1st Symposium on Foundations of Responsible Computing

FORC 2020, June 1, 2020, Harvard University, Cambridge, MA, USA (virtual conference)

Edited by

Aaron Roth

LIPIcs – Vol. 156 – FORC 2020 www.dagstuhl.de/lipics

Editors

Aaron Roth
University of Pennsylvania, Philadelphia, PA, USA
aaroth@cis.upenn.edu

ACM Classification 2012
Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Theory of database privacy and security; Theory of computation

ISBN 978-3-95977-142-9

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/978-3-95977-142-9.

Publication date
May, 2020

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://portal.dnb.de.

Digital Object Identifier: 10.4230/LIPIcs.FORC.2020.0

ISBN 978-3-95977-142-9 **ISSN 1868-8969** **https://www.dagstuhl.de/lipics**

LIPIcs – Leibniz International Proceedings in Informatics

LIPIcs is a series of high-quality conference proceedings across all fields in informatics. LIPIcs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

ISSN 1868-8969

https://www.dagstuhl.de/lipics

Contents

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Preface

With the rise of the consumer internet, algorithmic decision making became personal. Beginning with relatively mundane things like targeted advertising, machine learning was brought to bear to make decisions *about people* (e.g. which ad to show them), and was trained on enormous datasets of personal information that we increasingly generated unknowingly, as part of our every-day "digital exhaust". In the last several years, these technologies have been deployed in increasingly consequential domains. We no longer just use machine learning for targeting ads. We use it to inform criminal sentencing, to set credit limits and approve loans, and to inform hiring and compensation decisions. All of this means that it is increasingly urgent that our automated decision-making upholds social norms like "privacy" and "fairness" that we are accustomed to thinking about colloquially and informally, but are difficult to define precisely enough to encode as constraints on algorithms. It also means that we must grapple with strategic interactions, as changes in our algorithms lead to changes in the behavior of the users whose data the algorithms operate on.

It is exactly because the definitions are so difficult to get right that strong theoretical foundations are badly needed. We need definitions that have meaningful semantics, and we need to understand both the limits of our ability to design algorithms satisfying these definitions, and the tradeoffs involved in doing so. Foundations of Responsible Computing is a venue for developing this theory.

Our first program is a great example of the kind of work we aim to feature. We have 17 accepted papers, 10 of which appear in this proceedings (we allow authors to opt to instead have a one-page abstract appear on the website but not in the proceedings, to facilitate different publication cultures). The program includes formal proposals for how to reason about different kinds of privacy that fall short of differential privacy, but that we much reckon with because of legal or other practical realities. It includes work studying the implications of imposing fairness constraints in the presence of faulty data. It contains work aimed at making strong but to-date impractical fairness constraints more actionable. And it contains papers studying the strategic and game theoretic effects of deployed algorithms. This is all to say, our inaugural conference has much to say about the foundations of computation in the presence of pressing social concerns.

Finally, let me note that our program committee finished their work in the midst of a historic global pandemic, that has and continues to disrupt all of our lives. Despite this, they did a remarkable job. The ongoing pandemic will mean that we cannot meet in person for FORC 2020, but it will not lessen the impact of the work, now to be presented in a remote format.

Aaron Roth
Philadelphia, PA
April 12, 2020

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Efficient Candidate Screening Under Multiple Tests and Implications for Fairness

Lee Cohen

Tel Aviv University, Israel
leecohencs@gmail.com

Zachary C. Lipton

Carnegie Mellon University, Pittsburgh, PA, USA
Amazon AI, Palo Alto, CA, USA
zlipton@cmu.edu

Yishay Mansour

Tel Aviv University, Israel
Google Research, Tel Aviv, Israel
mansour.yishay@gmail.com

Abstract

When recruiting job candidates, employers rarely observe their underlying skill level directly. Instead, they must administer a series of interviews and/or collate other noisy signals in order to estimate the worker's skill. Traditional economics papers address screening models where employers access worker skill via a single noisy signal. In this paper, we extend this theoretical analysis to a multi-test setting, considering both Bernoulli and Gaussian models. We analyze the optimal employer policy both when the employer sets a fixed number of tests per candidate and when the employer can set a dynamic policy, assigning further tests adaptively based on results from the previous tests. To start, we characterize the optimal policy when employees constitute a single group, demonstrating some interesting trade-offs. Subsequently, we address the multi-group setting, demonstrating that when the noise levels vary across groups, a fundamental impossibility emerges whereby we cannot administer the same number of tests, subject candidates to the same decision rule, and yet realize the same outcomes in both groups. We show that by subjecting members of noisier groups to more tests, we can equalize the confusion matrix entries across groups, seemingly eliminating any disparate impact concerning outcomes.

2012 ACM Subject Classification Social and professional topics → Computing / technology policy; Mathematics of computing → Probabilistic inference problems; Computing methodologies → Unsupervised learning

Keywords and phrases algorithmic fairness, random walk, inference

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.1

Related Version A full version of the paper is available at [7], `https://arxiv.org/abs/1905.1136`.

Funding *Lee Cohen*: This work was supported in part by The Yandex Initiative for Machine Learning.
Zachary C. Lipton: This work was supported by the AI Ethics and Governance Fund.
Yishay Mansour: This work was supported in part by a grant from ISF.

1 Introduction

Consider an employer seeking to hire new employees. Clearly, the employer would like to hire the best employees for the task, but how will she know which are best fit? Typically, the employee will gather information on each candidate, including their education, work history, reference letters, and for many jobs, they will actively conduct interviews. Altogether, this information can be viewed as the *signal* available to the employer.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 1; pp. 1:1–1:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Suppose that candidates can be either *skilled* or *unskilled*. If the firm hires an "unskilled" candidate, it will incur a significant cost on account of lost productivity. For this reason, the employer would like to minimize the number of *False Positive* mistakes, instances where *unskilled* candidates are hired. On the other hand, the employer desires not to *overspend* on the hiring process, limiting the number of interviews per hired candidate (either on average, or absolutely). However, fewer interviews weakens the signal, causing the employer to make more mistakes. At the heart of our model is this inherent trade-off between the quality of the signal and the cost of obtaining the signal. This marks a departure from the classical economics literature, in which the signal is commonly regarded as a given.

Complicating matters, hiring efficiency is not the only desiderata at play. In society, candidates belong to various *demographic groups*, and we may strive to design policies that are in some sense *fair* vis-a-vis group membership. While *fairness* can be an elusive notion, regulators must translate it to concrete rules and laws. In the United States, a body of anti-discrimination law dating to the Civil Rights act of 1964, subjects decisions that result in disparate outcomes (as delineated by race, age, gender, religion, etc.) to extra scrutiny: employers must not only show that preference for some groups over others did not drive the decision (disparate treatment doctrine) but also justify that any observed disparities arise from a business necessity (disparate impact doctrine), whether or not those disparities were intentional.

In this paper, we seek to understand how a complex hiring process would interact with the requirements of fairness. We extend the theory on candidate screening and statistical discrimination, addressing the setting in which employers can subject employees to multiple tests, which we assume to be conditionally independent given the worker's skill level and group membership. To build intuition, most of our analysis focuses on a Bernoulli model of both worker skill and screening. Additionally, we extend the traditionally-studied Gaussian skill and screening models to the multi-test setting (Section 5).

Unlike the classical papers, in which an employer's hiring policy is given by a simple thresholding rule, our setting requires greater care to derive the optimal employer policy. In our setting, we imagine that the employer wishes to minimize the number of tests performed subject to a constraint upper-bounding the false positive rate. We characterize the optimal policy in this case as a randomized threshold policy. Subsequently, we show that this is not always an optimal policy and consider the setting in which employers can allocate tests dynamically. Namely, employers decide after each result whether to (i) hire the candidate; (ii) reject the candidate and move on to the next one; or (iii) administer a subsequent test. In the Bernoulli case, the optimal policy consists of administering tests until each candidate's posterior likelihood of being a high-skilled worker either dips below the prior or rises above a threshold determined by the tolerable false positive rate. We reduce the analysis of this process to a random walk over the log posterior odds and derive the solution via the corresponding Gambler's ruin problem.

We consider the ramifications for fairness within our model when employees, despite possessing similarly-distributed skills, are evaluated with differing noise levels. We show impossibility results, as well as, a solution to equalize confusion matrix entries by adjusting the number of tests according to group parameters. Finally, we present a simple way to estimate group parameters without knowing the true skill levels (i.e., unsupervised learning), and give bounds in terms of the number of candidates from a group for good estimation.

