{Z \sim \mathcal{D}_T}[Z|A_T = \textit{black}] \neq \Pr_{Z \sim \mathcal{D}_T}[Z|A_T = \textit{white}]$). In order for this to be a domain adaptation problem, we assume $\Pr_{Z \sim \mathcal{D}_S}[Z|A_S] \neq \Pr_{Z \sim \mathcal{D}_T}[Z|A_T]$. Note, this can be true even if $\mathcal{D}_S = \mathcal{D}_T$ but the distributions conditioned on A_S and A_T differ. We focus on binary classification tasks with label $Y \in \mathcal{Y}$, e.g. income classification is shared over both domains. For this task we can create a classifier by finding a hypothesis $g : \mathcal{D} \to \mathcal{Y}$ from a hypothesis space $\mathcal{H}$.

Let us assume that we can learn a "fair" classifier g for the source domain and task. If we use a small amount of data from the target domain, will the fairness from the source sensitive attribute A_S transfer to the target domain and sensitive attribute A_T? We can define the notion of a "fairness" distance – how far away the classifier is from perfectly fair – in a given domain S as Δ_{Fair_S}. Within this formulation we consider two definitions of fairness.

The first distance is *equality of opportunity* [103]. A classifier is said to be fair under equality of opportunity if the false positive rates (FPR) over sensitive attributes are equal. In other words if we have a binary sensitive attribute A, then equality of opportunity requires that $\Pr(\hat{Y} = 1|A = 0, Y = 0) = \Pr(\hat{Y} = 1|A = 1, Y = 0)$, where $\hat{Y}$ gives the outcome of classifier g. Thus, how far away a classifier g is from equal opportunity (or the fairness distance of equal opportunity) can be defined as

$$\Delta_{EOp_S}(g) \triangleq \left|\mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}[g(Z_0^0)] - \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}[g(Z_1^0)]\right|,$$

where $\mathcal{D}_{S_\alpha^l} = P_{Z \sim \mathcal{D}_S}[Z|A = \alpha, Y = l]$. In our running example $\Delta_{EOp_S}(g)$, where A_S is gender, is the difference between the likelihood that a low-income man is predicted to be high-income and the likelihood that a low-income woman is predicted to be high-income. A symmetric definition and set of analysis can be made for false negative rate (FNR).

The second definition of fairness which we consider is *equalized odds* [103]. A classifier is said to be fair under equalized odds if both the FPR *and* FNR over the sensitive attribute are equal: Similar to equal opportunity, we define the fairness distance of equalized odds as:

$$\begin{aligned}\Delta_{EO_S}(g) \triangleq &\left|\mathbb{E}_{Z_0^0 \sim \mathcal{D}_{S_0^0}}[g(Z_0^0)] - \mathbb{E}_{Z_1^0 \sim \mathcal{D}_{S_1^0}}[g(Z_1^0)]\right| \\ &+ \left|\mathbb{E}_{Z_0^1 \sim \mathcal{D}_{S_0^1}}[1 - g(Z_0^1)] - \mathbb{E}_{Z_1^1 \sim \mathcal{D}_{S_1^1}}[1 - g(Z_1^1)]\right|.\end{aligned}$$

Again using our running example, the distance of equalized odds in the source domain is given by the difference of expected FPRs between females and males (as above), plus the difference of expected FNRs (high-income predicted to be low-income) between females and males.

Given a classifier g that has a fairness guarantee in the source domain, the fairness distance in the target domain should be bounded by the fairness distance in the source domain:

$$\Delta_{Fair_T}(g) \leq \Delta_{Fair_S}(g) + \epsilon \tag{7.1}$$

The key question we hope to answer is: what is ϵ?

7.3 Bounds on Fairness in the Target Domain

To expand inequality (7.1) we need to start with some definitions. Given a hypothesis space $\mathcal{H}$ and a true labeling function $f(Z) : \mathcal{D} \to \mathcal{Y}$, we can define the error of a hypothesis $g \in \mathcal{H}$ as $\epsilon_S(g, f) = \mathbb{E}_{Z \sim \mathcal{D}_S}[|f(Z) - g(Z)|]$, the expectation of disagreement between the hypothesis g and the true label f. We can then define the ideal joint hypothesis that minimizes the combined error over both the source and target domains as $g^* = \arg\min_{g \in \mathcal{H}} \epsilon_S(g, f) + \epsilon_T(g, f)$.

Following Ben-David et al. [24] we define the $\mathcal{H}$-divergence between probability distributions as

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{g \in \mathcal{H}} \left| \Pr_{\mathcal{D}}[I(g)] - \Pr_{\mathcal{D}'}[I(g)] \right|, \tag{7.2}$$

where $I(g)$ is the set for which $g \in \mathcal{H}$ is the characteristic function ($Z \in I(g) \Leftrightarrow g(Z) = 1$). We can compute an approximation $\hat{d}_{\mathcal{H}}(\mathcal{D}, \mathcal{D}')$ by finding a hypothesis h that finds the largest difference between the samples from $\mathcal{D}$ and $\mathcal{D}'$ [23]. This divergence can be used to look at the differences in distributions, which is important when moving from a source domain to a target domain.

Additionally, following Ben-David et al. [24], we define the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$ as the set of hypotheses

$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(Z) = h(Z) \oplus h'(Z) \quad \text{for some } h, h' \in \mathcal{H}, \tag{7.3}$$

where $\oplus$ is the XOR function. The symmetric difference hypothesis space is used to find disagreements between a potential classifier g and a true labeling function f.

Theorem 7.1. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_0^0}$, $\mathcal{U}_{S_1^0}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^0}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^0}$, $\mathcal{D}_{S_1^0}$, $\mathcal{D}_{T_0^0}$, and $\mathcal{D}_{T_1^0}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\begin{aligned}\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0}) \\ &+ 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0,\end{aligned}$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Using both the definition of $\mathcal{H}$-divergence and symmetric difference hypothesis space, Theorem 7.1 provides a VC-dimension bound on the equal opportunity distance in the target domain given the equal opportunity distance in the source domain. Due to space limitations, full proofs for all theorems can be found in Appendix D.2.

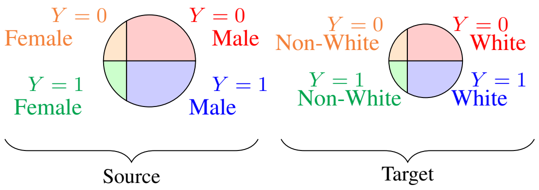


Figure 7.1: Both the source and target distributions can be split into four quadrants: 1) negative minority 2) negative majority 3) positive minority and 4) positive majority.

This theorem provides insights on when domain adaptation for fairness can be used. Firstly the $\hat{d}$ terms in the bound suggest that 1) the source and target distributions of negatively labeled items that have a sensitive attribute label of 0 should be close, and 2) the source and target distributions of the negatively labeled items that have a sensitive attribute label of 1 should be close. In Figure 7.1 the red quadrants should be close to the red quadrants while the orange quadrants should be close to the orange quadrants across domains. In traditional domain adaptation, ignoring fairness, the entire domains should be close (the entire circle), which means that if there are few minority data-points then the distance of the minority spaces will be ignored. The fairness bound instead puts equal emphasis on both the majority and minority.

Secondly, the λ terms become small when the hypothesis space contains a function g^* that has low error on both the source and target space on the two negative segments in each domain (the red and orange spaces in Figure 7.1). Since we are looking at equal opportunity, the function g^* only needs to have low error on the negative space for both the majority and minority. Therefore, we can use the trivial function $g^*(Z) = 0$ and the λ terms go to 0.

Lastly, Theorem 7.1 depends on the VC-dimension d. Since bounds with VC-dimensions explode with models like neural networks, we also provide bounds using Rademacher Complexity in Appendix D.1.

Equalized odds, while similar to equal opportunity, is a stricter fairness constraint. Theorem 7.2 provides a VC-dimension bound on the difference of equal odds in the target domain given the source domain.

Theorem 7.2. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S^l_\alpha}$ are samples of size m', each drawn from $\mathcal{D}_{S^l_\alpha}$ for all $\alpha \in \mathcal{A} = \{0,1\}$ and $l \in \mathcal{Y} = \{0,1\}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_0},\mathcal{U}_{S^0_0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_1},\mathcal{U}_{S^0_1}) \\ + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^1_0},\mathcal{U}_{S^1_0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^1_1},\mathcal{U}_{S^1_1}) + 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda^0_0 + \lambda^0_1 + \lambda^1_0 + \lambda^1_1$, and $\lambda^l_\alpha = \epsilon_{S^l_\alpha}(g^*, f) + \epsilon_{T^l_\alpha}(g^*, f)$.

The $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ terms suggest, that in order for equalized odds to transfer successfully then, 1) the source and target distributions of negatively labeled items on *both* sensitive attribute labels 0 and 1 should be close, 2) the source and target distributions of the positively labeled items on *both* sensitive attribute labels 0 and 1 should be close. In other words, all four quadrants of the source should individually be close to the respective four quadrants of the target in Figure 7.1.

Additionally, the λ term shows that there should be a hypothesis that performs well over *all* of these subspaces. This implication is intuitive given that equalized

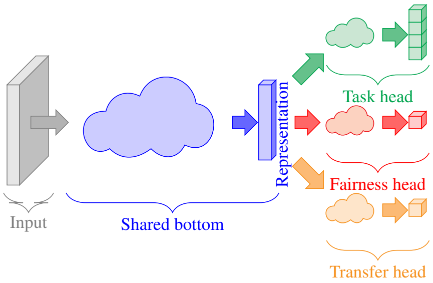


FIGURE 7.2: At a high level, our general framework combines a primary training objective, a fairness objective, and a transfer objective to improve fairness goals in a target domain. Table 7.1 provides mathematical details for different configurations.

odds, by definition, wants a classifier to perform well in both the negative and positive space across both groups.

7.4 Modeling to Transfer Fairness

With this theoretical understanding, how should we change our training? As motivated previously, we consider the case where we have a small amount of labelled data (both labels $\mathcal{Y}$ and sensitive attributes $\mathcal{A}$) in the target domain and a large amount of labelled data in the source domain.

As shown in the previous section, equality of opportunity will transfer *if* the distance between the respective distributions of source and target are close together as visually portrayed in Figure 7.1. Ganin et al. [92] proved that traditional domain adaptation can be framed as minimizing the distance between source and target with adversarial training. [30, 82, 146, 156] similarly have applied adversarial training to achieve fairness goals, and Madras et al. [162] proved that equality of odds can be optimized with adversarial training similar to domain adaptation.

Loss Term	Theorem 1	Adversarial (Eq. 7.4)	Regularization (Eq. 7.5)
Fairness head	$\Delta_{EOp_S}(g)$	$\lambda_{Fair} L_A \left(a(h(Z^0)), A\right)$	$\lambda_{Fair} L_{MMD} \left(a(h(Z^0)), A\right)$
Transfer head	$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_0}, \mathcal{U}_{S^0_0})$	$\lambda_{DA} L_d \left(d(h(Z^0_0)), d\right)$	$\lambda_{DA} L_{MMD} \left(d(h(Z^0)), d\right)$
	$\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_1}, \mathcal{U}_{S^0_1})$	$\lambda_{DA} L_d \left(d(h(Z^0_1)), d\right)$	

TABLE 7.1: Relationship between terms in Theorem 7.1 and Loss functions

We build on this intuition to design a learning objective for transferring equality of opportunity to a target domain. Adversarial training conceptually enables minimizing a $\hat{d}$ term from Theorem 7.1; and Δ_{Fair_S} can be optimized using [30, 162] or one of the other myriad of traditional fairness learning objectives. As such, we begin with the following loss:

$$\min \left[\sum_{Z \sim (\mathcal{D}_S \cup \mathcal{D}_T)} L_Y(g(h(Z)), f(Z)) + \sum_{(A, Z^0) \sim \mathcal{D}_{S0}} \lambda_{Fair} L_A \left(a(h(Z^0)), A\right) \right.$$
$$\left. + \sum_{(d, Z^0_0) \sim \left(\mathcal{D}_{S^0_0} \cup \mathcal{D}_{T^0_0}\right)} \lambda_{DA} L_d \left(d(h(Z^0_0)), d\right) + \sum_{(d, Z^0_1) \sim \left(\mathcal{D}_{S^0_1} \cup \mathcal{D}_{T^0_1}\right)} \lambda_{DA} L_d \left(d(h(Z^0_1)), d\right) \right], \tag{7.4}$$

where $L_Y(g(h(Z)), f(Z))$ is the loss function training $g(h(Z))$ over hidden representation $h(Z)$ to predict the task label $f(Z)$. To optimize Δ_{Fair_S}, $a(h(Z^0))$ tries to predict the sensitive attribute A from the source and $L_A \left(a(h(Z^0)), A\right)$ provides an adversarial loss that includes a negated gradient on h following [30]. For transfer, we minimize $\hat{d}$ terms by including another adversarial loss $L_d \left(d(h(Z^\alpha_l)), d\right)$, where $d(h(Z^\alpha_l))$ tries to predict whether a sample comes from the source or target domain. Each of these loss components maps to terms in Theorem 7.1 as laid out in Table 7.1.

Recently, Zhang et al. [234] used adversarial training on a one dimensional representation of the data (effectively the model's prediction). From this perspective, we can use a wide variety of losses over predictions to replace adversarial losses, such as [31, 232] minimizing the correlation between group and the one dimensional representation of the data. Like previous work, we find that these approaches to be more stable and still effective in comparison to adversarial training, despite not being provably optimal. In our experiments we use a MMD loss [40, 99, 154] over predictions:

$$\min \Bigg[\sum_{Z \in \mathcal{D}_S \cup \mathcal{D}_T} L_Y(f(Z), g(Z)) + \sum_{(A, Z^0) \sim \mathcal{D}_{S^0}} \lambda_{Fair} L_{MMD}\left(a(h(Z^0)), A\right) + \sum_{(d, Z^0) \sim \left(\mathcal{D}_{S^0} \cup \mathcal{D}_{T^0}\right)} \lambda_{DA} L_{MMD}\left(d(h(Z^0)), d\right) \Bigg], \tag{7.5}$$

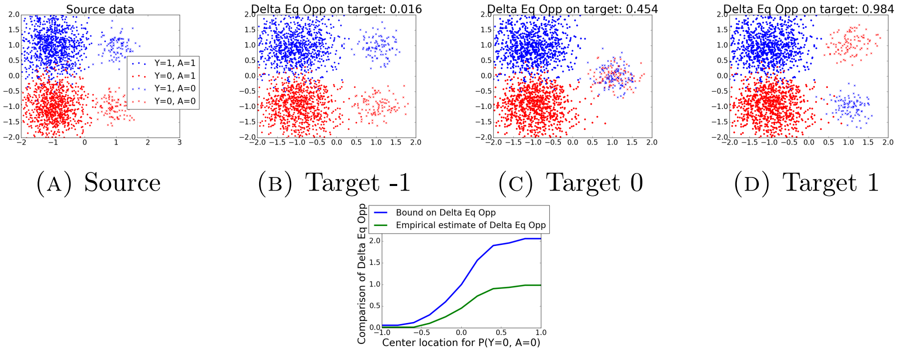
where $\lambda_{Fair} L_{MMD}\left(a(h(Z^0)), A\right)$ is the MMD regularization over the sensitive attributes in the source domain, $\lambda_{DA} L_{MMD}\left(d(h(Z^0)), d\right)$ is the MMD regularization over source/target membership. Again Table 7.1 maps the terms in Eq. 7.5 to those in Theorem 7.1.

Care must be taken when performing domain adaptation with regards to fairness. Either multiple transfer heads should be included in the loss for all necessary quadrants (See Figure 7.1 and Eq. 7.4), or balanced data – equally representing all necessary quadrants – should be used as in [162] and Eq. 7.5. Experiments in this chapter use the MMD regularization as in Eq. 7.5 and balanced data is used for both the fairness head as well as the transfer heads.

7.5 Experiments

To better understand the theoretical results presented above, we now present both synthetic and realistic experiments exploring tightness of our theoretical bound as well as the ability to improve the transfer of fairness across domains during model training.

7.5.1 Synthetic Examples



(A) Source (B) Target -1 (C) Target 0 (D) Target 1

(E) Fairness

FIGURE 7.3: Synthetic examples showing how distribution difference of $P(Z|Y, A = 0)$ in the target domain affects theoretical and empirical equality of opportunity (best viewed in color). In the title of each plot we give the equal opportunity distance $\Delta_{EOp_T}(g)$ in the target domain.

We show how well the theoretical bounds align with actual transfer of fairness. A synthetic dataset is used to examine how the distribution distance terms $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^{Y=0}_{A=0}}, \mathcal{U}_{S^{Y=0}_{A=0}})$ and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^{Y=0}_{A=1}}, \mathcal{U}_{S^{Y=0}_{A=1}})$ in Eq. (7.1) affect the fairness distance of equal opportunity $\Delta_{EOp_T}(g)$.

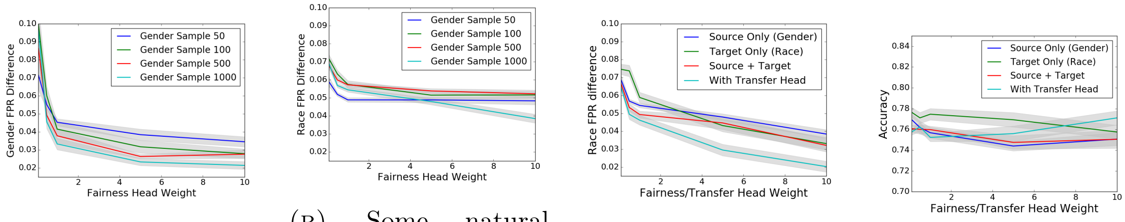
In this synthetic example, we generate data $Z \in \mathbb{R}^2$ using Gaussian distributions. As we can see in Figure 7.3a, the source domain consists of four Gaussians, with

$Y = 1$ largely lying above $Y = 0$ and $A = 1$ lying to the left of $A = 0$; $A = 1$ is the majority of the data ($\sigma = 0.5$ with 900 samples). For $A = 0$, the data is generated using $\sigma = 0.3$ with 100 samples. The target domain, like the source domain, consists of majority data with $A = 1$ and the data from $A = 1$ is generated from the same distribution in both domains: $\mathcal{U}_{T^{Y=0}_{A=1}} \sim \mathcal{N}([-1, -1], \sigma)$ and $\mathcal{U}_{T^{Y=1}_{A=1}} \sim \mathcal{N}([-1, 1], \sigma)$. However, in order to understand the transfer of fairness, we shift the distributions of $\mathcal{U}_{T^{Y=0}_{A=0}} \sim \mathcal{N}([1, c], \sigma)$ and $\mathcal{U}_{T^{Y=1}_{A=0}} \sim \mathcal{N}([1, -c], \sigma)$ in the target domain ($c = -1, 0, 1$ for 7.3b, 7.3c and 7.3d, respectively). By varying the overlap between these distributions, and their alignment with the source data, we are able to understand the relationship between the $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$ terms above and the fairness distance of equal opportunity $\Delta_{EOp_T}(g)$. For each setting, we train linear classifiers on the source domain and examine the performance in the target domain.

Qualitative Analysis We see in Fig. 7.3b that when the distribution $P(Z|Y = 0, A = 0)$ across domains is close, thus a smaller $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_0}, \mathcal{U}_{S^0_0})$, there is better transfer of fairness the source to the target domain, seen in the smaller $\Delta_{EOp_T}(g)$. As the distribution distance gets larger, the $\Delta_{EOp_T}(g)$ also increases. Consider the worst case of a sign flip for the minority $A = 0$, as shown in Fig. 7.3d: the FPR for the majority $A = 1$ is close to 0%, while the FPR for the minority $A = 0$ is close to 100%.

Quantitative Analysis In Figure 7.3e, we compare the derived bound of $\Delta_{EOp_T}(g)$ (Eq. 7.1) with its empirical estimate as we vary c[1]. As shown in Figure 7.3e, the theoretical bound on the equal opportunity distance is close to the observed equal

[1] As in [23], $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T^0_0}, \mathcal{U}_{S^0_0})$ is estimated by a linear classifier trained on samples $\mathcal{U}_{T^0_0}, \mathcal{U}_{S^0_0}$. The plot omits the VC term for simplicity, which is relatively small when sample size m' is large and VC-dimension d is low.



(A) Effect of fairness head: Improving $\Delta_{EOp_{\text{gender}}}$ with varying number of gender-balanced samples.

(B) Some natural transfer occurring without explicit transfer: $\Delta_{EOp_{\text{race}}}$ is improved with gender data.

(C) Effect of transfer head: better transfer from gender (1000 samples) to race (50 samples).

(D) Accuracy graph for transferring from gender (1000 samples) to race (50 samples).

FIGURE 7.4: Effect of fairness/transfer head on the UCI data. The shaded areas show the standard error of the mean across trials. Note the head weight (x-axis) starts from 0.1.

opportunity distance when the distance between the negative minority space across domains, $\hat{d}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0})$, is small. This suggests, minimizing the domain distance terms in Eq. 7.1 could lead to a better equal opportunity transfer.

7.5.2 Real Data

We now explore how and when our proposed modeling approach in Section 7.4 facilitates the transfer of fairness from the source to the target domain on two real-world datasets. Note, we use these datasets exclusively for understanding our theory and model, and *not* as a comment on when or if the proposed tasks and their application are appropriate, as in [2].

Dataset 1: The UCI Adult[2] dataset contains census information of over 40,000 adults from the 1994 Census, with the task of determining income brackets of $>$ $\$50,000$ or $\leq \$50,000$. We focus on two sensitive attributes: binary valued gender,

[2] https://archive.ics.uci.edu/ml/datasets/adult

and race, converted to binary values [‘white’, ‘non-white’] as done by Madras et al. [162].

Dataset 2: As in [2] we use ProPublica’s COMPAS recidivism data[3] to try to predict recidivism for over 10,000 defendants based on age, gender, demographics, prior crime count, etc. We again focus on two sensitive attributes: gender and race (binarized to [‘white’, ‘non-white’]).

Experiment Setup For both datasets, cross-validation is used to choose the hyper-parameters. Comparable baseline accuracy (around 84% for Dataset 1 and 80% for Dataset 2, see appendix D.4 for more details) is achieved with 64 embedding dimension for categorical features, single hidden layer with 256 shared hidden units, 512 batch size, 0.1 learning rate with Adagrad optimizer, and $10,000$ epochs for training. We perform 30 runs for each set of experiments and average over the results.

Sparsity Issues and Natural Transfer We examine the effectiveness of just the fairness heads in the proposed model. The amount of gender-balanced data created for the fairness head is varied to observe how applying the fairness head affects the FPR difference.

We examine how this procedure effects the FPR difference across genders (i.e., the FPR difference between “Female” and “Male” examples). Figure 7.4a shows that the fairness head works as expected: with sufficient data and a large enough weight, the fairness head is able to improve the FPR gap across genders. Further, we find

[3] https://github.com/propublica/compas-analysis

that with very few examples on which to apply the fairness head, the gender FPR gap does not close. This aligns with previous results found in [30, 31, 162].

Second, we examine how running the fairness head on gender affects the FPR gap across race. As shown in Figure 7.4b, there is a natural transfer of equal opportunity from gender to race – applying a fairness loss with respect to gender also improves the fairness of the model with respect to race. This highlights that sometimes there is a natural transfer of equal opportunity, presenting general value in improving the FPR gap with respect to gender, and no explicit transfer optimization is needed. (Similar to the transfer questions posed previously by Madras et al. [162] and Gupta et al. [102]).

Effectiveness of Transfer Head We now explore how adding the transfer head can further improve equality of opportunity in the target domain. We compare four different model arrangements: (1) **Source Only**: We only add a fairness head for the source domain; (2) **Target Only**: We only add a fairness head for the target domain; (3) **Source+Target**: We add two fairness heads, one for source and for target; (4) **Transfer**: We include three heads – both source and target fairness heads as well as the transfer head for equality of opportunity.

Experiment setting: As in typical transfer learning setting, we will focus on the case where we observe a large number of samples in the source domain (e.g., 1000 for each race "white" and "non-white"), but a smaller sample size in the target domain (e.g., 100 for each gender "male" and "female"), and the same for gender to race. We explore equality of opportunity with respect to FPR in the target domain, as we vary the weight on the fairness and transfer heads.

Results: Figure 7.4c shows that including the transfer head results in a better equal opportunity transfer, compared to the same setting without transfer (Figure 7.4b). Table 7.2 summarizes the full results on both datasets. We can see that including both the fairness heads and the transfer head consistently gives the best improvement in equal opportunity (FPR difference) in almost all cases.

Effect of Target Sample Size Last, we consider how the amount of data from the target domain affects our ability to improve equal opportunity there, as sample efficiency is a core challenge.

Experiment setting: We follow a similar experimental procedure as before with two modifications. First, we vary the number of samples we observe for each sensitive group in the target domain to be in $\{50, 100, 500, 1000\}$. We examine the efficacy of the four approaches depending on the amount of data available for debiasing in the target domain. Second, this analysis is performed for both transferring from race (source) to gender (target), as well as from gender (source) to race (target).

Results: Table 7.2 summarizes the results. Applying the fairness and transfer heads to the large amount of source data closes the FPR gap in the target domain. Increasing the amount of data in the target domain significantly helps the performance of the "Target Only" and the "Source+Target" models. This is intuitive since directly debiasing in the target domain is feasible with sufficient data. With sufficient data, the results converge to be approximately equivalent to the transfer model.

These experiments show that the transfer model is effective in decreasing the FPR gap in the target domain and is more sample efficient than previous methods.

			Smallest FPR difference achieved on Target (FPR-diff ± std. dev)			
	Source to Target	#Target Samples	Source only	Target only	Source + Target	With Transfer Head
Dataset 1	Gender to Race	50	0.038 ± 0.013	0.033 ± 0.019	0.032 ± 0.020	**0.020 ± 0.016**
		100	**0.038 ± 0.013**	**0.038 ± 0.021**	0.044 ± 0.024	0.040 ± 0.024
		500	0.038 ± 0.013	0.053 ± 0.010	0.043 ± 0.017	**0.025 ± 0.018**
		1000	0.038 ± 0.013	**0.027 ± 0.018**	**0.027 ± 0.019**	0.031 ± 0.021
	Race to Gender	50	0.061 ± 0.054	0.035 ± 0.015	0.020 ± 0.026	**0.008 ± 0.009**
		100	0.061 ± 0.054	0.028 ± 0.014	0.021 ± 0.015	**0.009 ± 0.011**
		500	0.061 ± 0.054	0.028 ± 0.013	0.019 ± 0.013	**0.014 ± 0.011**
		1000	0.061 ± 0.054	0.021 ± 0.012	**0.015 ± 0.014**	0.020 ± 0.014
Dataset 2	Gender to Race	50	0.027 ± 0.008	0.041 ± 0.006	0.009 ± 0.004	**0.001 ± 0.001**
		100	0.027 ± 0.008	0.036 ± 0.007	0.005 ± 0.005	**0.003 ± 0.001**
		500	0.027 ± 0.008	0.038 ± 0.008	0.003 ± 0.002	**0.001 ± 0.001**
		1000	0.027 ± 0.008	0.021 ± 0.005	0.006 ± 0.005	**0.002 ± 0.001**
	Race to Gender	50	0.040 ± 0.004	0.070 ± 0.005	0.035 ± 0.004	**0.019 ± 0.002**
		100	0.040 ± 0.004	0.055 ± 0.007	0.034 ± 0.003	**0.017 ± 0.002**
		500	0.040 ± 0.004	0.042 ± 0.008	0.027 ± 0.004	**0.019 ± 0.002**
		1000	0.040 ± 0.004	0.034 ± 0.011	0.028 ± 0.004	**0.018 ± 0.002**

TABLE 7.2: Comparison between the proposed model and the baselines. The numbers in bold indicate the smallest FPR difference achieved in the target domain w.r.t. varying number of target samples.

7.6 Conclusion

In this work we provide the first theoretical examination of transfer of machine learning fairness across domains. We adopt a general formulation of domain adaptation for fairness that covers a wide variety of fairness challenges, from proxies of sensitive attributes, to applying models in unanticipated settings. Within this general formulation, we have provided theoretical bounds on the transfer of fairness for equal opportunity and equalized odds using both VC-dimension and Rademacher Complexity. Based on this theory, we developed a new modeling approach to transfer fairness to a given target domain. In experiments we validate our theoretical results and demonstrate that our modeling approach is more sample efficient in improving fairness metrics in a target domain.

This Chapter takes a look at mitigating measurement bias found in the three level view of the world from Chapter 1. A combination of social injustice bias and measurement bias can lead to bias augmentation in models. The balanced learning method developed here helps deal with unbalanced data by focusing on equal opportunity.

7.7 Authors and Publication

This chapter was written by Candice Schumann (while working at Google), Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, Ed H. Chi. It is under submission to the International Joint Conference on Artificial Intelligence 2020. It was presented at the AI for Social Good Workshop at NeurIPS 2019, and the Ph.D. in Research Conference at Google 2019.

Chapter 8: Research Extensions

8.1 A Multi-Stage Human-Machine Framework for Mental Health Risk Assessment

Machine learning is beginning to have a large impact on the ways that people think about addressing problems in healthcare [159, 235] and mental health [7, 149, *inter alia*], just as it is having large impacts everywhere else. The ability to obtain data about people's day to day thoughts and experiences via social media—unobtrusive windows into what Coppersmith et al. [67] call the "clinical whitespace" between clinician encounters, in the form of social media posts, wearables data, etc.—is looking to be thoroughly disruptive, and the ability to engage with people via natural spoken interactions on all manner of electronic devices creates potential for even more

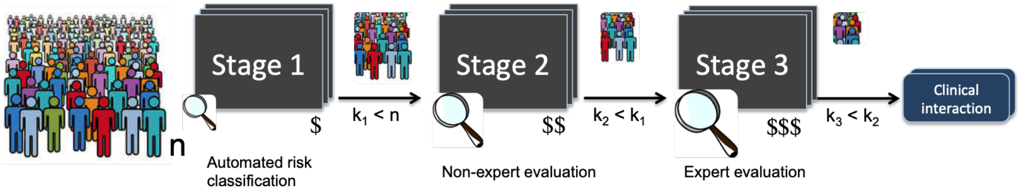


Figure 8.1: We apply a multi-armed bandit framework in mental health to identify at-risk individuals, progressing from automated analysis of social media posts, to risk evaluation by non-experts, to expert evaluation. The goal is to optimize the number of people at high risk who go on to receive detailed clinical attention, given limited resources.

windows into people's everyday thoughts and experiences, enhancing the ability to detect new problems earlier and monitor patients under treatment more effectively and at lower cost.

It is becoming clear that traditional approaches to these problems do not suffice. Franklin et al. [88], for example, conclude from a large meta-analysis that there has been no improvement in predictive ability for suicidal thoughts and behaviors over the last 50 years, and argue their findings "suggest the need for a shift in focus from *risk factors* to machine learning-based risk *algorithms*" (their emphasis). The technological community is increasingly aware of this problem space and enthusiastic about contributing [e.g. 155, 167, 236], with significant progress in ethical data collection [67, 176] and effective use of those data in predictive models [67, 71, 117, 121, 168].

In this work we introduce a concrete technological proposal for addressing this problem, involving a basic shift in the way we think about machine learning in mental health: the dominant paradigm of individual-level classification is not an end in itself; rather it provides components in a population-based framework involving both machines and humans, where limited resources give rise to a critical need for effective and appropriate ways to set priorities.

At the core of our technical approach is the recognition that the multi-armed bandit problem in machine learning is a good fit for the real-world scenario created by scaling up the application of technology for detection and monitoring in mental health: what is the best way to allocate limited resources among competing choices, given only limited information? We adopt a *tiered* multi-armed bandit formulation originally introduced with application to hiring or admissions decisions [198], where a succession of stages is applied to a population of applicants, each stage successively more expensive but also more informative, in order to optimize the value of the set of applicants who are chosen (See Figure 8.1).

To briefly summarize the model, we cast tiered decision making as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting [58]. Here, arms represents individuals with latent true risk profiles. The end goal is to select a subset for clinical interaction, after narrowing the pool over successive stages or tiers. In our current model we have three stages of assessment: (1) automated risk classification using an NLP model, (2) non-expert risk assessment, and (3) expert risk assessment.[1]

Our key insight is that, by replacing a population of potential hires with a population of people with potential mental health problems, and by replacing "value" with "risk", this tiered framework maps directly to a population-level formulation of the assessment problem. Using real data and human annotation, our simulations demonstrate the value of using this framework to combine (cheap, less accurate) automation with (more expensive, more accurate) human evaluation of social media in order to identify individuals within a population who are at high risk for a suicide attempt.

Our MAB approach outperforms all comparable baselines. On average, our MAB approach more than doubles the population sensitivity of the expert baseline for the same resource amount. These results are only a first step on the way to

[1] Although we approximate an intermediate stage of non-experts using crowdsourced judgments, the idea of true crowdsourcing, in the sense of Mechanical Turk and similar platforms, need not, and should not, be considered a part of the proposal. Rather, we use crowdsourcing to approximate an intermediate level of cost and expertise. Such intermediate levels exist in the real world, e.g. a social work trainee would have less expertise in suicidality assessment than than a trained crisis-line staffer or a specialist clinical psychologist.

practical deployment. To get the rest of the way there, further theoretical research and experimentation are required in order to expand the evidence base for this approach. Equally important, for this and any other proposal, careful consideration of the balance between privacy and prevention must continue and, crucially, that conversation needs to integrate the voices of (at least) technologists, in-the-trenches clinicians, policy makers, and those with lived experience of the conditions we are trying to help address.

8.1.1 Authors and Publication

This section was joint work by Samuel Dooley, Candice Schumann, Han-Chin Shing, John P. Dickerson, and Philip Resnik. The full version of the paper can be found in Appendix E. A full version of the paper is in submission to the 2020 Knowledge Discovery in Databases conference in the Applied Data Science Track.

8.2 Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

As mentioned in previous chapters there are multiple metrics and approaches to algorithmic fairness [32, 59, 87]. Indeed there are many camps in the machine learning fairness community and many definitions of machine learning fairness do not fit well within pre-existing legal and moral frameworks. The rapid expansion of this field makes it difficult for professions to keep up, let alone the general public. It is therefore extremely important that non-experts can understand various mathematical definitions of fairness sufficiently to provide opinions.

In this joint work we take a step toward addressing this issue by studying peoples' comprehension and perceptions of three definitions of ML fairness: *demographic parity*, *equal opportunity,* and *equalized odds* [103]. Specifically, we address the following research questions:

RQ1 When provided with an explanation intended for a non-technical audience, do non-experts comprehend each definition and its implications?

RQ2 Do demographics play a role in comprehension?

RQ3 How are comprehension and sentiment related?

RQ4 How do the different definitions compare in terms of comprehension?

We developed two online surveys to address these research questions. We presented participants with a simplified decision-making scenario and a accompanied *fairness rule* expressed in the scenario's context. We asked questions related to the participants' comprehension of and sentiment toward this rule. Tallying the number of correct responses to the comprehension questions gives us a *comprehension score* for each participant. In Study-1, we found that this comprehension score is a consistent and reliable indicator of understanding demographic parity.

Then, in Study-2, we used a similar approach to compare comprehension among all three definitions of interest. We find that 1) education is a significant predictor of rule understanding, 2) the counterintuitive definition of Equal Opportunity with False Negative Rate was significantly harder to understand than other definitions, and 3) participants with low comprehension scores tended to express less negative sentiment toward the fairness rule.

8.2.1 Authors and Publication

This section was joint work by Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, and Michael Carl Tschantz. The full version of the paper can be found in Appendix F. An inital abstract of this work was published in the 2020 Artificial Intelligence, Ethics, and Society conference. A full version of the paper is in submission to the 2020 International Conference on Machine Learning.

8.3 Student Evaluation

Projects such as the Diverse Cohort Selection Problem (Chapter 4), and Tiered Hiring (Chapter 5) require an underlying true utility for an applicant. This is extremely difficult to estimate since the concept of success is inherently subjective and could encompass a variety of factors. Additionally, what looks like success for one person is not necessarily the same for another. In past projects we have relied on just using admit vs. reject decisions as a proxy for student success. This seems insufficient and unsatisfactory since admittance to a graduate program does not necessarily imply future success in the program. Additionally, past admission decisions may have incorporated implicit biases of the graduate admissions committee. After working with the Department of Computer Science and the Graduate School we therefore decided to add an additional question to the graduate review process. Every year the department requires students and advisors to fill out a review of the past years work. A new feature allows advisors to score their student as "Outstanding", "Exceeds Expectations", "Meets Expectations", "Below Expectations", "Unsatisfactory", and “Inadequate opportunity to observe”. These new scores allow for a more detailed view of “success” and allows us to build better models of student utility. Note that this is still a proxy measurement of success which can still include implicit biases of advisors. This multi-year data collection project will provide more insight into the admissions processes and will allow for better evaluations of current and future algorithms.

Chapter 9: Future Work (or things I wish I had time to do)

9.1 Group Fair Bandit with switching sensitive attributes

In Chapter 6 I introduce at a contextual multi-armed bandit algorithm which learns a societal bias term for sensitive groups. This algorithm assumes that the arms are partitioned into groups (in the binary case n arms are partitioned into groups P_1 and P_2 where $P_1 \cup P_2 = [n]$ and $P_1 \cap P_2 = \varnothing$ and P_2 having some societal bias included in the reward when an arm is pulled). This is not necessarily realistic as in many real world uses of multi-armed bandits do not maintain sensitive attributes over time. Instead at time t the arms could be partitioned into $P_{1,t}$ and $P_{2,t}$ and at tie $t+1$ the arms could be partitioned into $P_{1,t+1}$ and $P_{2,t+1}$ where $P_{1,t} \neq P_{1,t+1}$ and $P_{2,t} \neq P_{2,t+1}$. An adaptable algorithm is needed for this situation.

9.2 Group Fair outcomes in Cohort Selections

In both Chapter 4 and Chapter 5 I look at a selecting a cohort of arms from a large pool of arms using pure exploration multi-armed bandit algorithms. In Chapter 6 I look at a contextual multi-armed bandit that takes societal bias into account when pulling arms. The clear next step in this work is to introduce fairness into cohort selections. This could either done by either introducing another societal bias term into the arm pulling process or by using an adjusted form of equal opportunity. Since we are now dealing with a cohort or yes/no decisions a false positive (or negative) rate can be equalized across groups.

9.3 Robust and Adverse Cohort Selection

The following two problems share the additional problem that arms selected for the final cohort may not actually appear in the final cohort. The Multi-Armed Bandit settings proposed in the diverse cohort selection problem (Chapter 4) and a multi-armed bandit approach to tiered interviewing (Chapter 5) both assume that all applicants selected will accept the offers and matriculate into the university or join the firm.

9.3.1 Matriculation

The matriculation problem follows a problem found currently in graduate and undergraduate admissions processes. Universities in the United States follow a similar timeline of accepting applications, reviewing applications, and sending admission decisions. When admissions decisions are sent out there is no guarantee of an applicant matriculating into the university. This is because an applicant will generally apply to more than one university and could get more than one acceptance. Let us assume that each university can only send out K acceptances for a given admissions process. Can we produce a Multi-Armed Bandit algorithm that selects K applicants while maximizing utility gained from matriculated students? In other words, we want an algorithm that returns

$$M^* = \underset{M \subset A ||M| = K}{\arg\max} \sum_{a \in M} m(a)u(a) \tag{9.1}$$

where $m(a)$ returns 1 if an applicant matriculates into the program and 0 otherwise, and $u(a) \in [0,1]$ is the true utility of an applicant.

A matriculation robust algorithm is needed to deal with this real world problem. Extending Chapter 5 to this setting is the step. This could be done by introducing probability of acceptance into the utility calculation.

9.3.2 Dueling Bandits

Moving away from university admissions and toward hiring, we move toward a more dynamic system where decisions are made over time. In this setting a company sees applicants arrive and leave over time. Good candidates may have a higher probability of leaving quickly due to other companies sending out offers. I would like to model this problem as a duling bandit problem with more than one agent making decisions over time. Each agent (company) has a different true utility function for applicants. There may, however, be some correlation between utility functions for each agent. My work so far in the diverse cohort selection problem (Chapter 4) and tiered interviewing (Chapter 5) use pure exploration mutli-armed bandit algorithms. In those situations we do not look at regret, instead we only care about the end cohort. In this problem, we move away from pure exploration and towards an exploration vs exploitation problem. We still want to maximize the overall utility of a cohort however, that cohort is selected over time.

9.4 Learning Diversity

A practical and efficient procedure for learning agents' submodular preferences on subsets of a ground set is needed. A direct application involves learning preferences of diverse group composition to facilitate assignment algorithms in matching markets (or those listed in Chapter 4 and Section 5). We can initially try to learn weights on a known diversity function such as

$$w(M) = \sum_{a \in A} u(a) + \lambda \sum_{i=1}^{L} \sqrt{\sum_{a \in P_i} u(a)} \tag{9.2}$$

where the goal is to find the λ that best suits an individuals preferences. Eventually we should attempt to learn which kind of diversity function is preferred by a human agent (for instance one person could think token diversity is enough, while other person could think that pure parity is needed).

The focus of these is to look at what people think is a good and diverse set of people. We could potentially use two different datasets.

9.4.1 Fantasy University

Participants will be asked to create a new university and select professors based on some rank as well as scraped demographic information. See https://planetterp.com/

for a potential dataset, or https://www.ratemyprofessors.com/. All personal information (like names and universities) should be abstracted away.

9.4.2 Graduate Admissions

We can potentially interview the graduate admissions chair to learn their λ. This could be compared to past graduate admissions chairs, or to chairs from other departments or schools.

9.5 A Bias Checker for Recommendation Letters

We know that individuals entering higher levels of education often go through some form of an application highlighting the reasons for their suitability for the job as part of the hiring process. If the contents of the application are taken at their face value, it is expected that they reflect the abilities and work ethic of the applicant in question. For the most part, this assumption holds because the applicant is in control of their application contents - the resume, cover letter, and interview. At the same time, the recommendation letter, another crucial element to the holistic review of individuals, is a different case, being out of the control of the applicant. Research has shown that recommendation letters not only reflect the work and skills of the subject, but also unconscious gender bias of the recommender. A study from Wayne State University found that letters written for female applicants for a medical faculty had a significantly higher percentage of doubt raisers – statements that question an applicant's aptness for a position – than those written for males [213]. The researchers also found that the most common possessives referring to female and male applicants ("her teaching", "his research") reinforce gender stereotyping. Another study from the University of Arizona revealed that recommenders used more standout adjectives, such as "outstanding" and "exceptional," to describe male applicants than female applicants for a chemistry/biochemistry faculty position at a large American research university [195]. In relation to how these evident manifestations of implicit gender bias affect hiring decision outcomes, a study conducted by Madera et al. [161] found that females applying for academic positions are at a disadvantage because their letters contain more doubt raisers than male applicant letters. This study also suggested future studies to look into developing methods for recommenders to eliminate doubt raisers.

Natural language processing (NLP) algorithms should be used to identify syntactic and topical differences between letters written for female versus male applicants to the University of Maryland graduate Computer Science programs. A classification model to categorize unseen letters as "writen for a female applicant" or "written for a male applicant" should be built. If the classifier shows the letter leaning strongly toward a certain gender, then the letter will be deemed "biased". The final goal of this study is to create a website where recommenders can copy and paste their own letter into a website and see whether and in what ways their letter is biased.

9.6 Building a more realistic graduate admissions classifier

In Chapter 4 and Chapter 5 I use a probabilistic classifier as the ground truth of acceptance. This classifier was built on potentially biased accept/reject decisions.

Additionally an accept decision does not necessarily mean that an individual will succeed in a graduate program. Instead, using the collected data from student evaluations (see Section 8.3) a new, hopefully more accurate classifier should be built. This will help in evaluation of past and future algorithms. Note that this will not eliminate all bias since emulators do have internal biases. This should get closer to building a classifier closer to the ground truth of success.

Chapter 10: Conclusion

At the beginning of this dissertation I stated that biases against groups of individuals found in machine learning models, whether it be from social injustice or measurement bias (See Figure 1.1), can be combated through the use of diversity and fairness constraints. I believe my work introduces new ways to do this. This thesis however, is just the beginning.

Indeed, a major field where we see bias impacts is found in hiring (see Chapter 3). Technologies being used in hiring are at risk at augmenting the already present societal bias. I suggest that instead of solely relying on algorithms to make decisions we should use technology to help increase diversity by helping assign human resources. The first steps at this idea where taken with Chapter 4 and Chapter 5. Moving these algorithms to a more realistic setting that takes matriculation and outside actors into account needs to be done. And with expansions into health and suicidality (see Section 8.1) the tiered multi-armed bandit framework shows promise outside of the hiring setting.