1.1 Related work

The classical economics literature on discrimination in employment can broadly be divided into two focuses. The *taste-based discrimination* model due to [4] models the market outcomes in a setting where employers express an explicit preference for hiring members of one group, acting as if an employee's demographic membership provides utility. This preference for certain groups induces a sorting of employees from the disadvantaged group towards those employers who discriminate the least with wages ultimately determined by the *marginal discriminator*. Subsequently, [17] suggested a statistical mechanism by which similarly-skilled employees from different groups might experience differential outcomes: the comparative difficulty of screening from one group vs. another. Many subsequent works extend this analysis, typically focusing on Gaussian models of worker quality and conditionally-Gaussian test scores [2, 1]. These papers consider the setting where workers are assessed via a single test characterized by a group-dependent noise level. Our work is differentiated from these by considering richer mechanisms for acquiring signal.

In the more recent literature on fairness in machine learning, researchers often focus on binary classification, with employees characterized by a protected characteristic (group membership), and other (non-protected) covariates [16, 13, 14]. There, the predictor is presumably used to guide a consequential decision, such as allocating some economic good (loans, jobs, etc.) [8] or assessing some penalty (e.g. risk scores to guide bail decisions) [6]. Papers then focus on various interventions for ensuring accurate prediction subject to various constraints such as demographic parity (outcomes independent of group membership), blindness (model cannot observe group membership), and equalized false negative and/or false positive rates [11]. Several simple impossibility results preclude simultaneously satisfying several combinations of these parities [5, 6, 15]. More recently, a number of papers have drawn inspiration from economic modeling, extending the literature on fairness in classification to consider longer-term dynamics, equilibria, and the emergence of feedback loops [12, 11, 9]. Finally, [3, 19] provide a survey of definitions from the algorithmic fairness literature.
Unrelated to fairness, [18] consider a model that is somehow resembles to ours in the context of A/B testing. They minimize the expected time per discovery (which can be viewed as hire) from an infinite pool of hypotheses (which can be viewed as candidates) with a bounded false discovery rate.

2 The Bernoulli Model

We formalize our problem as follows. An employer accesses an infinite pool of candidates (indexed by $i \in \mathbb{N}^+$), each of which has some (hidden) *skill level* $y_i \in \{0, +1\}$, which denote *unskilled* and *skilled*, respectively. Underlying worker skill levels y_i are sampled independently from a Bernoulli distribution with parameter p. An employer can access information about the i-th candidate through a sequence of τ tests, which are conditionally independent given y_i. Each *test result*, $\hat{y}_{i,j} \in \{0, +1\}$ disagrees with the ground truth skill with probability $\Pr[\hat{y}_{i,j} \neq y_i] = \frac{1-\sigma}{2}$, where $\sigma \in (0, 1)$, i.e., $\hat{y}_{i,j} = y_i \oplus Br(\frac{1-\sigma}{2})$[1]. For convenience, we denote the noise level as $\eta = \frac{1-\sigma}{2} \in (0, \frac{1}{2})$. We say that a test result $\hat{y}_{i,j}$ is *flipped* if $\hat{y}_{i,j} \neq y_i$, and the number of flipped results for a given candidate is denoted by Z^η_τ is $Z^\eta_\tau = \sum_{j=1}^{\tau} \mathbb{I}(\hat{y}_{i,j} \neq y_i)$, where $\mathbb{I}(\cdot)$ is the indicator function.

[1] $\oplus$ is the XOR operation between two binary random variables, and therefore $\hat{y}_{i,j}$ is also a random variable.

The employer decides weather or not to hire the current candidate, but unlike the secretary problem she can hire as many as she desires. A *selection criterion* is a mapping between test results of a single candidate to actions: $\texttt{Select}(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau_i}) \in \{0,1\}$, where 0 means *reject* and 1 means *accept* (hire). A *policy* π sets the selection criteria based on σ, p and other possible constraints such as probability to hire, error probability, etc. A *randomized threshold policy* is a policy π with parameters (τ,θ,r) such that $\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau_i}) = 1$ for $S_\tau := \sum_{i=1}^{\tau} \hat{y}_{i,j} > \theta$, $\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau_i}) = 0$ for $S_\tau < \theta$, and for $S_\tau = \theta$ the probability that $\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau_i}) = 1$ is r. We call a policy π a *threshold policy* if $r = 1$. In a *dynamic policy*, rather than setting a fixed number of tests per candidate, the employer may decide after each test whether to *accept*, *reject*, or to perform an additional test, i.e., $\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau_i}) \in \{0,1,\texttt{more}\}$. Note that for a dynamic policy, the number of tests τ is a random variable determined based on the tests' outcomes. When designing a policy, one must carefully consider the balance between the following desiderata:

1. **Minimize False Discovery Rate (FDR)** – the fraction of unskilled workers among the accepted candidates, i.e., $\mathrm{FDR} := \Pr[y_i = 0|\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = +1]$.
2. **Minimize False Omission Rate (FOR)** – the fraction of skilled workers among the rejected candidates, i.e., $\mathrm{FOR} := \Pr[y_i = +1|\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 0]$.
3. **Minimize False Negatives (FN)** – the amount of skilled workers classified as unskilled.
4. **Minimize False Positives (FP)** – the amount of unskilled workers classified as skilled.
5. **Ratio of accept probability and number of tests** – the number of tests performed per candidate hired, using a parameter $B > 1$, we have $\frac{\tau}{B} \leq \Pr[\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = +1]$.

For any fixed number of tests τ, increasing the threshold θ of a threshold policy decreases FDR while increasing FOR.

Loss: To handle the trade-off between the false positives, (i.e., when an unskilled candidate is accepted) and false negatives (i.e., when a skilled candidate is rejected), we introduce an α-loss, paramaterized by $\alpha \in [0,1]$ and defined as follows:

$$\ell_\alpha(b_1,b_2) = \alpha \cdot \mathbb{I}[b_1 = 0, b_2 = 1] + (1-\alpha)\cdot \mathbb{I}[b_1 = 1, b_2 = 0]$$

where $\mathbb{I}[\cdot]$ is the indicator function and $b_1, b_2 \in \{0,1\}$. The expected loss of a policy π is,

$$l_\alpha(\pi) = \mathbb{E}[\ell_\alpha(y_i, \pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}))] \tag{1}$$

where the expectation is over the type of the candidates y_i, the test results $\hat{y}_{i,j}$, and the decisions of π.

3 Analysis of the Bernoulli Model with One Group

To begin, we analyze this hiring model for a single group of candidates. The employer's goal is to minimize the expected loss, $l_\alpha(\pi)$, while maintaining a given acceptance probability. For brevity, we relegate all proofs to the Appendix.

3.1 The Simple Threshold Policy (Equal Number of Tests)

Consider the setting where the employer must subject all candidates to an equal number of tests τ and threshold θ (these parameters are chosen by the employer but thereafter constant across candidates). For a given threshold, we can relate the flip probability (error rate) of the test to the probability that a candidate is accepted as follows:

Recall that $\hat{y}_{i,j} = y_i \oplus Br(\eta)$, $S_\tau = \sum_{j=1}^{\tau} \hat{y}_{i,j}$, that $Z^\eta_\tau = \sum_{t=1}^{\tau} \mathbb{I}(\hat{y}_{i,j} \neq y_i)$, and that τ and θ are the only parameters of the threshold policy, π. Informally, S_τ is the number of passed tests and Z^η_τ is the number of flips (tests in error). The probability of hiring an unskilled candidate is given by:

$$\Pr[\pi(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1 | y_i = 0] = \Pr[S_\tau \geq \theta | y_i = 0] = \Pr[Z^\eta_\tau \geq \theta].$$

Since Z^η_τ is a binomial random variable with parameters τ and η, we can calculate this probability precisely as:

$$\Pr[\pi(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1 | y_i = 0] = \Pr[Z^\eta_\tau \geq \theta] = \sum_{k=\theta}^{\tau} \binom{\tau}{k} \eta^k (1-\eta)^{\tau-k},$$

and the probability of rejecting a skilled candidate is the probability that they encounter more than $\tau - \theta$ flips, thus:

$$\begin{aligned}\Pr[\pi(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 0 | y_i = +1] &= \Pr[S_\tau < \theta | y = +1] = \Pr[Z^\eta_\tau > \tau - \theta] \\ &= \sum_{k=\tau-\theta+1}^{\tau} \binom{\tau}{k} \eta^k (1-\eta)^{\tau-k}.\end{aligned}$$

Similarly, given a candidate's skill level, we can calculate the probability that they obtain exactly k positive tests out of τ, i.e,

$$\Pr[S_\tau = k | y_i = 0] = \Pr[Z^\eta_\tau = k] = \binom{\tau}{k} \eta^k (1-\eta)^{\tau-k}.$$

$$\Pr[S_\tau = k | y_i = +1] = \Pr[Z^\eta_\tau = \tau - k] = \binom{\tau}{k} \eta^{\tau-k} (1-\eta)^{k}.$$

Given these observations, we can now analyze the employer's choices.

Optimal solution for any ratio $\alpha \in (0, 1)$

The next theorem shows that for threshold policies, the expected loss $l_\alpha(\pi) = l_\alpha(\theta)$ is minimized at $\theta^*_{p,\alpha}$ such that $|\theta^*_{p,\alpha} - \tau/2| \leq \frac{\log(\frac{1}{p}) + \log(\frac{1}{\alpha})}{2\log(1+\frac{2\sigma}{1-\sigma})}$.

▶ **Theorem 1.** *The loss function $l_\alpha(\theta)$ is quasi-convex and a threshold of*

$$\theta^*_{p,\alpha} = \arg\min_\theta l_\alpha(\theta) = \left\lceil \frac{\tau}{2} - \frac{\log(\frac{1}{p} - 1) + \log(\frac{1}{\alpha} - 1)}{2\log(1 + \frac{2\sigma}{1-\sigma})} \right\rceil$$

minimizes loss for any values of $\alpha, p, \sigma \in (0, 1)$.

Next, we bound the number of tests required to guarantee that the probability of classification error by the majority decision rule (i.e., $\theta = \lceil \frac{\tau}{2} \rceil$) does not exceed a specified quantity δ.

▶ **Theorem 2.** *For every $\delta, p, \alpha \in (0, 1)$, performing $\tau = \Omega(\frac{\alpha+p-2p\alpha}{\sigma^2} \ln(\frac{1}{\delta}))$ tests per candidate and using majority as a decision rule (i.e., $\theta = \tau/2$) guarantees $l_\alpha(\pi) \leq \delta$.*

Equal cost for false positives and false negatives ($\alpha = \frac{1}{2}$)

Consider the simple loss consisting of the classification error rate (false positives and false negatives count equally), expressed via our loss function by setting $\alpha = \frac{1}{2}$. When skilled and unskilled candidates occur with equal frequency, i.e., $p = 1/2$, we can derive that the majority decision rule minimizes the classification error for any number of tests.

▶ **Corollary 3.** *Assume $p = 1/2$ and $\alpha = 1/2$. For any number of tests τ, the majority decision rule minimizes loss l_α. Namely, $\arg\min_\theta l_{\frac{1}{2}}(\theta) = \lceil \frac{1}{2}\tau \rceil$. In addition, for every $\delta \in (0,1)$, performing $\tau = \Omega(\frac{1}{\sigma^2}\ln(\frac{1}{\delta}))$ tests per candidate and using majority as a decision rule guarantees classification error with probability of at most δ.*

FDR minimization with limited number of tests per hire for balanced groups

Again, assuming balanced groups (i.e., $p = 1/2$), suppose that an employer would like to minimize the false discovery rate, subject to the constraint of lower bounding the hiring probability. We can model this optimization problem by introducing a budget parameter $B > 1$ to bound any predetermined (fixed) number of tests per hired candidate as follows:

$$\begin{aligned} &\arg\min_\pi && \mathrm{FDR}_\pi = \Pr[y_i = 0 | \Pr[\pi(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1] \\ &\text{subject to} && \frac{\tau_\pi}{\Pr[\pi(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1]} \leq B \end{aligned} \tag{2}$$

where τ_π is the number of tests π performs. The following theorem shows that the optimal policy is a randomized threshold policy.

▶ **Theorem 4.** *There exists a randomized threshold policy π which is an optimal solution for (2).*

3.2 The Dynamic Policy (Adaptively-Allocated Tests)

Recall that under a dynamic policy, the employer can decide after each test whether to accept, reject, or perform another test. In general, dynamic policies are more efficient than those that must set a fixed number of tests. To build intuition, consider a candidate that has passed 2 out of 3 tests. As seen above, under an optimally-constructed fixed-test policy, any candidate that fails a single test might be rejected.[2] However, the posterior probability that this candidate is in fact *skilled* may still be greater than that of a fresh candidate sampled from the pool. Thus we can improve on the fixed-test policy by dynamically allocating more tests to candidates until their posterior odds either dip below the prior odds or rise above the threshold for hiring. The following theorem formalizes this notion that it is better to administer more tests to a candidate that passed the majority of previous tests than to start afresh with a new candidate:

▶ **Theorem 5.** *For any p, σ, τ, a candidate i that passed $\theta > \frac{\tau}{2}$ out of τ tests is more likely to be a skilled than a freshly-sampled candidate i' for whom no test results are yet available, i.e., $\Pr[y_{i'} = +1] = p < \Pr[y_i = +1 | S_\tau = \theta]$.*

▶ Remark 6. If $\theta < \frac{\tau}{2}$, the inequality would have been reversed.

[2] For example, if $B = 18$ and $\eta = \frac{1}{3}$, the lowest false discovery rate is achieved by $\tau = \theta = 3$.

The Greedy Policy

We now present a greedy algorithm that continues to test a candidate so long as the posterior probability that $y_i = +1$ is greater than ϵ' and smaller than $1-\epsilon$, rejects a candidate whenever the posterior falls below ϵ' (absent fairness concerns, employers will set $\epsilon' = p$ for all groups), and accepts whenever the posterior rises above $1-\epsilon$. Given parameters $\epsilon, \epsilon' > 0$, we show that the greedy policy solves the optimization problem of minimizing the mean number of tests under these constraints, i.e.,

$$\begin{aligned}
&\underset{\tau}{\text{minimize}} && \mathbb{E}[\tau] \\
&\text{subject to} && \forall_i \pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 1 \text{ iff } \Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}] \geq 1-\epsilon \\
& && \forall_i \pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 0 \text{ iff } \Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}] < \epsilon'
\end{aligned}$$

Our analysis of this policy builds upon the observation that conditioned on a worker's skill, the posterior log-odds after each test perform a one-dimensional random walk, starting with the prior log-odds $\log(\frac{p}{1-p})$ and moving, after each test result, either left (upon a failed test) or right (upon a passed test). When (as in our model) the probability of a flip are equal for skilled and unskilled candidates, our random walk has a fixed step size. Moreover, our random walk has *absorbing barriers* corresponding to (when $\epsilon' = p$) falling below the prior log odds (on the left) and exceeding the hiring threshold (on the right). Owing to the fixed step size and absorbing barriers, our policy resembles the classic problem of Gambler's ruin, in which a gambler wins or loses a unit of currency at each step, and loses when crossing a threshold on the left (going bankrupt) or on the right (bankrupting the opponent). We formalize the random walk as follows where X_j is the position on the walk at time j:

1. X_0 is the prior log-odds of the candidate, i.e., $X_0 = \log\frac{p}{1-p}$.
2. After each test result, $\hat{y}_{i,j}$ is observed, $X_j = X_{j-1} + (2\hat{y}_{i,j} - 1)\cdot\log\left(\frac{\Pr[\hat{y}_{i,j}=+1|y_i=+1]}{\Pr[\hat{y}_{i,j}=+1|y_i=0]}\right)$.

Let π_{Greedy} be the policy that accepts a candidate if $\Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,j}] \geq 1-\epsilon$, rejects if $\Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,j}] < \epsilon'$, and otherwise conducts an additional test, i.e.,

$$\pi_{Greedy}(\hat{y}_{i,1},\ldots,\hat{y}_{i,j}) = \begin{cases} 0 & \text{if } \Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,j}] < \epsilon' \\ 1 & \text{if } \Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,j}] \geq 1-\epsilon\ . \\ \text{retest} & \text{else} \end{cases}$$

An employer will generally set the lower absorbing barrier to reject all candidates with posterior log odds less than p since a fresh candidate from the pool is expected to be better. However, when noise levels differ across groups, we may prefer *in the interest of fairness* to set ϵ' lower than p for members of the noisier group, allowing us to equalize the frequency of false negatives across groups (see Section 4).

▶ **Lemma 7.** *Let $\beta, \beta' \in \mathbb{R}$ be the parameters that satisfy $\frac{\beta}{\beta+1} = 1-\epsilon$ and $\frac{\beta'}{\beta'+1} = \epsilon'$ (i.e., $\beta = \frac{1-\epsilon}{\epsilon}$and $\beta' = \frac{\epsilon'}{1-\epsilon'}$). Then $X_\tau \geq \log\beta$ iff $\Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}] \geq 1-\epsilon$ (iff the candidate is accepted) and $X_\tau < \log\beta'$ iff $\Pr[y_i = +1|\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}] < \epsilon'$ (iff the candidate is rejected).*

▶ **Corollary 8.** *The policy π_{Greedy} can be described as follows.*

$$\pi_{Greedy}(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = \begin{cases} 0 & \textit{if } X_\tau < \log\frac{\epsilon'}{1-\epsilon'} \\ 1 & \textit{if } X_\tau \geq \log(\frac{1-\epsilon}{\epsilon}) \\ \text{retest} & \text{else} \end{cases}$$

Table 1 Confusion matrix for π_{greedy} assuming $\epsilon \leq 1/4$ and $\epsilon' \leq p \leq 1/2$.