Taking it a step further, including fairness into these systems is vitally important. Chapter 6 incorporates group fairness into a contextual multi-armed bandit setting. Learning to deal with a societal bias term not only helps to treat sensitive groups fairly, but also gives us insight into where the societal bias is coming from. This could help reduce bias outside of the algorithm. Moreover, learning to debias during learning is important. But it is also important to know how to transfer that debiasing knowledge over to a new, potentially more difficult domain. Chapter 7 provides the first theoretical analysis of this setting.

Figuring out how to incorporate fairness into algorithms is important, but understanding fairness definitions is needed to understand how debiasing is happening. It's important not only for the computer scientists creating and running the algorithms, but also important for the stakeholders and non-experts whom the algorithms are affecting. Section 8.2 delves into non-expert understanding of statistical fairness definitions. On the other hand, fairness is just one side of the story. Diversity is another aspect that needs to be understood. What does a person mean when they day they want to select a diverse set of applicants?

And finally we come back to my original question. How do we mitigate bias found the world as it is measured from our algorithms and models? Or at least, how do we mitigate bias found in graduate admissions (the original inspiration for this research)? A small part of the answer can be found here.

Appendix A: The Diverse Cohort Selection Problem

A.1 Table of Symbols

For ease of exposition and quick reference, Table A.1 lists each symbol used in Chapter 4, along with a brief description of that symbol.

A.2 CLUCB Algorithm

The Combinatorial Lower-Upper Confidence Bound (CLUCB) algorithm by Chen et al. [58] is shown in Algorithm 5. At the beginning of the algorithm, pull each arm once and initialize the empirical means with the rewards from that first arm pull. During iteration t of the algorithm, first find the set M_t using the Oracle. Then, compute the confidence radius for each arm. Find the worst case for each arm and compute a new set $\tilde{M}_t$ using the worst case estimates of the arms. If the utility of the initial set M_t and the worst case set $\tilde{M}_t$ are equal then output set M_t. Pull the most uncertain arm (the arm with the widest radius) from the symmetric difference of the two sets M_t and $\tilde{M}_t$. Update the empirical means.

Algorithm 5 Combinatorial Lower-Upper Confidence Bound (CLUCB)

```
Require: Confidence $\delta \in (0,1)$; Maximization oracle: $Oracle(\cdot) : \mathbb{R}^n \to \mathcal{M}$
Weak pull each arm $a \in [n]$ once.
Initialize empirical means $\bar{\mathbf{u}}_n$
$\forall a \in [n]$ set $T_n(a) \leftarrow 1$
for $t = n, n+1, \ldots$ do
    $M_t \leftarrow Oracle(\bar{\mathbf{u}}_t)$
    $\forall a \in [n]$ compute confidence radius $rad_t(a)$
    for $a = 1, \ldots, n$ do
        if $a \in M_t$ then $\tilde{u}_t(a) \leftarrow \bar{u}_t(a) - rad_t(a)$
        else $\tilde{u}_t(a) \leftarrow \bar{u}_t(a) + rad_t(a)$
    $\tilde{M}_t \leftarrow Oracle(\tilde{\mathbf{u}}_t)$
    if $\tilde{w}(\tilde{M}_t) = \tilde{w}(M_t)$ then
        Out $\leftarrow M_t$
        return Out
    $p_t \leftarrow \arg\max_{a \in (\tilde{M}_t \setminus M_t) \cup (M_t \setminus \tilde{M}_t)} rad_t(a)$
    Pull arm $p_t$
    Update empirical means $\bar{\mathbf{u}}_{t+1}$ using the observed reward
    $T_{t+1}(p_t) \leftarrow T_t(p_t) + 1$
    $T_{t+1} \leftarrow T_t(a)\ \forall a \neq p_t$
```

A.3 Proofs

Theorem A.2 (Chen et al. 2014)**.** *Given any* $\delta \in (0,1)$*, any decision class* $\mathcal{M} \subseteq 2^{[n]}$*, and any expected rewards* $\mathbf{u} \in \mathbb{R}^n$*, assume that the reward distribution* φ_a *for each arm* $a \in [n]$ *has mean* $u(a)$ *with an* σ*-sub-Gaussian tail. Let* $M_* = \arg\max_{M \in \mathcal{M}} w(M)$ *denote the optimal set. Set* $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4nt^3}{\delta}/T_t(a)\right)}$ *for all* $t > 0$ *and* $a \in [n]$*. Then, with probability at least* $1-\delta$*, the SWAP algorithm with only weak pulls returns the optimal set* $\texttt{Out} = M_*$ *and*

$$T \leq O\left(\sigma^2 width(\mathcal{M})^2 \mathbf{H} \log(nR^2\mathbf{H}/\delta)\right) \tag{A.1}$$

where T *denotes the number of samples used by the SWAP algorithm,* $\mathbf{H}$ *is defined in Eq.4.2.*

In this section, we formally prove the theorems discussed in Chapter 4. Some lemmas we show directly feed from Chen et al. [58]'s paper.

A.3.1 Strong Arm Pull Problem

The following maps to Lemma 8 in Chen et al. [58].

Lemma A.3. *Suppose that the reward distribution* φ_a *is a* σ*-sub-Gaussian distribution for all* $a \in [n]$*. And if, for all* $t > 0$ *and all* $a \in [n]$*, the confidence radius* $rad_t(a)$ *is given by*

$$rad_t(a) = \sigma\sqrt{\frac{2\log\left(\frac{4nt^3j^3}{\delta}\right)}{T_t(a)}}$$

where $T_t(a)$ is the number of samples of arm a up to round t. Since $s > 1$ the number of samples in a single strong pull will be s each with cost j. Then, we have

$$\Pr\left[\bigcap_{t=1}^{\infty} \xi_t\right] \geq 1 - \delta.$$

Proof. Fix any $t > 0$ and $a \in [n]$. Note that φ_a is a σ-sub-Gaussian tail distribution with mean $w(a)$ and $\bar{w}_t(a)$ is the empirical mean of φ_a from $T_t(a)$ samples.

$$\Pr\left[|\bar{w}_t(a) - w_t(a)| \geq \sigma\sqrt{\frac{2\log\left(\frac{4nt^3j^3}{\delta}\right)}{T_t(a)}}\right]$$

$$= \sum_{b=1}^{t-1} \Pr\left[|\bar{w}_t(a) - w_t(a)| \geq \sigma\sqrt{\frac{2\log\left(\frac{4nt^3j^3}{\delta}\right)}{bs}}, T_t(a) = bs\right] \tag{A.2a}$$

$$\leq \sum_{b=1}^{t-1} 2\exp\left(\frac{-bs\left(\sigma\sqrt{\frac{2\log\left(\frac{4nt^3j^3}{\delta}\right)}{bs}}\right)^2}{2\sigma^2}\right) \tag{A.2b}$$

$$= \sum_{b=1}^{t-1} \frac{\delta}{2nt^3j^3}$$

$$\leq \frac{\delta}{2nt^2j^3} \tag{A.2c}$$

where Eq.A.2a follows from the fact that $1 \leq T_t(a)/s \leq t-1$ and Eq.A.2b follows from Hoeffding's inequality. By a union bound over all $a \in [n]$, we see that $\Pr[\xi_t] \geq$

$1 - \frac{\delta}{2t^2j^3}$. Using a union bound again over all $t > 0$, we have

$$\begin{aligned}
\Pr\left[\bigcap_{t=1}^{\infty} \xi_t\right] &\geq 1 - \sum_{t=1}^{\infty} \Pr[\neg \xi_t] \\
&\geq 1 - \sum_{t=1}^{\infty} \frac{\delta}{2t^2j^3} \\
&= 1 - \frac{\pi^2}{12j^3}\delta \\
&\geq 1 - \delta
\end{aligned}$$

□

The rest of the lemmas in Chen et al. [58]'s paper hold. We can now prove Theorem A.4

Theorem A.4. *Given any $\delta \in (0,1)$, any decision class $\mathcal{M} \subseteq 2^{[n]}$, and any expected rewards $\mathbf{w} \in \mathbb{R}^n$, assume that the reward distribution φ_a for each arm $a \in [n]$ has mean $w(a)$ with an σ-sub-Gaussian tail. Let $M_* = \arg\max_{M \in \mathcal{M}} w(M)$ denote the optimal set. Set $rad_t(a) = \sigma\sqrt{2\log\left(\frac{4nt^3j^3}{\delta}/T_t(a)\right)}$ for all $t > 0$ and $a \in [n]$. Then, with probability at least $1 - \delta$, the CLUCB algorithm with only strong pulls where $j \geq 1$ and $s > j$ returns the optimal set $\texttt{Out} = M_*$ and*

$$T \leq O\left(\frac{\sigma^2 width(\mathcal{M})^2 \mathbf{H} \log(nj^3R^2\mathbf{H}/\delta)}{s}\right) \tag{A.3}$$

where T denotes the number of samples used by the CLUCB algorithm, $\mathbf{H}$ is defined in Eq.4.2.

Proof. Lemma A.3 indicates that the event $\xi \triangleq \bigcap_{t=1}^{\infty} \xi_t$ occurs with probability at least $1 - \delta$. In the rest of the proof, we shall assume that this event holds.

By using Lemma 9 from Chen et al. [58] and the assumption on ξ, we see that $\texttt{Out} = M_*$. Next, we focus on bounding the total number of T samples.

Fix any arm $a \in [n]$. Let $T(a)$ denote the total information gained from pulling arm $a \in [n]$. Let t_a be the last round which arm a is pulled, which means that $p_{t_a} = e$. It is easy to see that $T_{t_a}(a) = T(a) - s$. By Lemma 10 from chen et. al., we see that $rad_{t_a} \geq \frac{\Delta_a}{3width(\mathcal{M})}$. Using the definition of rad_{t_a}, we have

$$\frac{\Delta_a}{3width(\mathcal{M})} \leq \sigma\sqrt{\frac{2\log(4nt_a^3j^3/\delta)}{T(a)-s}} \leq \sigma\sqrt{\frac{2\log(4nT^3j^3/\delta)}{T(a)-s}}. \tag{A.4}$$

By solving Eq.A.4 for $T(a)$, we obtain

$$T(a) \leq \frac{18width(\mathcal{M})^2\sigma^2}{\Delta_a^2}\log(4nT^3j^3/\delta) + s \tag{A.5}$$

Define $\tilde{\mathbf{H}} = \max\{width(\mathcal{M})^2\sigma^2\mathbf{H}, 1\}$. Using similar logic to Chen et al. [58] and the fact that the information gained per pull is s, we show that

$$T \leq \frac{499\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} + 2n \tag{A.6}$$

Theorem 4.1 follows immediately from Eq. A.6.

If $n \geq \frac{1}{2}T$, then $T \leq 2n$ and Eq. A.6 holds. For the second case we assume $n < \frac{1}{2}T$. Since $T > n$, we write

$$T = \frac{C\,\tilde{\mathbf{H}}\log\left(4nj^3\,\tilde{\mathbf{H}}\,/\delta\right)}{s} + n, \text{ for some } C > 0. \tag{A.7}$$

If $C < 499$, then Eq. A.6 holds. Suppose, on the contrary, that $C > 499$. We know that $T = \frac{1}{s}\sum_{a\in[n]} T(a)$. Using this fact and summing Eq. A.5 for all $a \in [n]$, we have

$$
\begin{aligned}
T &\leq \frac{1}{s}\left(ns + \sum_{a\in[n]} \frac{18\,width(\mathcal{M})^2\sigma^2}{\Delta_a^2}\log(4nj^3T^3/\delta)\right) \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4nj^3T^3/\delta)}{s} \\
&= n + \frac{18\,\tilde{\mathbf{H}}\log(4nj^3/\delta)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(T)}{s} \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4nj^3/\delta)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(2C\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta))}{s} \qquad \text{(A.8)} \\
&= n + \frac{18\,\tilde{\mathbf{H}}\log(4nj^3/\delta)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(2C)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(\tilde{\mathbf{H}})}{s} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(2C)\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} + \frac{54\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} \qquad \text{(A.9)} \\
&= (126 + 54\log(2C))\frac{\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} \\
&< n + \frac{C\,\tilde{\mathbf{H}}\log(4nj^3\,\tilde{\mathbf{H}}\,/\delta)}{s} \qquad \text{(A.10)} \\
&= T, \qquad \text{(A.11)}
\end{aligned}
$$

where Eq. A.8 follows from Eq. A.7 and the assumption that $n < \frac{1}{2}T$; Eq. A.9 follows from $\tilde{\mathbf{H}} \geq 1$, $j \geq 1$, and $\delta < 1$; Eq. A.10 follows since $126 + 54\log(2C) < C$ for all $C > 499$; and Eq. A.11 is due to Eq. A.7. So Eq. A.11 is a contradiction. Therefore $C \leq 499$ and we have proved Eq. A.6. □

Corollary A.4.1. *SWAP with only strong pulls is equally or more efficient than SWAP with only weak pulls when* $s > 0$ *and* $0 < j \leq C^{\frac{s}{3}-\frac{1}{3}}$ *where* $C = 4n\tilde{\mathbf{H}}/\delta$.

Proof.

$$\begin{aligned} T_{strong} &\leq T_{weak} \\ \frac{499\tilde{\mathbf{H}}\log(4nj^3\tilde{\mathbf{H}}/\delta)}{s} + 2n &\leq 499\tilde{\mathbf{H}}\log(4nj^3\tilde{\mathbf{H}}/\delta) + 2n \\ \frac{\log(Cj^3)}{s} &\leq \log(C) \end{aligned} \tag{A.12}$$

Solving for Eq.A.12 we get $s > 0$ and $0 < j \leq C^{\frac{s}{3}-\frac{1}{3}}$. □

A.3.2 Strong Weak Arm Pull (SWAP)

The following corresponds to Lemma 8 in work by the Chen et al. [58].

Lemma A.5. *Suppose that the reward distribution φ_a is a σ_1-sub-Gaussian distribution for all $a \in [n]$. For all $t > 0$ and all $a \in [n]$, the confidence radius $rad_t(a)$ is given by*

$$rad_t(a) = \sigma_1\sqrt{\frac{2\log\left(\frac{4nCost_t^3}{\delta}\right)}{T_t(a)}}$$

where $T_t(a)$ is the number of samples of arm a up to round t. Since $s > 1$, the number of samples in a single strong pull are s each with cost j. Then, we have

$$\Pr\left[\bigcap_{t=1}^{\infty}\xi_t\right] \geq 1 - \delta.$$

Proof. Fix any $t > 0$ and $a \in [n]$. Note that φ_a is σ_1-sub-Gaussian tail distribution with mean $w(a)$ and $\bar{w}(a)$ is the empirical mean of φ_a from $T_t(a)$ samples. Then we have

$$\Pr\left[|\bar{w}_t(a) - w_t(a)| \geq \sigma_1 \sqrt{\frac{2\log\left(\frac{4nCost_t^3}{\delta}\right)}{T_t(a)}}\right] \tag{A.13}$$

$$= \sum_{b=1}^{t-1} \Pr\left[|\bar{w}_t(a) - w_t(a)| \geq \sigma_1 \sqrt{\frac{2\log\left(\frac{4nCost_t^3}{\delta}\right)}{Gain_b}}\right] \tag{A.14}$$

$$\leq \sum_{b=1}^{t-1} 2\exp\left(\frac{-Gain_b\left(\sigma_1\sqrt{\frac{2\log\left(\frac{4nCost_t^3}{\delta}\right)}{Gain_b}}\right)^2}{2R^2}\right) \tag{A.15}$$

$$= \sum_{b=1}^{t-1} \frac{\delta}{2nAvCost^3t^3}$$

$$\leq \frac{\delta}{2nt^2AvCost^3} \tag{A.16}$$

where $AvCost$ equal to the average cost until time t. Eq.A.14 follows from $1 \leq T_t(a)/Gain_t \leq t-1$ and Eq.A.15 follows from Hoeffding's inequality. By a union bound over all $a \in [n]$, we see that $\Pr[\xi_t] \geq 1 - \frac{\delta}{2t^2AvCost_t^3}$. Using a union bound again over all $t > 0$, we have

$$\begin{aligned}
\Pr\left[\bigcap_{t=1}^{\infty} \xi_t\right] &\geq 1 - \sum_{t=1}^{\infty} \Pr[\neg\xi_t] \\
&\geq 1 - \sum_{t=1}^{\infty} \frac{\delta}{2t^2AvCost^3} \\
&= 1 - \frac{\pi^2}{12AvCost^3}\delta
\end{aligned}$$

$$\geq 1 - \delta$$

□

Given that the rest of the lemmas in the Chen et al. [58] paper hold, we now prove the main theorem of Chapter 4.

Theorem A.6. *Given any $\delta_1, \delta_2, \delta_3 \in (0,1)$, any decision class $\mathcal{M} \subseteq 2^{[n]}$ and any expected rewards $\mathbf{w} \in \mathbb{R}^n$, assume that the reward distribution φ_a for each arm $a \in [n]$ has mean $w(a)$ with an σ_1-sub-Gaussian tail. Let $M_* = \arg\max_{M \in \mathcal{M}} w(M)$ denote the optimal set. Set $rad_t(a) = \sigma_1 \sqrt{2 \log \left(\frac{4nCost_t^3}{\delta} / T_t(a) \right)}$ for all $t > 0$ and $a \in [n]$, set $\epsilon_1 = \sigma_2 \sqrt{2 \log \left(\frac{1}{2} \delta_2 / T \right)}$, and set $\epsilon_2 = \sigma_3 \sqrt{2 \log \left(\frac{1}{2} \delta_3 / n \right)}$. Then, with probability at least $(1-\delta_1)(1-\delta_2)(1-\delta_3)$, the SWAP algorithm (Algorithm 1) returns the optimal set $\mathtt{Out} = M_*$ and*

$$T \leq O\left(\frac{R^2 width(\mathcal{M})^2 \mathbf{H} \log \left(nR^2 \left(\bar{X}_{Cost} - \epsilon_1 \right)^3 \mathbf{H} / \delta \right)}{\bar{X}_{Gain} - \epsilon_2} \right), \tag{A.17}$$

where T denotes the number of samples used by Algorithm 1, $\mathbf{H}$ is defined in Eq. 4.2 and $width(\mathcal{M})$ is defined by Chen et al. [58].

Proof. Lemma A.5 indicates that the event $\xi \triangleq \bigcap_{t=1}^{\infty} \xi_t$ occurs with probability at least $1 - \delta$. In the rest of the proof, we assume that this event holds.

Using Lemma 9 from Chen et al. [58] and the assumption on ξ, we see that $\mathtt{Out} = M_*$. Next, we bound the total number of T samples.

Fix any arm $a \in [n]$. Let $T(a)$ denote the total information gained from pulling arm $a \in [n]$. Let t_a be the last round which arm a is pulled, which means that

$p_{t_a} = a$. Trivially, $T_{t_a}(a) = T(a) - s$. By Lemma 10 from Chen et al. [58], we see that $rad_{t_a} \geq \frac{\Delta_a}{3width(\mathcal{M})}$. Using the definition of rad_{t_a}, we have

$$\frac{\Delta_a}{3width(\mathcal{M})} \leq R\sqrt{\frac{2\log(4nCost_{t_a}^3/\delta)}{T(e) - Gain_{t_a}}}$$
$$\leq R\sqrt{\frac{2\log(4nCost_T^3/\delta)}{T(a) - Gain_{t_a}}}. \tag{A.18}$$

Solving for $T(a)$ in Eq. A.18 we get

$$T(a) \leq \frac{18width(\mathcal{M})^2R^2}{\Delta_e^2}\log(4nCost_T^3/\delta) + Gain_{t_a} \tag{A.19}$$

Define $\bar{X}_{Cost} = \mathbb{E}[Cost]$ as the expected cost of pulling an arm. Since we strong pull an arm with probability $\alpha = \frac{s-j}{s-1}$, we know

$$\bar{X}_{Cost} = \mathbb{E}[Cost_T] = \alpha j + (1-\alpha). \tag{A.20}$$

Define X_{Cost_t} as the cost of pulling an arm at time t. Assuming that each random variable X_{Cost_t} is R_1-sub-Gaussian we can write the following using the Hoeffding inequality,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} C_{Cost_t} - \bar{X}_{Cost}\right| \geq \epsilon_1\right) \leq 2\exp\left(-\frac{T\epsilon_1^2}{2R_1}\right) \tag{A.21}$$

If we set $\epsilon_1 = R_1\sqrt{2log(\frac{1}{2}\delta_2)/T}$ then with probability $(1-\delta_2)$

$$\frac{Cost_T}{T} \in \left(\bar{X}_{Cost} - \epsilon_1, \bar{X}_{Cost} + \epsilon\right). \tag{A.22}$$

Combining Eq. A.19 and Eq. A.22 we get

$$T(e) \leq \frac{18width(\mathcal{M})^2 R^2}{\Delta_e^2} \log(4n(\bar{X}_{Cost} - \epsilon_1)^3 T^3/\delta) + Gain_{t_e} \tag{A.23}$$

Define $\bar{X}_{Gain} = E[Gain]$ as the expected information gain from pulling an arm. Since we pull an arm with probability α, we know that

$$\bar{X}_{Gain} = E[Gain] = \alpha s + (1 - \alpha) \tag{A.24}$$

Define X_{Gain_t} as the information gain of pulling an arm at time t. Assuming that each random variable X_{Gain_t} is R_2-sub-Gaussian we can write the following using the Hoeffding inequality.

$$\Pr\left(\left|\frac{1}{n}\sum_{e\in[n]} Gain_{t_e} - \bar{X}_{Gain}\right| \geq \epsilon_2\right) \leq 2\exp\left(\frac{-n\epsilon_2^2}{2R_2^2}\right) \tag{A.25}$$

If we set $\epsilon_2 = R_2\sqrt{2log(\frac{1}{2}\delta_3)/n}$ then with probability $(1 - \delta_2)$

$$\frac{\sum_{e\in[n]} Gain_{t_e}}{n} \in \left(\bar{X}_{Gain} - \epsilon_2, \bar{X}_{Gain} + \epsilon_2\right). \tag{A.26}$$

Similarly to the proof for Theorem 4.1, define $\tilde{\mathbf{H}} = \max\{width(\mathcal{M})^2 R^2 \mathbf{H}, 1\}$. In the rest of the proof we will show that

$$T \leq \frac{499\,\tilde{\mathbf{H}}\log\left(4n\left(\bar{X}_{Cost} + \epsilon_1\right)^3 \tilde{\mathbf{H}}/\delta\right)}{\bar{X}_G ain - \epsilon_2} + 2n \tag{A.27}$$

Notice that theorem follows immediately from Eq. A.27.

If $n \geq \frac{1}{2}T$, then Eq. A.27 holds. Let's then assume that $n < \frac{1}{2}T$. Since $T > n$, we can write

$$T = \frac{C\,\tilde{\mathbf{H}}\log(4n(X_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta}{\bar{X}_{Gain} - \epsilon_2} + n \tag{A.28}$$

If $C \leq 499$ then Eq. A.27 holds. Suppose then that $C > 499$. Notice that $T = \sum_{a\in[n]} T(a)/Gain_{t_a}$. By summing up Eq. A.23 for all $a \in [n]$ we have

$$\begin{aligned}
T &\leq n + \sum_{a\in[n]} \frac{18 width(\mathcal{M})^2 R^2 \log(4n(\bar{X}_{Cost} + \epsilon_1)T^3/\delta}{\Delta_a^2 Gain_{t_a}} \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3 T^3/\delta)}{\bar{X}_{Gain} - \epsilon_2} && \text{(A.29)} \\
&= n + \frac{18\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3/\delta)}{\bar{X}_{Gain} - \epsilon_2} + \frac{54\,\tilde{\mathbf{H}}\log(T)}{\bar{X}_{Gain} - \epsilon_2} \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3/\delta)}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(2c\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} - \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta))}{\bar{X}_{Gain} - \epsilon_2} && \text{(A.30)} \\
&= n + \frac{18\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3/\delta)}{\bar{X}_{Gain} - \epsilon_2} + \frac{54\,\tilde{\mathbf{H}}\log(2C)}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(\tilde{\mathbf{H}})}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2} \\
&\leq n + \frac{18\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(2C)\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2} \\
&\quad + \frac{54\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2} && \text{(A.31)} \\
&= n + (126 + 54\log(2C))\frac{\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost} + \epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain} - \epsilon_2}
\end{aligned}$$

$$
\begin{aligned}
&< \quad n + \frac{C\,\tilde{\mathbf{H}}\log(4n(\bar{X}_{Cost}+\epsilon_1)^3\,\tilde{\mathbf{H}}\,/\delta)}{\bar{X}_{Gain}-\epsilon_2} && \text{(A.32)} \\
&= \quad T, && \text{(A.33)}
\end{aligned}
$$

where Eq. A.29 follows from Eq. A.26; Eq. A.30 follows from Eq. A.28 and the assumption $n < \frac{1}{2}T$; Eq. A.31 follows from $\tilde{\mathbf{H}} \geq 1$, $\delta < 1$, and $\bar{X}_{Cost} + \epsilon \geq 1$; Eq. A.32 follows since $126 + 54\log(2C) < C$ for all $C > 499$; and Eq. A.33 is due to Eq. A.28. So Eq. A.33 is a contradiction. Therefore $C \leq 499$ and we have proved Eq. A.27. □

A.4 Additional Details about the Admissions Decisions

Classifier

To effectively model the graduate admissions process, we needed a way to accurately represent whether a particular applicant will be admitted to the program. Using 3 years of previous admissions data, including letters of recommendation, we built a classifier modeling the graduate chair's decision for a particular applicant. The classifier's accuracy can be found in Table A.2.

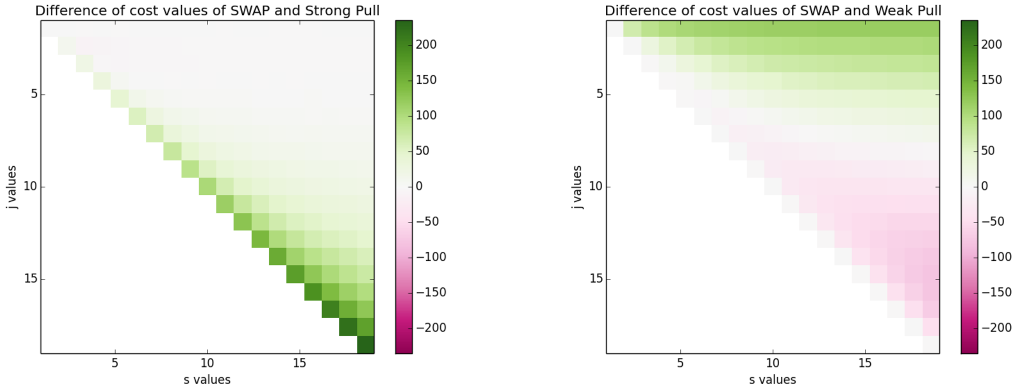
Some general features from the application are GPA, GRE scores, TOEFL scores, area of interest (Machine Learning, Theory, Vision, and so on), previous degrees, and universities attended. We included country of origin since the nature of applications may vary in different regions due to cultural norms. Another basic feature included was sex. We included this to check if the classifier picked up on any biased decision making (with sex and region).

Other features were generated from automatically processing the recommendation letters. Text from the letters was pulled from pdfs and OCR for scanned letters. We then cleaned the raw text with NLTK, removing stop words and stemming text [35]. One feature we chose was the length of recommendation letter, chosen after polling the admissions committee on what they thought would be important. Schmader et al. [195] used Latent Dirichlet Allocation (LDA) to find word groups in recommendation letters for Chemistry and Biochemistry students [36]. Their five word groups included standout words (excellen*, superb, outstanding etc.), ability words (talent*, intell*, smart*, skill*, etc.), grindstone words (hardworking, conscientious, depend*, etc.), teaching words (teach, instruct, educat*, etc.), and research words (research*, data, study, etc.). We found that these word groups translated well to Computer Science students. Important words for acceptance were research words, standout words, and ability words. Letters that only included words from the teaching word group indicated a less useful recommendation letter. We used counts of the various word groups as a feature in the classifier.

A.5 Additional Experimental Results

A.5.1 Gaussian Experiments

While running SWAP, we first compare where the general, varied-cost version of SWAP is better than SWAP with strong pulls only (Figure A.1a) and where it is better than SWAP with only weak pulls (Figure A.1b). We then noticed that there



(A) Heat map showing where SWAP is better than Strong Pull Only.

(B) Heat map showing where SWAP is better than Weak Pull Only.

FIGURE A.1: Differences between SWAP, Strong only, and Weak only.

should be an optimal zone where the general version of SWAP would perform better than both of the trivial cases.

Both graphs examine the symmetric difference between the average cost values of SWAP and either Strong or Weak Pull only with different parameter values of s and j.

A.5.2 Graduate Admissions Experiment

We ran SWAP over both Masters and Ph.D. students over various values of s (Figure A.2). The total cost of running these experiments aligns with the resources spent during the actual admissions decision process.

When running SWAP experiments to formally promote diversity, one experiment not listed in Chapter 4 was testing our diverse SWAP algorithm over an applicant's main choice of research area (Table A.3). In practice, the applicants accepted already

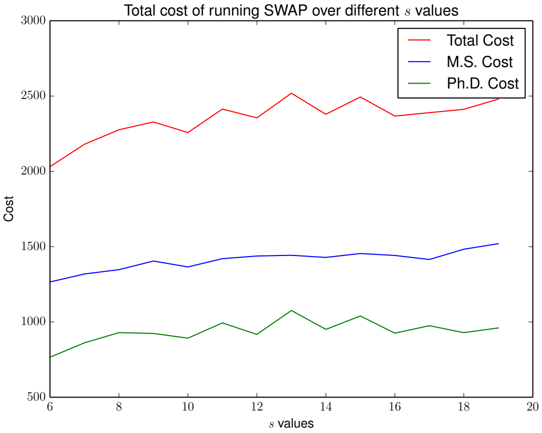


Figure A.2: Total cost of running SWAP over different s values

had a high diversity utility in regards to research area. SWAP slightly increased this diversity utility.

Variable	Summary
n	Number of applications
K	Size of cohort wanted
A	Set of applications
a_i	a single application with $i \in [n]$
$u(a_i)$	True utility of arm a_i where $u(a_i) \in [0, 1]$
$\mathbf{u}$	The set of true utilities.
$\hat{u}(a_i)$	Empirical estimate of utility of arm a_i
$rad(a_i)$	Uncertainty bound around arm a_i. The true utility $u(a_i)$ should lie with $\hat{u}(a_i) - rad(a_i)$ and $\hat{u}(a_i) + rad(a_i)$
$\mathcal{M}$	Decision class. Set of potential cohorts (subsets of arms).
w	Submodular and monotone function for total utility of a cohort. $w : \mathcal{M} \times \mathbb{R}^n \to \mathbb{R}$
$Oracle(\cdot)$	Maximization oracle
M^*	The optimal cohort given the true utilities $\mathbf{u}$ and total utility function w
Δ_a	Gap score for an arm a defined in Equation 4.1
$\mathbf{H}$	Hardness of a problem defined in Equation 4.2
$width(\mathcal{M})$	The smallest distance between any two sets in $\mathcal{M}$
j	Cost of a strong arm pull
s	Information gain of a strong arm pull (ie. the reward is counted s times and is pulled from a tighter distribution around the true utility of an arm)
$Cost_t$	Total cost of pulling arms up until time t
$T_t(a)$	Total information gain for arm a up until time t
M_t	Best cohort of arms at time t, given the empirical utilities
$\tilde{u}_t(a)$	Worst case empirical utility of arm a (See lines 9-10 of Algorithm 1)
$\tilde{M}_t$	Best cohort of arms at time t, given worst case empirical utilities
$\text{spp}(s, j)$	Strong pull policy probability function. See Equation 4.3 for an example
σ	We assume that each arm has a σ-sub-Gaussian tail
$\bar{X}_{Cost}$	Expected cost (expected j value)
$\bar{X}_{Gain}$	Expected information gain (expected s value)
δ	Probability that the algorithms output the best sets (See Theorem 4.1 and Theorem 4.2)
w_{DIV}	Diversity function
w_{TOP}	Top-K function. $\sqrt{w_{\text{TOP}}}$ is the square-root of the top-K function.

Table A.1: All symbols used in Chapter 4

Type	% Correct	Precision	Recall
Ph.D.	77.8%	61.1%	39.7%
Masters	89.2%	13.1%	55.3%
Total	85.5%	33.5%	42.0%

Table A.2: Current predictor results on the testing data

	General	Diversity
SWAP	8.3 (0.03)	32.5 (0.03)
Actual	8.6	27.4

Table A.3: SWAP's average gain in reported area of study diversity over our actual acceptances. The first column shows general fit utility and the second diversity utility The standard deviation over the experiments of SWAP can be found in parentheses.

Appendix B: Making the Cut: A Bandit-based Approach to Tiered Interviewing

B.1 Table of Symbols

In this section, for expository ease and reference, we aggregate all symbols used in Chapter 5 and give a brief description of their meaning and use. We note that each symbol is also defined explicitly in this dissertation; Table B.1 is provided as a reference.

B.2 Proofs

In this section, we provide proofs for the theoretical results presented in Chapter 5. Appendix B.2.1 gives proofs for CACO, defined as Algorithm 2 in Section 5.3. Appendix B.2.2 gives proofs for BRUTaS, defined as Algorithm 3 in Section 5.4.

B.2.1 CACO

Theorem 5.1 requires lemmas from Chen et al. [58]. We restate the theorem here for clarity and then proceed with the proof.

Theorem B.1. *Given any $\delta \in (0,1)$, any $\epsilon \in (0,1)$, any decision classes $\mathcal{M}_i \subseteq 2^{[n]}$ for each stage $i \in [m]$, any linear function w, and any expected rewards $u \in \mathbb{R}^n$,* *assume that the reward distribution* φ_a *for each arm* $a \in [n]$ *has mean* $u(a)$ *with a* σ*-sub-Gaussian tail. Let* $M_i^* = \arg\max_{M \in \mathcal{M}_i}$ *denote the optimal set in stage* $i \in [m]$*. Set* $rad_t(a) = \sigma\sqrt{2\log(\frac{4K_{i-1}Cost_{i,t}^3}{\delta})/T_{i,t}(a)}$ *for all* $t > 0$ *and* $a \in [n]$*. Then, with probability at least* $1 - \delta$*, the* CACO *algorithm (Algorithm 2) returns the set* $\texttt{Out}$ *where* $w(\texttt{Out}) - w(M_m^*) < \epsilon$ *and*

$$T \leq O\left(\sigma^2 \sum_{i\in[m]} \left(\frac{j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{1}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{1}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\right)\right).$$

Proof. Assume we are in some round i, and that we are at time t_a where some arm a is going to be pulled for the last time in round i. Set $rad_{i,t}(a) = \sigma\sqrt{\frac{2\log(4K_{i-1}Cost_{i,t}^3/\delta)}{T_{i,t}(a)}}$ Using Lemma 13 from Chen et al. [58] we know that $rad_{i,t} \geq \max\left\{\frac{\Delta_a}{6}, \frac{\epsilon}{2K_i}\right\}$. Before arm a is pulled the following must be true:

$$rad_{i,t_a} \geq \max\left\{\frac{\Delta_a}{6}, \frac{\epsilon}{2K_i}\right\} \tag{B.1}$$

$$\begin{aligned} rad_{i,t_a} &= \sigma\sqrt{\frac{2\log(4K_{i-1}Cost_{i,t_a}^3/\delta)}{T_i(a) - s_i}} \\ &\leq \sigma\sqrt{\frac{2\log(4K_{i-1}j_i^3 t_a^3/\delta)}{T_i(a) - s_i}}. \end{aligned} \tag{B.2}$$

Equation B.2 holds since $j_i > j_{i-1} > \cdots > j_0$. Given equations B.1 and equation B.2 we have,

$$\max\left\{\frac{\Delta_a}{6}, \frac{\epsilon}{2K_i}\right\} \leq \sigma\sqrt{\frac{2\log(2K_{i-1}j_i^3 t_a^3/\delta)}{T_i(a) - s_i}}$$

$$\leq \sigma \sqrt{\frac{2\log(2K_{i-1}j_i^3 T_i^3/\delta)}{T_i(a) - s_i}}$$

Solving for $T_i(a)$ we have,

$$\begin{aligned} T_i(a) \leq & \sigma^2 \min\left\{\frac{72}{\Delta_a^2}, \frac{16K_i^2}{\epsilon^2}\right\} \log(4K_{i-1}j^3T_i^3/\delta) \\ & + s_i \end{aligned} \tag{B.3}$$

Note that

$$T_i(a) \leq T(a)$$

$$\frac{j_i}{s_i} \sum_{a \in A_{i-1}} T_i(a) = T_i. \tag{B.4}$$

We will show later on in the proof

$$\begin{aligned} T_i \leq & 499 \frac{\sigma^2 j_i}{s_i} \left(\sum_{a \in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \log\left(\frac{4\sigma^2 j_i^4}{s_i\delta} \sum_{a \in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \\ & + 2j_iK_{i-1}. \end{aligned} \tag{B.5}$$

Summing up over equation B.5 we have

$$T \leq 499\sigma^2 \sum_{i \in [m]} \left(\frac{j_i}{s_i} \left(\sum_{a \in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \log\left(\frac{4\sigma^2 j_i^4}{s_i\delta} \sum_{a \in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\right) \tag{B.6}$$

which proves theorem 5.1.

Now we will go back to prove equation B.5. If $K_{i-1} \geq \frac{1}{2}T_i$, then we see that $T_i \leq 2K_{i-1}$ and therefore equation B.5 holds. Assume, then, that $K_{i-1} < \frac{1}{2}T_i$. Since $T_i > K_{i-1}$, we can write

$$T = C\frac{\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{2K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right). \tag{B.7}$$

If $C < 499$ then equation B.5 holds. Suppose then that $C > 499$. Using equation B.4 and summing equation B.6 for all active arms $a \in A_{i-1}$, we have

$$\begin{aligned}
T_i &\leq \frac{j_i}{s_i}\left(K_{i-1}s_i + \sum_{a\in A_{i-1}} \sigma^2 \min\left\{\frac{72}{\Delta_a^2}, \frac{16K_i^2}{\epsilon^2}\right\}\log\left(\frac{4K_{i-1}j_i^3T_i^3}{\delta}\right)\right)\\
&\leq K_{i-1}j_i + \frac{18\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}j_i^3T_i^3}{\delta}\right)\\
&= K_{i-1}j_i + \frac{18\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}j_i^3}{\delta}\right)\\
&\quad + \frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log(T)\\
&\leq K_{i-1}j_i + \frac{18\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}j_i^3}{\delta}\right)\\
&\quad + \frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{2C\sigma^2 j^4}{s_i\delta}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
& \log\left(\frac{4K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\Bigg)\Bigg) \qquad \text{(B.8)} \\
= \; & K_{i-1}j_i+\frac{18\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}j_i^3}{\delta}\right) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\log(2C) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{\sigma^2 j^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& \log\log\left(\frac{4K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
\leq \; & K_{i-1}j_i+\frac{18\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& \log\left(\frac{4K_{i-1}\sigma^2 j^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& \log(2C)\log\left(\frac{4K_{i-1}\sigma^2 j^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}\sigma^2 j^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right) \\
& +\frac{54\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)\log\left(\frac{4K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}}\min\left\{\frac{4}{\Delta_a^2},\frac{K_i^2}{\epsilon^2}\right\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= K_{i-1}j_i + (126 + 54\log(2C))\frac{\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \\
&\quad \log\left(\frac{4K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \\
&\leq K_{i-1}j_i + C\frac{\sigma^2 j_i}{s_i}\left(\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \\
&\quad \log\left(\frac{4K_{i-1}\sigma^2 j_i^4}{s_i\delta}\sum_{a\in A_{i-1}} \min\left\{\frac{4}{\Delta_a^2}, \frac{K_i^2}{\epsilon^2}\right\}\right) \qquad \text{(B.9)} \\
&= T_i \qquad \text{(B.10)}
\end{aligned}
$$

where equation B.8 follows from equation B.7 and the assumption that $K_{i-1} < \frac{1}{2}T_i$; equation B.9 follows since $136 + 54\log(2C) < C$ for all $C > 499$; and B.10 is due to B.7. Equation B.10 is a contradiction. Therefore $C \leq 499$ and we have proved equation B.5. □

B.2.2 BRUTaS

In order to prove Theorem 5.2, we first need a few lemmas.

Lemma B.2. *Let $\Delta_{(1)}, \ldots, \Delta_{(n)}$ be a permutation of $\Delta_1, \ldots \Delta_n$ (defined in Eq. (5.2)) such that $\Delta_{(1)} \leq \ldots \leq \Delta_{(n)}$. Given a stage $i \in [m]$, and a phase $t \in [\tilde{K}_i]$, we define random event $\tau_{i,t}$ as follows*

$$
\tau_{i,t} = \left\{\forall i \in [n] \setminus (A_t \cup B_t) \quad |\hat{u}_{i,t}(a) - u(a)| < \frac{\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_b - t+1)}}{6}\right\}. \qquad \text{(B.11)}
$$

Then, we have

$$\tau = \Pr\left[\bigcap_{i=1}^{m}\bigcap_{t=1}^{\tilde{K}_i}\tau_{i,t}\right] \geq 1 - n^2 \exp\left(-\frac{\sum_{b=1}^{m} s_b(\bar{T}_b - \tilde{K}_b)/(j_i\,\widetilde{\log}(\tilde{K}_b))}{72\sigma^2\tilde{\mathbf{H}}}\right). \tag{B.12}$$

Proof. In round i at phase t, arm a has been pulled $\bar{T}(a)$ times. Therefore, by Hoeffding's inequality, we have

$$\Pr\left[|\hat{u}_{i,t}(a) - u(a)| \geq \frac{\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_b - t+1)}}{6}\right] \leq 2\exp\left(-\frac{\bar{T}_{i,t}(a)\Delta^2_{(n-\sum_{b=0}^{i-1}\tilde{K}_b - t+1)}}{72\sigma^2}\right) \tag{B.13}$$

By using the definition of $\tilde{T}_{i,t}$, the quantity $\tilde{T}_{i,t}\Delta^2_{(n-\sum_{b=1}^{\tilde{K}_{i-1}}\tilde{K}_b - t+1)}$ on the right-hand side of Eq. B.13 can be further bounded by

$$\begin{aligned}
&\bar{T}_{i,t}\Delta_{(n-\sum_{b=1}^{\tilde{K}_{i-1}}\tilde{K}_b - t+1)} \\
&= (s_i\tilde{T}_{i,t} + \sum_{b=1}^{i-1} s_b\tilde{T}_{b,\tilde{K}_b+1})\Delta^2_{(n-\sum_{b=1}^{\tilde{K}_{i-1}}\tilde{K}_b - t+1)} \\
&\geq \left(\frac{s_i(\bar{T}_{i,t} - \tilde{K}_i)}{j_i\,\widetilde{\log}(\tilde{K}_i)(\tilde{K}_i - t + 1)} + \sum_{b=0}^{i-1}\frac{s_b(\bar{T}_{b,\tilde{K}_b+1} - \tilde{K}_b)}{j_b\,\widetilde{\log}(\tilde{K}_b)(\tilde{K}_b - \tilde{K}_b + 2)}\right)\Delta^2_{(n-\sum_{b=1}^{i-1}\tilde{K}_b - t+1)} \\
&\geq \sum_{b=1}^{i}\frac{s_b(\bar{T}_b - \tilde{K}_b)}{j_b\,\widetilde{\log}(\tilde{K}_b)\tilde{\mathbf{H}}},
\end{aligned}$$

where the last inequality follows from the definition of $\tilde{\mathbf{H}} = \max_{i\in[n]} i\Delta^{-2}_{(i)}$. By plugging the last inequality into Eq. B.13, we have

$$\Pr\left[|\hat{u}_{i,t}(a) - u(a)| \geq \frac{\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_b - t+1)}}{6}\right] \tag{B.14}$$

$$\leq 2\exp\left(-\frac{\sum_{b=1}^{i}\left(s_b(\bar{T}_b-\tilde{K}_b)/(j_b\,\widetilde{\log}(\tilde{K}_b)\right)}{72\sigma^2\tilde{\mathbf{H}}}\right) \tag{B.15}$$

Now, using Eq. B.14 and a union bound for all $i \in [m]$, all $t \in [\tilde{K}_i]$, and all $a \in [n] \setminus (A_{t,i} \cup B_{t,i})$, we have

$$\begin{aligned}
&\Pr\left[\bigcap_{i=1}^{m}\bigcap_{t=1}^{\tilde{K}_i}\tau_{i,t}\right]\\
&\geq 1-2\sum_{i=1}^{m}\sum_{t=1}^{\tilde{K}_i}(n-\sum_{b=0}^{i-1}\tilde{K}_b-t+1)\exp\left(-\frac{\sum_{b=1}^{i}\left(s_b(\bar{T}_b-\tilde{K}_b)/(j_b\,\widetilde{\log}(\tilde{K}_b))\right)}{72\sigma^2\tilde{\mathbf{H}}}\right)\\
&\geq 1-n^2\exp\left(-\frac{\sum_{b=1}^{m}s_b(T_b-K_b)/(j_i\,\widetilde{\log}(K_i))}{72\sigma^2\mathbf{H}}\right).
\end{aligned}$$

□

Lemma B.3. *Fix a stage $i \in [m]$, and a phase $t \in [\tilde{K}_i]$, suppose that random event $\tau_{i,t}$ occurs. For any vector $\mathbf{a} \in \mathbb{R}^n$, suppose that supp$(\mathbf{a}) \cap (A_{i,t} \cup B_{i,t} = \varnothing$, where supp$(\mathbf{a}) \triangleq \{i|a(i) \neq 0\}$ is the support of vector $\mathbf{a}$. Then, we have*

$$|\langle\tilde{\mathbf{u}}_{i,t},\mathbf{a}\rangle-\langle\mathbf{u}_{i,t},\mathbf{a}\rangle| < \frac{\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_i-t+1)}}{6}\|\mathbf{a}\|_1$$

Proof. Suppose that $\tau_{i,t}$ occurs. Then, we have

$$\begin{aligned}
&|\langle\tilde{\mathbf{u}}_{i,t},\mathbf{a}\rangle-\langle\mathbf{u}_{i,t},\mathbf{a}\rangle| \qquad\qquad (B.16)\\
&=|\langle\tilde{\mathbf{u}}_{i,t}-\mathbf{w},\mathbf{a}|\\
&=\left|\sum_{b=1}^{n}\left(\tilde{u}_{t,i}(b)-u(b)\right)a(b)\right|
\end{aligned}$$

$$\leq \left| \sum_{b \in [n] \setminus (A_{i,t} \cup B_{i,t})} \left(\tilde{u}_{i,t}(b) - u(b)\right) a(b) \right| \tag{B.17}$$

$$\leq \sum_{b \in [n] \setminus (A_{i,t} \cup B_{i,t})} \left| \left(\tilde{u}_{i,t}(b) - u(b)\right) a(b) \right|$$

$$\leq \sum_{b \in [n] \setminus (A_{i,t} \cup B_{i,t})} |\tilde{u}_{i,t}(b) - u(i)||a(i)|$$

$$< \frac{\Delta_{(n - \sum_{b=0}^{i-1} \tilde{K}_i - t + 1)}}{6} \sum_{b \in [n] \setminus (A_{i,t} \cup B_{i,t})} |a(b)| \tag{B.18}$$

$$= \frac{\Delta_{(n - \sum_{b=0}^{i-1} \tilde{K}_i - t + 1)}}{6} \|\mathbf{a}\|_1 \tag{B.19}$$

where Eq. B.17 follows from the assumption that $\mathbf{a}$ is supported on $[n] \setminus (A_{i,t} \cup B_{i,t})$; Eq. B.18 follows from the definition of $\tau_{i,t}$ (Eq. B.11). □

Lemma B.4. *Fix a stage $i \in [m]$, and a phase $t \in [\tilde{K_i}]$. Suppose that $A_{i,t} \subseteq M_*$ and $B_{i,t} \cap M_* = \varnothing$. Let M be a set such that $A_{i,t} \subseteq M$ and $B_{i,t} \cap M = \varnothing$. Let a and b be two sets satisfying $a \subseteq M \setminus M_*$, $b \subseteq M_* \setminus M$, and $a \cap b = \varnothing$. Then, we have*

$$A_{i,t} \subseteq (M \setminus a \cup b)$$

and

$$B_{i,t} \cap (M \setminus a \cup b) = \varnothing$$

and

$$(a \cup b) \cap (A_{i,t} \cup B_{i,t}) = \varnothing.$$

Lemma B.4 is due to Chen et al. [58].