	General ϵ'		**When $\epsilon' = p$**	
	Skilled ($y_i = +1$)	**Unskilled** ($y_i = 0$)	**Skilled**	**Unskilled**
accept	$\text{TPR} = \Theta\left(1 - \frac{\epsilon'}{p}(1-\sigma)\right)$	$\text{FPR} = \Theta\left(\epsilon(p - \epsilon' + \epsilon'\sigma)\right)$	$\Theta(\sigma)$	$\Theta(\epsilon p \sigma)$
reject	$\text{FNR} = \Theta\left(\frac{\epsilon'}{p}(1-\sigma)\right)$	$\text{TNR} = \Theta\left(1 - \epsilon(p - \epsilon' + \epsilon'\sigma)\right)$	$\Theta(1-\sigma)$	$\Theta(1 - \epsilon p \sigma)$

We use the following parameters in the next theorems:

$$a = \left\lceil \frac{\log(\frac{(1-\epsilon)(1-\epsilon')(1+\sigma)}{\epsilon\epsilon'(1-\sigma)})}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil \gg \frac{1}{\sigma} \quad \text{and} \quad z = \left\lceil \frac{\log(\frac{p(1-\epsilon')(1+\sigma)}{\epsilon'(1-p)(1-\sigma)})}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil$$

▶ **Theorem 9** (Expected number of tests per type). *The expected number of tests until a decision (namely accept or reject) for skilled candidates is* $\mathbb{E}[\tau_s] = \frac{1}{\sigma}\left(a \cdot \frac{1-(\frac{1-\sigma}{1+\sigma})^z}{1-(\frac{1-\sigma}{1+\sigma})^a} - z\right) \approx \frac{2a}{1+\sigma} - \frac{z}{\sigma}$ *and* $\mathbb{E}[\tau_u] = \frac{1}{\sigma}\left(z - a \cdot \frac{1-(\frac{1+\sigma}{1-\sigma})^z}{1-(\frac{1+\sigma}{1-\sigma})^a}\right) \approx \frac{z}{\sigma}$ *for unskilled candidates.*

For the probabilities of the candidates to be accepted or rejected, conditioned on their true skill level, we present the results in a form of confusion matrix in Table 1.

▶ **Theorem 10.** *The expected number of tests until deciding whether to accept or reject a candidate is* $\mathbb{E}[\tau | \pi(y_{i,\tau}) \in \{0,1\}] \approx \frac{ap}{\sigma}$, *where* $a \gg \frac{1}{\sigma}$.

4 Fairness Considerations in the Two-Group Setting

Two Groups – Threshold Policies

We now discuss the effects of a threshold policy when candidates belong to two groups, G_1 and G_2 whose skill level is distributed identically, but whose tests are characterized by different noise levels. Without loss of generality, we assume that $\eta_1 < \eta_2$, where η_i is the probability that a test result of a candidate from G_i is different from his skill level. To begin, we note the fundamental irreconcilability of equalizing either the false positive or the false negative rates across groups with subjecting candidates to the same policy.

▶ **Theorem 11** (Impossibility result). *When noise levels differ between two groups with identical skill level distribution, a single Threshold Policy* π *(with the same number of tests* τ *and the same threshold* θ *for both groups) cannot have equality in either the false negative rates or in the false positive rates across the groups. Particularly, there is a higher false positive rate in the noisier group, as an unskilled candidate from* G_2 *is more likely to be accepted by the threshold policy than an unskilled candidate from* G_1*:*

$$\text{FPR}^{\eta_1}_{\theta,\tau} = \Pr_{\eta_1}[\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 1 | y_i = 0] < \Pr_{\eta_2}[\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 1 | y_i = 0] = \text{FPR}^{\eta_2}_{\theta,\tau},$$

and also a higher false negative rate, as a skilled candidate from G_2 *is more likely to be rejected than a skilled candidate from* G_1*:*

$$\text{FNR}^{\eta_1}_{\theta,\tau} = \Pr_{\eta_1}[\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 0 | y_i = +1] < \Pr_{\eta_2}[\pi(\hat{y}_{i,1},\ldots,\hat{y}_{i,\tau}) = 0 | y_i = +1] = FNR^{\eta_2}_{\theta,\tau}.$$

Connection to Economics Literature. Aigner and Cain [1] discuss a similar case under a Gaussian screening model where the variance (noise level) of the single test differs across the two groups. Similarly, they note that qualified candidates fare worse in the noisy group but that unqualified candidates fare better in the noisier group. Our work differs from theirs in that we consider the effect of multiple tests and the ability to optimize over the number of tests.

Two Groups–Dynamic policy

We now consider the (dynamic) hiring policy in the setting when employees belong to two groups, G_1 and G_2 with identically-distributed skills but different noise levels $\eta_1 < \eta_2$. We note that there are two ingredients that explain the differences among the groups: (i) The step size, $\log\left(\frac{\Pr[\hat{y}_{i,j}=+1|y_i=+1]}{\Pr[\hat{y}_{i,j}=+1|y_i=0]}\right) = \log\left(\frac{1-\eta}{\eta}\right)$ of G_2 (the noisier group) is smaller than the step size of G_1. Thus these candidates must typically pass more tests before they are accepted; and (ii) Skilled candidates in group G_2 exhibit less drift to the right (they have a higher probability of failing a test). Consequently, when an employer (rationally) sets $\epsilon' = p$ for all groups, a skilled candidate from G_2 is more likely to be fail a test in step 1, at which point the dynamic policy summarily rejects them. These two facts explain both the higher false negative rates for G_2 and the longer expected duration until acceptance. By setting $\epsilon' < p$ for members of the noisier group, we can equalize false negative rates. Precisely, setting $\epsilon' = \frac{\eta_1}{\eta_2}p$ achieves the desired parity. The cost of this intervention is that it requires more tests for candidates from the noisier group. Here, our random walk analysis can be leveraged to determine exactly how many more. Once again, we cannot provide equality across the groups in all desired ways – the same acceptance criterion, the same expected number of tests, and the same false negative rates between groups – with the noise differs across groups.

5 Gaussian Worker Screening Model

In this section, we work out the analytic solutions for the conditional expectation of worker qualities given a series of conditionally independent tests $Y_1, ...Y_n$ s.t. $\forall i, j$, $Y_i \perp Y_j | Q$. We assume that the worker quality Q normally distributed with mean μ_Q and variance σ_Q^2, so instead of binary skill level we have continuous quality of candidates. Conditioned on $Q = q$, each test is generated according to the structural equation $y_i = q + \eta$, where η is a normally distributed noise term with mean 0 and variance σ_η^2. Equivalently, we can say that the conditional distribution for each test $P(Y|Q = q)$ is Gaussian with mean q and variance σ_η^2. We refer the reader to the full version [7] for further details.

We show that we can equalize conditional variance between the two groups by giving more interviews to noisier group, and that it yields the same conditional expectations.

▶ **Theorem 12.** *For two groups, G_1, G_2 with the same worker quality Q, that differ only in the variance of their noise $\sigma_{\eta_1}^2 < \sigma_{\eta_2}^2$, the variance can be equalized by using $n_2 = \frac{\sigma_{\eta_2}^2}{\sigma_{\eta_1}^2} n_1$ interviews (or tests) for G_2, where n_1 is the number of interviews for each candidate from G_1.*

▶ **Theorem 13.** *When equalizing conditional variances between G_1, G_2 by using $n_2 = \frac{\sigma_{\eta_2}^2}{\sigma_{\eta_1}^2} n_1$, we get the same conditional expectations, $\mathbb{E}_{\eta_1}[Q|Y_1, ..., Y_{n_1}] = \mathbb{E}_{\eta_2}[Q|Y_1, ..., Y_{n_2}]$.*

6 Unsupervised Parameter Estimation

Now, under the assumption of realizable case, we explain how one can estimate the parameters p and σ given tests results from a homogeneous population. Surprisingly, we discover that parameter recovery in this model does not require any ground truth labels indicating whether an employee is skilled or unskilled. We use Hoeffding's inequality to bound the absolute difference between the estimated parameters and the true parameters by choosing δ as the wanted upper bound and solving for the number of samples or ϵ.

▶ **Lemma 14** (Hoeffding's inequality). *Let $y_1, \ldots, y_m$ be σ^2-sub-gaussian random variables. Then, for any $\epsilon > 0$,*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} y_i - \mathbb{E}[y_i]\right| \geq \epsilon\right] \leq 2e^{-m\epsilon^2/2\sigma^2}.$$

If $y_1, \ldots, y_m$ are Bernoulli random variables with parameter p,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} y_i - p\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2}.$$

We start by estimating σ and then use it to derive an estimate for p. The estimated parameters are denoted by $\hat{\sigma}$ and $\hat{p}$. Notice that in order to have any information regarding the true value of σ, we need to have candidates with at least two tests. Hence, from now on we assume exactly that, i.e., $\forall_i \pi_{\text{Greedy}}(\hat{y}_{i,1}) = \mathit{more}$ for dynamic policies and $\tau \geq 2$ for fixed number of tests policies.

Now, in both policies we have showed that the optimal rule is to reject candidates that fail their first test. Therefore inconsistencies between the first two tests are seen only in cases where $\hat{y}_{i,1} = 1, \hat{y}_{i,2} = 0$.

Let c be the number of inconsistencies in the first two tests, i.e., $c = |\{(\hat{y}_{i,1}, \hat{y}_{i,2}) : y_{i,1} \neq y_{i,2}\}|$, and let m be the number of candidates with at least two tests. Since c is generated by sampling m times, the distribution $Br((\frac{1+\sigma}{2})(\frac{1-\sigma}{2})) = Br(\frac{1-\sigma^2}{4})$ and we can estimate σ as stated in the next theorem:

▶ **Theorem 15.** *If we have results from $m \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ candidates, by using $\hat{\sigma} = \sqrt{1 - 4\frac{c}{m}}$, then with probability $1 - \delta$ we have that $|\hat{\sigma} - \sigma| \leq \epsilon$.*

Having an estimation of the parameter $\hat{\sigma}$, we can calculate the estimated p as follows: Let $p_{\hat{y}_{*,1}=1} := \frac{\sum_i \mathbb{I}(\hat{y}_{i,1}=1)}{m}$ be the percentage of positive first tests. Since this number is generated by the distribution $Br(\frac{1}{2}(p(1+\sigma) + (1-p)(1-\sigma))) = Br(\frac{1}{2} + (2p-1)\frac{\sigma}{2})$, we can estimate $\hat{p}$ using the estimated value of $\hat{\sigma}$.

▶ **Theorem 16.** *If we have results from $m \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ candidates, by using $\hat{p} = \frac{2(p_{y_{*,1}=1}-1)+\hat{\sigma}}{\hat{\sigma}}$, we get that with probability $1 - \delta$ we have that $|\hat{p} - p| \leq 2\epsilon$.*

Under the Gaussian screening model, the parameter estimation is also straightforward (assuming realizability) without access to the true skill level of the employees. We start by looking at a single candidate, i. Each of his test results, $\hat{y}_{i,j}$ is generated from a conditional distribution $P(Y_i|Q_i = q_i)$ which is a Gaussian with mean q_i and variance σ_η^2. Since this variance is common among all the candidates, we can simply average the estimated variance of every candidate to get an approximation for σ_η^2. Suppose $\hat{y}_{i,1}, \ldots, \hat{y}_{i,n}$ is a sequence of n i.i.d tests of candidate i, and let $\boldsymbol{y_i} = \frac{1}{n}\sum_{j=1}^{n} y_{i,j}$ be the empirical mean of candidate i's tests.

The following theorem is a result from Hoeffding's Inequality, in which we use to bound the error of our estimated parameters.

▶ **Theorem 17.** *By using the following as estimators for Gaussian parameters* $\hat{\mu}_Q = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{y_i}$, $\hat{\sigma}_\eta^2 = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}(y_{i,j} - \boldsymbol{y_i})^2$ *and* $\hat{\sigma}_Q^2 = \frac{1}{m}\sum_{i=1}^{m}(\hat{\mu}_Q - \boldsymbol{y_i})^2$ *(notice that* $\mathbb{E}[\hat{\sigma}_\eta^2] = \sigma_\eta^2$ *and* $\mathbb{E}[\hat{\sigma}_Q^2] = \sigma_Q^2$*), the difference between each parameter and it's estimator is bounded by* $O(\sqrt{\frac{1}{m}\ln(\frac{1}{\delta})})$.

7 Discussion and Future Work

Consider two groups with identically-distributed skills and characterized by different noise levels in screening. Our results demonstrate that if a regulatory body (e.g., policymakers or a regulator) insists on the same number of tests and the same decision rule for both groups, this would yield higher false positive rates in any threshold policy. As a result, hired candidates from the noisier group would suffer higher rates of firing. In turn, this might lead employers to erroneously conclude that this group's skill level is lower than it actually is. This paper presents a policy that handles this problem by minimizing the false positive rates of both groups, in the form of a greedy policy. Moreover, the greedy policy is efficient, minimizing the expected number of tests per hire among all policies that achieve a specified false positive rate and continue testing every candidates that appear better than the a new one. However, the dynamic policy will still suffer (as does the simple threshold policy) from higher false negative rates for the noisier group, violating a notion of fairness dubbed *equality of opportunity* in the recent literature on fairness in machine learning [11]. We addressed this problem by modifying the greedy policy to reject candidate iff $\Pr[y_i = +1|\hat{y}_{i,1}\ldots\hat{y}_{i,\tau}] < \epsilon'$ by setting $\epsilon' < p$. Our greedy policy can be made forgiving and equalize false negative rates across groups.

Implications for Fairness

When it comes to "business justification", Civil Rights regulation in the United States might be open to more than one interpretation regarding group-based disparities. In disparate impact doctrine, the statistical disparity of interest, e.g., in the famous 4/5 test concerns the decisions itself. In our model, if one were to apply a uniform hiring policy, administering the same number of tests to all applicants and applying the same threshold, a disparate impact might emerge. By subjecting members of noisier groups to more tests, we can equalize the confusion matrix entries across groups, seemingly eliminating any disparate impact concerning outcomes.

However, in this case, both the number of tests administered, and the inferences drawn from the results depend explicitly on group membership, potentially raising concerns about disparate treatment and procedural fairness. Another interesting question might be to consider what disparate doctrine might have to say about disparities not in outcomes but in testing procedures.

Our setup motivates a new dimension to the discussion – even when members of the two groups have statistically identical outcomes, and even putting aside concerns about group-blindness, members of the more heavily-tested group may experience adversity. For example, perhaps these candidates, subject to more interviews, would not be able to interview with as many employers, thus lowering their overall likelihood of finding employment.

It would be interesting to introduce strategic behavior to our setting and understand the implications. For example, the candidates might have a utility that depends on whether they received the job, and disutility associated with how long their interview process was. Their overall utility can simply the difference between the two. Such a strategic model will cause some candidates not to apply, and the stream of candidates applying would have significant different characteristics than the overall population. Such a strategic setting would pose additional fairness challenges, since the mechanism would also control applies and not only who is hired.

References

1 Dennis J Aigner and Glen G Cain. Statistical theories of discrimination in labor markets. *ILR Review*, 30(2):175–187, 1977.

2 Kenneth Arrow et al. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.

3 Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL: `http://www.fairmlbook.org`.

4 Gary S Becker. The economics of discrimination chicago. *University of Chicago*, 1957.

5 Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

6 Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

7 Lee Cohen, Zachary C. Lipton, and Yishay Mansour. Efficient candidate screening under multiple tests and implications for fairness. *CoRR*, abs/1905.11361, 2019. `arXiv:1905.11361`.

8 Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*, 2018. `arXiv:1808.00023`.

9 Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency (FAT*)*, 2018.

10 William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

11 Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems (NeurIPS)*, 2016.

12 Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *World Wide Web Conference (WWW)*, 2018.

13 Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *International Conference on Data Mining (ICDM)*, 2010.

14 Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDM Workshops*, 2011.

15 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2017.

16 Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Knowledge Discovery in Databases (KDD)*, 2008.

17 Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, pages 659–661, 1972.

18 Sven Schmit, Virag Shah, and Ramesh Johari. Optimal testing in the experiment-rich regime. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

19 Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, 2018.

20 Ward Whitt. Uniform conditional stochastic order. *Journal of Applied Probability*, 17(1):112–123, 1980.

A Technical Proofs

A.1 Proofs from Section 3

Proof of Theorem 1. To prove the theorem, we show that the loss function $l_\alpha(\tau, \theta)$, as a function of θ is quasi-convex and achieves its minimum value at $\left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})}) \right\rceil$. Namely, we show that the loss is monotone increasing for $\left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})}) \right\rceil \leq \theta \leq \tau - 1$, i.e., increasing θ increases the loss: $l_\alpha(\theta) < l_\alpha(\theta+1)$.
Similarly, we show that for $1 \leq \theta \leq \left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})}) \right\rceil$, we have $l_\alpha(\theta) < l_\alpha(\theta - 1)$.
Indeed,

$$l_\alpha(\theta+1,\tau) - l_\alpha(\theta,\tau) = -\alpha \Pr[y=0, S_\tau = \theta] + (1-\alpha)\Pr[y=+1, S_\tau=\theta]$$

$$= -\alpha \Pr[S_\tau = \theta | y = 0]\Pr[y=0] + (1-\alpha)\Pr[S_\tau = \theta|y=+1]\Pr[y=+1]$$

Since $\Pr[y=0] = 1-p$ and $\Pr[y=+1] = p$, we have

$$l_{\frac{1}{2}}(\theta+1,\tau) - l_{\frac{1}{2}}(\theta,\tau) = -(1-p)\alpha\Pr[S_\tau=\theta|y=0] + p(1-\alpha)\Pr[S_\tau=\theta|y=+1].$$