Lemma B.5. *Fix any stage $i \in [m]$, and any phase $t \in [\tilde{K}_i]$ such that $\sum_{b=0}^{i-1} \tilde{K}_i + t > 0$. Suppose that event $\tau_{i,t}$ occurs. Also assume that $A_{i,t} \subseteq M_*$ and $B_{i,t} \cap M_* = \varnothing$. Let $a \in [n] \setminus (A_{i,t} \cup B_{i,t})$ be an active arm such that $\Delta_{(n-\sum_{b=0}^{i-1} K_i - t+1)} \leq \Delta_a$. Them, we have*

$$\tilde{u}_{i,t}(M_{i,t}) - \tilde{u}_{i,t}(\tilde{M}_{i,t,a}) > \frac{2}{3}\Delta_{(n-\sum_{b=0}^{i-1} \tilde{K}_i - t+1)}$$

Lemma B.5 is due to Chen et al. [58].

Lemma B.6. *Fix any stage $i \in [m]$, and any phade $t \in [K_i]$ such that $\sum_{b=0}^{i-1} K_i + t$ > 0. Suppose that event $\tau_{i,t}$ occurs. Also assume that $A_{i,t} \subseteq M_*$ and $B_{i,t} \cap M_* = \varnothing$. Suppose an active arm $a \in [n] \setminus (A_{i,t} \cap B_{i,t})$ satisfies that $a \in (M_* \cap \neg M_{i,t}) \cup (\neq M_* \cap M_{i,t})$. Then, we have*

$$\tilde{u}_{i,t}(M_{i,t}) - \tilde{u}_{i,t}(\tilde{M}_{i,t,a}) \leq \frac{1}{3}\Delta_{(n-\sum_{b=0}^{i-1} K_i - t+1)}$$

Lemma B.6 is due to Chen et al. [58].

Now we can prove Theorem 5.2, restated below for clarity.

Theorem B.7. *Given any $\bar{T}_i$s such that $\sum_{i\in[m]} \bar{T}_i = \bar{T} > n$, any decision class $\mathcal{M}_K \subseteq 2^{[n]}$, any linear function w, and any true expected rewards $\mathbf{u} \in \mathbb{R}^n$, assume that reward distribution φ_a for each arm $a \in [n]$ has mean $u(a)$ with a σ-sub-Gaussian tail. Let $\Delta_{(1)}, \ldots, \Delta_{(n)}$ be a permutation of $\Delta_1, \ldots, \Delta_n$ (defined in Eq. 5.2) such that $\Delta_{(1)} \leq \ldots \leq \Delta_{(n)}$. Define $\tilde{\mathbf{H}} \triangleq \max_{i\in[n]} i\Delta_{(i)}^{-2}$. Then, Algorithm 3 uses at most $\bar{T}_i$ samples per stage $i \in [m]$ and outputs a solution $\mathtt{Out} \in \mathcal{M}_K \cup \{\perp\}$ such that*

$$\Pr[\mathtt{Out} \neq M_*] \leq n^2 \exp\left(-\frac{\sum_{b=1}^{m} s_b(\bar{T}_b - \tilde{K}_b)/(j_b \widetilde{\log}(\tilde{K}_b))}{72\sigma^2\tilde{\mathbf{H}}}\right) \tag{B.20}$$

where $\widetilde{\log}(n) \triangleq \sum_{i=1}^{n} i^{-1}$, and $M_* = \arg\max_{M\in\mathcal{M}_K} w(M)$.

Proof. First we show that the algorithm takes at most $\bar{T}_i$ samples in every stage $i \in [m]$. It is easy to see that exactly one arm is pulled for $\tilde{T}_{i,1}$ times in stage i, one arm is pulled for $\tilde{T}_{i,2}$ times in stage i, ..., and one arm is pulled for $\tilde{T}_{i,\tilde{K}_i - 1}$ times in stage i. Therefore, the total number of samples used by the algorithm in stage $i \in [m]$ is bounded by

$$
\begin{aligned}
j_i \sum_{t=1}^{\tilde{K}_i} \tilde{T}_{i,t} &\leq j_i \sum_{t=1}^{\tilde{K}_i} \left(\frac{\bar{T}_i - \tilde{K}_i}{\widetilde{\log}(\tilde{K}_i) j_i (\tilde{K}_i - t + 1)} + 1 \right) \\
&= \frac{(\bar{T}_i - \tilde{K}_i) j_i}{\widetilde{\log}(\tilde{K}_i) j_i} \widetilde{\log}(\tilde{K}_i) + \tilde{K}_i \\
&= \bar{T}_i.
\end{aligned}
$$

By Lemma B.2, we know that the event τ occurs with probability at least $1 - n^2 \exp\left(- \frac{\sum_{b=1}^{m} s_b (\bar{T}_b - \tilde{K}_b)/(j_i \widetilde{\log}(\tilde{K}_i))}{72 \sigma^2 \tilde{\mathbf{H}}} \right)$. Therefore, we only need to prove that, under event τ, the algorithm outputs M_*. Assume that the event τ occurs in the rest of the proof.

We will use induction. Fix a stage $i \in [m]$ and phase $t \in [\tilde{K}_i]$. Suppose that the algorithm does not make any error before stage i and phase t, i.e. $A_{i,t} \subseteq M_*$ and $B_{i,t} \cap M_* = \varnothing$. We will show that the algorithm does not err at stage i, phase t.

At the beginning of phase t in stage i there are exactly $\sum_{b=0}^{i-1} \tilde{K}_i + t - 1$ inactive arms $|A_{i,t} \cup B_{i,t}| = \sum_{b=0}^{i-1} \tilde{K}_i + t - 1$. Therefore there must exist an active arm $e_{i,t} \in [n] \setminus (A_{i,t} \cup B_{i,t})$ such that $\Delta_{e_{i,t}} \geq \Delta_{(n - \sum_{b=0}^{i-1} \tilde{K}_i - t + 1)}$. Hence, by Lemma B.5, we have

$$
\tilde{w}_{i,t}(M_{i,t}) - \tilde{w}_{i,t}(M_{i,t,e_{i,t}}) \geq \frac{2}{3} \Delta_{(n - \sum_{b=0}^{i-1} \tilde{K}_i - t + 1)}. \tag{B.21}
$$

Notice that the algorithm makes an error in phase t in stage i if and only if it accepts an arm $p_{i,t} \notin M_*$ or rejects an arm $p_{i,t} \in M_*$. On the other hand, arm $p_{i,t}$ is accepted when $p_{i,t} \in M_{i,t}$ and is rejected when $p_{i,t} \notin M_{i,t}$. Therefore, the algorithm makes an error in phase t in stage i if and only if $p_t \in (M_* \cap \neg M_{i,t}) \cup (\neg M_* \cap M_{i,t})$.

Suppose that $p_t \in (M_* \cap \neg M_{i,t}) \cup (\neg M_* \cap M_{i,t})$. Using Lemma B.6, we see that

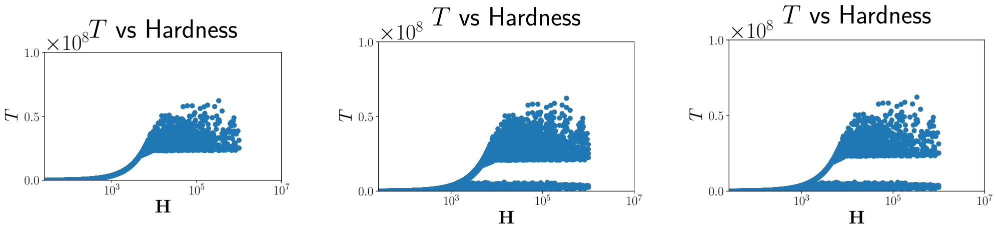
$$\tilde{w}_{i,t}(M_{i,t}) - \tilde{w}_{i,t}(\tilde{M}_{i,t,p_{i,t}}) \leq \frac{1}{3}\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_i-t+1)}. \tag{B.22}$$

By combining Eq. B.21 and Eq. B.22, we see that

$$\tilde{w}_{i,t}(M_{i,t}) - \tilde{w}_{i,t}(\tilde{M}_{i,t,p_{i,t}}) \tag{B.23}$$
$$\leq \frac{1}{3}\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_i-t+1)} \tag{B.24}$$
$$< \frac{2}{3}\Delta_{(n-\sum_{b=0}^{i-1}\tilde{K}_i-t+1)} \tag{B.25}$$
$$\leq \tilde{w}_{i,t}(M_{i,t}) - \tilde{w}_{i,t}(\tilde{M}_{i,t,e_{i,t}}) \tag{B.26}$$

However, Eq. B.23 is contradictory to the definition of $p_{i,t} \triangleq \arg\max_{e\in[n]\setminus(A_{i,t}\cup B_{i,t})} \tilde{w}_{i,t}(M_{i,t}) - \tilde{w}_{i,t}(M_{i,t,e})$. This proves that $p_t \notin (M_* \cap \neg M_{i,t}) \cup (\neg M_* \cap M_{i,t})$. This means that the algorithm does not err at phase t in stage i, or equivalently $A_{i,t+1} \subseteq M_*$ and $B_{i,t+1} \cap M_* = \varnothing$.

Hence we have $A_{m,\tilde{K}_{m+1}} \subseteq M_*$ and $B_{m,\tilde{K}_{m+1}} \subseteq \neg M_*$ in the final phase of the final stage. Notice that $|A_{m,\tilde{K}_{m+1}}| + |B_{m,\tilde{K}_{m+1}}| = n$ and $A_{m,\tilde{K}_{m+1}} \cap B_{m,\tilde{K}_{m+1}} = \varnothing$. This means that $A_{m,\tilde{K}_{m+1}} = M_*$ and $B_{m,\tilde{K}_{m+1}} = \neg M_*$. Therefore the algorithm outputs $\texttt{Out} = A_{m,\tilde{K}_{m+1}} = M_*$ after phase $\tilde{K}_m$ in stage m. □



(A) δ: 0.075, ϵ: 0.075

(B) δ: 0.3, ϵ: 0.075

(C) δ: 0.075, ϵ: 0.3

FIGURE B.1: Hardness (**H**) vs theoretical cost (T) as user-specified parameters to the CACO algorithm.

B.3 Visualization of CACO bound

Figure B.1 shows how the theoretical bound defined in Theorem 5.1 changes as parameters change vs. Hardness **H** defined in Equation 5.3.

B.4 Experimental Setup

The machines used for the experiments had 32GB RAM, 8 Intel SandyBridge CPU cores, and were initialized with Red Hat Enterprise Linux 7.3. A single run of SWAP over the graduate admissions data takes about 1 minute depending on the parameters. See Table B.2 for parameters used.

B.5 Additional Experimental Results

In this section, we present additional experimental results for CACO and BRUTaS. Table B.3 supports the Gaussian simulation experiments of Section 5.5.1, specifically, the comparison of CACO and BRUTaS to two baseline pulling strategies.

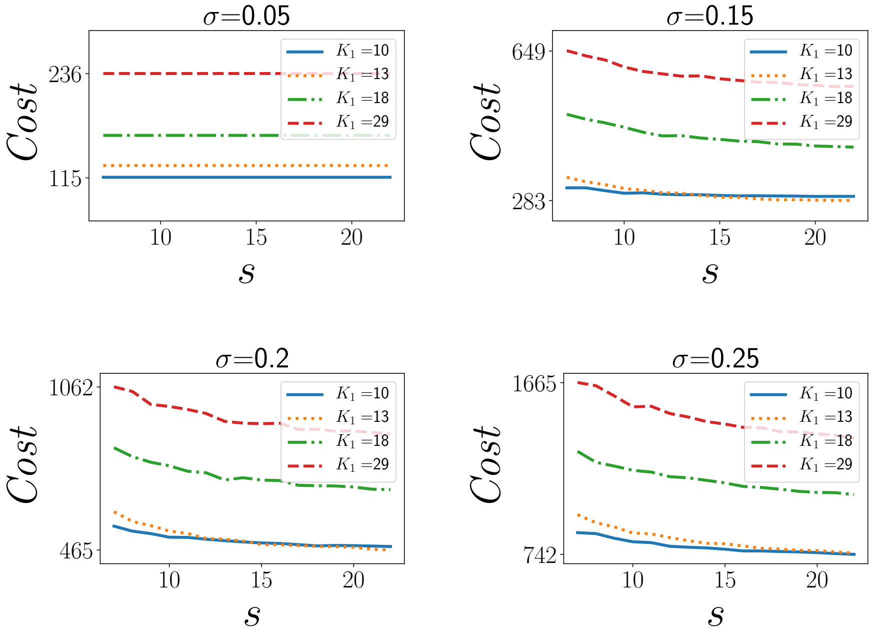


FIGURE B.2: Comparison of *Cost* over information gain (s) as σ increases for CACO. Here, $\delta = 0.05$ and $\epsilon = 0.05$.

Table B.4 also supports the Gaussian simulation experiments from Section 5.5.1. Here, we vary δ instead of ϵ, as was done in Figure 5.2 in Chapter 5. As expected, when δ increases, the cost decreases. However, the magnitude of the effect is smaller than the effect from decreasing ϵ or varying K_1. This is also expected, as discussed in the final paragraphs of Section 5.3, and shown in Figure 5.1.

Figure B.2 shows that, as the standard deviation σ of the Gaussian distribution from which rewards are drawn increases, so too does the total cost of running CACO. The qualitative behavior shown in, e.g., Figure 5.2 of Chapter 5 remains: as information gain s increases, overall cost decreases; as s increases substantially, we see a saturation effect; and, as final cohort size K increases, overall cost increses.

Figure B.3 shows the behavior of CACO for different arm initializations, representing different utilities and groupings. We chose 4 representative initializations. For most initializations, when $K_1 = 10$, higher values of s_2 do not result in gains. This is because with $K_1 = 10$ and $K = 7$, there are only 3 decisions to make on which arms to cut and the information gain from the initial pull of all arms in stage 2 grants enough information, thus no additional pulls need to be made and cost is uniform across s_2. However, if the problem of selecting from the short list is hard enough, additional resources must be spent to narrow the decisions down, as in the top left graph, where total costs decrease as s_2 increases for $K_1 = 10$ because additional pulls need to be made after the initial pulls of remaining arms in stage 2. This reflects real life well: usually, the short list can be cut down with one additional round of (more informative) interviews. However, in rare situations, some candidates are so close to each other that additional assessments need to be made about them. Another interesting result is that $K_1 = 10$ is not always the most cost effective option. If many of the initial candidates are close together in utility, it will be hard to narrow it down to a final 10 based on resume review alone: more candidates should be allowed to move onto the next round which has higher information gain. This can be seen in the bottom right graph.

B.6 Limitations

This experiment uses real data but is still a simulation. The classifier is not a true predictor of utility of an applicant. Indeed, finding an estimate of utility for an

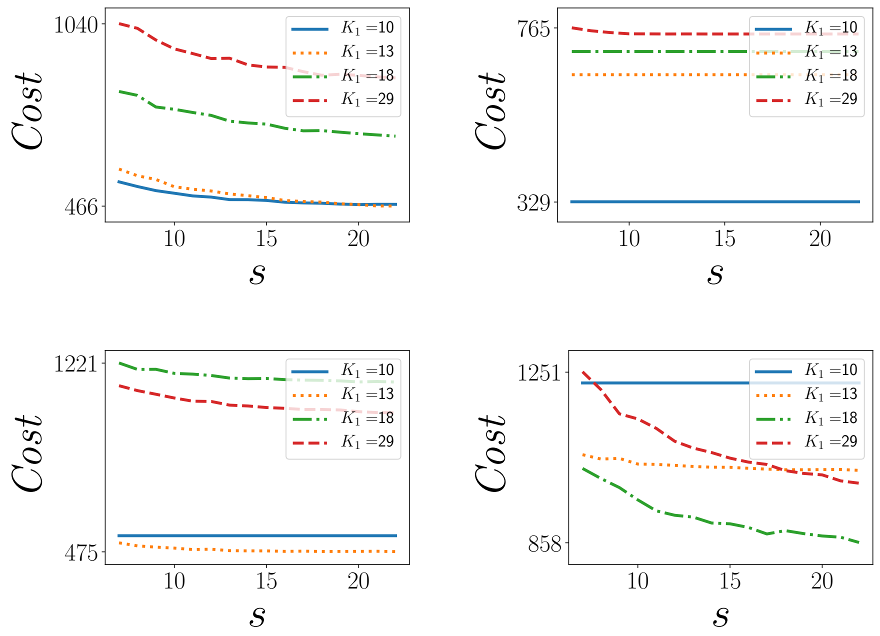


FIGURE B.3: Comparison of *Cost* over information gain (s) for different sets of arms for CACO. Here, $\delta = 0.075$, $\epsilon = 0.05$, $\sigma = 0.2$.

applicant is a nontrivial task. Additionally, the data that we are using incorporates human bias in admission decisions, and reviewer scores. This means that the classifier—and therefore the algorithms—may produce a biased cohort. Training a human committee or using quantitative methods to (attempt to) mitigate the impact of human bias in review scoring is important future work. Similarly, CACO and BRUTAS require an objective function to run; recent advances in human value judgment aggregation [89, 172] could find use in this decision-making framework. Additionally, although we were able to empirically show that both CACO and BRUTAS perform well using a submodular function w_{DIV}, there are no theoretical guarantees for submodular functions.

B.7 Structured Interviews for Graduate Admissions

The goal of the interview is to help judge whether the applicant should be granted admission. The interviewer asks questions to provide insight into the applicant's academic capabilities, research experience, perseverance, communication skills, and leadership abilities, among others.

Some example questions include:

- Describe a time when you have faced a difficult academic challenge or hurdle that you successfully navigated. What was the challenge and how did you handle it?
- What research experience have you had? What problem did you work on? What was most challenging? What did you learn most from the experience?

- Have you had any experiences where you were playing a leadership or mentoring role for others?
- What are your goals for graduate school? What do you want to do when you graduate?
- What concerns do you have about the program? What will your biggest challenge be? Is there anything else we should discuss?

The interviewer fills out an answer and score sheet during the interview. Each interviewer follows the same questions and is provided with the same answer and score sheet. This allows for consistency across interviews.

Symbol	Summary
n	Number of applicants/arms
A	Set of all arms (e.g., the set of all applicants)
a	An arm in A (e.g., an individual applicant)
K	Size of the required cohort
$\mathcal{M}_K(A)$	Decisions class or set of possible cohorts of size K
$u(a)$	True utility of arm a where $u(a) \in [0, 1]$
$\hat{u}(a)$	Empirical estimate of the utility of arm a
$rad(a)$	Uncertainty bound around the empirical estimate of the utility $\hat{u}(a)$ of arm a
w	Submodular and monotone objective function for a cohort where $w : \mathbb{R}^n \times \mathcal{M}_K(A) \rightarrow \mathbb{R}$
$Oracle$	Maximization oracle defined in Equation 5.1 and used by CACO
$COracle$	Constrained maximization oracle used by BRUTaS
M^*	Optimal cohort given the true utilities
Δ_a	The gap score of arm a defined in Equation 5.2
$\mathbf{H}$	The hardness of a problem defined in Equation 5.3
j_i	Cost of an arm pull at stage i
s_i	Information gain of an arm pull at stage i
m	Number of pulling stages (or interview stages)
K_i	Number of arms moving onto the next stage (stage $i + 1$)
A_i	The active arms that move onto the next stage (stage $i + 1$)
$T(a)$	Total information gain for arm a
$\tilde{u}(a)$	Worst case estimate of utility of arm a
$\tilde{A}_i$	Best cohort chosen by using the worst case estimates of utility
ϵ	We want to return a cohort with total utility bounded by $w(M^*) - \epsilon$ for Algorithm 2
δ	The probability that we are within ϵ of the best cohort for Algorithm 2
$\bar{T}_i$	Budget constraint for round i
$\bar{T}$	Total budget
T	Total Cost for CACO
σ	Property of the σ-sub-Gaussian tailed normal distribution
p	The arm with the greatest uncertainty in CACO
$\tilde{K}_i$	Number of decisions to make in round i
$\tilde{T}_{i,t}$	Budget for BRUTaS in stage i, round t
$M_{i,t}$	Best cohort chosen in BRUTaS stage i, round t, using empirical utilities
$\tilde{M}_{i,t,a}$	Pessimistic estimate in BRUTaS stage i, round t, for arm a
$p_{i,t}$	Arm which results in largest gap in BRUTaS stage i, round t
$\tilde{\mathbf{H}}$	Hardness for BRUTaS
$P(a)$	Probability of acceptance for an arm (candidate), estimated by Random Forest Classifier
g	Number of groups for submodular diversity function

Parameter	Range
δ	0.3,0.2,0.1,0.075,0.05
ϵ	0.3,0.2,0.1,0.075,0.05
σ	0.1,0.2
j	6
s	$7, \dots, 20$

Table B.2: Parameters for graduate admissions experiments

Algorithm	Cost	Utility
Random	2750	138.9 (5.1)
Uniform	2750	178.4 (0.2)
CACO	2609	231.0 (0.1)
BRUTaS	2750	244.0 (0.1)

Table B.3: Comparing CACO and BRUTaS to the baseline of Uniform and Random

δ	Cost			
	$K_1 = 10$	$K_1 = 13$	$K_1 = 18$	$K_1 = 29$
0.050	552.475	605.250	839.525	1062.725
0.075	542.425	582.675	827.025	1040.700
0.100	537.175	587.900	820.575	1078.975
0.200	503.650	568.300	801.525	1012.550

Table B.4: Cost for CACO over various δ, for $\epsilon = 0.05, \sigma = 0.20, s_2 = 7$

	Experiment	Cost	w_{TOP}	w_{DIV}	w_{DIV}
				over Gender	over Region
Actual	–	~2,000	60.9	10.1	17.9
Random	lower	1,359	40.2 (0.3)	9.7 (0.2)	16.9 (0.3)
	~equivalent	2,277	43.6 (0.5)	9.9 (0.1)	17.2 (0.2)
	higher	11,556	72.9 (4.9)	11.5 (0.1)	18.1 (3.5)
Uniform	lower	1,359	49.7 (0.3)	9.8 (0.1)	17.7 (0.1)
	~equivalent	2,277	54.7 (0.3)	9.9 (0.2)	18.3 (0.4)
	higher	11,556	79.5 (3.2)	11.9 (0.3)	19.6 (0.6)
SWAP	lower	1,400 1,500	58.7 (0.5)	10.1 (0.1)	19.0 (0.1)
	~equivalent	1,900–2,000	60.2 (0.4)	10.5 (0.1)	19.1 (0.1)
	higher	2,500–2,700	61.5 (0.5)	10.8 (0.2)	19.3 (0.1)
CACO	lower	1,400–1,460	61.1 (0.1)	10.1 (0.2)	18.9 (0.1)
	~equivalent	1,950–1,990	78.7 (0.2)	10.7 (0.1)	19.4 (0.2)
	higher	2,500–2,700	80.1 (0.4)	12.0 (0.3)	19.8 (0.3)
BRUTaS	lower	1,649	61.2 (0.2)	10.6 (0.1)	19.1 (0.2)
	~equivalent	2,038	79.3 (0.3)	10.7 (0.1)	19.8 (0.3)
	higher	2,510	80.2 (0.3)	12.0 (0.2)	19.9 (0.2)

Table B.5: Utility vs Cost over five different algorithms (Random, Uniform, SWAP, CACO, BRUTaS) and the actual admissions decisions made at our university. (Since CACO is a probabilistic method, the cost is given over a range of values.) For each of the algorithms, we give results assuming a cost/budget *lower*, roughly *equivalent*, and *higher* than that used by the real admissions committee. Both CACO and BRUTaS produce equivalent cohorts to the actual admissions process with lower cost, or produce high quality cohorts than the actual admissions process with equivalent cost. Our extension of SWAP to this multi-tiered setting also performs well relative to Random and Uniform, but performs worse than both CACO and BRUTaS across the board.

Appendix C: Group Fairness in Bandit Arm Selection

C.1 Additional Related Work

A closely related area to our work is the research into fairness in rankings [203], multi-stakeholder recommender systems [1], and item allocation [25, 26]. When algorithms return rankings for an individual to select from, one must pay attention to the ordering and the positioning of various groups [203]. One can see this as an application of the group fairness concept to the slates that are chosen for display. A particular aspect of recommendation systems that one needs to keep in mind is that often there are different stakeholders: the person receiving the recommendation, the company giving the recommendation, and the businesses that are the subjects of recommendation [1]. Finally, when goods are allocated, such as housing or subsidies one may need to observe both individual and group fairness [25]. Indeed, group fairness is specifically important in, e.g., Singapore, which has specifically enforced notions of group fairness when allocating public housing [26].

C.2 Naive Group Fairness

Much of the research on fairness in machine learning focuses on fairness in classification [81]. One popular definition of group fairness in classification is the Rawlsian notion of demographic parity [188]. Formally, given a protected demographic group A, we want:

$$\Pr(\hat{Y} = 1|A = 0) = \Pr(\hat{Y} = 1|A = 1), \tag{C.1}$$

where the probability of assigning a classification label $\hat{Y} = 1$ does not change based on the sensitive attribute class A. Demographic parity is important when ground truth classes Y are extremely noisy for sensitive groups due to some societal or measurement bias. Assume that we have a classifier that predicts whether an individual should receive a loan where our sensitive attribute A is binary gender. Demographic parity states that the probability of getting a loan should be the same for males ($A = 0$) and females ($A = 1$).

In converting this definition of demographic parity to the the multi-armed bandit setting, we alter the definition to be that the probability of pulling an arm a does not change based on group membership P_j:

$$\Pr(\textit{pull } a|a \in P_0) = \Pr(\textit{pull } a|a \in P_1). \tag{C.2}$$

Continuing our running example, assume we are a loan agency. The loan agency receives 4 applications at every timestep t: an applicant from a young female, an applicant from a young male, an applicant from a older female, an applicant from an older male; we must choose one application to grant at each timestep. After granting a loan the loan agency receives a down payment on that loan as reward. This reward is then used to update the estimates of whether or not a "good" loan application was received for the pulled arm. Assume that the loan agency wants to act fairly using the binary sensitive attribute of gender. Then, the probability that the loan agency chooses a female applicant at timestep t should be the same as the probability of choosing a male applicant.

C.2.1 A Motivating Example: Linear Regret

Algorithm 6 NaiveGroupFair

```
Require: $\delta$, $P_1$, $P_2$
for $t = 1 \dots T$ do
    $P \leftarrow$ Randomly choose group $P_1$ or $P_2$.
    Pull arm in $P$ based on TopInterval
```

A naive algorithm to enforce this definition of fairness is defined in Algorithm 6. We first pick from the groups uniformly at random, and then apply a regular CMAB algorithm like TopInterval[1] or ContextualThompsonSampling [3] to choose which arm to pull within the group. Using our running example, NaiveGroupFair would randomly pick between male or female, and then choose the best applicant between the younger and older pair.

Assume that NaiveGroupFair randomly chooses the male group during the first timestep and at this timestep the two best applicants are in the female group. Assume that by chance, this worst-case scenario happens at every timestep t. We can extend this argument to any constant number of groups, hence this shows that we have a linear regret for Algorithm 6.

[1] TopInterval is a variant of the contextual bandit LinUCB by Auer et al. [12].

We could, then, just focus on inner group regret,

$$R(t) = f(x_{i^*,t}) - f(x_{a,t}) \quad \text{where } i, a \in P_j,$$

instead of overall regret. In other words, we could focus on the regret of choosing between the younger and older applicant for both genders. This separates the arms into two CMAB problems. This is unsatisfying as it ignores and removes the interaction and differences between groups. We therefore suggest that a new definition of regret that includes a concept of societal bias is needed in this case.

C.3 Proofs

C.3.1 Two Groups

In order to prove Theorem 6.1, we first prove two lemmas.

Lemma C.1. *The following holds for any i at any time t, with probability at least $1-\delta$:*

$$\left|\hat{\beta}_{i,t} \cdot x_{i,t} - (\beta_i \cdot x_{i,t} + \mathbb{1}[i \in P_1]\psi_{P_1} \cdot x_{i,t})\right| \leq w_{i,t}. \tag{C.3}$$

Proof. There are two cases: $i \in P_1$ or $i \notin P_1$.

Focusing on the first case, inequality C.3 becomes:

$$\left|\hat{\beta}_{i,t} \cdot x_{i,t} - \beta_i \cdot x_{i,t}\right| \leq w_{i,t}.$$

By the standard properties of OLS estimators [135], $\hat{\beta}i, t \sim \mathcal{N}\left(\beta_i, \sigma^2(X_{i,t}^T, X_{i,t})^{-1}\right)$. Then, for any fixed $x_{i,t}$:

$$\hat{\beta}_{i,t} \cdot x_{i,t} \sim \mathcal{N}\left(\beta_i \cdot x_{i,t}, x_{i,t}^T \sigma^2(X_{i,t}^T, X_{i,t})^{-1} x_{i,t}\right).$$

Using the definition of the Quantile function and the symmetric property of the normal distribution, with probability at least $1 - \frac{\delta}{nT}$,

$$\hat{\beta}_{i,t} \sim \mathcal{N}\left(\beta_i, \sigma^2(X_{i,t}^T, X_{i,t})^{-1}\right).$$

Exploring the second case where $i \in P_1$, inequality C.3 can be replaced with

$$\left|\hat{\beta}_{i,t} \cdot x_{i,t} - C_i \cdot x_{i,t}\right| \leq w_{i,t}$$

where $C_i = \beta_i + \psi_{P_1}$. Again, by the standard properties of OLS estimators $\hat{\beta}i, t \sim \mathcal{N}\left(C_i, \sigma^2(X_{i,t}^T, X_{i,t})^{-1}\right)$, we have for any fixed $x_{i,t}$:

$$\hat{\beta}_{i,t} \cdot x_{i,t} \sim \mathcal{N}\left(C_i \cdot x_{i,t}, x_{i,t}^T \sigma^2(X_{i,t}^T, X_{i,t})^{-1} x_{i,t}\right).$$

This uses the definition of the Quantile function and the symmetric property of the normal distribution, with probability at least $1 - \frac{\delta}{nT}$.

Therefore, the probability that inequality C.3 fails to hold for any i at any timestep t is at most $nT \cdot \frac{\delta}{nT} = \delta$. □

Lemma C.2. *The following holds for any group P_j, any arm i, at any time t, with probability at least $1 - \delta$:*

$$\left| \hat{\psi}_{P_j,t} \cdot x_{i,t} - \bar{\psi}_{P_j} \cdot x_{i,t} \right| \leq b_{P_j,i,t}. \tag{C.4}$$

Proof. By the standard properties of OLS estimators $\hat{\psi}_{P_j,t} \sim \mathcal{N}\left(\bar{\psi}_{P_j}, \sigma^2(\mathcal{X}_{P_j,t}^T, \mathcal{X}_{P_j,t})^{-1}\right)$. For any fixed $x_{i,t}$,

$$\hat{\psi}_{i,t} \cdot x_{i,t} \sim \mathcal{N}\left(\bar{\psi}_{P_j} \cdot x_{i,t}, x_{i,t}^T \sigma^2 (\mathcal{X}_{P_j,t}^T, \mathcal{X}_{P_j,t})^{-1} x_{i,t}\right).$$

Using the definition of the quantile function and the symmetric property of the normal distribution, with probability at least $1 - \frac{\delta}{\frac{n}{|P_j|}T}$, inequality C.4 holds. Therefore, the probability that this fails to hold for any i at any timestep t is at most $\frac{n}{|P_j|}T \cdot \frac{\delta}{\frac{n}{|P_j|}T} = \delta$. □

With Lemma C.1 and Lemma C.2, we can now prove Theorem 6.1.

Proof. Regret for GroupFairTopInterval can be grouped into three terms for any $T_1 \leq T$:

$$\begin{aligned} R(T) = &\sum_{t\colon\, t \text{ is an explore round}} \mathit{regret}(t) \\ &+ \sum_{t\colon\, t \text{ is an exploit round and } t<T_1} \mathit{regret}(t) \\ &+ \sum_{t\colon\, t \text{ is an exploit round and } t\geq T_1} \mathit{regret}(t) \end{aligned} \tag{C.5}$$

Starting with the first term, define $p_t = \frac{1}{t^{2/3}}$ to be the probability that timestep t is an exploration round. Then, for any t,

$$\sum_{t'<t} p_{t'} = \Theta(t^{2/3}). \tag{C.6}$$

We now focus on the third term of Equation C.5, where t is an exploit round and $t > T_1$. Throughout the rest of the proof we assume Lemma C.1 and Lemma C.2. Fix a exploit timestep t where arm i^t is played. Then,

$$\begin{aligned} \mathit{regret}(t) &\leq 2w_{i^t,t} + 2b_{P_1,i^t,t} + 2b_{P_2,i^t,t} \\ &\leq 2 \max_i \left(w_{i,t} + b_{P_1,i,t} + b_{P_2,i,t}\right) \\ &\leq 2 \left(\max_i w_{i,t} + \max_i b_{P_1,i,t} + \max_i b_{P_2,i,t} \right). \end{aligned} \tag{C.7}$$

Note that:

$$w_{i,t} = Q_{\mathcal{N}\left(0, x_{i,t}\left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t}^T\right)} \left(\frac{\delta}{2nT} \right).$$

Similarly,

$$b_{P_j,i,t} = Q_{\mathcal{N}\left(0, x_{i,t}\left(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}\right)^{-1} x_{i,t}^T\right)} \left(\frac{\delta}{2\frac{n}{|P_j|}T} \right).$$

We first bound

$$\begin{aligned} x_{i,t} \left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t} &\leq ||x_{i,t}|| \lambda_{\max} \left(\left(X_{i,t}^T X_{i,t}\right)^{-1} \right) \\ &= ||x_{i,t}|| \frac{1}{\lambda_{\min} \left(X_{i,t}^T X_{i,t}\right)} \\ &\leq \frac{1}{\lambda_{\min} \left(X_{i,t}^T X_{i,t}\right)} \end{aligned} \tag{C.8}$$

where the last inequality holds since $||x_{i,t}|| \leq 1$ for all i and t. Using similar logic,

$$x_{i,t}\left(\mathcal{X}_{P_j,t}^T\mathcal{X}_{P_j,t}\right)^{-1}x_{i,t} \leq \frac{1}{\lambda_{\min}\left(\mathcal{X}_{P_j,t}^T\mathcal{X}_{P_j,t}\right)}. \tag{C.9}$$

Let $G_{i,t}$ be the number of observations of arm i with contexts drawn uniformly from the distribution for arm i prior to timestep t. Similarly, let $\mathcal{G}_{P_j,t}$ be the number of observations of group P_j with contexts drawn uniformly from the distribution for group P_j prior to timestep t. Let $L > \max_t \lambda_{\max}(x_{i,t}^T, x_{i,t})$. For any $\alpha \in [0,1]$, using the superaddivity of minimum eigenvectors for positive semidefinite matrices, we get

$$\mathbb{E}\left[\lambda_{\min}(X_{i,t}^T X_{i,t})\right] \geq \frac{G_{i,t}}{d}\lambda_{\min_{i,d}} \geq \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}. \tag{C.10}$$

Similarly,

$$\mathbb{E}\left[\lambda_{\min}(\mathcal{X}_{P_j,t}^T\mathcal{X}_{P_j,t})\right] \geq \frac{\mathcal{G}_{P_j,t}}{d}\lambda_{\min_{P_j,d}} \geq \left\lfloor \frac{\mathcal{G}_{P_j,t}}{d} \right\rfloor \lambda_{\min_{P_j,d}}. \tag{C.11}$$

Equation C.10 implies that

$$\Pr_{X_{i,t}}\left[\lambda_{\min}(X_{i,t}^T, X_{i,t}) \leq \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}\right]$$

$$\leq \Pr_{X_{i,t}}\left[\lambda_{\min}(X_{i,t}^T, X_{i,t}) \leq \alpha\mathbb{E}[\lambda_{\min}(X_{i,t}^T X_{i,t})]\right] \tag{C.12}$$

$$\leq \Pr_{X_{i,t}}\left[\lambda_{\min}(X_{i,t}^T, X_{i,t}) \leq \alpha\lambda_{\min}(\mathbb{E}[X_{i,t}^T X_{i,t}])\right] \tag{C.13}$$

$$\leq d\exp\left(\frac{-(1-\alpha)^2\lambda_{\min}(\mathbb{E}[X_{i,t}^T X_{i,t}])}{2L}\right) \tag{C.14}$$

$$\leq d\exp\left(\frac{-(1-\alpha)^2\mathbb{E}[\lambda_{\min}(X_{i,t}^T X_{i,t})]}{2L}\right) \tag{C.15}$$

$$\leq d \exp\left(\frac{-(1-\alpha)^2 \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}{2L}\right) \tag{C.16}$$

where Inequalities C.12 and C.16 are from equation C.10, Inequalities C.13 and C.15 are from Jensen's inequality [170], and Inequality C.14 uses a Matrix Chernoff Bound [170].

Using Inequality C.16 after rearranging with probability $1-\delta$:

$$\lambda_{\min}(X_{i,t}^T X_{i,t}) \geq \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}} \tag{C.17}$$

when

$$G_{i,t} \geq d\left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{i,d}}}\right)\left(\ln\frac{1}{\delta} + \ln d\right). \tag{C.18}$$

Using similar logic with probability $1-\delta$, we have

$$\lambda_{\min}(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}) \geq \alpha \left\lfloor \frac{\mathcal{G}_{P_j,t}}{d} \right\rfloor \lambda_{\min_{P_j,d}} \tag{C.19}$$

when

$$\mathcal{G}_{P_j,t} \geq d\left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{P_j,d}}}\right)\left(\ln\frac{1}{\delta} + \ln d\right). \tag{C.20}$$

Using a multiplicative Chernoff bound [170] for a fixed timestep t with probability $1-\delta'$, the number of exploitation rounds prior to rounds t will satisfy

$$\left| G_t - \sum_{t'<t} p_{t'} \right| \leq \sqrt{\ln\frac{2}{\delta'} \sum_{t<t'} p_{t'}} \tag{C.21}$$

For a fixed i and timestep t using a multiplicative Chernoff bound, with probability $1-\delta'$, the number of exploitation rounds for arm i prior to round t will satisfy

$$\left| G_{i,t} - \frac{G_t}{n} \right| \leq \sqrt{\ln \frac{2}{\delta'} \frac{G_t}{n}}. \tag{C.22}$$

Similarly, for a fixed group P_j and timestep t with probaility $1-\delta'$, the number of exploration rounds for group P_j prior to round t will satisfy

$$\left| \mathcal{G}_{i,t} - \frac{G_t}{|P_j|/n} \right| \leq \sqrt{\ln \frac{2}{\delta'} \frac{G_t}{n/|P_j|}} \tag{C.23}$$

where $|P_j|$ is the size of group P_j.

Combining equations C.21 and C.22 with probability at least $1-2\delta'$ for a fixed arm i and timestep t, if $\sum_{t'<t} P_{t'} \geq 36n \ln^2 \frac{2}{\delta'}$ we have

$$\left| G_{i,t} - \frac{\sum_{t'<t} p_{t'}}{n} \right| \leq \frac{\sum_{t'<t} p_{t'}}{2n}. \tag{C.24}$$

Similarly, combining equations C.21 and C.23 with probability at least $1-2\delta'$ for a fixed group P_j and timestep t:

$$\left| \mathcal{G}_{i,t} - \frac{\sum_{t'<t} P_{t'}}{n/|P_j|} \right| \leq \frac{\sum_{t'<t} p_{t'}}{2n}. \tag{C.25}$$

Therefore, equation C.17 holds with probability $1-\delta'$ when

$$\frac{\sum_{t'<t} p_t}{2n} \geq d \left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{i,d}}} \right) \left(\ln \frac{1}{\delta} + \ln d \right). \tag{C.26}$$

Similarly, equation C.19 holds with probability $1 - \delta'$ when

$$\frac{\sum_{t'<t} p_t}{2n/|P_j|} \geq d \left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{P_j,d}}} \right) \left(\ln \frac{1}{\delta} + \ln d \right). \tag{C.27}$$

Therefore, since $n/|P_j| < n$, the number of rounds after which we have sufficient samples such that the estimators are well-concentrated is

$$T_1 = \Theta \left(\min_a \left(\frac{dnL}{\lambda_{min_a,d}} \left(\ln^2 \frac{2}{\delta} + \ln d \right) \right)^{3/2} \right) \tag{C.28}$$

where $a \in [n] \cup P_1 \cup P_2$.

Also note that for any $t \geq T_1$ we have

$$\sum_{t'<t} p_{t'} = \Omega \left(\min_a \left(\frac{dnL}{\lambda_{\min_{a,d}}} \left(\ln^2 \frac{2}{\delta'} + \ln d \right) \right) \right). \tag{C.29}$$

We can now bound the third term in Equation C.5.

$$\begin{aligned} &\sum_{t:\ t\ is\ an\ exploit\ round\ and\ t \geq T_1} regret(t) \\ &\leq 2 \sum_{t \geq T_1} \left(\max_i w_{i,t} + \max_i b_{P_1,i,t} + \max_i b_{P_2,i,t} \right) \\ &\leq 2 \sum_{t \geq T_1} \left(\max_i Q_{\mathcal{N}\left(0, \lambda_{\max}((X_{i,t}^T X_{i,t}))^{-1}\right)} \left(\frac{\delta}{2nT} \right) \right. \\ &\quad + \max_i Q_{\mathcal{N}\left(0, \lambda_{\max}((\mathcal{X}_{P_1,t}^T \mathcal{X}_{P_1,t}))^{-1}\right)} \left(\frac{\delta}{2 \frac{n}{|P_1|} T} \right) \\ &\quad \left. + \max_i Q_{\mathcal{N}\left(0, \lambda_{\max}((\mathcal{X}_{P_2,t}^T \mathcal{X}_{P_2,t}))^{-1}\right)} \left(\frac{\delta}{2 \frac{n}{|P_2|} T} \right) \right) \end{aligned} \tag{C.30}$$

$$
\begin{aligned}
&\leq 2\sum_{t\geq T_1}\left(Q_{\mathcal{N}\left(0,\frac{1}{\min_i \lambda_{\min}((X_{i,t}^T X_{i,t}))^{-1}}\right)}\left(\frac{\delta}{2nT}\right)\right.\\
&\quad + Q_{\mathcal{N}\left(0,\frac{1}{\min_i \lambda_{\min}((\mathcal{X}_{P_1,t}^T \mathcal{X}_{P_1,t}))^{-1}}\right)}\left(\frac{\delta}{2\frac{n}{|P_1|}T}\right)\\
&\quad \left.+ Q_{\mathcal{N}\left(0,\frac{1}{\min_i \lambda_{\min}((\mathcal{X}_{P_2,t}^T \mathcal{X}_{P_2,t}))^{-1}}\right)}\left(\frac{\delta}{2\frac{n}{|P_2|}T}\right)\right)\\
&\leq 2\sum_{t\geq T_1}\left(Q_{\mathcal{N}\left(0,\frac{1}{\min_i \alpha\left\lfloor\frac{G_{i,t}}{d}\right\rfloor\lambda_{\min_{i,d}}}\right)}\left(\frac{\delta}{2nT}\right)\right.\\
&\quad + Q_{\mathcal{N}\left(0,\frac{1}{\alpha\left\lfloor\frac{\mathcal{G}_{P_1,t}}{d}\right\rfloor\lambda_{\min_{P_1,d}}}\right)}\left(\frac{\delta}{2\frac{n}{|P_1|}T}\right)\\
&\quad \left.+ Q_{\mathcal{N}\left(0,\frac{1}{\alpha\left\lfloor\frac{\mathcal{G}_{P_2,t}}{d}\right\rfloor\lambda_{\min_{P_2,d}}}\right)}\left(\frac{\delta}{2\frac{n}{|P_2|}T}\right)\right) + 3\delta' T \qquad \text{(C.31)}\\
&\leq 2\sum_{t\geq T_1}\left(\sqrt{\frac{\ln\frac{2nT}{\delta}}{\min_i \alpha\left\lfloor\frac{G_{i,t}}{d}\right\rfloor\lambda_{\min_{i,d}}}}\right.\\
&\quad + \sqrt{\frac{\ln\frac{2\frac{n}{|P_1|}T}{\delta}}{\min_i \alpha\left\lfloor\frac{\mathcal{G}_{P_1,t}}{d}\right\rfloor\lambda_{\min_{P_1,d}}}}\\
&\quad \left.+ \sqrt{\frac{\ln\frac{2\frac{n}{|P_2|}T}{\delta}}{\min_i \alpha\left\lfloor\frac{\mathcal{G}_{P_2,t}}{d}\right\rfloor\lambda_{\min_{P_2,d}}}}\right) + 6\delta' T \qquad \text{(C.32)}\\
&\leq 2\sum_{t\geq T_1}\left(3\sqrt{\frac{\ln\frac{2nT}{\delta}}{\min_i \alpha\left\lfloor\frac{G_{i,t}}{d}\right\rfloor\lambda_{\min_{i,d}}}}\right) \qquad \text{(C.33)}
\end{aligned}
$$

$$
\begin{aligned}
&= O\left(\sum_{t\geq T_1}\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i G_{i,t}\lambda_{\min_{i,d}}}} + \delta' T\right) \\
&= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t\geq T_1}\sqrt{\frac{1}{\min_i G_{i,t}}} + \delta' T\right) \\
&= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t\geq T_1}\sqrt{\frac{n}{\sum_{t'<t} p_{t'}}} + \delta' T\right) \\
&= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t\geq T_1}\sqrt{\frac{n}{t^{2/3}}} + \delta' T\right) \qquad (C.34) \\
&= O\left(\sqrt{dn\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t\in[T_1,T]}\frac{1}{t^{1/3}} + \delta' T\right) \\
&= O\left(\sqrt{dn\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}T^{2/3} + \delta' T\right) \qquad (C.35)
\end{aligned}
$$

where (C.30) is due to Equation C.7, (C.31) is due to Equations C.10 and C.11, (C.32) is due to Chernoff bounds, (C.33) is due to the fact that $\frac{n}{|P_j|} < n$ and $G_{P_j,t} > \min_i G_{i,t}$, and (C.34) is due to Equation C.6. Theorem 6.1 follows by combining Equations C.5, C.6, C.28, and C.35 and setting $\delta' = \min\left(\frac{1}{3nT}, \frac{1}{T^{1/3}}\right)$.

□

C.3.2 Multiple Groups

In in order to prove Theorem C.5, we first prove two lemmas.

Lemma C.3. *The following holds for any i at any time t, with probability at least $1-\delta$*

$$
\left|\hat{\beta}_{i,t}\cdot x_{i,t} - \left(\beta_i \cdot x_{i,t} + \sum_{j=1}^{m} \mathbb{1}\left[i \in P_j\right]\psi_{P_j}\cdot x_{i,t}\right)\right|
$$

Algorithm 7 GroupFairTopInterval (Multiple Groups)

```
Require: $\delta$, $(P_1, \dots, P_m)$, $\rho$
for $t = 1 \dots T$ do
    with probability $\frac{1}{t^{1/3}}$, play $i_t \in_R \{1, \dots, n\}$
    Else
        for $j = 1 \dots, m$ do
            Let $\hat{\psi}_{P_j,t} = \left(\mathcal{X}^T_{P_j,t}\mathcal{X}_{P_j,t}\right)^{-1} \mathcal{X}^T_{P_j,t}\mathcal{Y}_{P_j,t}$
        for $i = 1 \dots n$ do
            Let $\hat{\beta}_{i,t} = \left(X^T_{i,t}X_{i,t}\right)^{-1} X^T_{i,t}Y^T_{i,t}$
            Let $F_{i,t} = \mathcal{N}\left(0, \sigma^2 x_{i,t}\left(X^T_{i,t}X_{i,t}\right)^{-1} x^T_{i,t}\right)$
            Let $w_{i,t} = Q_{F_{i,t}}\left(\frac{\delta}{2nt}\right)$
            for $j$ where $i \in P_j$ do
                Let $\mathcal{F}_{P_j,i,t} = \mathcal{N}\left(0, \sigma^2 x_{i,t}\left(\mathcal{X}^T_{P_j,t}\mathcal{X}_{P_j,t}\right) x^T_{i,t}\right)$
                Let $b_{P_j,i,t} = Q_{\mathcal{F}_{P_j,i,t}}\left(\frac{\delta}{2\frac{n}{|P_j|}T}\right)$
                Let $\hat{u}_{i,t} = \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t} + \rho - \hat{\psi}_{P_j,t} \cdot x_{i,t} + b_{P_j,i,t}$
        Play $\arg\max_i \hat{u}_{i,t}$ and observe reward $y_{i,t}$
```

$$\leq w_{i,t} \tag{C.36}$$

Proof. Inequality C.36 can be replaced with

$$\left| \hat{\beta}_{i,t} \cdot x_{i,t} - C_i \cdot x_{i,t} \right| \leq w_{i,t}$$

where $C_i = \beta_i + \psi_{P_j}$ and $i \in P_j$. By the standard properties of OLS estimators $\hat{\beta}_{i,t} \sim N\left(C_i, \sigma^2 (X_{i,t}^T X_{i,t})^{-1}\right)$. For any fixed $x_{i,t}$:

$$\hat{\beta}_{i,t} \cdot x_{i,t} \sim N\left(C_i \cdot x_{i,t}, x_{i,t}^T \sigma^2 (X_{i,t}^T X_{i,t})^{-1} x_{i,t}\right)$$

Using the definition of the quantile function and the symmetric property of the normal distribution, with probability at least $1 - \frac{\delta}{nT}$, Inequality C.36 holds. Therefore, the probability that inequality C.36 fails to hold for any i at any timestep t is at most $nT \frac{\delta}{nT} = \delta$. □

Lemma C.4. *The following holds for any group P_j, any arm i, at any timestep t, with probability at least $1 - \delta$:*

$$\left| \hat{\psi}_{P_j,t} \cdot x_{i,t} - \psi_{P_j,t} \cdot x_{i,t} \right| \leq b_{P_j,i,t}. \tag{C.37}$$

Proof. By the standard properties of OLS estimators,

$$\hat{\psi}_{P_j,t} \sim N\left(\psi_{P_j}, \sigma^2 (\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t})^{-1}\right).$$

For any fixed $x_{i,t}$:

$$\hat{\psi}_{P_j,t} \cdot x_{i,t} \sim N\left(\psi_{P_j} \cdot x_{i,t}, x_{i,t}^T \sigma^2 (\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t})^{-1} x_{i,t}\right).$$

Using the definition of the quantile function and the symmetric property of the normal distribution, with probability of at least $1 - \frac{\delta}{\frac{n}{|P_j|}T}$ inequality C.37 holds. Therefore the probability this fails to hold for any i at timestep t is at most $\frac{n}{|P_j|}T\frac{\delta}{\frac{n}{|P_j|}T} = \delta$. □

Theorem C.5. *For m groups $P_1, \ldots, P_m$, where ρ is the expected average reward,* GroupFairTopInterval (Multiple Groups) *has regret*

$$R(T) = O\left(\sqrt{\frac{dn\ln\frac{2nT}{\delta}}{l}}T^{2/3} + \left(\frac{dnmL}{l}\left(\ln^2\frac{2nT}{\delta} + \ln d\right)\right)^{2/3}\right) \tag{C.38}$$

where $l = \min_i \lambda_{min_{i,d}}$ and $L > \max_t \lambda_{\max}(x_{i,t}^T x_{i,t})$.

We can now prove Theorem C.5.

Proof. Assume that both Lemma C.3 and Lemma C.4 hold for all arms i and all timesteps t.

Regret for GroupFairTopInterval (Multiple Groups) can be grouped into three terms for any $T_1 \le T$:

$$R(T) = \sum_{t:\ t \text{ is an explore round}} regret(t)$$

$$\begin{aligned} &+ \sum_{t:\ t \text{ is an exploit round and } t<T_1} \mathit{regret}(t) \\ &+ \sum_{t:\ t \text{ is an exploit round and } t\geq T_1} \mathit{regret}(t) \qquad \text{(C.39)} \end{aligned}$$

Starting with the first term in Equation C.39, define $p_t = \frac{1}{t^{1/3}}$ to be the probability that timestep t is an exploration round. Then, for any t,

$$\sum_{t'<t} p_{t'} = \Theta(t^{2/3}) \qquad \text{(C.40)}$$

Focusing on the third term of Equation C.39, fix an exploit timestep t where arm i_t is played. Then,

$$\begin{aligned} \mathit{regret}(t) &\leq 2w_{i_t,t} + \max_j(2b_{P_j,i_t,t}) \\ &\leq 2\max_{i,j}(w_{i,t} + b_{P_j,i,t}) \\ &\leq 2\left(\max_i w_{i,t} + \max_{i,j} b_{P_j,i,t}\right) \qquad \text{(C.41)} \end{aligned}$$

From Algorithm 7, note that

$$w_{i,t} = Q_{N\left(0, x_{i,t}\left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t}^T\right)}\left(\frac{\delta}{2nT}\right).$$

Similarly,

$$b_{P_j,i,t} = Q_{N\left(0, x_{i,t}\left(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}\right)^{-1} x_{i,t}^T\right)}\left(\frac{\delta}{2\frac{n}{|P_j|}T}\right)$$

We will first bound $x_{i,t}\left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t}^T$.

$$\begin{aligned} x_{i,t}\left(X_{i,t}^T X_{i,t}\right)^{-1} x_{i,t}^T &\leq ||x_{i,t}||\lambda_{\max}\left(\left(X_{i,t}^T X_{i,t}\right)^{-1}\right) \\ &= ||x_{i,t}||\frac{1}{\lambda_{\min}(X_{i,t}^T X_{i,t})} \\ &\leq \frac{1}{\lambda_{\min}(X_{i,t}^T X_{i,t})} \end{aligned} \tag{C.42}$$

where inequality C.42 is due to $||x_{i,t}|| \leq 1$ for all arms i and all timesteps t.

Using similar logic:

$$x_{i,t}\left(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}\right)^{-1} x_{i,t}^T \leq \frac{1}{\lambda_{\min}\left(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}\right)}. \tag{C.43}$$

Let $G_{i,t}$ be the number of observations of arm i with context i drawn uniformly from the distribution for arm i prior to timestep t. Similarly, let $\mathcal{G}_{P_j,t}$ be the number of observations of group P_j with context drawn uniformly from the distribution for group P_j prior to timestep t. Let $L > \max_t \lambda_{\max}\left(x_{i,t}^T x_{i,t}\right)$.

For any $\alpha \in [0,1]$, using the superadditivity of minimum eugenvectors for positive semi-definite matrices, we get:

$$\begin{aligned} \mathbb{E}\left[\lambda_{\min}(X_{i,t}^T X_{i,t})\right] &\geq \frac{G_{i,t}}{d}\lambda_{\min_{i,d}} \\ &\geq \left\lfloor\frac{G_{i,t}}{d}\right\rfloor. \end{aligned} \tag{C.44}$$

Similarly,

$$\mathbb{E}\left[\lambda_{min}(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t})\right] \geq \left\lfloor\frac{G_{P_j,t}}{d}\right\rfloor \lambda_{\min_{P_j.d}}. \tag{C.45}$$

Equation C.44 implies that:

$$\begin{aligned}
&\Pr_{x_{i,t}}\left[\lambda_{\min}(X_{i,t}X_{i,t}) \le \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}\right] \\
&\le \Pr_{x_{i,t}}\left[\lambda_{\min}(X_{i,t}^T X_{i,t}) \le \alpha \mathbb{E}\left[\lambda_{\min}(X_{i,t}^T X_{i,t})\right]\right] && \text{(C.46)} \\
&\le \Pr_{x_{i,t}}\left[\lambda_{\min}(X_{i,t}^T X_{i,t}) \le \alpha \lambda_{\min}\left(\mathbb{E}\left[X_{i,t}^T X_{i,t}\right]\right)\right] && \text{(C.47)} \\
&\le d \exp\left(\frac{-(1-\alpha)^2 \lambda_{\min}\left(\mathbb{E}\left[X_{i,t}^T X_{i,t}\right]\right)}{2L}\right) && \text{(C.48)} \\
&\le d \exp\left(\frac{-(1-\alpha)^2 \mathbb{E}\left[\lambda_{\min}\left(X_{i,t}^T X_{i,t}\right)\right]}{2L}\right) && \text{(C.49)} \\
&\le d \exp\left(\frac{-(1-\alpha)^2 \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}{2L}\right) && \text{(C.50)}
\end{aligned}$$

where inequality C.46 comes from inequality C.44, inequality C.47 is due to Jensen's inequality, inequality C.48 is due to a matrix Chernoff Bound, inequality C.49 is due to Jensen's inequality, and inequality C.50 is due to inequality C.44. After rearranging inequality C.50, with probability $1-\delta$,

$$\lambda_{min}(X_{i,t}^T X_{i,t}) \ge \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}} \tag{C.51}$$

when

$$G_{i,t} \ge d\left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{i,d}}}\right)\left(\ln\frac{1}{\delta} + \ln d\right). \tag{C.52}$$

Using similar logic with probability $1-\delta$, we have

$$\lambda_{\min}\left(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t}\right) \ge \alpha \left\lfloor \frac{G_{P_j,t}}{d} \right\rfloor \lambda_{\min_{P_j,d}} \tag{C.53}$$

when

$$\mathcal{G}_{P_j,t} \geq d\left(\frac{L}{(1-\alpha)^2\lambda_{\min_{P_j,d}}}\right)\left(\ln\frac{1}{\delta}+\ln d\right). \tag{C.54}$$

Using a multiplicative Chernoff bound for a fixed timestep t with probability $1-\delta'$, the number of exploitation rounds prior to rount t will satisfy

$$\left|G_t - \sum_{t'<t} p_{t'}\right| \leq \sqrt{\ln\frac{2}{\delta'}\sum_{t'<t} p_{t'}}. \tag{C.55}$$

For a fixed i and timestep t, using a multiplicative Chernoff bound for a fixed timestep t with probability $1-\delta'$, the number of exploitation rounds for arm i prior to round t will satisfy

$$\left|G_{i,t} - \frac{G_t}{n}\right| \leq \sqrt{\ln\frac{2}{\delta'}\frac{G_t}{n}} \tag{C.56}$$

Similarly, for a fixed group P_j and timestep t with probability $1-\delta'$, the number of exploration rounds for group P_j prior to round t will satisfy

$$\left|G_{P_j,t} - \frac{G_t}{n/|P_j|}\right| \leq \sqrt{\ln\frac{2}{\delta'}\frac{G_t}{n/|P_j|}} \tag{C.57}$$

where $|P_j|$ is the size of group P_j.

Combining inequality C.55 and inequality C.56, with probability $1-2\delta'$ for a fixed arm i and timestep t, if $\sum_{t'<t} p_{t'} \geq 36n\ln^2\frac{2}{\delta'}$ we have

$$\left|G_{i,t} - \frac{\sum_{t'<t} p_{t'}}{n}\right| \leq \frac{\sum_{t'<t} p_{t'}}{2n}. \tag{C.58}$$

Similarly, combining inequality C.55 and inequality C.57 with probability at least $1 - 2\delta'$ for a fixed group P_j and fixed timestep t:

$$\left| G_{i,t} - \frac{\sum_{t'<t} p_{t'}}{n/|P_j|} \right| \leq \frac{\sum_{t'<t} p_{t'}}{2n/|P_j|}. \tag{C.59}$$

Therefore inequality C.51 holds with probability $1 - \delta'$ when

$$\frac{\sum_{t'<t} p_{t'}}{2n} \geq d \left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{i,d}}} \right) \left(\ln \frac{1}{\delta} + \ln d \right). \tag{C.60}$$

Similarly, inequality C.53 holds with probability $1 - \delta'$ when

$$\frac{\sum_{t'<t} p_{t'}}{2n/|P_j|} \geq d \left(\frac{L}{(1-\alpha)^2 \lambda_{\min_{i,d}}} \right) \left(\ln \frac{1}{\delta} + \ln d \right). \tag{C.61}$$

Therefore, since $\frac{n}{|P_j|} < n$, the number of rounds after which we have sufficient samples such that the estimators are well-concentrated is

$$T_1 = \Theta \left(\min_a \left(\frac{dnmL}{\lambda_{\min_{a,d}}} \left(\ln^2 \frac{2}{\delta} + \ln d \right) \right)^{3/2} \right) \tag{C.62}$$

where $a \in [n] \cup P_1 \cup \cdots \cup P_m$.

Also note that for any $t > T_1$ we have:

$$\sum_{t'<t} p_{t'} = \Omega \left(\min_a \left(\frac{dnmL}{2 \min_{a,d}} \left(\ln^2 \frac{2}{\delta'} + \ln d \right) \right) \right). \tag{C.63}$$

Now we can bound the third term in equation C.39.

$$\sum_{t:\ t \text{ is an exploit round and } t>T_1} \mathit{regret}(t)$$

$$\leq 2 \sum_{t>T_1} \left(\max_i w_{i,t} + \max_{i,j} b_{P_j,i,t} \right) \tag{C.64}$$

$$\leq 2 \sum_{t>T_1} \left(\max_i Q_{N\left(0,\lambda_{\max}(X_{i,t}^T X_{i,t})^{-1}\right)} \left(\frac{\delta}{2nT}\right) + \max_j Q_{N\left(0,\lambda_{\max}(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t})^{-1}\right)} \left(\frac{\delta}{2\frac{n}{|P_j|}T}\right)\right)$$

$$\leq 2 \sum_{t>T_1} \left(Q_{N\left(0,\frac{1}{\min_i \lambda_{\min}(X_{i,t}^T X_{i,t})}\right)} \left(\frac{\delta}{2nT}\right) + Q_{N\left(0,\frac{1}{\min_j \lambda_{\min}(\mathcal{X}_{P_j,t}^T \mathcal{X}_{P_j,t})}\right)} \left(\frac{\delta}{2\frac{n}{|P_j|}T}\right)\right)$$

$$\leq 2 \sum_{t>T_1} \left(Q_{N\left(0,\frac{1}{\min_i \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}\right)} \left(\frac{\delta}{2nT}\right) + Q_{N\left(0,\frac{1}{\min_j \alpha \left\lfloor \frac{\mathcal{G}_{P_j,t}}{d} \right\rfloor \lambda_{\min_{P_j,d}}}\right)} \left(\frac{\delta}{2\frac{n}{|P_j|}T}\right)\right) + 3\delta' T \tag{C.65}$$

$$\leq 2 \sum_{t>T_1} \left(\sqrt{\frac{\ln \frac{2nT}{\delta}}{\min_i \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}} + \sqrt{\frac{\ln \frac{2\frac{n}{min_j |P_j|}T}{\delta}}{\min_i \alpha \left\lfloor \frac{\mathcal{G}_{P_j,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}} \right) + 6\delta' T \tag{C.66}$$

$$\leq 2 \sum_{t>T_1} \left(2\sqrt{\frac{\ln \frac{2nT}{\delta}}{\min_i \alpha \left\lfloor \frac{G_{i,t}}{d} \right\rfloor \lambda_{\min_{i,d}}}} \right) + 6\delta' T \tag{C.67}$$

$$= O\left(\sum_{t>T_1} \sqrt{d \frac{\ln \frac{2nT}{\delta}}{\min_i G_{i,t} \lambda_{\min_{i,d}}}} + \delta' T \right)$$

$$= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t>T_1}\sqrt{\frac{1}{\min_i G_{i,t}}} + \delta' T\right)$$

$$= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t>T_1}\sqrt{\frac{n}{\sum_{t'<t} p_{t'}}} + \delta' T\right)$$

$$= O\left(\sqrt{d\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t>T_1}\sqrt{\frac{n}{t^{2/3}}} + \delta' T\right) \tag{C.68}$$

$$= O\left(\sqrt{dn\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}\sum_{t\in[T_1,T]}\frac{1}{t^{1/3}} + \delta' T\right)$$

$$= O\left(\sqrt{dn\frac{\ln\frac{2nT}{\delta}}{\min_i \lambda_{\min_{i,d}}}}T^{2/3} + \delta' T\right) \tag{C.69}$$

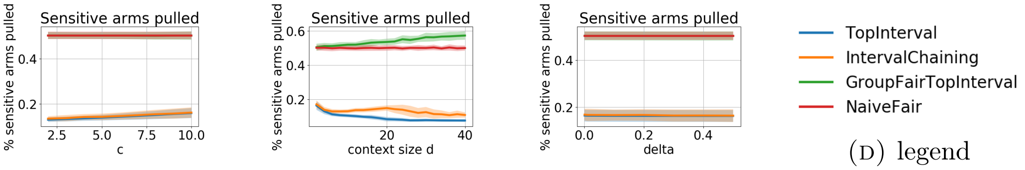
where inequality C.64 is due to equation C.41, inequality C.65 is due to equation C.44 and equation C.45, inequality C.66 is due to a Chernoff bound, inequality C.67 is due to the fact that $\frac{n}{\min_j |P_j|} < n$ and $\min_j \mathcal{G}_{P_j,t} \geq min_i G_{i,t}$, and equation C.68 is due to equation C.40.

Combining equation C.39, equation C.40, equation C.63, and equation C.69 and setting $\delta' = \min(\frac{1}{3nT}, \frac{1}{T^{1/3}})$ we get Theorem C.5. □

C.4 Additional Experiments

Additionally to the experiments found in Section 6.4.1, we ran the following experiments:

(a) Varying the range in which coefficients are chosen (between [0,c]) while setting the total budget $T = 1000$, the number of arms $n = 10$, the error mean $\mu = 10$, the number of sensitive arms equal to 5, and the context dimension $d = 2$ (Figures C.1a and C.2a).



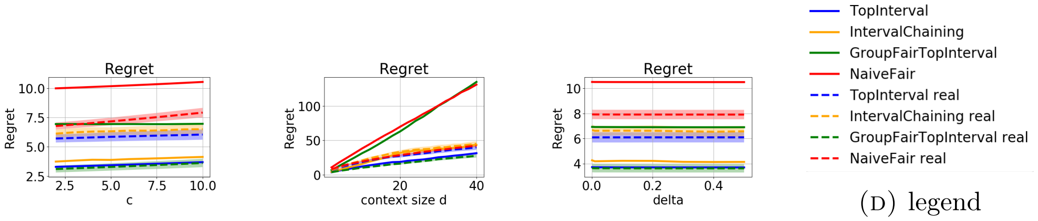
(D) legend

(A) $n = 10$, $\mu = 10$, number of sensitive arms $= 5$

(B) $T = 1000$, $\mu = 10$, number of sensitive arms $= 5$

(C) $n = 10$, $T = 1000$, number of sensitive arms $= 5$

FIGURE C.1: Percentage of total arm pulls that were pulled using sensitive arms.



(D) legend

(A) $n = 10$, $\mu = 10$, number of sensitive arms $= 5$

(B) $T = 1000$, $\mu = 10$, number of sensitive arms $= 5$

(C) $n = 10$, $T = 1000$, number of sensitive arms $= 5$

FIGURE C.2: Regret for synthetic experiments. The solid lines are regret given the rewards received from pulling the arms (including the group bias). The dashed lines is the true regret (without the group bias).

(b) Varying the context dimension while setting the total budget $T = 1000$, the number of arms $n = 10$, the error mean $\mu = 10$, and the number of sensitive arms equal to 5 (Figures C.1b and C.2b).

(c) Varying probability δ while setting the total budget $T = 1000$, the number of arms $n = 10$, the error mean $\mu = 10$, the number of sensitive arms equal to 5, and the context dimension $d = 2$ (Figures C.1c and C.2c).

Appendix D: Transfer of Machine Learning Fairness across Domains

D.1 Rademacher Complexity

We provide additional bounds dependent on Radmacher Complexity based on the following definition of data-driven empirical Rademacher Complexity

Definition 1. Given a hypothesis space $\mathcal{H}$, a sample $S \in \mathcal{X}^m$, the empirical Rademacher Complexity of $\mathcal{H}$ is defined as

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \frac{2}{m}\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} | \sum_{i=1}^{m} \sigma_i h(x_i)| \middle| S = (x_1, \ldots, x_m) \right].$$

The expectation is taken over $\sigma = (\sigma_1, \ldots, \sigma_m)$ where $\sigma_i \in \{-1, +1\}$ are uniform independent random variables. The Rademacher Complexity of a hypothesis space is defined as the expectation of $\hat{\mathfrak{R}}$ over all sample sets of size m

$$\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_S \left[\hat{\mathfrak{R}}_S(\mathcal{H}) \middle| |S| = m \right]. \tag{D.1}$$

Rademacher Complexity measures the ability of a hypothesis space to fit random noise. The empirical Rademacher Complexity function allows us to estimate the Rademacher Complexity using a finite sample of data. Rademacher Complexity bounds can lead to tighter bounds than those of VC-dimension, especially when analyzing neural network models.

When transitioning to Rademacher Complexity we need to change the binary labels from $\{0,1\}$ to $\{-1,1\}$. This means that the error of a hypothesis g is defined as

$$\epsilon_{S^l_\alpha}(g,f) = \mathbb{E}_{z^l_\alpha \sim D_{S^l_\alpha}}\left[\frac{|g(z^l_\alpha) - f(z^l_\alpha)|}{2}\right].$$

Additionally, we need new definitions of the equal opportunity and equalized odds distances over the new binary group membership. The equal opportunity distance is defined as

$$\Delta_{EOp_S}(g) \triangleq \mathbb{E}_{Z_0^{-1} \sim D_{S_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1} \sim D_{S_1^{-1}}}\left[\frac{1+g(z_1^{-1})}{2}\right],$$

while the equlized odds distance is defined as

$$\begin{aligned}\Delta_{EO_T}(g) \triangleq & \left|\mathbb{E}_{Z_0^{-1} \sim D_{T_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1} \sim D_{T_1^{-1}}}\left[\frac{1+g(z_1^{-1})}{2}\right]\right| \\ & + \left|\mathbb{E}_{Z_0^{1} \sim D_{T_0^{1}}}\left[\frac{1+g(z_0^{1})}{2}\right] - \mathbb{E}_{Z_1^{1} \sim D_{T_1^{1}}}\left[\frac{1+g(z_1^{1})}{2}\right]\right|.\end{aligned}$$

Using these new definitions Theorem D.1 provides a Rademacher Complexity bound of the equal opportunity distance in the target space. This closely resembles the VC-dimension bound in Theorem 7.1.

Theorem D.1. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ are samples of size m', each drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, and $\mathcal{D}_{T_1^{-1}}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in* *the target space is bounded by*

$$\begin{aligned}\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^{-1}},\mathcal{U}_{S_0^{-1}}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^{-1}},\mathcal{U}_{S_1^{-1}}) \\ &+ 2\left(\mathfrak{R}_{U_{T_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{T_1^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_1^{-1}}}(\mathcal{H})\right) \\ &+ 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},\end{aligned}$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

The proof also follows a similar logic to the sketch given for Theorem 7.1 with the additional step of using a modification of Corollary 7 given by Mansour et al. [164].

Similarly, Theorem D.2 provides a Rademacher Complexity bound of the equalized odds distance in the target space.

Theorem D.2. *Let* $\mathcal{H}$ *be a hypothesis space. If* $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ $\mathcal{U}_{S_0^1}$, $\mathcal{U}_{S_1^1}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^1}$ *are samples of size* m'*, each drawn from* $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, $\mathcal{D}_{T_1^{-1}}$, $\mathcal{D}_{S_0^1}$, $\mathcal{D}_{S_1^1}$, $\mathcal{D}_{T_0^1}$, *and* $\mathcal{D}_{T_1^1}$ *respectively, then for any* $\delta \in (0, 1)$*, with probability at least* $1-\delta$ *(over the choice of samples), for every* $g \in \mathcal{H}$ *(where* $\mathcal{H}$ *is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$\begin{aligned}\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) &+ \frac{1}{2}\Big(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}},\mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}},\mathcal{U}_{T_1^{-1}}) \\ &+\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^1},\mathcal{U}_{T_0^1}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^1},\mathcal{U}_{T_1^1})\Big) \\ &+ 2\Big(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \\ &+ \hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) +\hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H})\Big)\end{aligned}$$

$$+ 12\sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \lambda_{EO},$$

where $\lambda_{EO} = \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1$*, and* $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Given either the Rademacher Complexity bounds or the VC-dimension bounds, the implications stay the same. In order for a successful transfer of fairness the two (or four) subspace domains should be close across the source and target domains. Additionally, there should be a hypothesis in the hypothesis space that performs well over all of the relevant subspaces.

D.2 Proofs

Lemma D.1. *(From Ben-David et al. [24]) For any hypotheses* $h, h' \in \mathcal{H}$*,*

$$|\epsilon_S(h, h') - \epsilon_T(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T).$$

Lemma D.2. *(From [23, 73]) For any labeling functions* f_1*,* f_2*, and* f_3*, we have*

$$\epsilon(f_1, f_2) \leq \epsilon(f_1, f_3) + \epsilon(f_2, f_3).$$

D.2.1 VC-dimension bounds

Lemma D.3. *(From Ben-David et al. [24]) Let* $\mathcal{H}$ *be a hypothesis space on* $\mathcal{Z}$ *with VC-dimension* d*. If* $\mathcal{U}$ *and* $\mathcal{U}'$ *are samples of size* m *from* $\mathcal{D}$ *and* $\mathcal{D}'$ *respectively and* $\hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}')$ *is the empirical* $\mathcal{H}$*-divergence between samples, then for any* $\delta \in (0, 1)$*,* *with probability at least* $1-\delta$*,*

$$d_{\mathcal{H}}(\mathcal{D},\mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') + 4\sqrt{\frac{d\log(2m)+\log(\frac{2}{\delta})}{m}}.$$

Theorem D.4. *Let* $\mathcal{H}$ *be a hypothesis space of VC dimension* d*. If* $\mathcal{U}_{S_0^0}$*,* $\mathcal{U}_{S_1^0}$*,* $\mathcal{U}_{T_0^1}$*,* $\mathcal{U}_{T_1^0}$ *are samples of size* m' *each, drawn from* $\mathcal{D}_{S_0^0}$*,* $\mathcal{D}_{S_1^0}$*,* $\mathcal{D}_{T_0^0}$*, and* $\mathcal{D}_{T_1^0}$ *respectively, then for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *(over the choice of samples), for every* $g \in \mathcal{H}$ *(where* $\mathcal{H}$ *is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\begin{aligned}\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0},\mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0},\mathcal{U}_{S_1^0}) \\ &+ 8\sqrt{\frac{2d\log(2m')+\log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0,\end{aligned}$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*,f) + \epsilon_{T_\alpha^l}(g^*,f)$*.*

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^0\sim D_{S_0^0}} \geq \mathbb{E}_{Z_1^0\sim D_{S_1^0}}$. Then we can rewrite $\Delta_{EOp_S}(g)$ as follows:

$$\begin{aligned}\Delta_{EOp_S}(g) &= \mathbb{E}_{Z_0^0\sim\mathcal{D}_{S_0^0}}\left[g(Z_0^0)\right] - \mathbb{E}_{Z_1^0\sim\mathcal{D}_{S_1^0}}\left[g(z_1^0)\right] \\ &= \mathbb{E}_{Z_0^0\sim\mathcal{D}_{S_0^0}}\left[g(Z_0^0)\right] + \mathbb{E}_{Z_1^0\sim\mathcal{D}_{S_1^0}}\left[1-g(z_1^0)\right] - 1 \\ &= \epsilon_{S_0^0}(g,f) + \epsilon_{S_1^0}(1-g,f) - 1,\end{aligned}$$

where the last line follows from the fact that equal opportunity only cares about the error on the false data-points.

We now have the tools to find an upper-bound on $\Delta_{EOp_T}(g)$.

$$\begin{aligned}
\Delta_{EOp_T}(g) =&\epsilon_{T_0^0}(g,f) + \epsilon_{T_1^0}(1-g,f) - 1 \\
&\le \epsilon_{T_0^0}(g,g^*) + \epsilon_{T_0^0}(f,g^*) + \epsilon_{T_1^0}(1-g,g^*) + \epsilon_{T_1^0}(f,g^*) - 1 \qquad \text{(D.2)} \\
&= \epsilon_{T_0^0}(g^*,f) + \epsilon_{T_0^0}(g,g^*) + \epsilon_{T_1^0}(g^*,f) + \epsilon_{T_1^0}(1-g,g^*) - 1 \\
&= \epsilon_{T_0^0}(g^*,f) + \epsilon_{T_0^0}(g,g^*) + \epsilon_{S_0^0}(g,g^*) - \epsilon_{S_0^0}(g,g^*) \\
&\quad + \epsilon_{T_1^0}(g^*,f) + \epsilon_{T_1^0}(1-g,g^*) + \epsilon_{S_1^0}(1-g,g^*) - \epsilon_{S_1^0}(1-g,g^*) - 1 \\
&\le \epsilon_{T_0^0}(g^*,f) + \epsilon_{S_0^0}(g,g^*) + \Big|\epsilon_{T_0^0}(g,g^*) - \epsilon_{S_0^0}(g,g^*)\Big| \\
&\quad + \epsilon_{T_1^0}(g^*,f) + \epsilon_{S_1^0}(1-g,g^*) + \Big|\epsilon_{T_1^0}(1-g,g^*) - \epsilon_{S_1^0}(1-g,g^*)\Big| - 1 \\
&\le \epsilon_{T_0^0}(g^*,f) + \epsilon_{S_0^0}(g,g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&\quad + \epsilon_{T_1^0}(g^*,f) + \epsilon_{S_1^0}(1-g,g^*) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \qquad \text{(D.3)} \\
&\le \epsilon_{T_0^0}(g^*,f) + \epsilon_{S_0^0}(g,f) + \epsilon_{S_0^0}(g^*,f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&\quad + \epsilon_{T_1^0}(g^*,f) + \epsilon_{S_1^0}(1-g,f) + \epsilon_{S_1^0}(g^*,f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \\
&\qquad \text{(D.4)} \\
&= \epsilon_{S_0^0}(g,f) + \epsilon_{T_0^0}(g^*,f) + \epsilon_{S_0^0}(g^*,f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&\quad + \epsilon_{S_1^0}(1-g,f) + \epsilon_{T_1^0}(g^*,f) + \epsilon_{S_1^0}(g^*,f) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) - 1 \\
&= \epsilon_{S_0^0}(g,f) + \epsilon_{S_1^0}(1-g,f) - 1 + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \\
&\quad + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) + \lambda_0^0 + \lambda_1^0 \qquad \text{(D.5)} \\
&= \Delta_{EOp_S}(g) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) + \lambda_0^0 + \lambda_1^0 \\
&\le \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0}) \\
&\quad + 8\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_0^0 + \lambda_1^0, \qquad \text{(D.6)}
\end{aligned}$$

Where inequality D.2 is due to lemma D.2, inequality D.3 is due to lemma D.1 and the fact that $\mathcal{H}$ is a symmetric hypothesis space, inequality D.4 is due to lemma D.2, equality D.5 is due to the definition of λ_α^l, and inequality D.6 is due to lemma D.3. □

Theorem D.5. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If $\mathcal{U}_{S_\alpha^l}$ are samples of size m' each, drawn from $\mathcal{D}_{S_\alpha^l}$ for all $\alpha \in \Omega_A = \{0,1\}$ and $l \in \Omega_\mathcal{Y} = 0, 1$, then for any $\delta \in (0,1)$, with probability at least $1 - \delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$\begin{aligned}\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0},\mathcal{U}_{S_0^0}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0},\mathcal{U}_{S_1^0}) \\ &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^1},\mathcal{U}_{S_0^1}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^1},\mathcal{U}_{S_1^1}) \\ &+ 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO},\end{aligned}$$

where $\lambda_{EO} = \lambda_0^0 + \lambda_1^0 + \lambda_0^1 + \lambda_1^1$, and $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. WLOG assume $\mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] \geq \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[g]$ and $\mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] \geq \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[g]$. Then,

$$\begin{aligned}\Delta_{EO_S} &= \mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] - \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[g] + \mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] - \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[g] \\ &= \mathbb{E}_{Z_0^0 \sim D_{S_0^0}}[g] + \mathbb{E}_{Z_1^0 \sim D_{S_1^0}}[1-g] + \mathbb{E}_{Z_0^1 \sim D_{S_0^1}}[g] + \mathbb{E}_{Z_1^1 \sim D_{S_1^1}}[1-g] - 2 \\ &= \epsilon_{S_0^0}(g, f) + \epsilon_{S_1^0}(1-g, f) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_1^1}(1-g, f) - 2.\end{aligned}$$

Using this and the previous lemmas we have:

$$
\begin{align}
\Delta_{EO_T}(g) &= \epsilon_{T_0^0}(g, f) + \epsilon_{T_1^0}(1 - g, f) + \epsilon_{T_0^1}(g, f) + \epsilon_{T_1^1}(1 - g, f) - 2 \notag \\
&\leq \epsilon_{T_0^0}(g, g^*) + \epsilon_{T_0^0}(f, g^*) + \epsilon_{T_1^0}(1 - g, g^*) + \epsilon_{T_1^0}(f, g^*) \notag \\
&\quad + \epsilon_{T_0^1}(g, g^*) + \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_1^1}(1 - g, g^*) + \epsilon_{T_1^1}(f, g^*) - 2 \tag{D.7} \\
&= \epsilon_{T_0^0}(g^*, f) + \epsilon_{T_0^0}(g, g^*) + \epsilon_{S_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*) \notag \\
&\quad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{T_1^0}(1 - g, g^*) + \epsilon_{S_1^0}(1 - g, g^*) - \epsilon_{S_1^0}(1 - g, g^*) \notag \\
&\quad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{T_0^1}(g, g^*) + \epsilon_{S_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*) \notag \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{T_1^1}(1 - g, g^*) + \epsilon_{S_1^1}(1 - g, g^*) - \epsilon_{S_1^1}(1 - g, g^*) - 2 \notag \\
&\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \Big|\epsilon_{T_0^0}(g, g^*) - \epsilon_{S_0^0}(g, g^*)\Big| \notag \\
&\quad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, g^*) + \Big|\epsilon_{T_1^0}(1 - g, g^*) - \epsilon_{S_1^0}(1 - g, g^*)\Big| \notag \\
&\quad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, g^*) + \Big|\epsilon_{T_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*)\Big| \notag \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1 - g, g^*) + \Big|\epsilon_{T_1^1}(1 - g, g^*) - \epsilon_{S_1^1}(1 - g, g^*)\Big| - 2 \notag \\
&\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \notag \\
&\quad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) \notag \\
&\quad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) \notag \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1 - g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 \tag{D.8} \\
&\leq \epsilon_{T_0^0}(g^*, f) + \epsilon_{S_0^0}(g, f) + \epsilon_{S_0^0}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) \notag \\
&\quad + \epsilon_{T_1^0}(g^*, f) + \epsilon_{S_1^0}(1 - g, f) + \epsilon_{S_1^0}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0}) \notag \\
&\quad + \epsilon_{T_0^1}(g^*, f) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_0^1}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) \notag
\end{align}
$$

$$
\begin{aligned}
&+ \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(1-g, f) + \epsilon_{S_1^1}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 \qquad \text{(D.9)}\\
&= \lambda_0^0 + \epsilon_{S_0^0}(g, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0})\\
&\quad + \lambda_1^0 + \epsilon_{S_1^0}(1-g, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})\\
&\quad + \lambda_0^1 + \epsilon_{S_0^1}(g, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1})\\
&\quad + \lambda_1^1 + \epsilon_{S_1^1}(1-g, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2\\
&= \Delta_{EO_S}(g) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^0}, D_{S_0^0}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^0}, D_{S_1^0})\\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) + \lambda_{EO}\\
&\leq \Delta_{EO_S}(g) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^0}, \mathcal{U}_{S_0^0}) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^0}, \mathcal{U}_{S_1^0})\\
&\quad + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^1}, \mathcal{U}_{S_0^1}) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^1}, \mathcal{U}_{S_1^1})\\
&\quad + 16\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda_{EO}, \qquad \text{(D.10)}
\end{aligned}
$$

where inequality D.7 is due to lemma D.2, inequality D.8 is due to lemma D.1 and the fact that $\mathcal{H}$ is a symmetric hypothesis space, inequality D.9 is due to lemma D.2, and inequality D.10 is due to lemma D.3. □

D.2.2 Rademacher Complexity Bounds

Lemma D.6. *(A modification of Corollary 7 from Mansour et al. [164]) Let $\mathcal{H}$ by a hypothesis set of classifiers mapping the feature space X to the labels $\{-1, 1\}$. Let $\mathcal{U}$ and $\mathcal{U}'$ be the set of samples each of size m sampled from $\mathcal{D}$ and $\mathcal{D}'$ respectively.*

Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{U}$ and $\mathcal{U}'$:

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') + 4\left(\mathfrak{R}_{\mathcal{U}}(\mathcal{H}) + \mathfrak{R}_{\mathcal{U}}(\mathcal{H})\right) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Theorem D.7. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ are samples of size m' each, drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, and $\mathcal{D}_{T_1^{-1}}$ respectively, then for any $\delta \in (0, 1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equal opportunity in the target space is bounded by*

$$\begin{aligned}\Delta_{EOp_T}(g) \leq \Delta_{EOp_S}(g) &+ \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_0^{-1}},\mathcal{U}_{S_0^{-1}}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{T_1^{-1}},\mathcal{U}_{S_1^{-1}}) \\ &+ 2\left(\mathfrak{R}_{U_{T_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_0^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{T_1^{-1}}}(\mathcal{H}) + \mathfrak{R}_{U_{S_1^{-1}}}(\mathcal{H})\right) \\ &+ 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},\end{aligned}$$

where $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}} \geq \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}$. Then we can rewrite Δ_{EOp_S} as follows.

$$\begin{aligned}\Delta_{EOp_S}(g) &= \mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}\left[\frac{1+g(z_1^{-1})}{2}\right] \\ &= \mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}\left[1 - \frac{1+g(z_1^{-1})}{2}\right] - 1 \\ &= \mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}\left[\frac{1-g(z_1^{-1})}{2}\right] - 1\end{aligned}$$

$$= \mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}}\left[\frac{g(z_0^{-1}) - f(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}\left[\frac{g(z_1^{-1}) + f(z_1^{-1})}{2}\right] - 1 \tag{D.11}$$

$$= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) - 1,$$

where D.11 is due to the fact that $f(z_0^{-1}) = -1$ by definition.