The above expression is positive iff

$$(1-p)\alpha\Pr[S_\tau = \theta|y=0] < p(1-\alpha)\Pr[S_\tau=\theta|y=+1] \tag{3}$$

Since $\Pr[S_\tau = \theta|y=0]$ is the probability of exactly θ flips, and $\Pr[S_\tau = \theta|y=+1]$ is the probability of exactly $\tau - \theta$ flips, we can calculate those probabilities as follows:

$$\Pr[S_\tau = \theta|y=0] = \binom{\tau}{\theta}(\frac{1-\sigma}{2})^\theta(\frac{1+\sigma}{2})^{\tau-\theta}$$

$$\Pr[S_\tau = \theta|y=+1] = \binom{\tau}{\tau-\theta}(\frac{1-\sigma}{2})^{\tau-\theta}(\frac{1+\sigma}{2})^{\theta}$$

Substituting expression in (3), we get

$$(1-p)\alpha\binom{\tau}{\theta}(\frac{1-\sigma}{2})^\theta(\frac{1+\sigma}{2})^{\tau-\theta} < p(1-\alpha)\binom{\tau}{\tau-\theta}(\frac{1-\sigma}{2})^{\tau-\theta}(\frac{1+\sigma}{2})^{\theta}.$$

Rearranging, we get

$$(\frac{1-\sigma}{1+\sigma})^{2\theta} < (\frac{1-\sigma}{1+\sigma})^{\tau}(\frac{p}{1-p})(\frac{1-\alpha}{\alpha}).$$

Applying log on both sides gets us

$$2\theta\log(\frac{1-\sigma}{1+\sigma}) < \tau\log(\frac{1-\sigma}{1+\sigma}) + \log(\frac{p}{1-p}) + \log(\frac{1-\alpha}{\alpha}).$$

Solving for θ, we find that the inequality holds if

$$\theta > \frac{\tau\log(\frac{1-\sigma}{1+\sigma}) + \log(\frac{p}{1-p}) + \log(\frac{1-\alpha}{\alpha})}{2\log(\frac{1-\sigma}{1+\sigma})} = \left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})}) \right\rceil$$

For $\theta \geq \left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})}) \right\rceil$, we have

$$(1-p)\alpha\Pr[S_\tau=\theta|y=0] < p(1-\alpha)\Pr[S_\tau=\theta|y=+1],$$

and for $\theta \leq \left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})})\right\rceil$, we have

$$\alpha(1-p)\Pr[S_\tau = \theta | y = 0] > (1-\alpha)p\Pr[S_\tau = \theta | y = +1].$$

This implies that the maximum is $\theta^*_{p,\alpha} = \left\lceil \frac{1}{2}(\tau - \frac{\log(\frac{1}{p}-1)+\log(\frac{1}{\alpha}-1)}{\log(1+\frac{2\sigma}{1-\sigma})})\right\rceil$. ◀

Proof of Theorem 2. We start with a skilled candidate. The expected number of tests that a skilled candidate passes is $\mathbb{E}[S_\tau | y = +1] = \tau(\frac{1+\sigma}{2}) > \frac{\tau}{2}$.

By using Hoeffding's inequality for Bernoulli distributions, for every $\epsilon > 0$,

$$\Pr[\mathbb{E}[S_\tau] - S_\tau \geq \epsilon | y = +1] = \Pr[\tau(\frac{1+\sigma}{2}) - S_\tau \geq \epsilon | y = +1] \leq e^{-2\epsilon^2\tau} < \delta.$$

Choosing $\epsilon = \frac{\sigma}{2}$ yields $S_\tau \leq \frac{\tau}{2} < \lceil \frac{\tau}{2} \rceil$ (as τ is odd), which holds iff a majority threshold policy would predict that this is an unskilled candidate (false negative). Solving for τ, we get $\tau > \frac{1}{\sigma^2}\ln(\frac{1}{\delta})$.

We now repeat the process for an unskilled candidate. The expected number of tests that an unskilled candidate passes is $\mathbb{E}[S_\tau | y = 0] = \tau(\frac{1-\sigma}{2}) < \frac{\tau}{2}$.

By using Hoeffding's inequality again, we have

$$\Pr[S_\tau - \mathbb{E}[S_\tau] \geq \epsilon | y = 0] = \Pr[S_\tau - \tau(\frac{1-\sigma}{2}) \geq \epsilon | y = 0] \leq e^{-2\epsilon^2\tau} < \delta$$

Choosing $\epsilon = \frac{\sigma}{2}$ yields $S_\tau > \frac{\tau}{2}$, which holds iff a majority threshold falsely predicts that this is a skilled candidate (false positive). Solving for τ again, we get $\tau > \frac{1}{\sigma^2}\ln(\frac{1}{\delta})$.
Overall, $\tau > \frac{\alpha(1-p)}{\sigma^2}\ln(\frac{1}{\delta}) + \frac{p(1-\alpha)}{\sigma^2}\ln(\frac{1}{\delta}) = \Omega(\frac{\alpha+p-2p\alpha}{\sigma^2}\ln(\frac{1}{\delta}))$. ◀

Proof of Theorem 4. Let π' be any optimal policy for (2) (not necessarily threshold) with a fixed number of tests, τ. We will show, in two steps, how to transform it into an optimal randomized threshold policy. The first step is to symmetrize π'. Let $r_k = \Pr[\pi(\hat{y}) = 1 | S_\tau = k]$. Define a policy π'', which performs τ tests, and accepts with probability r_k where $k = S_\tau$. Clearly, both π' and π'' have the same accept probability. In addition, since condition on $S_\tau = k$, any sequence of outcomes is equally likely. Furthermore, and the probability that $y = 1$ given any sequence of outcomes with $S_\tau = k$, is identical. (Technically, S_τ is a sufficient statistics.) This implies that the false discovery rate is also unchanged.

This yields that π with the randomization vector r is also optimal.

The second step is to suppose – for sake of contradiction – that π'' is not a randomized threshold policy. We will show that we can improve the FDR of π'' while keeping the probability of acceptance unchanged. This will contradict the hypothesis that π' is optimal.

If π'' is not a randomized threshold policy, then there is no θ and k, such that

$$r_k = \Pr[\pi(\hat{y}_{i,1}, \ldots, \hat{y}_{i,\tau}) = 1 | S_\tau = k \neq \theta] = \begin{cases} 0, & \text{if } k < \theta \\ 1 & \text{if } k > \theta \end{cases}.$$

Now, let k be the minimal value such that $r_k > 0$ and let $0 < i < \tau - k$ be the minimal value for which $0 < r_{k+i} < 1$. Clearly, the FDR is lower at $S_\tau = k + i$ than at $S_\tau = k$. Intuitively, we can shift some probability mass, $\epsilon_k > 0$ from r_k to r_{k+i} in a way that maintains the acceptance probability of π and decreases the false positive rates.

Let $\epsilon_{k+i} > 0$ be such that $\epsilon_k \cdot r_k = \epsilon_{k+i} \cdot r_{k+i}$. Let r' be a modified randomization vector for π such that $r'_k = r_k(1 - \epsilon_k), r'_{k+i} = r_{k+i}(1 + \epsilon_{k+i})$ and for every $l \notin \{k, k+i\}$ $r'_l = r_l$. Since $\Pr[\pi(\hat{y}_{i,1}, \ldots, \hat{y}_{i,\tau}) = 1] = \sum_{l=1}^{\tau} r_l = \sum_{l \notin \{k,k+i\}} r_l + r'_k + r'_{k+i}$, the acceptance probability remains the same. As for the false discovery rate, since $\Pr[y_i = 0 | S_\tau = k + i] <$

$\Pr[y_i = 0|S_\tau = k]$, $\Pr[S_\tau = k+i]$ is higher with r' than with r, $\Pr[S_\tau = k]$ is lower with r' than with r and for any $l \notin \{k, k+i\}$, $\Pr[S_\tau = l]$ with r' is the same as with r, the false discovery rate with r' is lower, which contradicts the optimality of π with r as the randomization vector. ◀

Proof of Theorem 5. Using Bayes' theorem, the conditional probability can be decomposed as

$$\Pr[y_i = +1|S_\tau = \theta] = \frac{\Pr[y_i = +1]\Pr[S_\tau = \theta|y_i = +1]}{\Pr[S_\tau = \theta]} =$$

$$\frac{p\binom{\tau}{\theta}(\frac{1-\sigma}{2})^{\tau-\theta}(\frac{1+\sigma}{2})^{\theta}}{p\binom{\tau}{\theta}(\frac{1-\sigma}{2})^{\tau-\theta}(\frac{1+\sigma}{2})^{\theta} + (1-p)\binom{\tau}{\tau-\theta}(\frac{1+\sigma}{2})^{\tau-\theta}(\frac{1-\sigma}{2})^{\theta}}.$$