We now have the tools to find an upper bound on $\Delta_{EOp_T}(g)$.

$$\begin{aligned}
\Delta_{EOp_T}(g) &= \epsilon_{T_0^{-1}}(g, f) + \epsilon_{T_1^{-1}}(-g, f) - 1 \\
&\leq \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{T_1^{-1}}(f, g^*) - 1 \qquad \text{(D.12)} \\
&= \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{S_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*) \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*) - 1 \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, g^*) + |\epsilon_{T_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*)| \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, g^*) + |\epsilon_{T_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*)| - 1 \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) - 1 \qquad \text{(D.13)} \\
&\leq \epsilon_{T_0^{-1}}(g^*, f) + \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_0^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(g^*, f) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_1^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) - 1 \\
&\qquad \text{(D.14)} \\
&= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) - 1 + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + \lambda_0^{-1} + \lambda_1^{-1} \\
&= \Delta_{EOp_S}(g) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + \lambda_0^{-1} + \lambda_1^{-1} \\
&\qquad \text{(D.15)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \Delta_{EOp_S}(g) + \lambda_0^{-1} + \lambda_1^{-1} \\
&\quad + \frac{1}{2}\left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + 4\left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H})\right) + 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right) \\
&\quad + \frac{1}{2}\left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + 4\left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H})\right) + 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right)
\end{aligned}
\tag{D.16}
$$

$$
\begin{aligned}
&= \Delta_{EOp_S}(g) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + 2\left(\hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H})\right) \\
&\quad + 6\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1},
\end{aligned}
$$

where Eq. D.12 is due to Lemma D.2, Eq. D.13 is due to Lemma D.1, Eq. D.14 is due to Lemma D.2, Eq. D.15 is due to the definition of $\Delta_{EOp_S}(g)$, and Eq. D.16 is due to Lemma D.6. □

Theorem D.8. *Let $\mathcal{H}$ be a hypothesis space. If $\mathcal{U}_{S_0^{-1}}$, $\mathcal{U}_{S_1^{-1}}$, $\mathcal{U}_{T_0^{-1}}$, $\mathcal{U}_{T_1^{-1}}$ $\mathcal{U}_{S_0^1}$, $\mathcal{U}_{S_1^1}$, $\mathcal{U}_{T_0^1}$, $\mathcal{U}_{T_1^1}$ are samples of size m' each, drawn from $\mathcal{D}_{S_0^{-1}}$, $\mathcal{D}_{S_1^{-1}}$, $\mathcal{D}_{T_0^{-1}}$, $\mathcal{D}_{T_1^{-1}}$,$\mathcal{D}_{S_0^1}$, $\mathcal{D}_{S_1^1}$, $\mathcal{D}_{T_0^1}$, and $\mathcal{D}_{T_1^1}$ respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of samples), for every $g \in \mathcal{H}$ (where $\mathcal{H}$ is a symmetric hypothesis space) the distance from equalized odds in the target space is bounded by*

$$
\begin{aligned}
\Delta_{EO_T}(g) \leq \Delta_{EO_S}(g) &+ \frac{1}{2}\Big(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}}, \mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}}, \mathcal{U}_{T_1^{-1}}) \\
&+ \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^1}, \mathcal{U}_{T_0^1}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^1}, \mathcal{U}_{T_1^1})\Big) \\
&+ 2\Big(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H})
\end{aligned}
$$

$$
\begin{aligned}
&+ \hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H})\Big) \\
&+ 12\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \lambda_{EO},
\end{aligned}
$$

where $\lambda_{EO} = \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1$, *and* $\lambda_\alpha^l = \epsilon_{S_\alpha^l}(g^*, f) + \epsilon_{T_\alpha^l}(g^*, f)$.

Proof. Without loss of generality assume $\mathbb{E}_{Z_0^{-1}\sim D_{S_0^{-1}}} \geq \mathbb{E}_{Z_1^{-1}\sim D_{S_1^{-1}}}$ and $\mathbb{E}_{Z_0^1\sim D_{S_0^1}} \geq \mathbb{E}_{Z_1^1\sim D_{S_1^1}}$. Then we can rewrite Δ_{EOp_S} as follows.

$$
\begin{aligned}
\Delta_{EO_T}(g) &= \mathbb{E}_{Z_0^{-1}\sim D_{T_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] - \mathbb{E}_{Z_1^{-1}\sim D_{T_1^{-1}}}\left[\frac{1+g(z_1^{-1})}{2}\right] \\
&\quad + \mathbb{E}_{Z_0^1\sim D_{T_0^1}}\left[\frac{1+g(z_0^1)}{2}\right] - \mathbb{E}_{Z_1^1\sim D_{T_1^1}}\left[\frac{1+g(z_1^1)}{2}\right] \\
&= \mathbb{E}_{Z_0^{-1}\sim D_{T_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{T_1^{-1}}}\left[1-\frac{1+g(z_1^{-1})}{2}\right] - 1 \\
&\quad + \mathbb{E}_{Z_0^1\sim D_{T_0^1}}\left[\frac{1+g(z_0^1)}{2}\right] + \mathbb{E}_{Z_1^1\sim D_{T_1^1}}\left[1-\frac{1+g(z_1^1)}{2}\right] - 1 \\
&= \mathbb{E}_{Z_0^{-1}\sim D_{T_0^{-1}}}\left[\frac{1+g(z_0^{-1})}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{T_1^{-1}}}\left[\frac{1-g(z_1^{-1})}{2}\right] \\
&\quad + \mathbb{E}_{Z_0^1\sim D_{T_0^1}}\left[\frac{1+g(z_0^1)}{2}\right] + \mathbb{E}_{Z_1^1\sim D_{T_1^1}}\left[\frac{1-g(z_1^1)}{2}\right] - 2 \\
&= \mathbb{E}_{Z_0^{-1}\sim D_{T_0^{-1}}}\left[\frac{|g(z_0^{-1}) - f(z_0^{-1})|}{2}\right] + \mathbb{E}_{Z_1^{-1}\sim D_{T_1^{-1}}}\left[\frac{|g(z_1^{-1}) + f(z_1^{-1})|}{2}\right] \\
&\quad + \mathbb{E}_{Z_0^1\sim D_{T_0^1}}\left[\frac{|g(z_0^1) - f(z_0^1)|}{2}\right] + \mathbb{E}_{Z_1^1\sim D_{T_1^1}}\left[\frac{|g(z_1^1) + f(z_1^1)|}{2}\right] - 2 \\
&= \epsilon_{T_0^{-1}}(g,f) + \epsilon_{T_1^{-1}}(-g,f) + \epsilon_{T_0^1}(g,f) + \epsilon_{T_1^1}(-g,f) - 2
\end{aligned}
$$

Using this and previous lemmas we have

$$
\begin{aligned}
\Delta_{EO_T}(g) &= \epsilon_{T_0^{-1}}(g,f) + \epsilon_{T_1^{-1}}(-g,f) + \epsilon_{T_0^1}(g,f) + \epsilon_{T_1^1}(-g,f) - 2 \\
&\leq \epsilon_{T_0^{-1}}(g,g^*) + \epsilon_{T_0^{-1}}(f,g^*) + \epsilon_{T_1^{-1}}(-g,g^*) + \epsilon_{T_1^{-1}}(f,g^*)
\end{aligned}
$$

$$
\begin{aligned}
&+ \epsilon_{T_0^1}(g, g^*) + \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_1^1}(-g, g^*) + \epsilon_{T_1^1}(f, g^*) - 2 && \text{(D.17)} \\
&= \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{T_0^{-1}}(g, g^*) + \epsilon_{S_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*) \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{T_1^{-1}}(-g, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*) \\
&\quad + \epsilon_{T_0^1}(f, g^*) + \epsilon_{T_0^1}(g, g^*) + \epsilon_{S_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*) \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{T_1^1}(-g, g^*) + \epsilon_{S_1^1}(-g, g^*) - \epsilon_{S_1^1}(-g, g^*) - 2 \\
&\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, g^*) + \left|\epsilon_{T_0^{-1}}(g, g^*) - \epsilon_{S_0^{-1}}(g, g^*)\right| \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) + \left|\epsilon_{T_1^{-1}}(-g, g^*) - \epsilon_{S_1^{-1}}(-g, g^*)\right| \\
&\quad + \epsilon_{T_0^1}(f, g^*) + \epsilon_{S_0^1}(g, g^*) + \left|\epsilon_{T_0^1}(g, g^*) - \epsilon_{S_0^1}(g, g^*)\right| \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(-g, g^*) + \left|\epsilon_{T_1^1}(-g, g^*) - \epsilon_{S_1^1}(-g, g^*)\right| - 2 \\
&\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + \epsilon_{T_0^1}(f, g^*) + \epsilon_{S_0^1}(g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(-g, g^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 && \text{(D.18)} \\
&\leq \epsilon_{T_0^{-1}}(f, g^*) + \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_0^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) \\
&\quad + \epsilon_{T_1^{-1}}(f, g^*) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_1^{-1}}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + \epsilon_{T_0^1}(f, g^*) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_0^1}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) \\
&\quad + \epsilon_{T_1^1}(f, g^*) + \epsilon_{S_1^1}(-g, f) + \epsilon_{S_1^1}(g^*, f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) - 2 && \text{(D.19)} \\
&= \epsilon_{S_0^{-1}}(g, f) + \epsilon_{S_1^{-1}}(-g, f) + \epsilon_{S_0^1}(g, f) + \epsilon_{S_1^1}(-g, f) - 2 \\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1
\end{aligned}
$$

$$
\begin{aligned}
&= \Delta_{EO_S}(g) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) \\
&\quad + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1 \\
&\le \Delta_{EO_S}(g) + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1 \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^{-1}}, D_{S_0^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^{-1}}, D_{S_1^{-1}}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_0^1}, D_{S_0^1}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right) \\
&\quad + \frac{1}{2} \left(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_{T_1^1}, D_{S_1^1}) + 4 \left(\hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H}) \right) + 6 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \right)
\end{aligned}
\tag{D.20}
$$

$$
\begin{aligned}
&= \Delta_{EO_S}(g) + \frac{1}{2} \Big(\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^{-1}}, \mathcal{U}_{T_0^{-1}}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^{-1}}, \mathcal{U}_{T_1^{-1}}) \\
&\quad + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_0^1}, \mathcal{U}_{T_0^1}) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_{S_1^1}, \mathcal{U}_{T_1^1}) \Big) \\
&\quad + 2 \Big(\hat{\mathfrak{R}}_{U_{S_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^{-1}}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^{-1}}}(\mathcal{H}) \\
&\quad + \hat{\mathfrak{R}}_{U_{S_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_0^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{S_1^1}}(\mathcal{H}) + \hat{\mathfrak{R}}_{U_{T_1^1}}(\mathcal{H}) \Big) \\
&\quad + 12 \sqrt{\frac{\log \frac{2}{\delta}}{2m}} + \lambda_0^{-1} + \lambda_1^{-1} + \lambda_0^1 + \lambda_1^1,
\end{aligned}
$$

where Eq. D.17 is due to Lemma D.2, Eq. D.18 is due to Lemma D.1, Eq. D.19 is due to Lemma D.2, and D.20 is due to Lemma D.6. □

D.3 Experimental setup

For the UCI adult dataset we used all 14 features as provided in https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names. The original train/test split is used. For the COMPAS dataset we used the features provided in https://github.com/propublica/compas-analysis/blob/master/compas-scores.csv, and predict the risk of recidivism (decile_score) for each row.

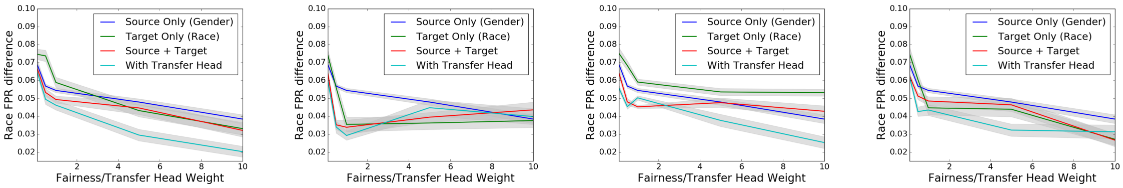
We did 10-fold cross-validation and choose the hyperparameters with the best performance on the validation data. 64 dimension embedding is used for categorical features and 256 hidden units are used in the model. We did parameter search and found 10K steps yields a good balance of runtime and accuracy. Each run takes about 1hr for UCI data and 0.5hrs for COMPAS on a single CPU with 2GB RAM. Increasing learning rate speeds up experiments but also hurts accuracy slightly (e.g., ~2pp decrease on UCI).

For range of parameters, we have considered the following: (1) batch size: $[64, 128, 256, 512]$; (2) learning rate: $[0.01, 0.1, 1.0]$; (3) number of hidden units: $[64, 128, 256, 512]$; (4) embedding dimension: $[32, 64, 128]$. (5) number of steps: $[5000, 10000, 20000, 50000]$.

D.4 Experiments

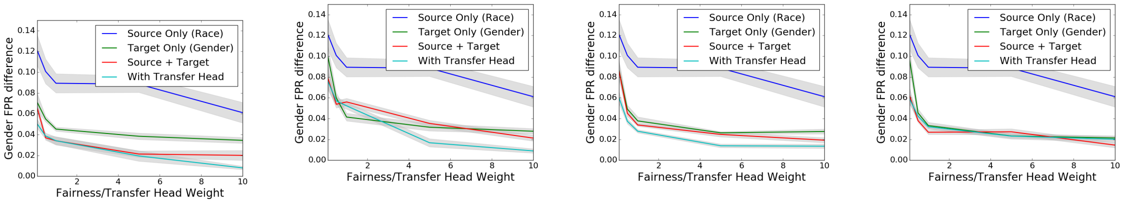
D.4.1 Experiment Results for fairness on UCI and COMPAS

Figure D.1 depicts the results of the analysis for transferring from gender to race, while Figure D.2 shows the results for transferring from race to gender, on the UCI



(A) 50 race samples. (B) 100 race samples. (C) 500 race samples. (D) 1000 race samples.

FIGURE D.1: Gender → Race on the UCI dataset. Comparison of FPR difference on sensitive attribute *race*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).



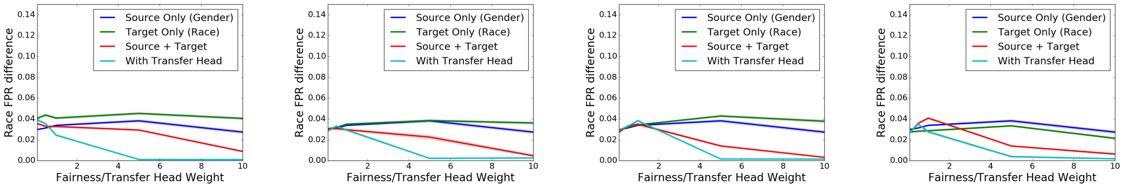
(A) 50 gender samples. (B) 100 gender samples. (C) 500 gender samples. (D) 1000 gender samples.

FIGURE D.2: Race → Gender on the UCI dataset. Comparison of FPR difference on sensitive attribute *gender*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).

dataset. Figure D.3 and Figure D.4 show the results on the COMPAS dataset. The line and the shaded areas show the mean and the standard error of the mean across 30 trials. These experiments show that the Transfer model is effective in decreasing the FPR gap in the target domain and is more sample efficient than previous methods.

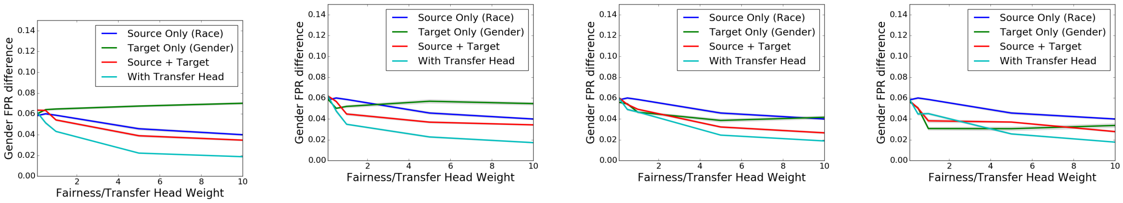
D.4.2 Accuracy vs. Fairness/Transfer Head Weight

In this section we further add the comparison on accuracy with respect to the weight of the fairness/transfer head. Fig. D.5 and Fig. D.6 show the results comparing the



(A) 50 race samples. (B) 100 race samples. (C) 500 race samples. (D) 1000 race samples.

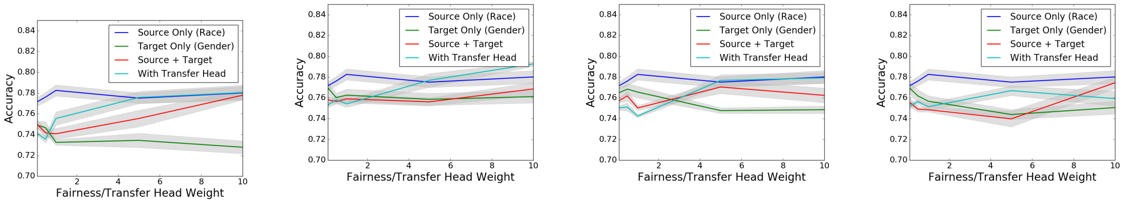
FIGURE D.3: Gender → Race on the COMPAS dataset. Comparison of FPR difference on sensitive attribute *race*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).



(A) 50 gender samples. (B) 100 gender samples. (C) 500 gender samples. (D) 1000 gender samples.

FIGURE D.4: Race → Gender on the COMPAS dataset. Comparison of FPR difference on sensitive attribute *gender*, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).

Transfer model with the baselines, by transferring *race* to *gender*, and *race* to *gender*, respectively. Fig. D.7 and Fig. D.8 show the results on COMPAS.



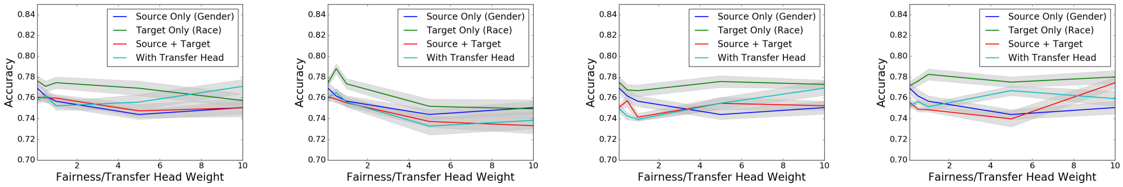
(A) 50 gender samples.

(B) 100 gender samples.

(C) 500 gender samples.

(D) 1000 gender samples.

FIGURE D.5: Comparison of accuracy on the UCI data for Race → Gender, by transferring from the source domain (1000 samples for each race) to the target domain (varying samples for each gender as indicated in the caption).



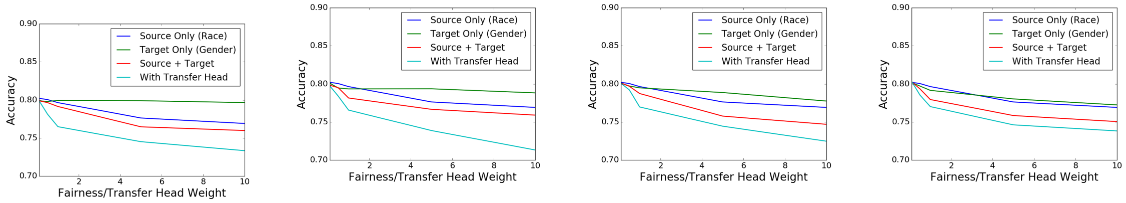
(A) 50 race samples.

(B) 100 race samples.

(C) 500 race samples.

(D) 1000 race samples.

FIGURE D.6: Comparison of accuracy on the UCI data for Gender → Race, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).



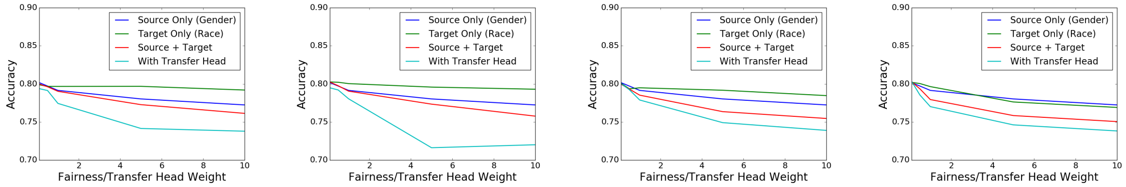
(A) 50 gender samples.

(B) 100 gender samples.

(C) 500 gender samples.

(D) 1000 gender samples.

Figure D.7: Comparison of accuracy on COMPAS for Race → Gender, by transferring from the source domain (1000 samples for each race) to the target domain (varying samples for each gender as indicated in the caption).



(A) 50 race samples.

(B) 100 race samples.

(C) 500 race samples.

(D) 1000 race samples.

Figure D.8: Comparison of accuracy on COMPAS for Gender → Race, by transferring from the source domain (1000 samples for each gender) to the target domain (varying samples for each race as indicated in the caption).

Appendix E: A Multi-Stage Human-Machine Framework for Mental Health Risk Assessment

There is a growing body of research on using automated classification to identify individuals with mental health issues through social media data. However, little work has been done looking at what it would mean to integrate such systems into a mental health ecosystem where traditionally assessment is a costly process involving clinical interviews, tests, or assessments of behavior. How can one best take advantage of available resources to ensure the largest number of people in need get attention? We introduce a multi-armed bandit method to identify individuals that are most at risk, within a given budget, by combining machine and human effort in a multi-stage framework. We examine our proposed framework in the context of suicide risk in a dataset of Reddit users, demonstrating via simulations that our

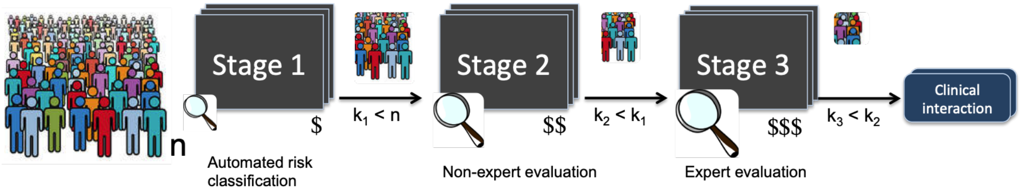


Figure E.1: We apply a multi-armed bandit framework in mental health to identify at-risk individuals, progressing from automated analysis of social media posts, to risk evaluation by non-experts, to expert evaluation. The goal is to optimize the number of people at high risk who go on to receive detailed clinical attention, given limited resources.

model doubles the performance of realistic baselines operating at similar budgets. Our discussion includes key insights, improvements, and ethical implications for real-world deployment.

E.1 Introduction

Machine learning is beginning to have a large impact on the ways that people think about addressing problems in healthcare [159, 235] and mental health [7, 149, *inter alia*], just as it is having large impacts everywhere else. The ability to obtain data about people's day to day thoughts and experiences via social media—unobtrusive windows into what Coppersmith et al. [67] call the "clinical whitespace" between clinician encounters, in the form of social media posts, wearables data, etc.—is looking to be thoroughly disruptive, and the ability to engage with people via natural spoken interactions on all manner of electronic devices creates potential for even more windows into people's everyday thoughts and experiences, enhancing the ability to detect new problems earlier and monitor patients under treatment more effectively and at lower cost.

This is no small matter, because mental illness is one of the most significant problems in healthcare. Considering both direct and indirect costs, mental illness exceeds cardiovascular diseases in the projected 2011-2030 economic toll of non-communicable diseases ($16.3T worldwide) and that total is more than the cost of cancer, chronic respiratory diseases, and diabetes *combined* [37]. Schizophrenia ranks higher in costs than congestive heart failure and stroke [115]. The personal and societal toll is also enormous. In 2016 suicide became the second leading cause of death in the U.S. among those aged 10-34 [108] and is a major contributor to mortality among those with schizophrenia and depression.

It is becoming clear that traditional approaches to these problems do not suffice. Franklin et al. [88], for example, conclude from a large meta-analysis that there has been no improvement in predictive ability for suicidal thoughts and behaviors over the last 50 years, and argue their findings "suggest the need for a shift in focus from *risk factors* to machine learning-based risk *algorithms*" (their emphasis). The technological community is increasingly aware of this problem space and enthusiastic about contributing [e.g. 155, 167, 236], with significant progress in ethical data collection [67, 176] and effective use of those data in predictive models [67, 71, 117, 121, 168].

Moving machine learning out of the lab will raise new challenges, however, because the mental health ecosystem is highly resource-limited. As detection of potential problems gets easier and more widespread, effective and scalable methods will be needed so that cases can be prioritized in terms of the attention needed, and so appropriate interventions can be offered across the entire range of severity.

In this paper we introduce a concrete technological proposal for addressing this problem, involving a basic shift in the way we think about machine learning in mental health: the dominant paradigm of individual-level classification is not an end in itself; rather it provides components in a population-based framework involving both machines and humans, where limited resources give rise to a critical need for effective and appropriate ways to set priorities.

At the core of our technical approach is the recognition that the multi-armed bandit problem in machine learning is a good fit for the real-world scenario created by scaling up the application of technology for detection and monitoring in mental health: what is the best way to allocate limited resources among competing choices, given only limited information? We adopt a *tiered* multi-armed bandit formulation originally introduced with application to hiring or admissions decisions [198], where a succession of stages is applied to a population of applicants, each stage successively more expensive but also more informative, in order to optimize the value of the set of applicants who are chosen. Our key insight is that, by replacing a population of potential hires with a population of people with potential mental health problems, and by replacing "value" with "risk", this tiered framework maps directly to a population-level formulation of the assessment problem. Using real data and human annotation, our simulations demonstrate the value of using this framework to combine (cheap, less accurate) automation with (more expensive, more accurate) human evaluation of social media in order to identify individuals within a population who are at high risk for a suicide attempt.

E.2 Problem Formulation

Let there be a population of individuals where each individual has some potential risk in a given mental health scenario, e.g. veterans at risk for suicide, or college students at risk for onset of schizophrenia. We assume a characterization of risk on a four-point scale (low, no, moderate, or severe). These labels are inherently context based and will depend upon the particular condition, but we will assume that they are derived by clinical experts and agreed upon for the given population (e.g. see [69, 168, 202]).

Given such a population, we take as our goal the identification of as many severe-risk individuals as possible, so they can receive more thorough assessment and appropriate intervention or treatment. In a world of infinite resources, this could be achieved by going straight to regular clinical interaction with every individual in a population. However, that world does not exist, and the mental health ecosystem is dramatically under-resourced; for example, fully a third of the U.S. population live in federally designated mental healthcare provider shortage areas.[1] This makes it essential to to improve our ability to *prioritize* clinicians' time and caseload, but in a way that minimizes the chance of missing at-risk individuals.

One promising direction is in the increasing ability to tap into what may be happening with individuals in an ongoing way via their social media, using machine learning methods for classification. Research into the efficacy of these inferences is ongoing, e.g. [41, 83, 101, 236], and see Section E.7 for discussion of ethical considerations, but such approaches show significant promise. For example, Coppersmith et al. [67] demonstrate an ability to predict suicide attempts based on social media that is much better than typical performance of clinicans based on traditional in-person evaluation, and Milne et al. [168] show that machine risk classification can greatly improve response latency by moderators on a peer-support forum.

At the same time, human review of individuals' social media content is also increasingly taking place, including, for example, by non-clinicians within Facebook's operations [98] and moderators in peer support forums [168], and we have also done initial work looking at the evaluation of social media content by trained personnel [131, 202]. This raises the possibility of exploring intermediate points between

[1] https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas-hpsas

inexpensive fully automated methods and expensive clinical interactions—and, in particular the idea that by combining different forms of evaluation, it may be possible to optimize the combination of machine and human effort in a way that produces the best outcome possible given the resources available.

E.3 Approach

We propose that mental health risk assessment should be viewed as a population-oriented, multi-stage problem, where subsets of individuals (who have opted in appropriately with informed consent) progress from less costly stages (that are also less informative, e.g. automated predictive models), to intermediate stages that require more resources but also provide potentially better information (for example, non-expert human judgments), to more costly forms of assessment, such as evaluation by a trained expert or a qualified clinician. Ultimately the goal is, within given resource limitations, to have as many people as possible who are actually at high risk progress through the entire pipeline to the highly limited and resource-intensive process of traditional, interactive clinical assessment; see Figure E.1 for the full pipeline.

We extended the recent budgeted multi-armed bandit (MAB) framework [198] named BRUTaS to our mental health framework. To briefly summarize the model, we cast tiered decision making as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting [58]. Here, arms represents individuals with latent true risk profiles, where S is the population of arms with $|S| = n$ (e.g., the cohort of all n individuals or the first group of people in Figure E.1). The end goal is to select a subset of $k \leq n$ (the final, and smallest, group of people in

Figure E.1) for clinical interaction, after narrowing the pool over successive stages or tiers. Each arm (or individual) $a \in S$ has an associated unknown true risk $u(a)$, and an empirical risk $\hat{u}(a)$ that the algorithm estimates and uses to make decisions. Each analysis stage i has an associated strength of arm pull defined as information gain s_i—a further generalization of earlier work [197]. The strength correlates with the confidence of the signal generated as well as the cost of performing an arm pull. For example, if we compare the signal generated from an expert reviewer (Stage 3 in Figure E.1) and a non-expert (Stage 2 in Figure E.1), one would be much more confident in the signal from the expert compared to the non-expert. Additionally, each analysis stage i has a cost j_i associated with it. Successive stages increase in both cost j_i and information gain s_i.

In our current model we have three stages of assessment: (1) automated risk classification using an NLP model, (2) non-expert risk assessment, and (3) expert risk assessment.[2] In that 3-stage setting, the goal is to select a final subset of size k out of the full cohort S. After each stage, the pool is narrowed (that is, for some subset of the remaining cohort, intervention decisions are fixed permanently). During stage i, k_i individuals move on to the next stage (i.e., we decide not to pursue a deeper intervention with $k_{i-1}-k_i$ individuals), where $n = k_0 > k_1 > k_2 > k_3 = k$).[3]

[2] Although we approximate an intermediate stage of non-experts using crowdsourced judgments, the idea of true crowdsourcing, in the sense of Mechanical Turk and similar platforms, need not, and should not, be considered a part of the proposal. Rather, we use crowdsourcing to approximate an intermediate level of cost and expertise. Such intermediate levels exist in the real world, e.g. a social work trainee would have less expertise in suicidality assessment than than a trained crisis-line staffer or a specialist clinical psychologist.

[3] Note that although in this paper we focus on the importance of getting as many of the right people as possible through to the end of the pipeline, this multi-stage architecture introduces new possibilities for intermediate outcomes, rather than a choice between a clinical interaction or nothing at all. For example, a low-cost intervention for people who reach Stage 2 or Stage 3 might be to send a caring contact [63] or information about help lines or peer support, and encouragement to reach out. Such interventions and their evaluation are a topic for future work.

Variable	Description
S	The population of individuals to evaluate.
n	The number of individuals in a population ($\|S\| = n$).
a	A single individual or arm ($a \in S$).
$u(a)$	The true risk of an individual.
$\hat{u}(a)$	The empirical risk of an individual.
k	The number of individuals chosen for clinical intervention.
k_i	The number of individuals to move on to stage $i+1$.
s_i	The information gain of evaluation in stage i.
j_i	The cost of evaluation in stage i.
T	The total budget.
T_i	The budget for stage i.

TABLE E.1: List of variables used in our approach.

Therefore, each stage i could be considered a selection problem where k_i individuals need to be selected in order to maximize the total empirical risk of the chosen individuals. More concretely, at each stage i a cohort M_i is chosen where $|M_i| = k_i$ where M_i is chosen as follows: $M_i = \arg\max_M \sum_{a \in M} \hat{u}(a)$. Finally, at each stage i, there is a budget T_i associated with how much information gathering can be performed at that stage, leading to a total budget of $T = \sum_{i=1}^{3} T_i$. Thus, there are a few hyperparameters to tune before running the algorithm: the number individuals to move on to each next stage k_i, budgets for each stage T_i, information gain for each stage s_i, and the cost for each stage j. Table E.1 presents a full list of variables and other symbols used throughout.

We propose that mental health assessment should be viewed as a population-oriented, multi-stage problem, where subsets of individuals progress from less costly stages (that are also less informative, e.g. automated predictive models of the kind emphasized in current mental health machine learning literature), to intermediate stages that require more resources but also provide better information (including traditional methods like requesting self-report scales, as well as new concepts such as automated interviews or clinician review of automated predictions), and ultimately to the most costly forms of assessment, such as in-person evaluation by a qualified clinician. Crucially, this does not obviate the need for individual-level predictive modeling, where significant advances have been achieved over the past several years by us and others [7, 50, 67, 70, 149, 236].[4] Rather, the individual level predictive models are re-cast as crucial components within the multi-stage framework.

We used the recent multi-armed bandit (MAB) framework [198]. To briefly summarize the model, we cast tiered decision making as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting [58]. Here, arms represents individuals with latent true risk profiles. The goal is to select a subset of $k \leq n$ arms S, with $|S| = n$ (e.g., the cohort of all n individuals), after narrowing the pool over successive stages or tiers. Each analysis stage has an associated strength of arm pull—a further generalization of Schumann et al. [197]. The strength determines the confidence of the signal generated (e.g., by the expert reviewer or clinician) as well as the cost of performing an arm pull.

Assume m stages of assessment. Then, in that m-stage setting, the goal is to select a final subset of size k_m of the full cohort S, with $|S| = n$. After each stage, the pool is narrowed (that is, for some subset of the remaining cohort, intervention decisions are fixed permanently). In other words, during each stage k_i individuals move on to the next stage (i.e., we decide not to pursue a deeper intervention with

[4] See Alonso et al. [7], Calvo et al. [50] for broader discussion of computational language analysis for mental health more generally, and Linthicum et al. [149] for a broader review of machine learning in suicide science).

$k_{i-1} - k_i$ individuals), where $n = k_0 > k_1 > \cdots > k_{m-1} > k_m = k$). Therefore, each stage i could be considered a selection problem where k_i individuals need to be selected in order to maximize some objective function. Finally, at each stage, there is a budget associated with how much information gathering can be performed at that stage In our pilot setting, we follow the intuition that individuals can and often are evaluated by an NLP-based system, by a crowd, or by an expert. These different evaluations provide signals about the potential, e.g., suicide risk of an individual, or more generally risk of a negative mental health event.

E.4 Experiments

E.4.1 Data

The intent of the framework is extremely general, and its potential will ultimately need to be evaluated across a wide range of mental health conditions and scenarios. Here we work with the UMD Reddit Suicidality Dataset [202], derived from Reddit, a collection of online communities discussing an enormous range of topics in which participants post anonymously. The dataset includes more than 1.5M posts across Reddit subcommunities, from 11,129 users who posted to the SuicideWatch community and a corresponding set of control users who never posted to SuicideWatch. The dataset includes human assessments of suicide risk on a four-point scale (no, low, moderate, and severe risk) based on SuicideWatch posts for a randomly selected subset of 242 of the users who posted to SuicideWatch. Four experts provided ratings, with good inter-rater reliability (Krippendorff's $\alpha = 0.81$). Crowdsource worker judgments based on SuicideWatch posts for the same 242 individuals, plus an additional 621 individuals, were also obtained, achieving moderate inter-rater reliability (Krippendorff's $\alpha = 0.55$). The end result is a unique dataset that contains data involving people's outreach for help (posts on SuicideWatch), along with high quality expert assessments of risk, moderate quality crowdsourcer assessments, and large-volume weak positive evidence for more than 10K people (by virtue of their having posted to SuicideWatch.

In order to facilitate the comparison of our multi-armed bandit approach to existing baselines, we compute average cost for expert and the crowdsourced reviews. From the UMD Reddit Suicidality Dataset metadata, we computed that the average crowdworker cost \$0.09 per evaluation of an individual. For Stage 3, discussion with experts suggest that an estimated cost of \$5.35 per individual is a reasonable first approximation. (All figures are in USD.) In the absence of a well-founded way to measure information gain at this point, we have assumed that the information gain of each stage is ten times that of the previous, which is within the range of parameters explored in Schumann et al. [198]; further exploration of this parameter is an important subject for future work.

E.4.2 Baselines

Recall, our goal is to identify the at-risk individuals from a population. In our setup, we have a population of 242 individuals where 42 of them are at risk (as defined by having an expert consensus risk label of *severe*). An individual is determined to be at risk by a consensus of four experts. We now outline several baseline approaches, reporting the cost of each approach, the number of individuals it evaluates, and the performance statistics.

Each of these baselines was evaluated on the UMD Reddit Suicidality Dataset with results reported in Table E.3. For those baselines with an element of randomness, for instance, selecting only 100 individuals to evaluate, the simulation of the baseline was performed 10,000 times. The mean and two standard deviations are reported.

E.4.2.1 Expert Baselines

The first set of baselines involve only *experts*. The most naïve approach to evaluate the population would be to have every expert evaluate every individual (**4Experts**). This would be the most expensive with $242 \cdot 4 = 968$ evaluations at a total cost of $968 \cdot \$5.35 = \$5{,}178.8$. However, this will yield the best results. It would have perfect predictive power, by the definition of how we have defined the at-risk individuals.

Another, less expensive option would be to have each individual only be evaluated by one Expert (**1Expert**). For instance, for each individual, randomly sample an expert to perform an evaluation, and use that evaluation as the prediction. This would only take 242 evaluations at a total cost of $1,294.7 and has slightly lower performance than **4Experts**; the population sensitivity of the former is 0.91 compared to 1.0 of the latter. The performance loss captures the noise in the evaluations of the experts. This baseline emulates likely real-world scenarios in which evaluations are distributed across a team of reviewers; it is similar, for example, to what happens to calls when they come in to a crisis line.

Yet a different approach would be to sample a cohort of the population and have experts perform evaluations only on that subset. Say we sample a cohort of 100 individuals and then have either all four experts evaluate each person in the cohort

(**4Experts-Sub**), or, for each individual in the cohort, randomly assign an expert to evaluate them (**1Expert-Sub**). The former baseline does 400 evaluations at a cost of $2,140, and the latter does 100 evaluations at a cost of $535.

E.4.2.2 NLP Baselines

Another set of baseline approaches involve using a classifier based on natural language processing (NLP). With these systems, each evaluation is very inexpensive. We assume that the cost of an evaluation by an algorithm is negligible, even though this is not strictly true. In reality, for a mental health provider, there is likely a cost to integrate and run the technology, which we do not estimate or factor into our analysis. Nevertheless, each individual machine evaluation is certainly very cheap, with sunk costs amortized over time, and so performing an evaluation on the entire population is very feasible. To do this, we have the NLP system evaluate each individual in the population and consider the predicted class (the argmax of the output probability vector) for each individual (**NLP-Full**). For comparison to the last two expert baselines, we also establish **NLP-Sub** which also first randomly selects a cohort and then runs the algorithm only on that cohort. The final pure NLP baseline would be to run the algorithm across all individuals in the population, and then only take the top k most confident severe individuals (**NLP-Top-k**). This particular baseline will always perform worse than **NLP-Full**, but we include it for comparisons.

We employ a state of the art NLP approach: a three layer Hierarchical Attention Network [229, 3HAN]. A hierarchical attention layer is composed of a GRU [15] followed by attention mechanism that learns to pay attention to different parts of the input sequence to derive the output. In 3HAN, it is used to aggregate a sequence of word vectors to a sentence vector, a sequence of sentence vectors to a document vector, and finally a sequence of document vectors to a individual vector for making the prediction. See Figure E.2. We provide additional details for reproducibility in Section E.9.2.

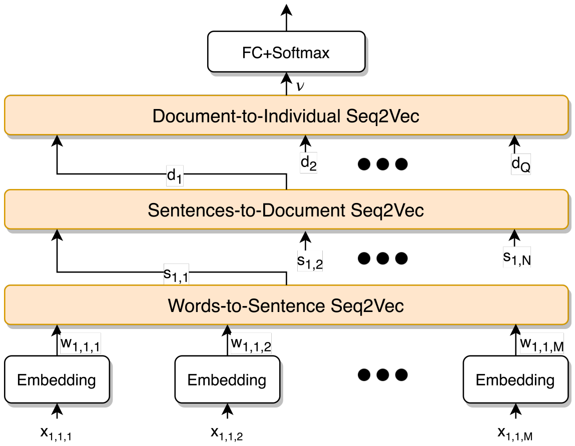


FIGURE E.2: Three-level Hierarchical Attention Network (3HAN)

The NLP system is first pre-trained on the weak supervision signal of individuals posting on the SuicideWatch forum, versus the control group of individuals who never posted on any mental health related forums. The NLP system is then further fine-tuned with the 621 individuals with moderate quality crowdsourcer assessments. No further tuning is done on the high quality expert assessments, and the set of individuals used for training are disjoint from the 242 individuals used for the MAB experiment.

E.4.2.3 Combination Baseline

Finally, we could combine two baselines together, like an NLP with an expert. This combination will have the algorithm evaluate every individual in a population, then take the top k individuals with highest confidence of being most severe, and then give that cohort to experts to evaluate. This aligns with a naive two-tiered system, though not using the multi-armed bandit approach that we propose. The most meaningful combination of these baselines is **NLP-Top-100** + **1Expert-Sub**.

E.4.3 MAB Experiments

For our main experiments, we use the MAB framework discussed in Section E.3 with the UMD Reddit Suicidality Dataset. We translate this data (with subsets of individuals rated by crowdsourcers and clinical experts), and the state of the art NLP classifier, into a three-stage evaluation process, where Stage 1 is an NLP evaluation, Stage 2 is a non-expert evaluation (simulated using the crowdsource labels), and Stage 3 is an expert evaluation.

E.4.3.1 Overall Experiment

The overarching experiment aims to investigate if a three-tiered MAB approach outperforms the most realistic baselines above for given fixed budgets. The most realistic scenarios for clinician screenings are those with a limited budget, such as **1Expert-Sub** and **1Expert**. Therefore, through these experiments, we report overall performance for the best models for budgets of \$553, \$1,300, or \$2,200. The first offers a comparison to **1Expert-Sub** baseline, the middle to **1Expert**, and the last to **4Experts-Sub**. Results are reported in Table E.3.

E.4.3.2 Hyperparameter Experiments

We conduct other experiments that support our overall experiment, like hyperparameter tuning. Recall from Section E.3 that there are many hyperparameters to this model, such as: budget (T) and budget allocation at each stage (T_i), cohort size transferred to each stage (k_1, k_2, k_3), output cohort size (k), and information gain and cost at each stage (s_i, j_i). We set the information gain and costs associated with each successive stage in our model using the calculations described in Section E.4.1. For Stage 1, we assume that review by an NLP system has negligible cost.

To start, we fixed total budget, T, at \$553, \$1,300, or \$2,200. We then can divide that total budget among the different stages, T_1, T_2, and T_3. We can do this division in two main ways: (1) adjusting the cohort sizes $\{k_1, k_2, k_3\}$, or (2) directly changing the number of evaluations at each stage. For (1), we performed a simple grid search over combinations of k_1 and k_2, and k_3 (results visualized in Figure E.3).

For (2), we studied how budget division across the different stages impacts performance. With a fixed T, we could vary the division of that budget to each stage. Recall that we are assuming that the cost for the first stage (NLP) is negligible. Therefore, we can allocate T to the crowdsourcer and expert stages, T_2 and T_3 respectively. Intuitively, we could (1) allocate most of the money to the expert reviews in Stage 3 (**More 3**), (2) allocate most of the money to the crowd reviews in Stage 2 (**More 2**), or (3) equally split it between Stages 2 and 3 (**Equal Split**); we detail overall budget values used in Table E.2, in real USD. Note that at $T = \$553$, there is only enough budget for one pull for every 100 individuals in the final cohort and a few pulls for each crowdsourcer. Therefore, there we have no degrees of freedom to allocate the budget to the stages in these settings.

	$1,300		$2,200	
	Stage 2	Stage 3	Stage 2	Stage 3
More 3	$200	$1,100	$300	$1,900
More 2	$765	$535	$1,500	$700
Equal	$620	$680	$1,100	$1,100

TABLE E.2: Budgets for allocation schemes distributing between Stages 2 and 3 for two budgets: $1,300 and $2,200. Stage 1 has no cost.

We report the results for $T = \$2{,}200$ in Table E.2. We carry out an experiment with $k_1 = 200$, $k_2 = 100$ and $k_3 \in \{1, 2, \ldots, 100\}$ with results reported in Figure E.3. Note that other values for the T produced similar results. This experiment will provide answers to the important questions: *Given a fixed budget, how do we best allocate that budget across the stages?* and *how does that change depending on how many individuals we can serve in our final cohort?*

E.4.3.3 Risk Encoding Experiment

This dataset has four ordinal rating levels: no, low, moderate, and severe risk. Our framework maximizes a numeric objective. Thus, we tried several different encoding schemes for these discrete classes, including: Binary method (**Bin**) where [No, Low, Moderate, Severe] maps to [0,0,0,1]; Linear method (**Lin**) where [No, Low, Moderate, Severe] maps to $[0, \frac{1}{3}, \frac{2}{3}, 1]$; and Exponential method (**Exp**) where [No, Low, Moderate, Severe] maps to $[0, \frac{1}{7}, \frac{3}{7}, 1]$.

E.4.4 Evaluation

Our ultimate metric is population sensitivity of a system. Any mental health evaluation tool will invariably recommend some number of individuals for further clinical attention, which we assume will be some form of clinical interaction. Depending on the form that interaction takes, we should be more tolerant of providing it (if is not prohibitively dangerous, intrusive, or expensive, e.g. an in-office clinical evaluation) with someone who is not at risk (false positives) than for not providing an intervention with someone who is at risk (false negatives). We consider a positive example to be an individual rated as at severe risk, i.e., the highest risk classification in the UMD Suicidality Dataset.

We report average statistics for each model in Table E.3. We report the sensitivity on the population and the cohort level. Since some of these baselines only evaluate a cohort of individuals, all those individuals in the population that were not in the cohort, are treated as negatives. Therefore, we report numbers at both the cohort and population level. To illustrate, in Table E.3, **1Expert-Sub** only evaluates 100 individuals. While the sensitivity on those 100 individuals is high, it also misses many at-risk individuals in the population, which increases the number of false negatives at the population level and decreases the population sensitivity.

E.5 Results

E.5.1 Overall Experiment

Simulating a real-world scenario in which resources are very limited, our MAB approach outperforms all comparable baselines. The most resource constrained approach (**1Expert-Sub**) for the Experts, evaluates only 100 individuals and achieves a population sensitivity of 0.34. For the NLP baselines, the approach that also only evaluates 100 individuals (**NLP-Sub**) achieves 0.27, while the NLP baseline which

Approaches	Budget	Number of Individuals Evaluated	Population Sensitivity	Cohort Sensitivity	Precision	Specificity	Cohort TP	Cohort FP	Cohort FN	Cohort TN
NLP-Full	-	242	0.64	0.64	0.24	0.59	27	82	15	118
NLP-Sub	-	100	0.27 ± 0.11	0.64 ± 0.18	0.25 ± 0.10	0.59 ± 0.08	11 ± 4.8	34 ± 7.3	6 ± 3.1	49 ± 7.7
NLP-Top-100	-	242	0.55	1	0.23	0	23	77	0	0
NLP-Top-100 + 1Expert-Sub	$535	242	0.49 ± 0.06	0.90 ± 0.11	0.71 ± 0.10	0.89 ± 0.05	21 ± 2.7	9 ± 3.9	2 ± 2.7	68 ± 3.9
MAB	$553	242	0.77 ± 0.12	0.77 ± 0.12	0.33 ± 0.05	0.66 ± 0.01	33 ± 5.3	67 ± 5.3	9 ± 5.3	132 ± 5.3
MAB	$1,300	242	0.85 ± 0.08	0.85 ± 0.08	0.36 ± 0.04	0.67 ± 0.04	36 ± 2.0	64 ± 2.0	6 ± 2.0	136 ± 2.0
MAB	$2,200	242	0.84 ± 0.03	0.84 ± 0.03	0.35 ± 0.03	0.67 ± 0.01	36 ± 4.0	64 ± 4.0	6 ± 4.0	136 ± 4.0
1Expert-Sub	$535	100	0.34 ± 0.03	0.91 ± 0.05	0.66 ± 0.14	0.92 ± 0.05	14 ± 2.3	7 ± 4.3	2 ± 2.3	77 ± 4.3
1Expert	$1,295	242	0.91 ± 0.08	0.91 ± 0.08	0.67 ± 0.08	0.91 ± 0.03	38 ± 3.4	18 ± 6.6	4 ± 3.4	182 ± 6.6
4Experts-Sub	$2,140	100	0.43 ± 0.10	1	1	1	18 ± 4	0	0	82 ± 4
4Experts	$5,179	242	1	1	1	1	42	0	0	200

TABLE E.3: Main experimental results. Comparisons based on budget should be made across sections in the table; strongest differences are at the lowest budget. For approaches with an element of randomness, means and two standard deviations are reported.

evaluates every individual (**NLP-Full**) achieves 0.64. (Note, **NLP-Full** outperforms **1Expert-Sub** because the former sees the entire population whereas the latter only sees the cohort subpopulation. The former's cohort sensitivity is lower than the latter's.) Finally, the combination baseline (**NLP-Top-100** + **1Expert-Sub** performs at 0.49. The MAB approach with the same resources of \$553 achieves an average population sensitivity of 0.77.

On average, our MAB approach more than doubles the population sensitivity of the expert baseline for the same resource amount. At \$553, our approach averaged 9 false negatives in the population and 33 true positives. In comparison to the expert baseline with \$535, it achieved an average of 26 false negatives in the population and 14 true positives.

E.5.2 Hyperparameter Experiments

We now present results from our hyperparameter experiments. Recall our first line of inquiry focuses on the hyperparameters k_1, k_2, and k_3. These values indicate the size of the cohort that moves to each successive stage in the MAB framework. We present these results in Figure E.3 for a budget of \$2,200. We use the higher budget in these experiments to draw out the nuances in the grid search over the k_i. In the figure, we report several slices of cube for eight values of k_3, where we plot k_1 on the x-axis and k_2 on the y-axis.

We observe two main points from these data: (1) as k_3 increases, the population sensitivity increases, and (2) higher values of k_1 correlate to poorer population sensitivity. This first result is intuitive since there are only 42 severe risk individuals in the population, the sensitivity will be low with low k_3. More interestingly, this

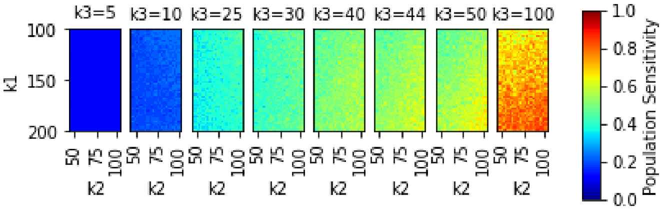


FIGURE E.3: *Grid test over* $\{k_1, k_2, k_3\}$ *for constant budget of $1,300. Y-axis: first-stage cohort* $k_1 \in [100, 200]$*; x-axis: second-stage cohort* $k_2 \in [50, 100]$*; left-to-right: final cohort size* $k_3 \in \{5, 10, 25, 30, 40, 44, 50, 100\}$*. Population sensitivity is reported.*

positive correlation between k_3 and population sensitivity holds for all combinations of k_1 and k_2. This reveals that for any fixed combination of cohort sizes k_1 and k_2, any increase in k_3 will lead to an increase in population sensitivity. Put in another, more policy prescriptive way, we suggest that it is always advantageous to include more individuals in the final output cohort, if budget permits.

Our second claim from this hyperparameter result in Figure E.3 is that higher values of k_1 correlate to poorer population sensitivity. This is qualitatively evident by the figure, and also supported as statistically significant with a simple linear regression between k_1 and population sensitivity with t-value 135.75 and p-value $p \ll 0.001$. What this indicates is that moving a smaller cohort to the crowdsourcers leads to worse population sensitivity. Put another way, the model performs worse when the NLP makes more discriminative decisions about individuals. Therefore, we conclude that while the MAB system benefits from the inclusion of the NLP system, the NLP provides a useful signal of risk, but over-reliance on the NLP system to remove individuals from the pipeline is not advisable. When adjusting hyperparameters, we must balance the power each stage has to remove individuals from the pipeline with the overall predictive power of that stage.

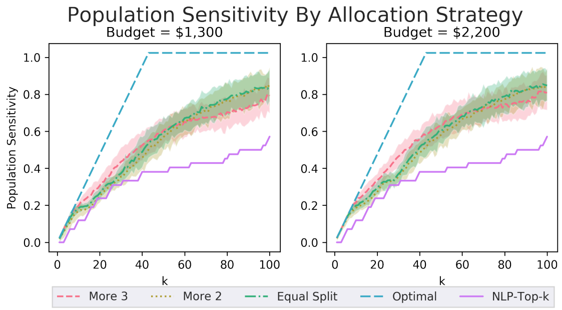


FIGURE E.4: Budget allocation plots for a MAB model with $k_1 = 200$, $k_2 = 100$, and $k_3 \in [1, 100]$. The **More 3**, **More 2**, and **Equal** allocation strategies are described in Section E.4.3.2.

We also conducted a similar analysis for k_2; a simple linear regression between k_2 and population sensitivity with t-value 136.02 and p-value $p \ll 0.001$. We find similar results from this regression analysis which indicate that there is a negative correlation between k_2 and population sensitivity. Again, this implies that with k_1 and k_3 fixed, the performance of the system improves with lower k_2. We can deduce from this analysis that the crowdsourcers provide useful signal to the MAB framework, but are not helpful in removing individuals from the pipeline. We will add that this conclusion is consistent with what clinical practitioners have conveyed to the researchers about crowdsourced evaluations.

Additionally, we analyzed the allocation strategy for a fixed budget among the different stages. For this experiment, we vary k_3 from 1 to 100 and keep k_1 and k_2 fixed at 200 and 100 respectively, as suggested by the results from the grid search. Our initial baseline against which to compare is the "omniscient" method that always returns a size-capped cohort with as many at-risk individuals as possible. Given our evaluation metric, the optimal baseline (**Opt**) is one which achieves the highest possible sensitivity for the dataset. Since there are 42 ground truth severe risk individuals in the dataset, if $k \leq 42$, then the best a model could do would be to choose k severe risk individuals and achieve a sensitivity of $k/42$. For $k > 44$, the best possible would choose 42 severe risk individuals and $k - 42$ others, which would result in a sensitivity of 1. This optimal baseline can also be thought of as only having experts evaluate the entire population, without any cap on budget.

In Figure E.4, we see that there are no significant differences between **More 2** and **Equal** which indicates that allocating more budget to the crowdsourcers does not improve the population sensitivity for various final cohort sizes k. This, again aligns with intuition provided by the clinical practitioners about the crowdsourced evaluations. However, we note that the analysis of strategy **More 3** is more nuanced. For low final cohort values, **More 3** outperforms the other two allocation strategies. This flips for higher k. We also note that when comparing the magnitude of this difference between budgets of $1,300 and $2,200, the magnitude is slightly more pronounced in the former. This suggests that in resource constrained settings, the allocation strategy matters more. Further, the allocation strategy that one would choose for a given scenario would depend upon the final cohort size. For example, if the final cohort is constrained to be 30 individuals, then More 3 outperforms the other two methods. However, this does not hold for larger k_3.

E.5.3 Risk Encoding Experiments

Using the same settings of k_1, k_2, and k_3 as in the budget allocation experiment: we varied the three encoding schemes of the ordinal variables into numeric values: **Linear**, **Binary**, and **Exponential**. Results from this experiment are reported in Figure E.5. We found no significant difference between the encoding schemes.

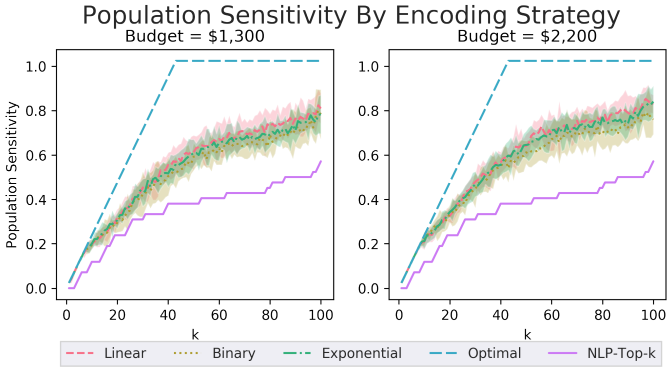


FIGURE E.5: Risk encoding plots of population sensitivity for a MAB model with $k_1 = 200$, $k_2 = 100$, and $k_3 \in [1, 100]$. The **Linear**, **Binary**, **Exponential** encoding methods are described in Section E.4.3.3

There is a rough ordering over the different encodings: Linear performs better than Exponential, which performs better than Binary.

E.6 Related Work

Multi-Armed Bandits. The main model on which our approach relies is derived from Schumann et al. [198]. They introduce the concept of tiers to the extant literature on multi-armed bandits. Bubeck et al. [45] provide an excellent overview on the history of the field. Historically, MAB work has been focused on selecting the best arm from a population, but works recently have moved to selecting the best cohort [46, 58]. There has been extensive research into the objective functions that get used in these models. Lin and Bilmes [148] introduced a monotone submodular function as a method for balancing individual utility and diversity of a set of items; this has been adapted to MAB models [197]. Additional work has been done on optimization algorithms for these types of functions [11, 134]. Ding et al. [79] and

Xia et al. [228] looked at a regret minimization MAB problem in which, when an arm is pulled, a random reward is received and a random cost is taken from the budget. [197] introduced a concept of “weak” and “strong” pulls in the Strong Weak Arm Pull (SWAP) algorithm. Taken together, this body of literature provides the theoretical backbone for the appropriateness and functionality of our approach.

Mental Health Datasets. The data we used lacks ground truth on whether or not the individual attempted suicide. Such information is extremely difficult to obtain, and it is even rarer to see datasets linking clinical and social media data. As a result, most work analyzing social media for mental health relies on non-ground-truth evidence such as online self-report [e.g. 65, 66, 160] or group membership participation and changes [e.g. 76, 77], though see [83] for important limitations of such proxy diagnostic signals. As one notable exception, Coppersmith et al. [67] report strong predictive results using a dataset that contains outcome data on suicide attempts, collected using `ourdatahelps.org`, an innovative platform for consented data donation. As another, Padrez et al. [176] pursued an innovative strategy for obtaining linked social media and clinical data, approaching more than 5,000 people in an urban emergency department to obtain consent. The promising news is that, of the subset who had Facebook or Twitter accounts (about half), nearly 40% were willing to share their social media and EMR data for research purposes.

Prediction of Risk using Machine Learning. Recently there has been a significant uptick in research activity in NLP and machine learning for mental health. A 2019 suicide risk prediction exercise using (an earlier version of) the UMD Reddit Suicidality Dataset took place in which an international set of 15 teams participated [236]; a number of other related shared tasks have also taken place [155, 167].

In real-world settings, automated prediction of mental health crisis has improved speed of response [168] and has been used to trigger interventions that substantially increase the likelihood that a person in acute distress will seek crisis services [121].

E.7 Ethical Considerations

This research underwent appropriate IRB review and its conduct has been informed by the ethical guidelines in Benton et al. [27].

The idea of actually deploying a system of the kind envisioned here raises questions, and potentially obstacles, requiring careful consideration. Even in trying to help a population, one can actually hurt individuals in that population [84]. One set of ethical questions involves the broader socio-technical problem of social media data use in mental healthcare [e.g. 27, 64, 75, 149, 166]. Privacy is of course a central consideration, and taking the wrong approach can undermine the larger goals; for example, well intentioned but insufficiently thought out applications of technology have in some cases already caused backlash [112, 142].

Other questions are more specific to our multi-stage framework. Bias, a general issue in machine learning, may manifest in our scenario when some populations present differently than the majority and could be filtered out too early [171, 199]. Our work also surfaces questions about resource allocation, introducing new degrees of freedom in budget allocation (e.g. Table E.2). Simulations can help evaluate alternatives, but ultimately decisions about technological deployment, staffing, and then then ensuing adjustments in clinical assessment and intervention, will involve considerations well outside the scope of any optimization strategy.

There is also the question of impact on the human beings engaged in the process, for example clinicians assessing social media of people they are not themselves treating, and how that relates to professional codes of ethics [e.g. 8], particularly the duties to warn and to inform [193]. How far do those codes extend in the context of this framework, and where would these responsibilities lie?

Finally, if approaches like ours were to be integrated into the mental health ecosystem, there could be large impacts on the labor and economy of both mental health professionals and non-experts. Even if the ultimate goal is to improve the efficacy and efficiency of the system, the most well-implemented changes can have net negative impacts, and along with the health and well-being of the potentially at-risk population, the well-being of the humans in the loop needs to be considered, as well [55, 116, 163].

E.8 Discussion, Obstacles, & Insights

Current technological research in mental health tends to treat machine-derived and human evaluations very differently. Our simulations support the claim that integrating these separate kinds of evaluation in a process of population based prioritization can dramatically increase the likelihood of an at-risk individual successfully being identified as requiring attention, while keeping resource levels the same. Concretely, we showed that—to the extent our assumptions and abstraction of the problem are reasonable—we can more than double the number of at-risk individuals identified, for the population in our dataset.

Section E.7 noted ethical challenges that need to be considered; let us also consider here the technical challenge of interpretability for the algorithmic elements of our approach.

The MAB framework is interpretable insofar as a decision to omit an individual from a successive stage can be interrogated by examining the individual's human or computer ratings. In that sense, it does not inject any *more* obfuscation of decision making than is already present at each of the tiers. Say, for example, we want to understand what happened for those, on average, 9 individuals in the $553 setting of the MAB framework who were not identified as being at risk. We observe that those individuals that were excluded after the NLP round were more likely to have been rated by the NLP as being of 'No' risk; the average probability of 'No' Risk for those excluded after the first round was 71% versus 8% for those that progressed. This included some individuals that were truly at high risk. This tells us that the MAB framework is (1) behaving like we would expect, and (2) it is only as good as the individual evaluations. In this case, we see that false negatives in the MAB model are a direct result of the NLP false negatives.

This ability to track from the MAB system's decisions to the component evaluations is one of the reasons we selected the hierarchical attention network approach [229], for our classifier: its hierarchical attention mechanism has greater potential for interpretation than many other models.[5] In general, we find it imperative that any evaluator (both human and machine) used in our proposed ecosystem is well trained and able to explain why evaluations were made. We are hopeful that, as we

[5] Although there is controversy over the relationship between attention and interpretability [119], that has generally been in the context of non-hierarchical networks. Our experience, though at this point only anecdotal, is that network attention in the hierarchical setting does tend to highlight evidence that is subjectively relevant. We plan to explore this further in future work.

progress in developing and validating the model, the properties of the MAB setting with regard to interpretability will increase the likelihood that policy makers will want to engage with our proposed solution [233].

These results are only a first step on the way to practical deployment. To get the rest of the way there, further theoretical research and experimentation are required in order to expand the evidence base for this approach. Equally important, for this and any other proposal, careful consideration of the balance between privacy and prevention must continue and, crucially, that conversation needs to integrate the voices of (at least) technologists, in-the-trenches clinicians, policy makers, and those with lived experience of the conditions we are trying to help address.

E.9 Reproducibility

E.9.1 Data

The UMD Reddit Suicidality dataset is available to researchers. Owing to the sensitive nature of the data, even though it is anonymous, access to the dataset is governed by a process developed and run in collaboration with suicide prevention experts at the American Association of Suicidology. See `umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html` for more information. Once a dataset access request has been approved, the data are delivered with two important files: (1) the expert and crowdsourced ratings per user, and (2) the text data per user.

E.9.2 NLP System Training Details

The 3HAN NLP model is built using AllenNLP [93]. Tokenization and sentence splitting are done using spaCy [111].

Training Details. The word embedding layer of 3HAN is initialized and fixed with the 200-dimensional Glove embedding trained on Twitter [181]. 3HAN is then pretrained on the binary Weak Supervision dataset from the weak supervision signal of whether the individuals posted on SuicideWatch. The model is then further fine-tuned on the moderate quality four-class Crowdsource dataset by transferring the weights (except the last fully-connected prediction layer) over. The Crowdsource dataset is split into a training set (80%) and a validation set (20%) during model development. Cross validation on the training set is used for hyperparameter tuning. We did not test on the Expert dataset until all parameters of the models were fixed. For 3HAN, we used ADAM with learning rate 0.003, trained for 100 epochs with early stopping on the validation dataset, with early stopping patience set to 30.

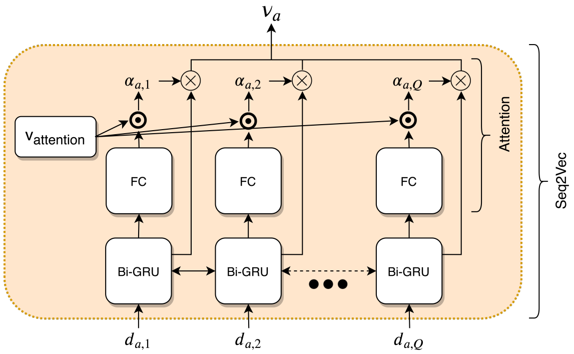


Figure E.6: Seq2Vec with Attention

Seq2Vec layers implementation. 3HAN's Seq2Vec layers use bi-directional GRU with attention [229]. For the purpose of reproducibility, we detail our implementation of the hierarchical attention layer in the context of aggregating a sequence of document vectors to an individual's vector, though the three layers are the same. See Figure E.6 for an illustration. For an individual a, the $|Q|$ Document vectors $\{d_{a,q}\}_{q=1}^{Q}$ representing the $|Q|$ documents of the individual are first passed through a bi-directional GRU layer. The outputs, after passing through a fully-connected layer and a non-linear layer, are then compared to a learnable attention vector, $v_{\text{attention}}$. Specifically,

$$g_{a,q} = \text{Bi-GRU}(d_{a,q}) \tag{E.1}$$

$$r_{a,q} = \tanh\left(W g_{a,q} + b\right) \tag{E.2}$$

$$\alpha_{a,q} = \frac{e^{r^{\top}_{a,q} v_{\text{attention}}}}{\sum_{q'=1}^{Q} e^{r^{\top}_{a,q'} v_{\text{attention}}}} \tag{E.3}$$

$$v_a = \sum\nolimits_{q=1}^{Q} \alpha_{a,q} g_{a,q} \tag{E.4}$$

The word-to-sentence layer has input dimension of 200, hidden dimension of 50, and output dimension of 100, since the bi-direction. The sentence-to-document and document-to-individual layer, similarly, has input dimension of 100, hidden dimension of 50, and output dimension of 100. Hyperparameters were selected using cross validation on the training set split of Crowdsource dataset.

E.9.3 Running the MAB simulation

The code to run the BRUTaS algorithm is written in python and can be found here: https://github.com/principledhiring/TieredHiring.

At the end of the above steps, there should be one file with all the human evaluations in them (‘human.csv’) and one with the machine predictions for each user (‘machine.prediction’). To load these files and create a list of arms, one can run this code:

```
from RSD import load_RSD
arms = load_RSD(human_labels=‘human.csv’,
            NLP_labels=‘machine.prediction’)
```

To run the BRUTaS algorithm, first set up your hyperparameters:

```
import oracles
S = [1,10,100]
J = [1,90,5350]
K = [200,100,50]
T = [2,2,2]
oracle = oracles.c_top_k_oracle
utility = oracles.top_k_utility
```

Then we can instantiate a BRUTaS object and run the algorithm:

```
from brutas import BRUTaS
b = BRUTaS(arms, T, K, S, J, oracle,
      utility, oracle_args=[])
b.run_alg()
```

To evaluate, find those arms that made it to the final stage; to do this, one can execute `b.arm_stage == 3`. This will facilitate the user to compute any statistics required using their favorite packages.

Appendix F: Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics

Bias in machine learning has manifested injustice in several areas, such as medicine, hiring, and criminal justice. In response, computer scientists have developed myriad definitions of *fairness* to correct this bias in fielded algorithms. While some definitions are based on established legal and ethical norms, others are largely mathematical. It is unclear whether the general public agrees with these fairness definitions, and perhaps more importantly, whether they *understand* these definitions. We take initial steps toward bridging this gap between ML researchers and the public, by addressing the question: *does a lay audience understand a basic definition of ML fairness?* We develop a metric to measure comprehension of three such definitions–demographic parity, equal opportunity, and equalized odds. We evaluate this metric using an online survey, and investigate the relationship between comprehension and sentiment, demographics, and the definition itself.

F.1 Introduction

Research into algorithmic fairness has grown in both importance and volume over the past few years, driven in part by the emergence of a grassroots Fairness, Accountability, Transparency, and Ethics (FATE) in Machine Learning (ML) community. Different metrics and approaches to algorithmic fairness have been proposed, many of which are based on prior legal and philosophical concepts, such as disparate impact and disparate treatment [32, 59, 87]. However, definitions of ML fairness do not always fit well within pre-existing legal and moral frameworks. The rapid expansion of this field makes it difficult for professionals to keep up, let alone the general public. Furthermore, misinformation about notions of fairness can have significant legal implications.[1]

Computer scientists have largely focused on developing mathematical notions of fairness, and incorporating them into ML systems. A much smaller collection of studies have measured public perception of bias and (un)fairness in algorithmic decision-making. However, as both the academic community and society in general continue to discuss issues of ML fairness, it remains unclear how to ensure that non-experts can understand various mathematical definitions of fairness sufficiently to provide opinions and critiques.

Our Contributions. We take a step toward addressing this issue by studying peoples' comprehension and perceptions of three definitions of ML fairness: *demographic parity*, *equal opportunity*, and *equalized odds* [103]. Specifically, we address the following research questions:

[1] https://www.cato.org/blog/misleading-veritas-accusation-google-bias-could-result-bad-law

RQ1 When provided with an explanation intended for a non-technical audience, do non-experts comprehend each definition and its implications?

RQ2 Do demographics play a role in comprehension?

RQ3 How are comprehension and sentiment related?

RQ4 How do the different definitions compare in terms of comprehension?

We developed two online surveys to address these research questions. We presented participants with a simplified decision-making scenario and a accompanied *fairness rule* expressed in the scenario's context. We asked questions related to the participants' comprehension of and sentiment toward this rule. Tallying the number of correct responses to the comprehension questions gives us a *comprehension score* for each participant. In Study-1, we found that this comprehension score is a consistent and reliable indicator of understanding demographic parity.

Then, in Study-2, we used a similar approach to compare comprehension among all three definitions of interest. We find that 1) education is a significant predictor of rule understanding, 2) the counterintuitive definition of Equal Opportunity with False Negative Rate was significantly harder to understand than other definitions, and 3) participants with low comprehension scores tended to express less negative sentiment toward the fairness rule.

F.2 Related Work

In response to many instances of bias in fielded artificial intelligence (AI) and machine learning (ML) systems, ML fairness has received significant attention from the computer-science community. Notable examples include gender bias in job-related ads [74], racial bias in evaluating names on resumes [48], and racial bias in predicting criminal recidivism [9].

To correct biased behavior, researchers have proposed several mathematical and algorithmic notions of fairness.

Most algorithmic fairness definitions found in literature are motivated by the philosophical notion of individual fairness (e.g., see [188]), and legal definitions of disparate impact/treatment (e.g., see [19]). Several ML-specific definitions of fairness have been proposed which claim to uphold these philosophical and legal concepts. These definitions of "ML fairness" fall loosely into three categories (for a review, see [60]). *Statistical Parity* posits that in a *fair* outcome, individuals from different protected groups have the same chance of receiving a positive (or negative) outcome. Similarly, *Predictive Parity* [103] asserts that the predictive accuracy should be similar across different protected groups–often measured by the false positive rate (FPR) or false negative rate (FNR) in binary classification settings. Myriad other definitions have been proposed, based on concepts such as calibration [183] and causality [136]. Of course, all of these definitions make limiting assumptions; no concept of fairness is perfect [103]. The question remains, *which* of these fairness definitions are appropriate, and in *what context?* There are two important components to answering this question: *communicating* these fairness definitions to a general audience, and *measuring their perception* of these definitions in context.

Communicating ML-related concepts is an active and growing research area. In particular, *interpretable ML* focuses on communicating the decision-making process and results of ML-based decisions to a general audience [150]. Many tools have been developed to make ML models more interpretable, and many demonstrably improve understanding of ML-based decisions [114, 190]. These models often rely on concepts from probability and statistics–teaching these concepts has long been an active area of research. Batanero et al. [20] provide an overview of teaching probability and how students learn probability; our surveys use their method of communicating probability, which relies on proportions. We draw on several other concepts from this literature for our study design; for example avoiding numerical and statistical representations [94, 95], which can be confusing to a general audience. Instead we provide relatable examples, accompanied by examples and graphics [109].

Effectively communicating ML concepts is necessary to achieve our second goal of understanding peoples' perceptions of these concepts. One particularly active research area focuses on how people perceive bias in algorithmic systems. For example, Woodruff et al. [225] investigated perceptions of algorithmic bias among marginalized populations, using a focus group-style workshop;Grgic-Hlaca et al. [100] studies the underlying factors causing perceptions of bias, highlighting the importance of selecting appropriate features in algorithmic decision-making; Plane et al. [182] look at perceptions of discrimination of online advertising. A related body of work studied how people perceive algorithmic decision-makers. Lee [139] studies perceptions of fairness, trust, and emotional response of algorithmic decision-makers — as compared to human decision-makers. Similar work studies perception of fairness in the context of splitting goods or tasks [140, 141]. Binns [32] studies how different explanation styles impact perceptions of algorithmic decision-makers.

This substantial body of prior research provided inspiration and guidance for our work. Prior work has studied both the effective communication of, and perceptions of, ML-related concepts. We hypothesize that these concepts are in fact related; to that end, we design experiments to simultaneously study peoples' *comprehension* of and *percpetions* of common ML fairness definitions.

F.3 Methods

To study perceptions of ML fairness, we conducted two online surveys where participants were presented with a hypothetical decision-making scenario. The participants were then presented with a "rule" for enforcing fairness. We then asked each participant several questions on their comprehension and perceptions of this fairness rule. We first conducted Study-1 to validate our methodology; we then conducted the larger and broader Study-2 to address our main research questions. Both studies were approved by our organization's standard ethical review process.

F.3.1 Study-1

In Study-1 we tested three different decision-making scenarios based on real-world decision problems: hiring, giving employee awards, and judging a student art project. However, we observed no difference in participant responses between these scenarios; for this reason we discuss only the the *hiring* decision scenario, which was also the subject of Study-2. Please see Section F.7 for a description of these scenarios and survey results. In Study-1, we chose (what we believe is) the simplest definition of ML fairness–demographic parity. In short, this rule requires that the fraction of one group who receives a *positive* outcome (i.e., an award or job offer) is equal for both groups.

F.3.1.1 Survey Design

Here we provide a high-level discussion of the survey design; the full text of each survey can be found in Section F.7. The participant is first presented with a consent form (see Section F.8). If consent is obtained, the participant sees a short paragraph explaining the decision-making scenario. To make demographic parity accessible to a non-technical audience, and to avoid bias related to algorithmic decision-making, we frame this notion of fairness as a *rule* that the decision-maker must follow to be fair. In the hiring scenario, we framed this decision rule as follows: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, the fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.*

We then ask two questions concerning participant evaluation of the scenario, nine comprehension questions about the fairness rule, two self-report questions on participant understanding and use of the rule, and four free response questions on comprehension and sentiment. For example, one comprehension question is: *Is the following statement TRUE OR FALSE: This hiring rule always allows the hiring manager to send offers exclusively to the most qualified applicants.* Finally, we collect demographic information (age, gender, race/ethnicity, education level, and expertise in a number of relevant fields).

We conducted in-person cognitive interviews [104] to pilot our survey, leading to several improvements in the question design. Most notably, because some cognitive interview participants appeared to use their own personal notions of fairness rather than our provided rule, we added questions to assess this compliance issue.

F.3.1.2 Recruitment and Participants

We recruited participants using the online service Cint [61], which allowed us to loosely approximate the 2017 U.S. Census distributions [47] for ethnicity and education level, allowing for broad representation. We required that participants be 18 years of age or older, and fluent in English. Participants were compensated using Cint's rewards system; according to a Cint representative: "[Participants] can choose to receive their rewards in cash sent to their bank accounts (e.g. via PayPal), online shopping opportunities with one of multiple online merchants, or donations to a charity."

In total 147 participants were included in the Study-1 analysis, including 75 men (51.0%), 71 women (48.3%), and 1 (0.7%) preferring not to answer. The average age was 46 years (SD = 16). Ethnicity and educational attainment are summarized in Table F.1. On average, participants completed the survey in 14 minutes.

Table F.1 summarizes the ethnicity and education level of participants in both Study-1 and Study-2.

F.3.1.3 Recruitment and Participants

We again used the Cint service to recruit participants. Because our initial sample (intended to target education, ethnicity, gender and age distributions approximating the U.S. census) skewed more highly educated than we had hoped, we added a second round one week later primarily targeting participants without bachelor's degrees. Hereafter, we report on both samples together.

In total 349 participants were included in the Study-2 analysis, including 142 men (40.7%), 203 women (58.2%), 1 other (0.3%), and 3 (0.9%) preferring not to answer.

	Percent of Sample		
	Census	**Study-1**	**Study-2**
Ethnicity			
AI or AN	0.7	0.7	0.9
Asian or NH or PI	5.7	1.4	2.3
Black or AA	12.3	10.2	15.8
Hispanic or Latinx	18.1	12.2	7.7
Other	2.6	2.7	1.4
White	60.6	72.8	71.9
Education Level			
Less than HS	12.1	6.1	6.9
HS or equivalent	27.7	29.9	24.9
Some post-secondary	30.8	30.6	24.9
Bachelor's and above	29.4	33.3	42.7

Table F.1: Participant demographics across ethnicity and education level, compared to the 2017 U.S. Census. AI = American Indian, AN = Alaska Native, NH = Native Hawaiian, PI = Pacific Islander, AA = African American. Note that in Study-2, two participants did not report their education level.

The average age was 45 years (SD = 15). Ethnicity and educational attainment are summarized in Table F.1. On average, participants completed the survey in 16 minutes.

F.3.2 Data Analysis

Free response questions were qualitatively coded for statistical testing. In Study-1, one question was coded by a single researcher for simple correctness (see Section F.6.2.1), and the other was independently coded by three researchers (resolved to 100%) to capture sentiment information (see Section F.6.2.3). In Study-2, both questions were independently coded by 2-3 researchers (resolved to 100%). Participants who provided nonsensical answers, answers not in English, or other nonresponsive answers to free response questions were excluded from all analysis.

The following methods were used for all statistical analyses unless otherwise specified. Correlations with nonparamentric ordinal data were assessed using Spearman's rho. Omnibus comparisons on nonparametric ordinal data were performed with a Kruskal–Wallis (K-W) test, and relevant post-hoc comparisons with Mann–Whitney U (M-WU) tests. Post-hoc p-values were adjusted for multiple comparisons using Bonferroni correction. χ^2 tests were used for comparisons of nominal data. Boxplots show median and first and third quartiles; whiskers extend to $1.5*$IQR (interquartile range), with outliers indicated by points.

F.3.3 Limitations

As with all surveys, our study has certain limitations. We recruited a demographically broad population, but web panels are generally more tech-savvy than the broader population [189]. We consider this acceptable for a first effort. Some participants may be satisficing rather than answering carefully. We mitigate this by disqualifying participants with off-topic or non-responsive free-text responses. Further, this limitation can be expected to be consistent across conditions, enabling reasonable comparison. Finally, better or clearer explanations of the fairness definitions we explored are certainly possible; we believe our explanations were sufficient to allow us to investigate our research questions, especially because they were designed to be consistent across conditions.

F.4 Results

In this section we first discuss the preliminary findings from Study-1 (see §F.4.1). These findings were used as hypotheses for further exploration and testing in Study-2; we discuss those results second (see §F.4.2).

F.4.1 Study-1

We analyze survey responses for Study-1 and make several observations. We first validate our comprehension score as a measure of participant understanding; we then generate hypotheses for further exploration in Study-2.

F.4.1.1 Our Survey Effectively Captures Rule Comprehension

We find that we can measure comprehension of the fairness rule. The comprehension score was calculated as the total correct responses out of a possible 9. All questions were weighted equally. The relevant questions included 2 multiple choice, 4 true/false, and 3 yes/no questions. The average score was 6.2 (SD=2.3).

We validate our comprehension score using two methods: internal validity testing, and correlation against two self-report and one free response question included in our survey (see Section F.6.2.1 for further details).

Internal Validity Cronbach's α and item-total correlation were used to assess internal validity of the comprehension score. Both measures met established thresholds [85, 173]: Cronbach's $\alpha = 0.71$, and item-total correlation for 8 of the 9 items (all but Q5) > 0.3.

Question Correlation We find that self-reported rule understanding and use are reflected in comprehension score. First, we compared comprehension score to self-reported rule understanding (Q13): “I am confident I know how to apply the award rule described above,” rated on a five-point Likert scale from strongly agree (1) to strongly disagree (5). The median response was “agree” (Q1 = 1, Q3 = 3). Higher comprehension scores tended to be associated with greater confidence in understanding (Spearman’s $\rho = 0.39$, $p < 0.001$), supporting the notion that comprehension score is a valid measure of rule comprehension.

Next, we compared comprehension score to a self-report question about the participant’s use of the rule (Q14), with the following options: (a) “I applied the provided award rule only,” (b) “ “I used my own ideas of what the correct award decision should be rather than the provided award rule,” or (c) “I used a combination of the provided award rule and my own ideas of what the correct award decision should be.” We find that participants who claimed to use only the rule scored significantly higher (mean 7.09) than those who used their own notions (4.68) or a combination (4.90) (post-hoc M-WU, $p < 0.001$ for both tests; corrected $\alpha = 0.05/3 = 0.017$). This further corroborates our comprehension score.

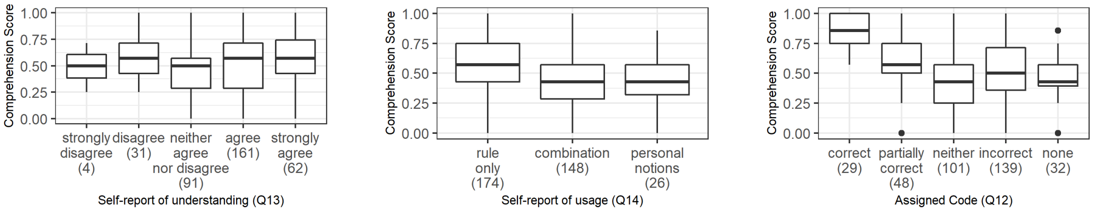
Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as one of five categories – **Correct**: describes rule correctly; **Partially correct**: description has some errors or is somewhat vague; **Neither**: vague description of purpose of the rule rather than how it works, or pure opinion; **Incorrect**: incorrect or irrelevant; and **None**: no answer, or expresses confusion. Participants whose responses were either correct (mean comprehension score = 7.71) or partially correct (7.03) performed significantly better on our survey than those responding with neither (5.13) or incorrect (4.24) (post-hoc M-WU, $p < 0.001$ for these four comparisons, corrected $\alpha = 0.005$). These findings further validate our comprehension score. Additional details of these results and the associated statistical tests can be found in Section F.6.2.1.

F.4.1.2 Hypotheses Generated

We analyzed the data from Study-1 in an exploratory fashion intended to generate hypotheses that could be tested in Study-2. We highlight here three key hypotheses that emerged from the data.

Education Influences Comprehension We used poisson regression models to explore whether various demographic factors were associated with differences in comprehension. We found that a model including education as a regressor had greater explanatory power than a model without (see Section F.6.2.2 for further details).

Disagreement with the Rule is Associated with Higher Comprehension Scores We asked participants for their opinion on the presented rule in a free response question (Q15). These responses were qualitatively coded to capture participant sentiment toward the rule in one of five categories – **Agree**: generally positive sentiment towards rule; **Depends**: describes both pros and cons of the given rule; **Disagree**: generally negative sentiment towards rule; **Not understood**: expresses confusion about rule; **None**: no answer, or lacks opinion on appropriateness of the rule. Participants who expressed disagreement with the rule performed better (mean comprehension score = 7.02) than those who expressed agreement (5.50), did not understand the rule (4.44), or provided no response (5.09) to the question (post-hoc M-WU, $p < 0.005$ for these three comparisons; corrected $\alpha = 0.05/10 = 0.005$). Section F.6.2.3 provides further details.



(A) Grouped by response to Q13

(B) Grouped by response to Q14.

(C) Grouped by coded response to Q12.

FIGURE F.1: Comprehension scores grouped by questions. In (a), self-reported understanding of the rule was not related to comprehension score. X-axis is reversed for figure and correlation test. In (b), rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from this figure and the relevant analysis. Finally, in (c), participants who provided either correct or partially correct responses tended to perform better.

Non-Compliance is Associated with Lack of Understanding We were interested in understanding why some participants failed to adhere to the rule, as measured by their self-report of rule usage in Q14. We labeled those who responded with either having used their own personal notions of fairness ($n = 29$) or some combination of their personal notions and the rule ($n = 28$) as “non-compliant" (NC), with the remaining $n = 89$ labeled as “compliant" (C). One participant who did not provide a response was excluded from this analysis, conducted using χ^2 tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (see Fig. F.12). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (see Fig. F.13).

This fits with the overall strong relationship we observed among comprehension scores, self-reported understanding, ability to explain the rule, and compliance.

Further, negative participant sentiment towards the rule (Q15) also appears to be associated with greater compliance (see Fig. F.14). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it. Refer to Section F.6.2.3 for further details.

F.4.2 Study-2

We first confirm the validity of our comprehension score, then compare comprehension across definitions and examine the hypotheses generated in Study-1.

F.4.2.1 Score Validation

We validated our metric using the same approach used in Study-1, i.e., assessing both internal validity and correlation with self-report and free-response questions. We report the results of this assessment here.

Internal Validity We again used Cronbach's α and item-total correlation to assess internal validity of the comprehension score. An initial assessment using all 349 responses yielded Cronbach's $\alpha = 0.38$, and item-total correlation > 0.3 for only four of the nine comprehension questions. Since both measures performed below established thresholds [85, 173], we investigated further and repeated these measurements individually for each fairness-definition condition (DP, FNR, FPR, EO). This procedure showed stark differences in Cronbach's α based on definition: DP $= 0.64$, FNR $= 0.39$, FPR $= 0.49$, EO $= 0.62$. Item-total correlations followed a similar pattern: best in DP, worst in FNR. Based on these differences, we iteratively removed problematic questions from the score on a per-definition basis until all remaining questions achieved an item-total correlation of > 0.3 [85]. By removing poorly performing questions, we increase our confidence that the measured comprehension scores are meaningful for further analysis. Table F.2 specifies which questions were retained for analysis in each definition.

	Questions								
	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
DP	X	X			X	X	X	X	X
FNR	X	X	X		X				
FPR	X	X	X	X		X		X	X
EO	X		X		X	X	X	X	X

TABLE F.2: Questions that were used for downstream analysis after iterative removal of questions with poor item-total correlation.

Because questions were dropped on a per-definition basis, the range of the resulting scores varied from 4-7 depending on the definition, rather than being a uniform 9. We normalized this treating comprehension score as a percentage of the maximum for each condition rather than a raw score. We report this *adjusted score* in the remainder of §F.4.2.

Question Correlation As in Study-1, we compare comprehension scores with responses to self-report and free response questions included in our survey.

First, we compared comprehension score to self-reported rule understanding (Q13), as described in §F.4.1.1. The median response was "agree" (Q1 = 2, Q3 = 3). We assess the correlation between these responses and comprehension score using

Spearman's rho (appropriate for ordinal data). Unlike in Study-1, there was no relationship between self-reported understanding and comprehension score (Fig. F.1a).

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14), as described in §F.4.1.1. A K-W test revealed a relationship between self-reported rule usage and comprehension score ($p < 0.001$). We find that participants who claimed to use only the rule tended to score higher (mean comprehension score = 0.60) than those who used their own notions (0.47) or a combination (0.45) thereof (post-hoc M-WU, $p < 0.01$ for both tests; corrected $\alpha = 0.05/3 = 0.017$). This suggests that participants are answering at least somewhat honestly: when they try to apply the rule, comprehension scores improve (see Fig. F.1b).

Finally, we asked participants to explain the rule in their own words (Q12). Each response was then qualitatively coded as one of five categories, as described in SF.4.1.1. These results can be seen in Fig. F.1c. A K-W test revealed a relationship between comprehension score and coded responses to Q12 ($p < 0.001$). Correct (mean comprehension score = 0.86) responses were associated with higher comprehension scores than partially correct (0.60), neither (0.44), incorrect (0.52), and none (0.46) responses ($p < 0.001$ for all); partially correct responses were also associated with higher comprehension scores than neither and none responses ($p < 0.001$ for both). No other differences were found (post-hoc M-WU; corrected $\alpha = 0.05/10 = 0.005$). These findings support our claim that our comprehension score is a valid measure of fairness-rule comprehension.

Covariate	**Est.**	**95% CI**	p
Education			
HS	0.02	[-0.08, 0.12]	0.720
Post-secondary, no BS	0.09	[-0.01, 0.19]	0.091
Bachelor's and above	0.17	[0.08, 0.27]	< 0.001
Definition			
EO	-0.04	[-0.11, 0.02]	0.218
FPR	-0.05	[-0.11, 0.02]	0.138
FNR	-0.14	[-0.20, -0.07]	< 0.001

Table F.3: Regression table for the best fit model, with two covariates: education (baseline: no HS) and definition (baseline: DP). Est. = estimate, CI = confidence interval.

F.4.2.2 Education and Definition are Related to Comprehension Score

One hypothesis generated by Study-1 was that comprehension score is positively correlated with education level. We investigated this hypothesis using linear regression models.