Since $\tau - \theta < \theta$ and $\binom{\tau}{\theta} = \binom{\tau}{\tau-\theta}$, we get

$$\frac{p(1+\sigma)^{2\theta-\tau}}{p(1+\sigma)^{2\theta-\tau} + (1-p)(1-\sigma)^{2\theta-\tau}} = \frac{p(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau}}{p(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau} + 1 - p}.$$

Since $(\frac{1+\sigma}{1-\sigma}) > 1$ it holds that $(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau} > 1$,

$$(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau}(1-p) > 1-p.$$

So,

$$(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau} > p(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau} + 1 - p,$$

And finally,

$$\Pr[y_{i'} = +1] = p < \frac{p(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau}}{p(\frac{1+\sigma}{1-\sigma})^{2\theta-\tau} + 1 - p} = \Pr[y_i = +1|S_\tau = \theta].$$ ◀

Proof of Lemma 7. Let $S'_\tau = \sum_{j=1}^{\tau}(2\hat{y}_{i,j} - 1)$, and let $s_\tau \in \{-\tau, \dots, \tau\}$ be any of the possible values of S'_τ. Note that

$$\frac{\Pr[\hat{y}_{i,j} = 1|y_i = 1]}{\Pr[\hat{y}_{i,j} = 1|y_i = 0]} = \frac{1+\sigma}{1-\sigma}.$$

Since the $\hat{y}_{i,j}$ are i.i.d., we have

$$\begin{aligned} X_\tau =& X_0 + \sum_{j=1}^{\tau}(2\hat{y}_{i,j} - 1)\cdot\log(\frac{\Pr[\hat{y}_{i,j} = +1|y_i = +1]}{\Pr[\hat{y}_{i,j} = +1|y_i = 0]}) \\ =& \log(\frac{p}{1-p}) + S_\tau \log(\frac{1+\sigma}{1-\sigma}) \\ =& \log((\frac{p}{1-p})(\frac{1+\sigma}{1-\sigma})^{S_\tau}). \end{aligned}$$

Since

$$\frac{\Pr[S_\tau = s_\tau|y_i = 1]}{\Pr[S_\tau = s_\tau|y_i = 0]} = (\frac{1+\sigma}{1-\sigma})^{s_\tau},$$

we have

$$X_\tau = \log((\frac{p}{1-p})(\frac{\Pr[S_\tau = s_\tau | y_i = 1]}{\Pr[S_\tau = s_\tau | y_i = 0]})). \tag{4}$$

Since

$$\Pr[S_\tau = s_\tau | y_i = 1] = \frac{\Pr[S_\tau = s_\tau] \cdot \Pr[y_i = 1 | S_\tau = s_\tau]}{\Pr[y_i = 1]}$$

and

$$\Pr[S_\tau = s_\tau | y_i = 0] = \frac{\Pr[S_\tau = s_\tau] \cdot \Pr[y_i = 0 | S_\tau = s_\tau]}{\Pr[y_i = 0]},$$

assigning $\Pr[y_i = 0] = 1 - p$ and $\Pr[y_i = 1] = p$, we get

$$\frac{\Pr[S_\tau = s_\tau | y_i = 1]}{\Pr[S_\tau = s_\tau | y_i = 0]} = \frac{(1-p) \cdot \Pr[y_i = 1 | S_\tau = s_\tau]}{p \cdot \Pr[y_i = 0 | S_\tau = s_\tau]}. \tag{5}$$

Applying (5) in (4) and adding $X_\tau \geq \log \beta$ gives us

$$X_\tau = \log\left(\frac{\Pr[y_i = 1 | S_\tau = s_\tau]}{\Pr[y_i = 0 | S_\tau = s_\tau]}\right) = \log\left(\frac{\Pr[y_i = 1 | S_\tau = s_\tau]}{1 - \Pr[y_i = 1 | S_\tau = s_\tau]}\right) \geq \log \beta$$

$$\frac{\Pr[y_i = 1 | S_\tau = s_\tau]}{1 - \Pr[y_i = 1 | S_\tau = s_\tau]} \geq \beta$$

$$\Pr[y_i = 1 | S_\tau = s_\tau] \geq \beta(1 - \Pr[y_i = 1 | S_\tau = s_\tau])$$

$$\Pr[y_i = 1 | S_\tau = s_\tau] \geq \frac{\beta}{1+\beta}$$

Applying (5) in (4) and adding $X_\tau < \log \beta'$ gives us

$$\frac{\Pr[y_i = 1 | S_\tau = s_\tau]}{1 - \Pr[y_i = 1 | S_\tau = s_\tau]} < \beta'$$

Hence

$$\Pr[y_i = 1 | S_\tau = s_\tau] < \frac{\beta'}{1+\beta'}$$ ◀

Proof of Theorem 9. First recall that given a skilled candidate, for every test j,

$$\Pr[\hat{y}_{i,j} = +1 | y_i = +1] = \frac{1+\sigma}{2}$$

$$\Pr[\hat{y}_{i,j} = 0 | y_i = +1] = \frac{1-\sigma}{2}$$

Hence

$$\Pr[\hat{y}_{i,j} = 0 | y_i = 1] - \Pr[\hat{y}_{i,j} = +1 | y_i = 1] = -\sigma.$$

The lower absorbing barrier is reached when a candidate's posterior skill level is lower than the prior of the skill level, i.e.,

$$\log \frac{\epsilon'}{1-\epsilon'} - \log\left(\frac{1+\sigma}{1-\sigma}\right)$$

and the starting point is just one step away from the lower absorbing barrier:

$$X_0 = \log \frac{p}{1-p}.$$

According to Corollary 8, the upper absorbing barrier is in

$$\log(\frac{1-\epsilon}{\epsilon}).$$

To derive the results for the expected duration of the random walk for skilled and unskilled candidates, we shift the locations of the absorbing points so that the lower barrier would be in 0 and also divide them by a step size (so now we have that every step is of size 1). The new upper absorbing barrier is at

$$a = \left\lceil \frac{\log(\frac{1-\epsilon}{\epsilon}) - (\log \frac{\epsilon'}{1-\epsilon'} - \log(\frac{1+\sigma}{1-\sigma}))}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil = \left\lceil \frac{\log(\frac{(1-\epsilon)(1-\epsilon')(1+\sigma)}{\epsilon\epsilon'(1-\sigma)})}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil .$$

And we also shift the starting point:

$$z = \left\lceil \frac{\log \frac{p}{1-p} - (\log \frac{\epsilon'}{1-\epsilon'} - \log(\frac{1+\sigma}{1-\sigma}))}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil = \left\lceil \frac{\log(\frac{p(1-\epsilon')(1+\sigma)}{\epsilon'(1-p)(1-\sigma)})}{\log(\frac{1+\sigma}{1-\sigma})} \right\rceil$$

As stated in [10], the expected duration of a random walk with absorbing barriers of 0 and a from $z = 1$ is (equation 3.4, chapter XIV [page 348]):

$$\mathbb{E}[\tau_s] = \mathbb{E}[D_{z=1}] = \frac{1}{q-p}\left(z - a \cdot \frac{1-(\frac{q}{p})^z}{1-(\frac{q}{p})^a}\right) = \frac{1}{-\sigma}\left(z - a \cdot \frac{1-(\frac{1-\sigma}{1+\sigma})^z}{1-(\frac{1-\sigma}{1+\sigma})^a}\right).$$

Hence,

$$\mathbb{E}[\tau_s] = \frac{1}{\sigma}\left(a \cdot \frac{1-(\frac{1-\sigma}{1+\sigma})^z}{1-(\frac{1-\sigma}{1+\sigma})^a} - z\right).$$

As for unskilled candidates, the absorbing points and the starting point are the same, the only difference is that

$$\Pr[\hat{y}_{i,j} = +1 | y_i] = \frac{1-\sigma}{2}$$

and

$$\Pr[\hat{y}_{i,j} = 0 | y_i = +1] = \frac{1+\sigma}{2}.$$

Therefore,

$$\Pr[\hat{y}_{i,j} = 0 | y_i = 0] - \Pr[\hat{y}_{i,j} = +1 | y_i = 0] = \sigma$$

and we deduce

$$\mathbb{E}[\tau_u] = \frac{1}{\sigma}\left(z - a \cdot \frac{1-(\frac{1+\sigma}{1-\sigma})^z}{1-(\frac{1+\sigma}{1-\sigma})^a}\right). \qquad ◀$$