Eleven models were tested, regressing different combinations of demographics (ethnicity, gender, education, and age) and condition (fairness definition). Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection [6]. Comparison by AIC revealed that the model using just education (edu) and fairness definition (def) as regressors was the model of best fit. In this model, having a Bachelor's degree or above resulted in a score increase of 0.17, and the FNR condition caused a score decrease of -0.14 ($p < 0.001$ for both; corrected $\alpha = 0.05/11 = 0.0045$). A regression table of the best fit model is below, in Table F.3.

AIC results of each of the eleven models, along with the relevant regressors, can be seen in Table F.4 in Section F.6.3.1. Comprehension score as a function of education and fairness definition can be seen in Figs. F.2 and F.3.

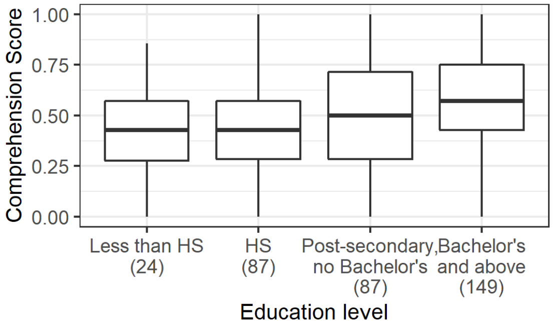


FIGURE F.2: Comprehension score grouped by education level. Higher education was associated with higher comprehension scores. Note that two participants who did not report their education level were removed from this figure and the relevant analysis.

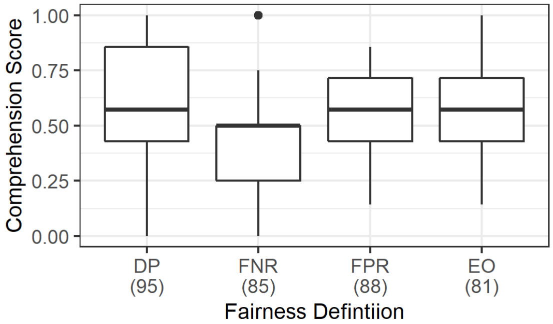


FIGURE F.3: Comprehension score grouped by fairness definition. The FNR condition was associated with lower comprehension sore.

F.4.2.3 Greater Negative Sentiment Toward the Rule is Associated with Higher Comprehension Scores

In Study-1, we found a relationship between participant sentiment towards the rule and comprehension score. To better interrogate this phenomenon, in Study-2 we added two more questions to the survey to directly address the issue of sentiment, rather than relying on a free response question. One (Q15) asks, “To what extent do you agree with the following statement: I like the hiring rule?", and is evaluated on a five-point Likert scale from “strongly agree" (1) to “strongly disagree" (5). The other (Q16) asks, “To what extent do you agree with the following statement: I

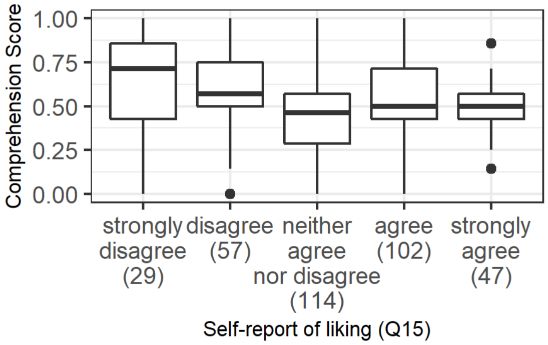


FIGURE F.4: Comprehension score grouped by response to Q15. Disliked of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

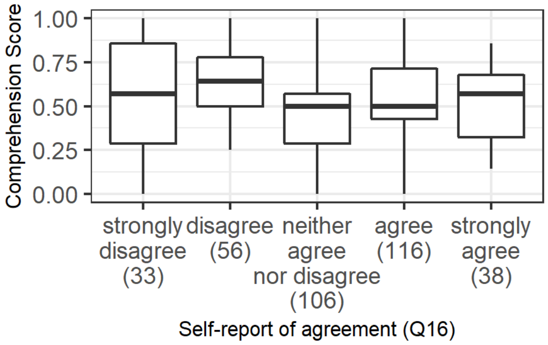


FIGURE F.5: Comprehension score grouped by response to Q16. Rule agreement was not correlated with comprehension score. X-axis is reversed for figure and correlation test.

agree with the hiring rule?", and is also evaluated on a five-point Likert scale from “strongly agree" (1) to “strongly disagree" (5).

Using Spearman’s rho, we assessed the correlation between responses to these two questions and comprehension score. A minor correlation was found between liking the rule and comprehension score, in that those who disliked the rule were more likely to have higher comprehension scores ($\rho = -0.15, p < 0.01$; see Fig. F.4). No correlation was found between agreeing with the rule and comprehension score (see Fig. F.5).

F.4.2.4 Non-Compliance is Associated with Lack of Understanding

A final hypothesis generated in Study-1 involves non-compliance: i.e., why do participants who report *not* using the rule to answer the comprehension questions behave this way? In Study-1, we found that this was due to the fact that non-compliant participants were less able to *understand* the rule, rather than because they did not *like* it. We also observed this in our results form Study-2: compliant participants exhibited higher self-reported understanding of the rule ($p < 0.001$, Fig. F.16), were more likely to correctly explain the rule ($p < 0.001$,Fig. F.17), and were significantly more likely to dislike the rule ($p < 0.05$, Fig. F.18). Refer to Section F.6.4 for more details. As with comprehension score, we observed no relationship between compliance and agreement with the rule (Fig. F.19).

F.5 Discussion

Bias in machine learning is a growing threat to justice; to date, ML bias has been documented in both commercial and government applications, in sectors such as medicine, criminal justice, and employment. In response, ML researchers have proposed various notions of *fairness* to correct these biases. Most ML fairness definitions are purely mathematical, and require some knowledge of machine learning. While they are intended to benefit the general public, it is unclear whether the general public agrees with — or even understands — these notions of ML fairness.

We take an initial step to bridge this gap by asking *do people understand the notions of fairness put forth by ML researchers?* To answer this question we develop a short questionnaire to assess understanding of three particular notions of ML fairness (demographic parity, equal opportunity, and equalized odds). We find that our comprehension score (with some adjustments for each definition) appears to be a consistent and reliable indicator of understanding the fairness metrics. The comprehension score demonstrated in this work lays a foundation for many future studies exploring other fairness definitions.

We do find, however, that comprehension is lower for equal opportunity, false negative rate than other definitions. In general, comprehension scores for equal opportunity (both FNR and FPR) were less internally consistent than other fairness rules, suggesting participant responses were also more "noisy" for equal opportunity. This is somewhat intuitive: equal opportunity is difficult to understand, as it only involves one type of error (FNR or FPR) rather than both. Furthermore, FNR participants had the lowest comprehension scores *and* the lowest consistency of all conditions. We believe this finding also matches intuition: FNR is a strange notion in the context of hiring, as it concerns only those who were *not* hired or offered jobs. Indeed, in free-response questions several participants mentioned that they do not understand why qualified candidates are *not* hired. We believe many participants fixated on this strange setting, impacting their comprehension scores. This finding is potentially problematic, as equal opportunity definitions are increasingly used in practice. Indeed, major fairness tools such as Google What-If tool [222] and the IBM AI Fairness 360 [22] specifically focus on equal opportunity. Further work should be put into making descriptions of nuanced fairness metrics more accessible.

Our analysis also identified other issues that should be considered when thinking about mathematical notions of fairness. First, we find that education is a strong predictor of comprehension. This is especially troubling, as the negative impacts of biased ML are expected to disproportionately impact the most marginalized [18] and displace employment opportunities for those with the least education [90]. Lack of understanding may hamper these groups' ability to effectively advocate for themselves. Designing more accessible explanations of fairness should be a top research priority.

Second, we find that those with the weakest comprehension of fairness metrics also express the least negative sentiment toward them. When fairness is a concern, there are always trade-offs — between accuracy and equity, or between different stakeholders, and so on. Balancing these trade-offs is an uncomfortable dilemma often lacking an objectively correct solution. It is possible that those who comprehend this dilemma *also* recognize the precarious trade-off struck by any mathematical definition of fairness, and are therefore dissatisfied with it. From another perspective, this finding is more insidious. If those with the weakest understanding of AI bias are also least likely to protest, then major problems in algorithmic fairness may remain uncorrected.

F.6 Extra Information

F.6.1 Methods

F.6.1.1 Cognitive Interviews

We recruited 9 participants from a large metropolitan area using Craigslist. We required participants to be over 18 years of age and fluent in English. Participants ranged between the ages of 20 and 66. These interviews took place on our organization's campus and lasted about 1 hour. All participants signed a written consent form prior to the interview, and were paid $30 for their time.

During these interviews, participants completed a preliminary version of the survey used in Study-1. After each survey question, we asked the participants several interview questions related to their comprehension of and feelings toward the survey. We found that some participants tended to use their own personal notions of fairness when answering comprehension questions rather than using the definition we provided. We were concerned that this would limit our ability to effectively measure comprehension. To address this problem, we rewrote several parts of our survey and added two new questions (Q14 and Q15).

F.6.1.2 Non-Expert Verification

We designed this study to assess *non-expert* understanding and opinions of ML fairness metrics. To this end, we asked respondents to self-rate their level of expertise in a variety of fields, including ML, at the end of the survey (see Section F.7.3). A number of participants did report having "expert" level experience in ML ($n = 2$ out of 147 in Study-1, and $n = 15$ out of 349 in Study-2). We considered removing these participants from the analyses, but ultimately did not because there was no relationship between self-reported ML expertise and comprehension score (Spearman's rho, for both studies).

F.6.2 Study-1: Detailed Results

F.6.2.1 Our Survey Effectively Captures Rule Comprehension

We find that our survey is internally consistent, and effectively measures participant comprehension of demographic parity. The former we evaluated using Cronbach's α and item-total correlation (discussed in §F.4.1.1), and the latter using two self-report measures and one free response question.

See Fig. F.6 for participant performance per question.

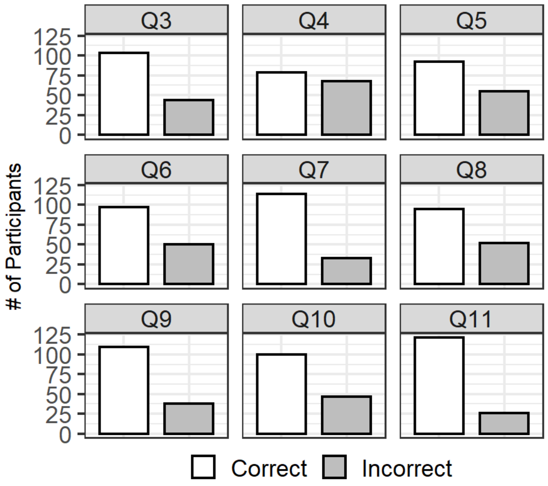


FIGURE F.6: Number of participants answering each question correctly. Each panel contains all 147 participants.

Self-reported rule understanding and use are reflected in comprehension score First, we compared comprehension score to self-reported rule understanding (Q13). Higher comprehension scores were associated with greater confidence in understanding (Spearman's rho), suggesting that participants were accurately assessing their ability to apply the rule (see Fig. F.7).

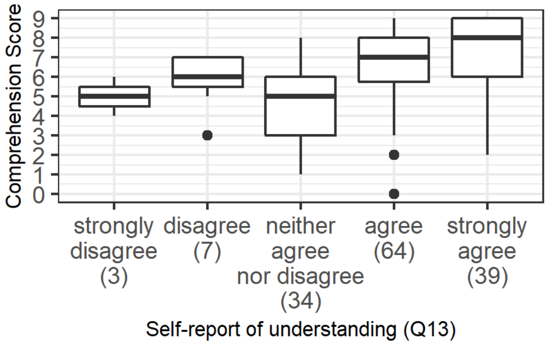


Figure F.7: Comprehension score grouped by response to Q13. Self-reported understanding of the rule was associated with higher comprehension scores. X-axis is reversed for figure and correlation test.

Next, we compared comprehension score to a self-report question about the participant's use of the rule (Q14) Participants who claimed to use only the rule tended to score higher than those who used their own notions of fairness or a combination thereof (K-W test, and post-hoc M-WU), suggesting that participants are answering somewhat honestly: when they try to apply the rule, comprehension scores improve (see Fig. F.8).

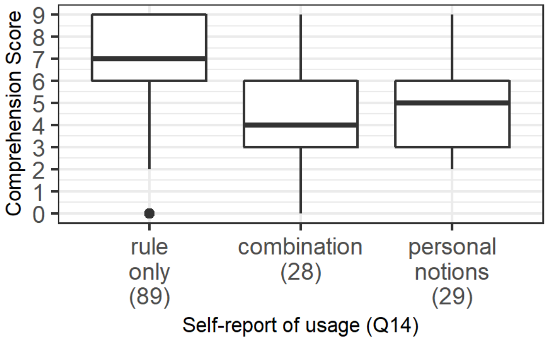


Figure F.8: Comprehension score grouped by response to Q14. Rule compliance (leftmost on the x-axis) was associated with higher comprehension scores. One participant who did not provide a response was excluded from the figure and relevant analysis.

Participants with higher comprehension scores are better able to explain the rule To further validate our comprehension score, we asked participants to explain the rule in their own words (Q12). Responses were qualitatively coded as one of five categories: **correct**, **partially correct**, **neither**, **incorrect**, or **none** (as discussed in §F.4.1.1). The results of this coding can be seen can be seen in Fig. F.9. Participants providing correct explanations of the rule attained higher comprehension scores (k-W test, and post-hoc M-WU), further corroborating our claim that our comprehension score is a valid measure of fairness rule comprehension.

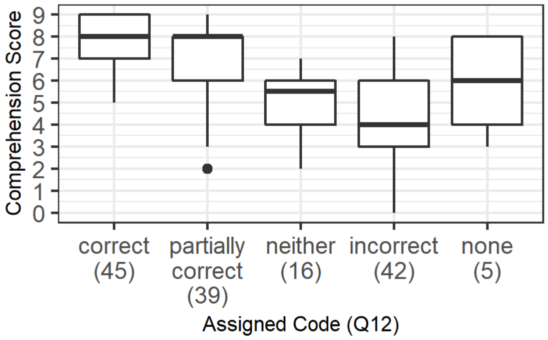


FIGURE F.9: Comprehension score grouped by code assigned to Q12 response. Participants who provided either correct or partially correct responses tended to perform better.

F.6.2.2 Education Influences Comprehension

During the cognitive interview phase, we observed a possible trend of comprehension scores being lower for older participants and those with less educational attainment. If true, this would suggest that fairness explanations should be carefully validated to ensure they can be used with diverse populations. We investigated this hypothesis, in an exploratory fashion, using poisson regression models.

Three models were tested. The first regressed score against all four demographic categories as predictors (gender, age, ethnicity, and education), the second omitted education, and the third tested only education. Models were compared using Akaike information criterion (AIC), a standard method of evaluating model quality and performing model selection [6]. Comparison by AIC revealed that model 1 (all four categories) was a better predictor for comprehension score than models 2 or 3 (AIC = 643.3, 651.2, and 660.5, respectively; difference = 0.0, 7.9, and 17.1). In model 1, only education showed correlation with comprehension score (effect size = 1.40, $p < 0.05$). Further work is needed to confirm this exploratory result.

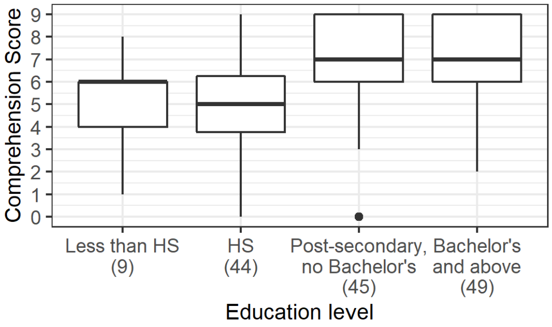


FIGURE F.10: Comprehension score grouped by education level. Higher education level was associated with higher comprehension scores.

F.6.2.3 Disagreement with the Rule is Associated with Higher Comprehension Scores

Participants were asked for their opinion on the presented rule in another free response question (Q15). These responses were then qualitatively coded to capture participant sentiment towards the rule as one of five categories: **agree**, **depends**, **disagree**, **not understood**, or **none** (as discussed in §F.4.1.2).

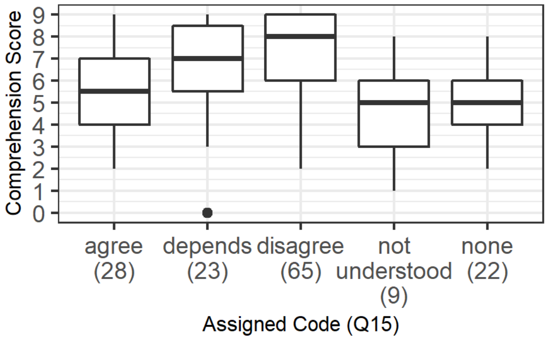


FIGURE F.11: Comprehension score grouped by code assigned to Q15 response. Participants who exhibited negative sentiment toward the rule responses tended to perform better.

This question was added based on the cognitive interviews (see Section F.6.1.1), where perception seemed to influence compliance. The results of coding Q15 can be seen in Fig. F.11. Participants who expressed disagreement with the rule performed better than those who expressed agreement, did not understand the rule, or provided no response to the question (K-W test, post-hoc M-WU). Note that this result should not be interpreted as an overall finding on the appropriateness of demographic parity. Instead we anticipate the perceptions of appropriateness of any fairness definition will be highly context-dependent.

Non-Compliance is Associated with Lack of Understanding We were interested in understanding why some participants failed to adhere to the rule, as measured by their self-report of rule usage in Q14. After labeling participants as either “non-compliant" (NC, $n = 57$) or “compliant" (C, $n = 89$), we conducted a series of χ^2 tests to investigate this phenomenon.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (see Fig. F.12). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (see Fig. F.13). Further, negative participant sentiment towards the rule (Q15) also appears to be associated with greater compliance (see Fig. F.14). Thus, non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it.

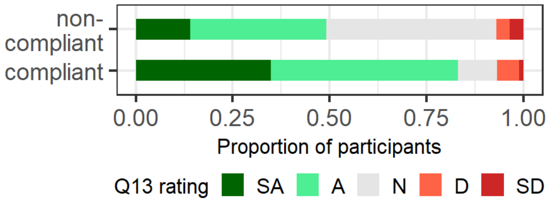


FIGURE F.12: Self-report of understanding (Q13) split by compliance (Q14). NC participants tend to report less confidence in their ability to apply the rule. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

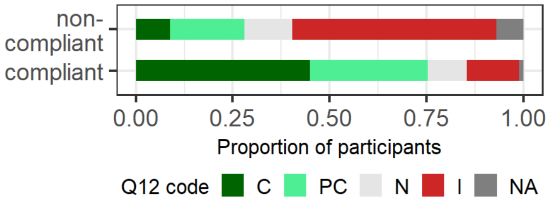


FIGURE F.13: Correctness of rule explanation (Q12) split by compliance (Q14). NC participants tend to be less able to explain the presented rule in their own words. C = correct, PC = partially correct, N = neither, I = incorrect, NA = none.

F.6.2.4 Decision Scenarios

For Study-1 we designed three decision-making scenarios to test whether the perceived importance or realism of a particular scenario influenced comprehension score. They are as follows:

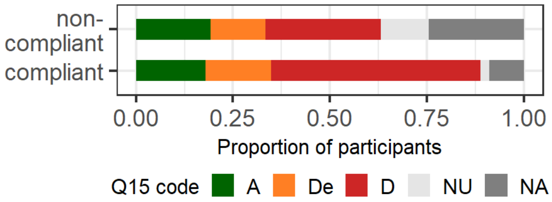


FIGURE F.14: Participant agreement with rule (Q15) split by compliance (Q14). NC participants tend to harbor less negative sentiment towards the rule. A = agree, De = depends, D = disagree, NU = not understood, NA = none.

- **Art Project (AP):** distributing awards for art projects to primary school students,
- **Employee Awards (EA):** distributing employee awards at a sales company, and
- **Hiring (HR):** distributing job offers to applicants.

In each scenario the students/employees/applicants are partitioned into two groups (parents' occupation for the first scenario, and binary gender for the other two scenarios). We use a between-subjects design: participants are randomly partitioned into three conditions, one for each scenario (AP, EA, or HR). For each condition we define the *fairness rule* in the context of the decision-making scenario (see Section F.7 for the full surveys).

Next we describe our main conclusion related to the different decision-making scenarios in Study-1: the scenario does not influence comprehension score.

Scenario does not Influence Comprehension Scores (RQ4) We were concerned that less important and/or realistic scenarios would cause participants to take the survey less seriously, and therefore perform more poorly. To test this, participants were randomly assigned to a scenario, resulting in the following distribution: AP = 41, EA = 49, HR = 57.

A K-W test revealed no differences between scenarios in terms of comprehension score (mean comprehension scores: AP = 6.0, EA = 6.74, HR = 5.86). However, differences did exist between scenarios in terms of importance (assessed in Q2), measured in hours of effort deemed necessary to make the relevant decision (K-W, $p < 0.001$). Post-hoc M-WU revealed that participants believed making a decision in the AP scenario merited fewer hours of effort (mean = 3.15hrs) than in the EA (13.52hrs, $p < 0.001$) or HR (15.23hrs, $p < 0.001$) scenarios (corrected $\alpha = 0.05/3 = 0.017$). See Fig. F.15 for distributions of responses.

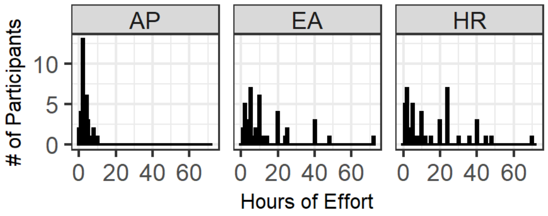


Figure F.15: Importance of a scenario by proxy of hours of effort necessary to make a decision in each scenario. AP merited less hours of effort than both EA and HR.

Of note, it is possible that perceived realism, assessed in Q1 on a five-point Likert scale, was also influenced by scenario (K-W, $p = 0.051$), but we may need larger sample sizes to confirm this. Regardless, while the nature of a scenario does influence participant perception in terms of importance and (possibly) realism, it does not appear to influence comprehension (at least for the scenarios we chose). For this reason, we chose to test a single scenario (HR) in Study-2.

F.6.3 Study-2: Detailed Results

F.6.3.1 Model Selection

In §F.4.2.2 we assessed eleven linear regression models for predicting comprehension scores. The best fit model, determined by model selection via AIC, included only education (edu) and fairness definition (def) as regressors. The results of model selection are below in Table F.4.

Model regressors	**AIC**	**dAIC**
edu + def	-51.0	0.0
edu	-39.1	12.0
gender + edu	-36.2	14.9
gender + age + eth + edu + def	-33.8	17.2
age + edu	-30.5	20.5
gender + age + edu	-27.6	23.4
def	-25.7	25.4
gender + age + eth + edu	-23.8	27.3
gender + age + def	-11.1	39.9
gender + age + eth + def	-8.4	42.6
gender + age + eth	1.1	52.1

Table F.4: Models tested in §F.4.2.2, sorted by best to least fit. The first model in the table (edu + def) is the model of best fit. dAIC = difference from model with lowest AIC value.

F.6.4 Non-Compliance

In §F.4.2.4 we sought to further investigate the findings of Study-1 with regards to compliance (Q14). To do so, we labeled those who responded (in Study-2) with either having used their own personal notions of fairness ($n = 26$) or some combination of their personal notions and the rule ($n = 148$) as "non-compliant" (NC), with the remaining $n = 174$ labeled as "compliant" (C). One participant who did not provide a response was excluded from this analysis, conducted using KW and χ^2 tests.

Non-compliant participants were less likely to self-report high understanding of the rule in Q13 (KW test, $p < 0.001$, see Fig. F.16). Moreover, non-compliance also appears to be associated with a reduced ability to correctly explain the rule in Q12 (χ^2 test, $p < 0.001$, see Fig. F.17). This fits with the overall strong relationship we observed among comprehension scores, ability to explain the rule, and compliance.

Further, greater dislike towards the rule (Q15) also appears to be associated with greater compliance (KW test, $p < 0.05$, see Fig. F.18). However, there was no relationship between disagreement towards the rule (Q16) and compliance (see Fig. F.19).

These results largely corroborate the notion that non-compliant participants appear to behave this way because they do not *understand* the rule, rather than because they do not *like* it.

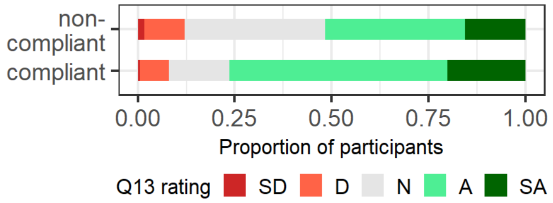


FIGURE F.16: Self-report of understanding (Q13) split by compliance (Q14). NC participants tend to report less confidence in their ability to apply the rule. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

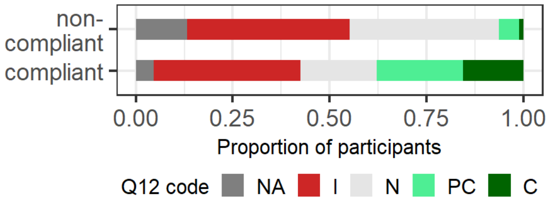


FIGURE F.17: Correctness of rule explanation (Q12) split by compliance (Q14). NC participants tend to be less able to explain the presented rule in their own words. C = correct, PC = partially correct, N = neither, I = incorrect, NA = none.

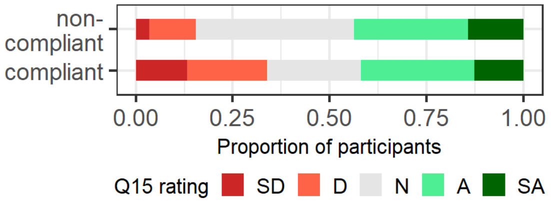


FIGURE F.18: Participant liking for rule (Q15) split by compliance (Q14). NC participants tend to dislike the rule less than C participants. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

F.7 Surveys

F.7.1 Study-1 Survey

Each of the surveys are split into four main sections. The first section is the consent form which can be found in Appendix F.8. The second section describes the scenario and asks questions about the given scenario (§F.7.1.1). The third section describes the fairness metric, defined as the rule, used (in this case it is demographic parity) and asks specific questions about the metric (§F.7.1.2). Finally the last section asks for demographic information (§F.7.3).

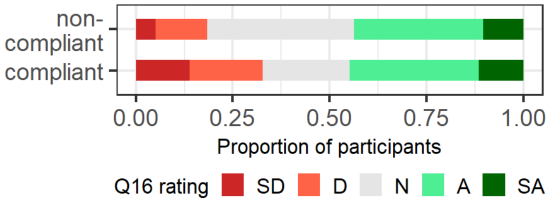


Figure F.19: Participant agreement with rule (Q16) split by compliance (Q14). No differences were found between NC and C participants. SA = strongly agree, A = agree, N = neither agree nor disagree, D = disagree, SD = strongly disagree.

F.7.1.1 Scenario descriptions and questions

The following is shown to each participant:

It is very important that you read each question carefully and think about your answers. The success of our research relies on our respondents being thoughtful and taking this task seriously.

□ I have read the above instructions carefully.

We then introduce one of three different decision making scenarios, described below, followed by two questions. Words that vary across scenario in the questions are shown as <art project, employee awards, hiring>.

Art project A fourth grade teacher is reviewing 20 student art projects. They will award lollipops to the top 4 students who put the most effort into their projects. The teacher knows that some of the students have artists as parents, who might have helped their children with their art project. The teacher's goal is to give out lollipops only based on the amount of effort that the student *themselves* put into their projects.

The teacher uses the following criteria to decide who should get a lollipop:

- Elaborateness of each project.

- Creativity of each project.

About 50% of the students have artists as parents, and 50% do not.

In the past, students with artists as parents typically put more effort into their projects.

In this group of students there is a wide range of project quality (as measured by elaborateness and creativity). However, this range of quality is about the same between students with artists as parents and those without.

The teacher wants to make sure that they award lollipops in a fair way, no matter whether the students' parents are artists or not.

Employee awards A manager at a sales company is deciding which of their 100 employees should receive each of 10 mid-year awards. The manager's goal is to give awards to employees who *will* have high net sales at the end of the year.

The manager uses the following criteria to decide who should get an award:

- Recent performance reviews

- Mid-year net sales

- Number of years on the job

About 50% of the employees are men, and 50% are women.

In the past, men have achieved higher end-of-year net sales than women.

In this group of employees, there is a wide range of qualifications (as measured by performance reviews, mid-year net sales, and number of years on the job). However, this range of qualifications is about the same between male and female employees.

The manager wants to make sure that this awards process is fair to the employees, no matter their gender.

Hiring A hiring manager at a new sales company is reviewing 100 new job applications. Each applicant has submitted a resume, and has had an interview. The manager will send job offers to 10 out of the 100 applicants. Their goal is to make offers to applicants who will have high net sales after a year on the job.

The manager will use the following to decide which applicants should receive job offers:

- Interview scores
- Quality of recommendation letters
- Number of years of prior experience in the field

About 50% of the applicants are men, and 50% are women.

In the past, men have achieved higher net sales than women, after one year on the job.

In this applicant pool there is a wide range of applicant quality (as measured by interview scores, recommendation letters, and years of prior experience in the field). However, the range of quality is about the same for both male and female applicants.

The hiring manager wants to make sure that this hiring process is fair to applicants, no matter their gender.

Questions

1. To what extent do you agree with the following statement: a scenario similar to the one described above might occur in real life.

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

2. How much effort should the <teacher, manager, hiring manager> put in to make sure this decision is fair? [short answer - number of hours]

F.7.1.2 Rule descriptions and questions

Unless otherwise noted the rule description is shown above each of the questions for reference. Correct answers are noted in red.

Art project The teacher uses the following award rule to distribute lollipops: *The fraction of students who receive lollipops that have artist parents should equal the fraction of students in the class that have artist parents. Similarly, the fraction of students who receive lollipops that do not have artist parents should equal the fraction of students in the class that do not have artist parents.*

Example 1: If 10 out of the 20 students in the class have artist parents, then 2 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 2 would be awarded to students without artist parents).

Example 2: If 5 out of the 20 students in the class have artist parents, then 1 out of the 4 lollipops would be awarded to students with artist parents (and the remaining 3 would be awarded to students without artist parents).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

3. Suppose a different teacher is considering awarding lollipops to the whole 4th grade. There are 100 students with artist parents, and 200 students without artist parents. The teacher decides to award 10 lollipops to students with artist parents. **Assuming the teacher is required to use the award rule above**, how many students without artist parents need to receive lollipops?

    (a) 10

    (b) 20

    (c) 40

    (d) 50

4. **Assuming the teacher is required to use the award rule above**, in which of these cases can a teacher award more lollipops to students without artist parents than to students with artist parents?

(a) When the students without artist parents have higher-quality projects (i.e., more elaborate and more creative) than those with artist parents.

(b) When there are more students without artist parents than those with artist parents.

(c) When students without artist parents have more creative projects than those with artist parents.

(d) This cannot happen under the award rule.

5. **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: Even if a student with artist parents has a project that is of the same quality (i.e., equally elaborate and equally creative) as another project by a student without artist parents, they can be treated differently (ie., only one of the students might get a lollipop).

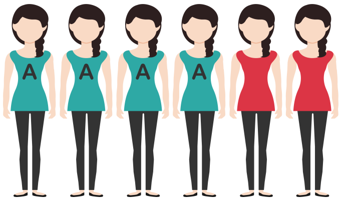
6. **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: If all students without artist parents have low-quality projects (i.e., low elaborateness and low creativity), but the teacher awards lollipops to some of them, then any lollipops awarded to students with artist parents must be awarded to those who have low-quality projects.

7. **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: Suppose the teacher is distributing 10 lollipops amongst a pool of students that includes students with and without artist parents. Even if all students with artist parents have low-quality (i.e.,

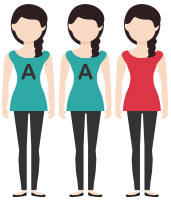
low elaborateness and low creativity) projects, some of them must still receive lollipops.

8. **Assuming the teacher is required to use the award rule above**, is the following statement TRUE OR FALSE: This award rule always allows the teacher to award lollipops exclusively to the students who have the highest quality (i.e., most elaborate and most creative) projects.

In the two examples above there are 20 students. Consider a different scenario, with **6 students – 4 with artist parents and 2 without, as illustrated below**. The next three questions each give a potential outcome for all six students (i.e., which of the 6 students receive awards). Please indicate which of the outcomes follow **the award rule above**.

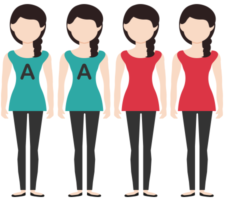


9. Alternative scenario 1:



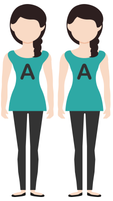
Does this distribution of awards obey the **award rule**? Yes

10. Alternative scenario 2:



Does this distribution of awards obey the **award rule**? No

11. Alternative scenario 3:



Does this distribution of awards obey the **award rule**? No

12. In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)

13. To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

   (a) I applied the provided award rule only.

   (b) I used my own ideas of what the correct award decision should be rather than the provided award rule.

   (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.

15. What is your opinion on the award rule? Please explain why. [short answer]

16. Suppose that you are the teacher whose job it is to distribute lollipops to students based on the criteria listed above (i.e., elaborateness of each project, creativity of each project). How would you ensure that this process is fair? [short answer]

17. Was there anything about this survey that was hard to understand or answer? [short answer]

Employee awards The manager uses the following award rule to distribute awards: *The fraction of employees who receive awards that are female should equal the fraction of employees that are female. Similarly, fraction of employees who receive awards that are male should equal the fraction of employees that are male.*

Example 1: If there are 50 female employees out of 100, then 5 out of the 10 awards should be awarded to female employees (and the remaining 5 would be made to male employees).

Example 2: If there are 30 female employees out of 100, then 3 out of the 10 awards should be awarded to female employees (and the remaining 7 would be made to male employees).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above award rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

3. Suppose a different manager is considering employees for a different award. There are 100 male employees and 200 female employees, and they decide to give awards to 10 male employees. **Assuming the manager is required to use the award rule above**, how many female employees do they need to give awards to?

   (a) 10

   (b) 20

   (c) 40

   (d) 50

4. **Assuming the manager is required to use the award rule above**, in which of these cases can a manager give more awards to female employees than to male employees?

   (a) When there are more well-qualified female employees than well-qualified male employees (i.e., more women have better performance reviews, higher mid-year net sales, and more years on the job).

(b) When there are more female employees than male employees.

(c) When female employees receive higher performance reviews than male employees.

(d) This cannot happen under the award rule.

5. **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: Even if a male employee's qualifications look similar to a female employee's (in terms of performance reviews, mid-year net sales, and years on the job), he can be treated differently (i.e., only one of the employees gets an award).

6. **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: If all female employees are unqualified (i.e., have low performance reviews, low mid-year net sales, and few years on the job), but you give awards to some of them, then awards given to male employees must be made to unqualified male employees.

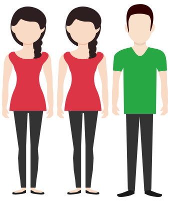
7. **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: Suppose the manager is distributing 10 awards amongst a pool that includes both male and female employees. Even if all male employees are unqualified for an award (i.e., have low performance reviews, low mid-year net sales, and few years on the job), some of them must still receive awards.

8. **Assuming the manager is required to use the award rule above**, is the following statement TRUE OR FALSE: This award rule always allows the

manager to distribute awards exclusively to the most qualified employees (i.e., employees with better performance reviews, high mid-year net sales, and high number of years on the job).

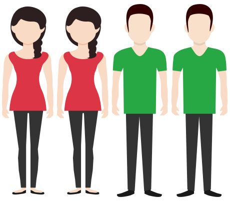
In the two examples above there are 100 employees. Consider a different scenario, with **6 employees– 4 female and 2 male, as illustrated below**. The next three questions each give a potential outcome for all six employees (i.e., which of the 6 employees receive awards). Please indicate which of the outcomes follow **the award rule above**.

9. Alternative scenario 1:



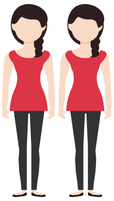
Does this distribution of awards obey the **award rule**? Yes

10. Alternative scenario 2:



Does this distribution of awards obey the **award rule**? No

11. Alternative scenario 3:

    Does this distribution of awards obey the **award rule**? No

12. In your own words, explain the **award rule**. [short answer] (The rule is not shown above this question)

13. To what extent do you agree with the following statement: I am confident I know how to **apply the award rule described above**?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided award rule only.

    (b) I used my own ideas of what the correct award decision should be rather than the provided award rule.

    (c) I used a combination of the provided award rule and my own ideas of what the correct award decision should be.

15. What is your opinion on the award rule? Please explain why. [short answer]

16. Suppose that you are the manager whose job it is to distribute mid-year awards to employees based on the criteria listed above (i.e., recent performance reviews, mid-year net sales, number of years on the job). How would you ensure that this process is fair? [short answer]

17. Was there anything about this survey that was hard to understand or answer? [short answer]

Hiring The hiring manager uses the following hiring rule to send out offers: *The fraction of applicants who receive job offers that are female should equal the fraction of applicants that are female. Similarly, fraction of applicants who receive job offers that are male should equal the fraction of applicants that are male.*

Example 1: If there are 50 female applicants out of the 100 applicants, then 5 out of the 10 offers would be made to female applicants (and the remaining 5 would be made to male applicants).

Example 2: If there are 30 female applicants out of the 100 applicants, then 3 out of the 10 offers would be made to female applicants (and the remaining 7 would be made to male applicants).

In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

3. Suppose a different hiring manager is considering applicants for a different job. There are 100 male applicants and 200 female applicants, and they decide to send offers to 10 male applicants. **Assuming the hiring manager is required to use the hiring rule above**, how many female applicants do they need to send offers to?

   (a) 10

   (b) 20

   (c) 40

   (d) 50

4. **Assuming the hiring manager is required to use the hiring rule above**, in which of these cases can a hiring manager make more job offers to female applicants than to male applicants?

   (a) When there are more well-qualified female applicants than well-qualified male applicants (i.e., more women have higher interview scores, higher quality recommendation letters, and more years of prior experience in the field).

   (b) When there are more female applicants than male applicants.

   (c) When female applicants receive better interview scores than male applicants.

(d) This cannot happen under the hiring rule.

5. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: Even if a male applicant's qualifications look similar to a female applicant's (in terms of interview scores, recommendation letters, and years of prior experience in the field), he can be treated differently (i.e., only one of the applicants will receive a job offer).

6. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: If all female applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), but you send job offers to some of them, then any job offers made to male applicants must be made to unqualified male applicants.

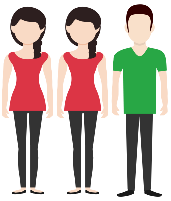
7. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: Suppose the hiring manager is sending out 10 job offers to a pool that includes male and female applicants. Even if all male applicants are unqualified (i.e., have low interview scores, low-quality recommendation letters, and few years of prior experience in the field), some of them must still receive job offers.

8. **Assuming the hiring manager is required to use the hiring rule above**, is the following statement TRUE OR FALSE: This hiring rule always

allows the hiring manager to send offers exclusively to the most qualified applicants (i.e., applicants with high interview scores, high quality recommendation letters, and high number years of prior experience in the field).

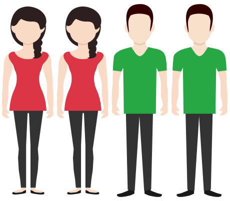
In the two examples above there are 100 applicants. Consider a different scenario, with **6 applicants – 4 female and 2 male, as illustrated below**. The next three questions each give a potential outcome for all 6 applicants (i.e., which of the 6 applicants receive job offers). Please indicate which of the outcomes follow **the hiring rule above**.

9. Alternative scenario 1:



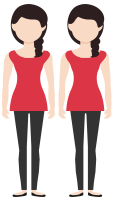
Does this distribution of job offers obey the **hiring rule**? Yes

10. Alternative scenario 2:



Does this distribution of job offers obey the **hiring rule**? No

11. Alternative scenario 3:



Does this distribution of job offers obey the **hiring rule**? No

12. In your own words, explain the **hiring rule**. [short answer] (The rule is not shown above this question)

13. To what extent do you agree with the following statement: I am confident I know how to **apply the hiring rule described above**?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided hiring rule only.

    (b) I used my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

    (c) I used a combination of the provided hiring rule and my own ideas of what the correct hiring decision should be.

15. What is your opinion on the hiring rule? Please explain why. [short answer]

16. Suppose that you are the hiring manager whose job it is to send job offers to applicants based on the criteria listed above (i.e., interview scores, quality of recommendation letters, number of years of prior experience in the field). How would you ensure that this process is fair? [short answer]

17. Was there anything about this survey that was hard to understand or answer? [short answer]

F.7.2 Study-2: Survey

Each of the surveys are split into four main sections. The first section is the consent form which can be found in Appendix F.8. The second section describes the hiring scenario and asks questions about it (§F.7.2.1). The third section describes the fairness metric, defined as the rule, used (in this case it is demographic parity) and asks specific questions about the metric (§F.7.2.2). Finally the last section asks for demographic information (§F.7.3).

F.7.2.1 Scenario description and questions

The following is shown to each participant (note that Step 3 is not shown to participants with the DP definition):

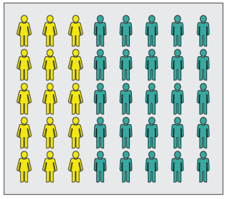
It is very important that you read each question carefully and think about your answers. The success of our research relies on our respondents being thoughtful and taking this task seriously.

□ I have read the above instructions carefully.

A company, Sales-a-lot, is reviewing their hiring process. They want to hire applicants who are high performing, and they also want to make sure that their hiring process is fair to their applicants, no matter their gender. To do this, Sales-a-lot employs an external firm, Recruit-a-matic, which keeps track of all applicants. This review will take place over one year.

For clarity at each stage of the hiring process we use images to represent the hiring pool.

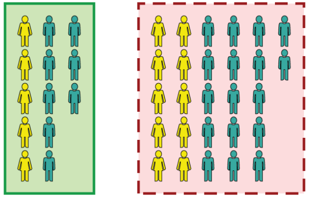
Step 1: Applicant Pool. At the beginning of the year, Sales-a-lot reviews all job applicants, and sends job offers to some of them. The initial applicant pool is shown with a gray background. For example, the following image shows an applicant pool with 15 female applicants and 25 male applicants:



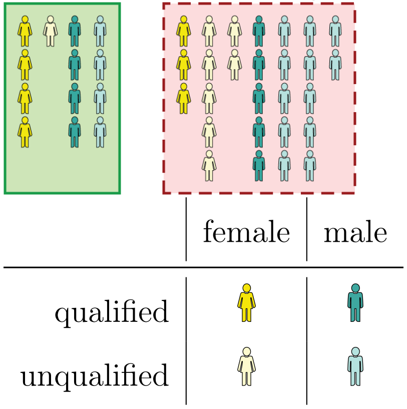
Step 2: Sending Job Offers. Next, Sales-a-lot sends job offers to some of these applicants, using the following criteria:

- Interview scores
- Quality of recommendation letters
- Number of years of prior experience in the field

Suppose that Sales-a-lot sends offers to 5 female applicants and 8 male applicants (so 10 female and 17 male applicants didn't receive offers). In the following image, applicants who received a job offer are shown on the left (with a green background) and applicants who didn't receive a job offer are shown on the right, with a red background):



Step 3: Applicant Evaluation. For the rest of the year, Recruit-a-matic (the external firm) keeps track of all applicants in the initial pool, whether they received job offers or not. At the end of the year, Rectruit-a-matic finds out which applicants were high performers, i.e. qualified (shown in dark), and which applicants were low performers, i.e. unqualified (shown in light). For example, the following image shows that most of the high performers received job offers, but some did not.



Questions

1. To what extent do you agree with the following statement: a scenario similar to the one described above might occur in real life.

    - Strongly agree

    - Agree

- Neither agree nor disagree
- Disagree
- Strongly disagree

2. How much effort, in hours, should Sales-a-lot put in to make sure these decisions were fair? [short answer - number of hours]

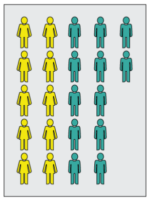
F.7.2.2 Rule descriptions and questions

The following sections provide fairness definitions (presented to participants as *rules*) for Demographic Parity, Equal Opportunity (FNR and FPR), and Equalized Odds. Unless otherwise noted the rule description is shown above each of the questions for reference. Correct answers are noted in red.

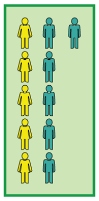
Demographic Parity. Recruit-a-matic uses the following rule to determine whether Sales-a-lot's hiring decisions were fair:

The fraction of male candidates who receive job offers should equal the fraction of female candidates who receive job offers.

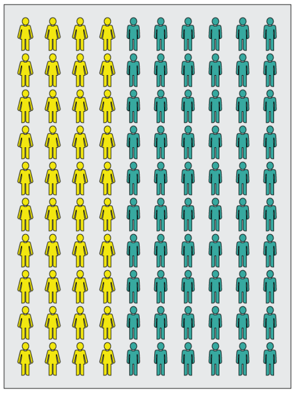
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following applicants (10 female and 12 male).



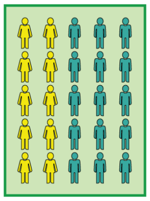
If Sales-a-lot sent job offers to the following number of applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 applicants as follows (40 female and 60 male).



If Sales-a-lot sent job offers to the following number of applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

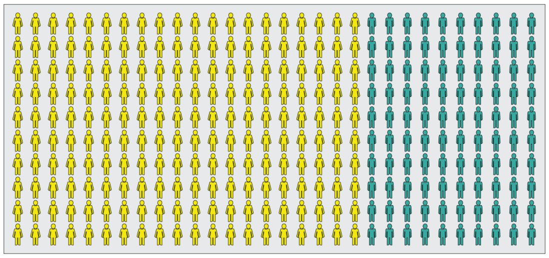


In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

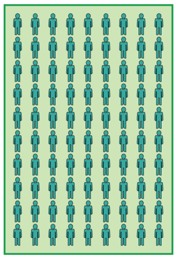
Please note that we ask you to apply and use ONLY the above hiring

rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

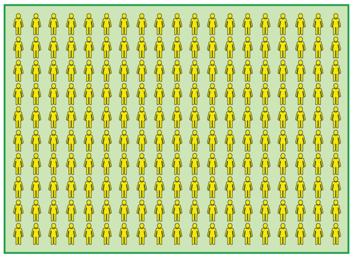
3. Suppose a different company considered applicants for a different job. There were 200 female applicants and 100 male applicants,

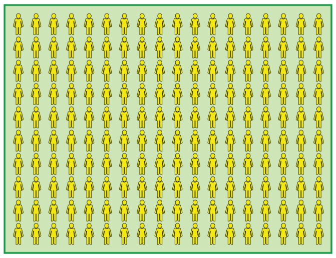
   

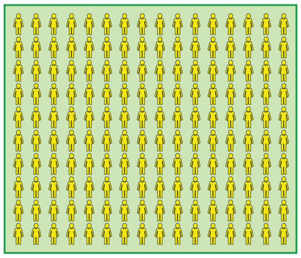
   and they did send job offers to 90 male applicants.

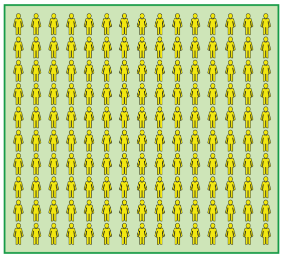
   

   Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many female applicants should have received job offers?

   (a) 190 

   (b) 180 

(c) 160 

(d) 150 

4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more qualified female applicants than qualified male applicants?

   (a) When there are more qualified female applicants than qualified male applicants (i.e., more women had low net sales at the end of the year).

   (b) When there are more female applicants than male applicants.

   (c) When female applicants receive worse interview scores than male applicants.

   (d) This cannot happen under the hiring rule.

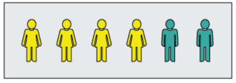
5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement TRUE OR FALSE: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.

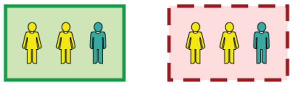
7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement TRUE OR FALSE: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 applicants (i.e., which of the 6 applicants do receive job offers). Please indicate which of the outcomes follow the hiring rule above.

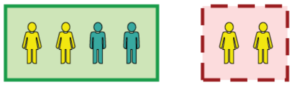


9. Sales-a-lot makes the following hiring decisions.



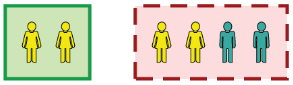
Do these decisions obey the hiring rule? Yes

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided hiring rule only.

    (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.

(c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

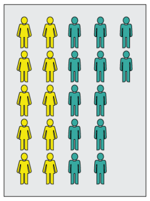
17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

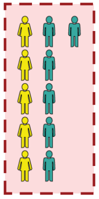
Equal Opportunity - FNR. Recruit-a-matic uses the following rule to determine whether Sales-a-lot's hiring decisions were fair:

The fraction of qualified male candidates who do not receive job offers should equal the fraction of qualified female candidates who do not receive job offers.

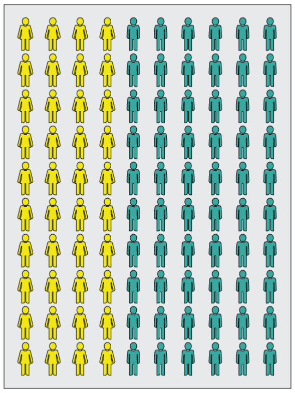
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following qualified applicants (10 female and 12 male).



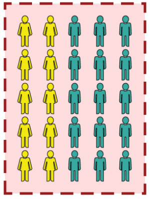
If Sales-a-lot did not send job offers to the following number of qualified applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 qualified applicants as follows (40 female and 60 male).



If Sales-a-lot did not send job offers to the following number of qualified applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

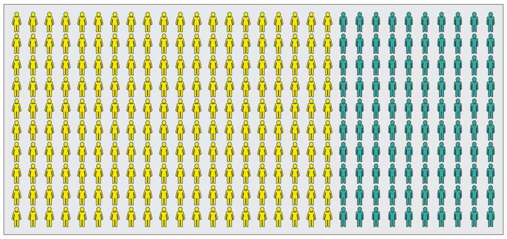


Note that in the above examples the remaining qualified applicants received job offers, but are not displayed here.

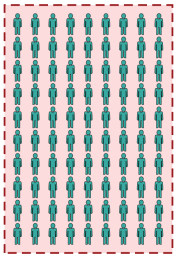
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

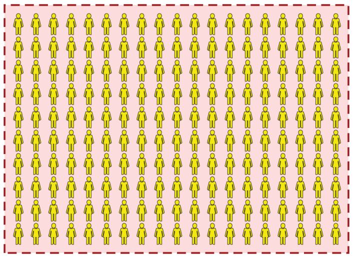
3. Suppose a different company considered applicants for a different job. There were 200 qualified female applicants and 100 qualified male applicants,

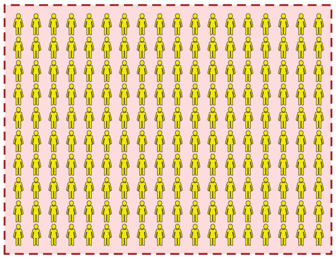


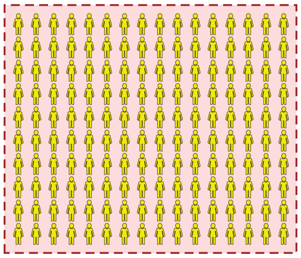
and they did not send job offers to 90 qualified male applicants.

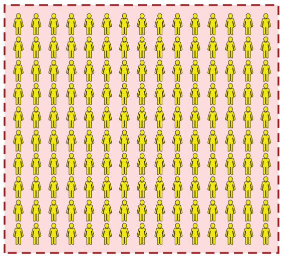


Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many qualified female applicants should not have received job offers?

(a) 190 

(b) 180 

(c) 160 

(d) 150 

4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have rejected more qualified female applicants than qualified male applicants?

   (a) When there are more qualified female applicants than qualified male applicants (i.e., more women had low net sales at the end of the year).

   (b) When there are more female applicants than male applicants.

   (c) When female applicants receive worse interview scores than male applicants.

   (d) This cannot happen under the hiring rule.

5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement TRUE OR FALSE: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

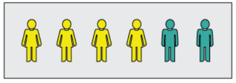
6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.

7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that even if all male applicants were unqualified

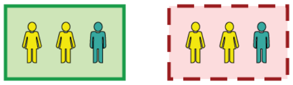
(they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement TRUE OR FALSE: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 qualified applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 qualified applicants (i.e., which of the 6 applicants do not receive job offers). Please indicate which of the outcomes follow the hiring rule above.

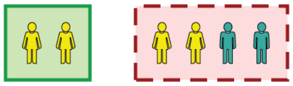


9. Sales-a-lot makes the following hiring decisions.



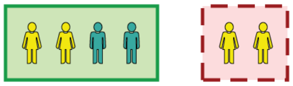
Do these decisions obey the hiring rule? Yes

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided hiring rule only.

    (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.

    (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

    - Strongly agree
    - Agree

- Neither agree nor disagree
- Disagree
- Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

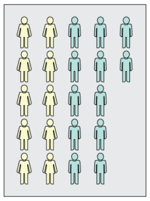
17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

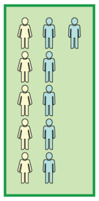
Equal Opportunity - FPR. Recruit-a-matic uses the following rule to determine whether Sales-a-lot's hiring decisions were fair:

The fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.

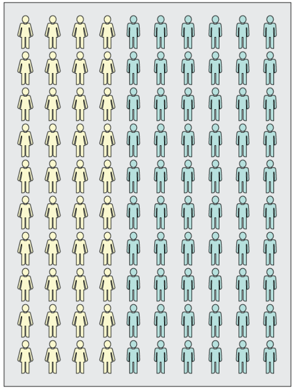
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following unqualified applicants (10 female and 12 male).



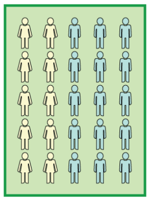
If Sales-a-lot sent job offers to the following number of unqualified applicants (5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



Example 2: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot reviewed a total of 100 unqualified applicants as follows (40 female and 60 male).



If Sales-a-lot sent job offers to the following number of unqualified applicants (10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

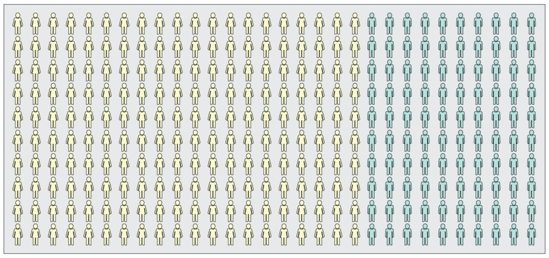


Note that in the above examples the remaining unqualified applicants did not receive job offers, but are not displayed here.

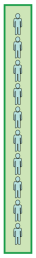
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

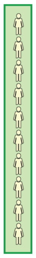
3. Suppose a different company considered applicants for a different job. There were 200 unqualified female applicants and 100 unqualified male applicants,

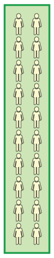
   

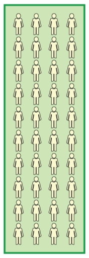
   and they did send job offers to 10 unqualified male applicants.

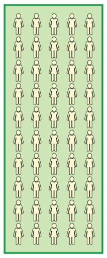
   

Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many unqualified female applicants should have received job offers?

(a) 10 

(b) 20 

(c) 40 

(d) 50 

4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more unqualified female applicants than unqualified male applicants?

(a) When there are more unqualified female applicants than unqualified male applicants (i.e., more women had low net sales at the end of the year).

(b) When there are more female applicants than male applicants.

(c) When female applicants receive worse interview scores than male applicants.

(d) This cannot happen under the hiring rule.

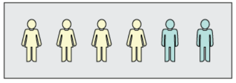
5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement TRUE OR FALSE: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.

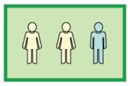
7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that even if all male applicants were unqualified (they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement TRUE OR FALSE: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 unqualified applicants – 4 female and 2 male, as illustrated below. The next three questions each give a different potential outcome for all 6 applicants (i.e., which of the 6 applicants receive job offers). Please indicate which of the outcomes follow the hiring rule above.

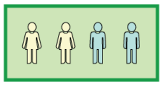


9. Sales-a-lot makes the following hiring decisions.

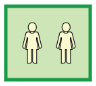
   Do these decisions obey the hiring rule? Yes

10. Sales-a-lot makes the following hiring decisions.

    Do these decisions obey the hiring rule? No

11. Sales-a-lot makes the following hiring decisions.

    Do these decisions obey the hiring rule? No

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided hiring rule only.

    (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.

    (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

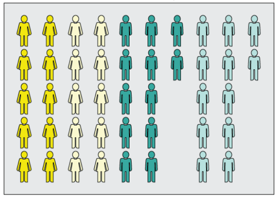
17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

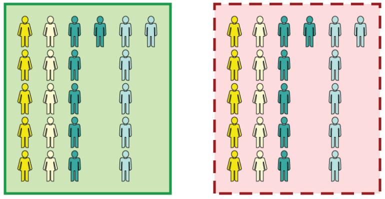
Equalized Odds. Recruit-a-matic uses the following rule to determine whether Sales-a-lot's hiring decisions were fair:

The fraction of qualified male candidates who do not receive job offers should equal the fraction of qualified female candidates who do not receive job offers. Similarly, the fraction of unqualified male candidates who receive job offers should equal the fraction of unqualified female candidates who receive job offers.

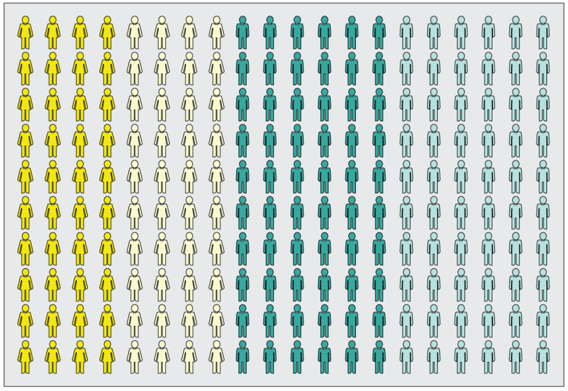
Example 1: Suppose that over the past year, Recruit-a-matic finds that Sales-a-lot received the following qualified applicants (10 female and 12 male) and unqualified applicants (10 female and 12 male).



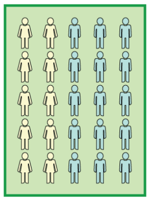
If Sales-a-lot did send offers to the following number of unqualified applicants (left, 5 female and 6 male), and did not send job offers to the following number of qualified applicants (right, 5 female and 6 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).

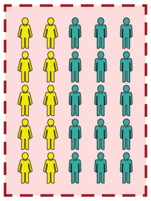


Example 2: Suppose that over the past year, Recruit-a-lot finds that Sales-a-lot reviewed a total of 100 qualified applicants (40 female and 60 male) and 100 unqualified applicants (40 female and 60 male).



If Sales-a-lot did send offers to the following number of unqualified applicants (left, 10 female and 15 male), and did not send job offers to the following number of qualified applicants (right, 10 female and 15 male), then this would be fair according to the hiring rule (note that there are other possible outcomes that are fair according to the hiring rule).



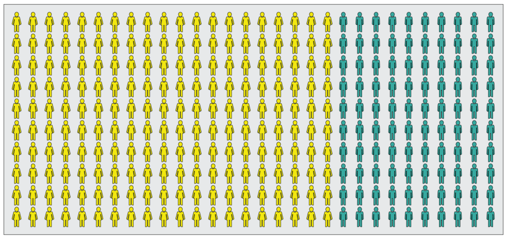


Note that in the above examples the remaining unqualified applicants did not receive job offers, but are not displayed here. Similarly, the remaining qualified applicants received job offers, but are not displayed here.

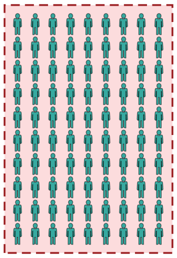
In the next section, we will ask you some questions about the information you have just read. Please note that this is not a test of your abilities. We want to measure the quality of the description you read, not your ability to take tests or answer questions.

Please note that we ask you to apply and use ONLY the above hiring rule when answering the following questions. You will have an opportunity to state your opinions and feelings on the rule later in the survey.

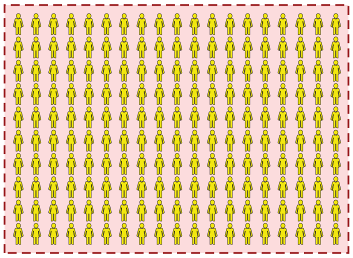
3. Suppose a different company considered applicants for a different job. There were 200 qualified female applicants and 100 qualified male applicants,

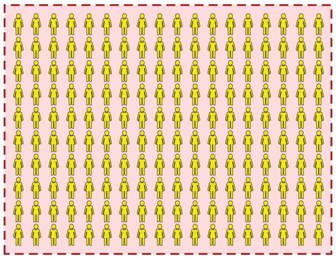


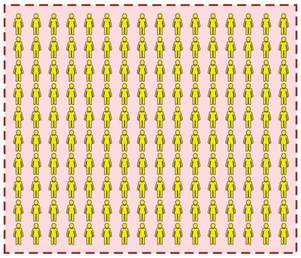
and they did not send job offers to 90 qualified male applicants.

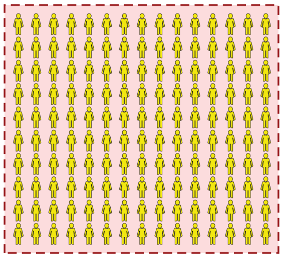


Assuming that Recruit-a-matic reviews their decisions using the hiring rule above, how many qualified female applicants should not have received job offers?

(a) 190 

(b) 180 

(c) 160 

(d) 150 

4. Assuming Recruit-a-matic reviews decisions using the hiring rule above, in which of these cases could Sales-a-lot have accepted more unqualified female applicants than unqualified male applicants?

   (a) When there are more unqualified female applicants than unqualified male applicants (i.e., more women had low net sales at the end of the year).

   (b) When there are more female applicants than male applicants.

   (c) When female applicants receive worse interview scores than male applicants.

   (d) This cannot happen under the hiring rule.

5. Consider one male applicant and one female applicant, both of whom are similarly qualified for the job (they achieve about the same net sales at the end of their first year). Is the following statement TRUE OR FALSE: The hiring rule above allows Sales-a-lot to make a job offer to one of these applicants and not the other.

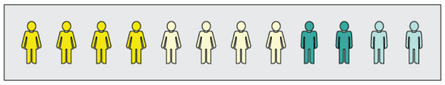
6. Consider a situation where all female applicants were unqualified (they all achieve low net sales at the end of their first year), but some of them received job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that some job offers made to male applicants must have been made to unqualified male applicants.

7. Suppose Sales-a-lot received 100 male and 100 female applicants, and eventually made 10 job offers. Is the following statement TRUE OR FALSE: The hiring rule above requires that even if all male applicants were unqualified

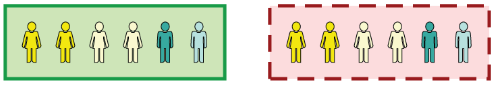
(they all achieve low net sales at the end of their first year), some of the unqualified males must have received job offers.

8. Is the following statement TRUE OR FALSE: The hiring rule above always allows Sales-a-lot to send job offers only to the most qualified applicants (those who achieve high net sales at the end of their first year).

Consider a different scenario than the two examples above, with 6 qualified applicants – 4 female and 2 male; and 6 unqualified applicants – 4 female and 2 male. The next three questions each give a different potential outcome for the applicants (i.e., which of the applicants did or did not receive job offers). Please indicate which of the outcomes follow the hiring rule above.

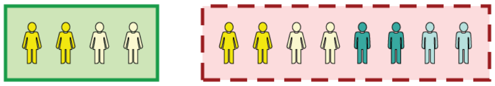


9. Sales-a-lot makes the following hiring decisions.



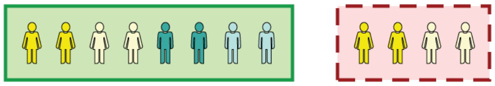
Do these decisions obey the hiring rule? Yes

10. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

11. Sales-a-lot makes the following hiring decisions.



Do these decisions obey the hiring rule? No

12. In your own words, explain the hiring rule. [short answer] [The rule is not shown above this question]

13. To what extent do you agree with the following statement: I am confident I know how to apply the hiring rule described above?

    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

14. Please select the choice that best describes your experience: When I answered the previous questions...

    (a) I applied the provided hiring rule only.

    (b) I used a combination of the provided hiring rule and my own ideas of what the correct hiring rule should be.

    (c) I used only my own ideas of what the correct hiring decision should be rather than the provided hiring rule.

15. To what extent do you agree with the following statement: I like the hiring rule?

    - Strongly agree
    - Agree

- Neither agree nor disagree
- Disagree
- Strongly Disagree

16. To what extent do you agree with the following statement: I agree with the hiring rule?
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly Disagree

17. Please explain your opinion on the hiring rule. [short answer]

18. Was there anything about this survey that was hard to understand or answer? [short answer]

F.7.3 Demographic Information

1. Please specify the gender with which you most closely identify:
    - Male
    - Female
    - Other
    - Prefer not to answer

2. Please specify your year of birth

3. Please specify your ethnicity (you may select more than one):

   - White
   - Hispanic or Latinx
   - Black or African American
   - American Indian or Alaska Native
   - Asian, Native Hawaiian, or Pacific Islander
   - Other

4. Please specify the highest degree or level of school you have completed:

   - Some high school credit, no diploma or equivalent
   - High school graduate, diploma or the equivalent (for example: GED)
   - Some college credit, no degree
   - Trade/technical/vocational training
   - Associate's degree
   - Bachelor's degree
   - Master's degree
   - Professional or doctoral degree (JD, MD, PhD)

5. How much experience do you have in each of the following areas? (1 - no experience, 2 - limited experience, 3 - significant experience, 4 - expert)

   (a) Human resources (making hiring decisions)

(b) Management (of employees)

(c) Education (teaching)

(d) IT infrastructure/systems administration

(e) Computer science/programming

(f) Machine learning/data science

We will maintain privacy of the information you have provided here. Your information will only be used for data analysis purposes.

F.8 Consent

F.8.1 Online Survey Consent Form

F.8.1.1 Project Title

Fairness Evaluation and Comprehension

F.8.1.2 Purpose of the Study

This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above 18. The purpose of this research project is to understand lay comprehension of different fairness metrics.

F.8.1.3 Procedures

The procedures will start with reading a brief description of a decision-making scenario. You will then be asked to answer some comprehension questions about the scenario. The questions will look like the following: What are the pros and cons of the notion of fairness described above?

Finally, you will be asked some demographics questions. The entire survey will take approximately 20 minutes or less.

F.8.1.4 Potential Risks and Discomforts

There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

F.8.1.5 Potential Benefits

There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

F.8.1.6 Confidentiality

Any potential loss of confidentiality will be minimized by storing all data (including information such as MTurk IDs and demographics) will be stored securely (a) in a password-protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected (MTurk IDs, IP addresses, cookies) will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the Principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

F.8.1.7 Compensation

You will receive $3. You will be responsible for any taxes assessed on the compensation.

If you will earn $100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over $100 only your name and address will be collected to receive compensation.

F.8.1.8 Right to Withdraw and Questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

F.8.1.9 Participant Rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

[Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to the [Blinded] IRB procedures for research involving human subjects.

F.8.1.10 Statement of Consent

By agreeing below you indicate that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. Please ensure you have made a copy of the above consent form for your records.

Pease ensure you have made a copy of the above consent form for your records. A copy of this consent form can be found here [link to digital copy].

□ I am age 18 or older

□ I have read this consent form

□ I voluntarily agree to participate in this research study

F.8.2 Cognitive Interview Consent Form

F.8.2.1 Project Title

Fairness Cognitive Interview

F.8.2.2 Purpose of the Study

This research is being conducted by [Blinded] at [Blinded]. We are inviting you to participate in this research project because you are above the age of 18, and fluent in English. The purpose of this research project is to understand lay comprehension of different fairness metrics.

F.8.2.3 Procedures

The procedure involves completing an interview. The full procedure will be approximately 1 hour in duration.

During the interview you will be audio recorded, if you agree to be recorded. You will be asked to first read a brief description of a decision-making scenario. You will then be asked to fill out a survey about the scenario. While answering questions you will be asked verbal questions related to how you reached your answer in the survey.

Sample survey question: Is the following statement true or false? This hiring rule allows the hiring manager to send offers exclusively to the most qualified applicants.

Sample interview question: How did you reach your answer to that survey question?

F.8.2.4 Potential Risks and Discomforts

There are several questions to answer over the course of this study, so you may find yourself growing tired towards the end. Outside of this, there are minimal risks to participating in this research study. All data collected in this study will be maintained securely (see Confidentiality section) and will be deleted at the conclusion of the study.

However, if at any time you feel that you wish to terminate your participation for any reason, you are permitted to do so.

F.8.2.5 Potential Benefits

There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of fairness metrics and their applications.

F.8.2.6 Confidentiality

Any potential loss of confidentiality will be minimized by storing all data (including information such as demographics) securely (a) in a password protected computer located at [Blinded] or (b) using a trusted third party (Qualtrics). Personally identifiable information that is collected will be deleted upon study completion. All other data gathered will be stored for three years post study completion, after which it will be erased. The only persons that will have access to the data are the principle Investigator and the Co-Investigators.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of [Blinded] or governmental authorities if you or someone else is in danger or if we are required to do so by law.

F.8.2.7 Compensation

You will receive $30. You will be responsible for any taxes assessed on the compensation.

If you will earn $100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over $100 only your name and address will be collected to receive compensation.

F.8.2.8 Right to Withdraw and Questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator: [Blinded]

F.8.2.9 Participant Rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact: [Blinded]

For more information regarding participant rights, please visit: [Blinded]

This research has been reviewed according to [Blinded] IRB procedures for research involving human subjects.

F.8.2.10 Statement of Consent

Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form.

Please initial all that apply (you may choose any number of these statements):

□ I agree to be audio recorded

□ I agree to allow researchers to use my audio recording in research publications and presentations.

□ I do not agree to be audio recorded

If you agree to participate, please sign your name below.

Bibliography

[1] Himan Abdollahpouri and Robin Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158*, 2019.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 60–69, 2018.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 127–135, 2013.

[4] Faez Ahmed, John P. Dickerson, and Mark Fuge. Diverse weighted bipartite b-matching. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[5] Faez Ahmed, John P. Dickerson, and Mark Fuge. Diverse weighted bipartite b-matching. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[6] H Akaike. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, 1974.

[7] Susel Góngora Alonso, Isabel de la Torre-Díez, Sofiane Hamrioui, Miguel López-Coronado, Diego Calvo Barreno, Lola Morón Nozaleda, and Manuel Franco. Data Mining Algorithms and Techniques in Mental Health: A Systematic Review, 2018.

[8] American Psychological Association. Ethical Principles of Psychologists and Code of Conduct. URL https://www.apa.org/ethics/code/.

[9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[10] Richard D Arvey and James E Campion. The employment interview: A summary and review of recent research. *Personal Psychology*, 35(2):281–322, 1982.

[11] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. Optimal greedy diversity for recommendation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1742–1748, 2015.

[12] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[13] H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence Journal*, 2018.

[14] Haris Aziz. A rule for committee selection with soft diversity constraints. *arXiv preprint arXiv:1803.11437*, 2018.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[16] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, pages 5802–5804, 2018.

[17] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3–11, 2019.

[18] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 2016.

[19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning. fairmlbook. org, 2018. *URL: http://www. fairmlbook. org*, 2018.

[20] Carmen Batanero, Egan J Chernoff, Joachim Engel, Hollylynne S Lee, and Ernesto Sánchez. Research on teaching and learning probability. In *Research on teaching and learning probability*. Springer, Cham, 2016.

[21] Douglas Belkin. SAT to give students 'adversity score'to capture social and economic background. *The Wall Street Journal*, 2019.

[22] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta,

A Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 2019.

[23] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[24] 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.

[25] Nawal Benabbou, Mithun Chakraborty, Xuan-Vinh Ho, Jakub Sliwinski, and Yair Zick. Diversity constraints in public housing allocation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 973–981, 2018.

[26] Nawal Benabbou, Mithun Chakraborty, Edith Elkind, and Yair Zick. Fairness towards groups of agents in the allocation of indivisible items. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 95–101, 2019.

[27] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

[28] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59(1):17–31, 2011.

[29] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

[30] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2017.

[31] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *Artificial Intelligence, Ethics, and Society*, 2019.

[32] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, 81:1–11, 2017.

[33] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 514–524, 2020.

[34] Christian Bird, Adrian Bachmann, Eirik Aune, John Duffy, Abraham Bernstein, Vladimir Filkov, and Premkumar Devanbu. Fair and balanced? bias in bug-fix datasets. In *Proceedings of the 7th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering*, pages 121–130, 2009.

[35] Steven Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pages 69–72, 2006.

[36] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[37] DE Bloom, ET Cafiero, E Jané-Llopis, S Abrahams-Gessel, LR Bloom, S Fathima, AB Feigl, T Gaziano, A Hamandi, M Mowafi, A Pandya, K Prettner, L Rosenberg, B Seligman, AZ Stein, and C. Weinstein. The global economic burden of noncommunicable diseases. *Geneva: World Economic Forum*, 2011.

[38] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[39] Miranda Bogen. All the ways hiring algorithms can introduce bias. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias [2019-08-22], 2019.

[40] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 343–351, 2016.

[41] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. *JMIR Mental Health*, 2016.

[42] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, March 2016.

[43] James A Breaugh and Mary Starke. Research on employee recruitment. *Journal of Management*, 2000.

[44] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. Multiwinner elections with diversity constraints. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[45] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, pages 1–122, 2012.

[46] Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265, 2013.

[47] U.S. Census Bureau. 2017 us census demographics, 2017. URL https://data.census.gov/cedsci.

[48] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356 (6334):183–186, 2017.

[49] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination

prevention. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3992–4001, 2017.

[50] Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 2017.

[51] Wei Cao, Jian Li, Yufei Tao, and Zhize Li. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1036–1044, 2015.

[52] Robert A Carter and Graham Philip. *Beyond the Ubaid: transformation and integration in the late prehistoric societies of the Middle East.* Number 63. Oriental Institute of the University of Chicago, 2010.

[53] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, page 160–169, 2019.

[54] Andrew Chamberlain. How long does it take to hire? https://www.glassdoor.com/research/time-to-hire-in-25-countries/, 2017.

[55] Adrian Chen. The laborers who keep dick pics and beheadings out of your facebook feed. *Wired*, 2014.

[56] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

[57] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT**, pages 339–348. ACM, 2019.

[58] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 379–387, 2014.

[59] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[60] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018. URL http://arxiv.org/abs/1810.08810.

[61] Cint. Cint. URL https://www.cint.com/.

[62] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Reinforcement learning with fairness constraints for resource distribution in human-robot teams. *arXiv preprint arXiv:1907.00313*, 2019.

[63] Katherine Anne Comtois, Amanda H. Kerbrat, Christopher R. DeCou, David C. Atkins, Justine J. Majeres, Justin C. Baker, and Richard K. Ries. Effect of Augmenting Standard Care for Military Personnel With Brief Caring Text Messages for Suicide Prevention. *JAMA Psychiatry*, 2019.

[64] Mike Conway and Daniel O'Connor. Social media, big data, and mental health: Current advances and ethical implications, 2016.

[65] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[66] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, 2014.

[67] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, 2018.

[68] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[69] Darcy J. Corbitt-Hall, Jami M. Gauthier, and Wendy Troop-Gordon. Suicidality Disclosed Online: Using a Simulated Facebook Task to Identify Predictors of Support Giving to Friends at Risk of Self-harm. *Suicide and Life-Threatening Behavior*, 2019.

[70] Cheryl M. Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C. Javitt, Carrie E. Bearden, and Guillermo A. Cecchi. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 2018.

[71] Cheryl M. Corcoran, Caridad Benavides, and Guillermo Cecchi. Natural language processing: Opportunities and challenges for patients, providers, and hospital systems. 2019.

[72] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Honolulu, HI, USA*, 2019.

[73] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

[74] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[75] Munmun De Choudhury. Opportunities of social media in health and wellbeing. *XRDS: Crossroads, The ACM Magazine for Students*, 2014.

[76] Munmun De Choudhury. Opportunities of social media in health and wellbeing. *XRDS: Crossroads, The ACM Magazine for Students*, 2015.

[77] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, 2016.

[78] Pierre Desrochers. Local diversity, human creativity, and technological innovation. *Growth and Change*, pages 369–394, 2001.

[79] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[80] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *available at: www. aies-conference. com/wp-content/papers/main/AIES_ 2018_ paper_ 9. pdf (accessed 6 August 2018).[Google Scholar]*, 2018.

[81] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[82] 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.

[83] Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2019.

[84] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018.

[85] BS Everitt and A Skrondal. *The Cambridge Dictionary of Statistics.* Cambridge University Press, 4th edition, 2010.

[86] Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. Multiwinner voting: A new challenge for social choice theory. *Trends in Computational Social Choice*, 74, 2017.

[87] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[88] Joseph C. Franklin, Jessica D. Ribeiro, Kathryn R. Fox, Kate H. Bentley, Evan M. Kleiman, Xieyining Huang, Katherine M. Musacchio, Adam C. Jaroszewski, Bernard P. Chang, and Matthew K. Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 2017.

[89] Rachel Freedman, J Schaich Borg, Walter Sinnott-Armstrong, J Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. In *AAAI*, 2018.

[90] Carl Benedikt Frey and Michael A Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.

[91] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, 2016. URL http://arxiv.org/abs/1609.07236.

[92] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[93] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

[94] Gerd Gigerenzer and Adrian Edwards. Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal*, 2003.

[95] Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M Schwartz, and Steven Woloshin. Helping doctors and patients make sense of health statistics. *Psychological science in the public interest*, 2007.

[96] Martin Gilbert. *The holocaust: A history of the Jews of Europe during the Second World War*. Macmillan, 1987.

[97] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In D. D.

Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2415–2423. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6316-satisfying-real-world-goals-with-dataset-constraints.pdf.

[98] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. Ethics and Artificial Intelligence: Suicide Prevention on Facebook, 2018.

[99] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. In *The Journal of Machine Learning Research*, 2012.

[100] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912. International World Wide Web Conferences Steering Committee, 2018.

[101] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 2017.

[102] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *CoRR*, abs/1806.11212, 2018. URL http://arxiv.org/abs/1806.11212.

[103] Moritz Hardt, Eric Price, , and Nati Srebro. Equality of opportunity in supervised learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3315–3323, 2016.

[104] Margaret C Harrell and Melissa A Bradley. Data collection methods. semi-structured interviews and focus groups. Technical report, Rand National Defense Research Inst santa monica ca, 2009.

[105] Michael M Harris. Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personal Psychology*, 42 (4):691–726, 1989.

[106] David Haussler and Manfred Warmuth. The probably approximately correct (pac) and other learning models. In *Foundations of Knowledge Acquisition*, pages 291–312. Springer, 1993.

[107] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, and Bushra Anjum. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog*, 2018.

[108] Holly Hedegaard, Sally C Curtin, and Margaret Warner. Suicide rates in the united states continue to increase. *NCHS Data Brief No. 309*, 2018.

[109] Robin M Hogarth and Emre Soyer. Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition*, 2015.

[110] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 600. ACM, 2019.

[111] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[112] Eric Horvitz and Deirdre Mulligan. Data, privacy, and the greater good. *Science*, 2015.

[113] Vivian Hunt, Dennis Layton, and Sara Prince. Diversity matters. *McKinsey & Company*, 2015.

[114] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst.*, 2011.

[115] Thomas R. Insel. Assessing the economic costs of serious mental illness. *American Journal of Psychiatry*, 2008.

[116] Lilly Irani. The hidden faces of automation. *XRDS: Crossroads, The ACM Magazine for Students*, 2016.

[117] Dan Iter, Jong Yoon, and Dan Jurafsky. Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, 2018.

[118] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

[119] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL-HLT*, 2019.

[120] Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, and Y Narahari. An incentive compatible multi-armed-bandit crowdsourcing mechanism with quality assurance. *arXiv preprint arXiv:1406.7157*, 2014.

[121] Adam C. Jaroszewski, Robert R. Morris, and Matthew K. Nock. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. *Journal of Consulting and Clinical Psychology*, 2019.

[122] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016.

[123] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 325–333, 2016.

[124] Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[125] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2444–2453, 2018.

[126] Nathan Kallus and Angela Zhou. Assessing disparate impacts of personalized interventions: Identifiability and bounds. *arXiv preprint arXiv:1906.01552*, 2019.

[127] Julian Katz-Samuels and Clayton Scott. Feasible arm identification. In *ICML*, 2018.

[128] Julian Katz-Samuels and Clayton Scott. Top feasible arm identification. In *AISTATS*, 2019.

[129] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

[130] Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments. *arXiv preprint arXiv:1905.10607*, 2019.

[131] Deanna L. Kelly, Max Spaderna, Vedrana Hodzic, Suraj Nair, Christopher Kitchen, Anne Werkheiser, Megan Powell, Stephanie Feldman, Fang Liu, Carol Espy-Wilson, Glen Coppersmith, and Philip Resnik. Blinded Clinical Ratings of Social Media Data are Correlated with In-Person Clinical Ratings

in Participants Diagnosed with Either Depression, Schizophrenia, or Healthy Controls. in preparation.

[132] Julia D. Kent and Maureen Terese McCarthy. *Holistic Review in Graduate Admissions*. Council of Graduate Schools, 2016.

[133] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Proceesings of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, 2016.

[134] Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability*. 2014.

[135] Chung-Ming Kuan. Classical least squares theory, 2004. URL http://homepage.ntu.edu.tw/~ckuan/pdf/et_ch3_Fall2009.pdf.

[136] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[137] Tallai Andherbertrobbins Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[138] Chao Lan and Jun Huan. Discriminatory transfer. *CoRR*, 2017. URL http://arxiv.org/abs/1707.00780.

[139] Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 2018.

[140] Min Kyung Lee and Su Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017.

[141] Min Kyung Lee, Anuraag Jain, Hae Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. In *Proc. ACM Hum.-Comput. Interact., 3, CSCW*. ACM, 2019.

[142] Naomi Lee. Trouble on the radar. *The Lancet Technology*, 2014.

[143] Julia Levashina, Christopher J Hartwell, Frederick P Morgeson, and Michael A Campion. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, 67(1):241–293, 2014.

[144] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670. ACM, 2010.

[145] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms.

In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

[146] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*, 2018.

[147] Jing Wu Lian, Nicholas Mattei, Renee Noble, and Toby Walsh. The conference paper assignment problem: Using order weighted averages to assign indivisible goods. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[148] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–520. Association for Computational Linguistics, 2011.

[149] Kathryn P. Linthicum, Katherine Musacchio Schafer, and Jessica D. Ribeiro. Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*, 37(3):214–222, may 2019. ISSN 0735-3936.

[150] Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[151] Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *RecSys*, 2014.

[152] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML)*, 2017.

[153] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning (ICML)*, 2016.

[154] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

[155] David E. Losada, Fabio Crestani, and Javier Parapar. Overview of eRisk: Early risk prediction on the internet. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Verlag, 2018.

[156] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[157] Paul E Lovejoy. *Transformations in slavery: a history of slavery in Africa*, volume 117. Cambridge University Press, 2011.

[158] Thomas Lux, Randall Pittman, Maya Shende, and Anil Shende. Applications of supervised learning techniques on undergraduate admissions data. In *Proceedings of the ACM International Conference on Computing Frontiers*, pages 412–417, 2016.

[159] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1910–1919, 2018.

[160] Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, 2018.

[161] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

[162] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *International Conference on Machine Learning (ICML)*, 2018.

[163] Suvradip Maitra. Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2020.

[164] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

[165] Jérémie Mary, Romaric Gaudel, and Philippe Preux. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data*, pages 325–336, 2015.

[166] Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using Twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics*, 2016.

[167] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, 2016.

[168] David N. Milne, Kathryn L. McCabe, and Rafael A. Calvo. Improving moderator responsiveness in online peer support through automated triage. *Journal of Medical Internet Research*, 2019.

[169] Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

[170] Michael Mitzenmacher and Eli Upfal. *Probability and computing: randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[171] Torin Monahan. Editorial: surveillance and inequality. Technical report, 2008. URL http://www.surveillance-and-society.org.

[172] Ritesh Noothigattu, Snehalkumar Neil S. Gaikwad, Edmond Awad, Sohan D'Souza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In *AAAI*, 2018.

[173] JC Nunnally. *Psychometric Theory.* McGraw-Hill, 2nd edition, 1978.

[174] Executive Office of the President, Cecilia Munoz, Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). *Big data: A report on algorithmic systems, opportunity, and civil rights.* Executive Office of the President, 2016.

[175] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2017.

[176] Kevin A. Padrez, Lyle Ungar, Hansen Andrew Schwartz, Robert J. Smith, Shawndra Hill, Tadas Antanavicius, Dana M. Brown, Patrick Crutchley, David A. Asch, and Raina M. Merchant. Linking social media and medical record data: A study of adults presenting to an academic, urban emergency department. *BMJ Quality and Safety*, 2016.

[177] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[178] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *arXiv preprint arXiv:1907.10516*, 2019.

[179] M Scott Peck. *The road less traveled: A new psychology of love, traditional values, and spiritual growth.* Simon and Schuster, 2002.

[180] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[181] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[182] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 935–951, 2017.

[183] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

[184] Richard A Posthuma, Frederick P Morgeson, and Michael A Campion. Beyond employment interview validity: A comprehensive narrative review of recent

research and trends over time. *Personal Psychology*, 55(1):1–81, 2002.

[185] Lijing Qin and Xiaoyan Zhu. Promoting diversity in recommendation by entropy regularizer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[186] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 784–791, 2008.

[187] Anshuka Rangi and Massimo Franceschetti. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1345–1352, 2018.

[188] J. Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[189] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343. IEEE, 2019.

[190] 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.

[191] Dan-Olof Rooth. Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3):523–534, 2010.

[192] Francesca Rossi and Nicholas Mattei. Building ethically bounded ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9785–9789, 2019.

[193] Mark A. Rothstein and Gil Siegal. Health Information Technology and Physicians' Duty to Notify Patients of New Medical Developments. *Houston Journal of Health Law & Policy*, 2012.

[194] Cynthia Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine*, 2013.

[195] Toni Schmader, Jessica Whitehead, and Vicki H. Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, page 509–514, 2007.

[196] Neal Schmitt. Social and situational determinants of interview decisions: Implications for the employment interview. *Personal Psychology*, 29(1):79–101, 1976.

[197] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. The diverse cohort selection problem. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 601–609, 2019.

[198] Candice Schumann, Zhi Lang, Jeffrey S. Foster, and John P. Dickerson. Making the cut: A bandit-based approach to tiered interviewing. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[199] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P. Dickerson. Group fairness in bandit arm selection. In *NeurIPS Workshop on Machine Learning and Causal Inference for Improved Decision Making*, 2019.

[200] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[201] Chaofeng Sha, Xiaowei Wu, and Junyu Niu. A framework for recommending relevant and diverse items. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3868–3874, 2016.

[202] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, 2018.

[203] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228. ACM, 2018.

[204] Adish Singla, Eric Horvitz, Pushmeet Kohli, and Andreas Krause. Learning to hire teams. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015.

[205] Adish Singla, Sebastian Tschiatschek, and Andreas Krause. Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[206] Adam Smith. The wealth of nations. *New York: The Modern Library*, 1776.

[207] Society for Human Resource Management. Human capital benchmarking report, 2016.

[208] Keivan G Stassun, Susan Sturm, Kelly Holley-Bockelmann, Arnold Burger, David J Ernst, and Donna Webb. The Fisk-Vanderbilt Master's-to-PhD Bridge Program: Recognizing, enlisting, and cultivating unrealized or unrecognized potential in underrepresented minority students. *American Journal of Physics*, 79(4):374–379, 2011.

[209] Richard S. Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2017.

[210] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[211] Wei Tang and Chien-Ju Ho. Bandit learning with biased human feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1324–1332, 2019.

[212] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[213] Frances Trix and Carolyn Psenka. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14 (2):191–220, 2003.

[214] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

[215] New York Civil Liberties Union. Stop-and-frisk data, 2018.

[216] United States Bureau of Labor Statistics. Job openings and labor turnover. https://www.bls.gov/news.release/pdf/jolts.pdf, 2018.

[217] Pelin Vardarlier, Yalcin Vural, and Semra Birgun. Modelling of the strategic recruitment process by axiomatic design principles. pages 374–383, 09 2014.

[218] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[219] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 864–872, 2019.

[220] Austin Waters and Risto Miikkulainen. Grade: Machine learning support for graduate admissions. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1479–1486, 2013.

[221] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 2016.

[222] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 2019.

[223] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, 2019.

[224] Laura Gollub Williamson, James E Campion, Stanley B Malos, Mark V Roehling, and Michael A Campion. Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology*, 82(6):900, 1997.

[225] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[226] Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, pages 433–441, 2015.

[227] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262, 2016.

[228] Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[229] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.

[230] H Peyton Young. *Equity: in theory and practice.* Princeton University Press, 1995.

[231] Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2483–2491, 2011.

[232] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 962–970, 2017.

[233] Baobao Zhang and Allan Dafoe. U.s. public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, 2020.

[234] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018. URL http://arxiv.org/abs/1801.07593.

[235] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2487–2495, 2019.

[236] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, 2019.