Deviations for the confusion matrix (Table 1). We split the claim in the confusion matrix (Table 1) into two parts. First, using equation (2.4) from chapter XIV [page 345] in [10], we get

$$\mathrm{FNR} = \Pr[\pi_{\mathrm{Greedy}}(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 0 | y_i = +1] = \frac{(\frac{1-\sigma}{1+\sigma})^a - (\frac{1-\sigma}{1+\sigma})^z}{(\frac{1-\sigma}{1+\sigma})^a - 1}$$

and

$$\mathrm{TNR} = \Pr[\pi_{\mathrm{Greedy}}(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 0 | y_i = 0] = \frac{(\frac{1+\sigma}{1-\sigma})^a - (\frac{1+\sigma}{1-\sigma})^z}{(\frac{1+\sigma}{1-\sigma})^a - 1}.$$

The second part follows from the fact the gambler's ruin must end in case of absorbing barriers.

$$\mathrm{TPR} = \Pr[\pi_{\mathrm{Greedy}}(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1 | y_i = +1] = 1 - \frac{(\frac{1-\sigma}{1+\sigma})^a - (\frac{1-\sigma}{1+\sigma})^z}{(\frac{1-\sigma}{1+\sigma})^a - 1} =$$

$$\frac{(\frac{1-\sigma}{1+\sigma})^z - 1}{(\frac{1-\sigma}{1+\sigma})^a - 1} = \frac{\frac{\epsilon'(1-p)(1-\sigma)}{p(1-\epsilon')(1+\sigma)} - 1}{\frac{\epsilon'\epsilon(1-\sigma)}{(1-\epsilon')(1-\epsilon)(1+\sigma)} - 1} = \frac{\frac{\mu(1-p)}{p} - 1}{\frac{\epsilon\mu}{(1-\epsilon)} - 1} = \frac{(1-\epsilon)(\mu(1-p)-p)}{p(\epsilon\mu - (1-\epsilon))},$$

Where $\mu := \frac{\epsilon'(1-\sigma)}{(1-\epsilon')(1+\sigma)}$. For $\epsilon \leq 1/4$ and $p < 1/2$ we get $0 \leq \mu \leq 1/3$ and $\mu = \Theta(\epsilon'(1-\sigma))$, therefore

$$\mathrm{TPR} = \Theta\left(\frac{p-\mu}{p}\right) = \Theta\left(1 - \frac{\epsilon'}{p}(1-\sigma)\right).$$

Hence $\mathrm{FNR} = \Theta(\frac{\epsilon'}{p}(1-\sigma))$.

$$\mathrm{FPR} = \Pr[\pi_{\mathrm{Greedy}}(\hat{y}_{i,1}, \dots, \hat{y}_{i,\tau}) = 1 | y_i = 0] = \frac{(\frac{1+\sigma}{1-\sigma})^z - 1}{(\frac{1+\sigma}{1-\sigma})^a - 1} = \frac{\frac{p(1-\epsilon')(1+\sigma)}{(1-p)\epsilon'(1-\sigma)} - 1}{\frac{(1-\epsilon')(1-\epsilon)(1+\sigma)}{\epsilon'\epsilon(1-\sigma)} - 1} =$$

$$= \frac{\frac{p}{(1-p)\mu} - 1}{\frac{(1-\epsilon)}{\epsilon\mu} - 1} \frac{\epsilon(p-(1-p)\mu)}{(1-p)(1-\epsilon-\epsilon\mu)} = \Theta\left(\epsilon(p-\mu)\right) = \Theta\left(\epsilon(p-\epsilon'+\epsilon'\sigma)\right)$$

Hence $\mathrm{TNR} = \Theta\left(1 - \epsilon(p-\epsilon'+\epsilon'\sigma)\right)$. ◀

Proof of Theorem 10.

$\mathbb{E}[\tau] = \mathbb{E}[\tau_s]p + \mathbb{E}[\tau_u](1-p) =$

$$= \frac{1}{\sigma}\left(a \cdot \frac{1 - (\frac{1-\sigma}{1+\sigma})^z}{1 - (\frac{1-\sigma}{1+\sigma})^a} - z\right)p + \frac{1}{\sigma}\left(z - a \cdot \frac{1 - (\frac{1+\sigma}{1-\sigma})^z}{1 - (\frac{1+\sigma}{1-\sigma})^a}\right)(1-p) =$$

$$\approx \frac{1}{\sigma}\left(a \cdot (1 - \frac{\epsilon'}{p}(1-\sigma)) - z\right)p + \frac{1}{\sigma}(z - a(\epsilon(p-\epsilon'+\epsilon'\sigma)))(1-p) \approx \frac{ap}{\sigma}$$ ◀

A.2 Proofs from Section 4

The next lemma aids in the proof of Theorem 11.

▶ **Lemma 18.** *Let Z_n^η be a Binomial random variable with parameters $n \in \mathbb{N}$ and $\eta \in (0,1)$. Given a number of successes, $k \in \{0,\dots,n\}$, we know that the probability mass function of Z_n^η is $f_k(\eta) := \Pr[Z_n^\eta = k] = \binom{n}{k}\eta^k(1-\eta)^{n-k}$. Let $\mathcal{L}(\eta|k)$ be the likelihood function of the event $Z_n^\eta = k$. Then the maximum likelihood of $f_k(\eta)$ is $\eta = \frac{k}{n}$. I.e.,*

$$\mathcal{L}(\eta|k) = argmax_\eta f_k(\eta) = argmax_\eta \binom{n}{k}\eta^k(1-\eta)^{n-k} = \frac{k}{n}.$$

Proof of Lemma 18. We notice that $\binom{n}{k}$ does not depend on η, thus

$$\text{argmax}_\eta f_k(\eta) = \text{argmax}_\eta \binom{n}{k}\eta^k(1-\eta)^{n-k} = \text{argmax}_\eta \eta^k(1-\eta)^{n-k}$$

The log-likelihood is particularly convenient for maximum likelihood estimation. Logarithms are strictly increasing functions, as a result, maximizing the likelihood is equivalent to maximizing the log-likelihood, i.e.,

$$\text{argmax}_\eta \eta^k(1-\eta)^{n-k} = \text{argmax}_\eta \ln(\eta^k(1-\eta)^{n-k}) = \text{argmax}_\eta k\ln(\eta) + (n-k)\ln(1-\eta)$$

Differentiating (with respect to η) and comparing to zero we get

$$\frac{d\ln(f_k(\eta))}{d\eta} = \frac{k}{\eta} - \frac{n-k}{1-\eta} = 0.$$

And after refactoring,

$$k(1-\eta) = (n-k)\eta$$

The function $\ln(f_k(\eta))$ is a strictly concave as its second derivative is negative,

$$\frac{d^2\ln(f_k(\eta))}{d\eta^2} = -\frac{k}{\eta^2} - \frac{n-k}{(1-\eta)^2} < 0,$$

And since the derivative of a strictly concave function is zero at $\frac{k}{n}$, then $\hat{\eta} = \frac{k}{n}$ is a global maximum. Therefore, $\hat{\eta} = \frac{k}{n}$ obtains absolute maximum in $f_k(\eta)$. ◀

Proof of Theorem 11. Let $Z_\tau^{\eta_i}$ be a random variable that represents the number of flips out of a τ-tests sequence with a noise level of η_i, i.e., $Z_\tau^{\eta_i}$ is the number of times when $y_j \neq y$ for $1 \le j \le \tau$. We use $Z_\tau^{\eta_i}$ to express $\Pr[\pi(\hat{y}_{i,1},\dots,\hat{y}_{i,\tau}) = 1|y_i = 0, \eta = \eta_i]$ as the probability that at least θ flips,

$$\Pr[\pi(\hat{y}_{i,1},\dots,\hat{y}_{i,\tau}) = 1|y_i = 0, \eta = \eta_i] = \Pr[Z_\tau^{\eta_i} \ge \theta]$$

and the probability of $\Pr[\pi(\hat{y}_{i,1},\dots,\hat{y}_{i,\tau}) = 1|q = +1, \eta = \eta_i]$ as at most $\tau - \theta$ flips, thus

$$\Pr[\pi(\hat{y}_{i,1},\dots,\hat{y}_{i,\tau}) = 1|y_i = +1, \eta = \eta_i] = \Pr[Z_\tau^{\eta_i} \le \tau - \theta].$$

From Lemma (18) and since probability density function (pdf) are is monotone increasing, we derive that the pdf of $Z_n^{\eta_2}$ satisfies *monotone likelihood ratio property* over the pdf of $Z_n^{\eta_1}$. This implies that the pdf of $Z_n^{\eta_2}$ also has *first-order stochastic dominance* over $Z_n^{\eta_1}$ by Theorem 1.1 in [20]. From *stochastic dominance*, we can derive the desired inequalities

$$FP_{\theta,\tau}^{\eta_1} = \Pr[\theta \le Z_n^{\eta_1}] < \Pr[\theta \le Z_n^{\eta_2}] = FP_{\theta,\tau}^{\eta_2}$$

and

$$FN_{\theta,\tau}^{\eta_1} = \Pr[Z_n^{\eta_1} \le \tau - \theta] < \Pr[Z_n^{\eta_2} \le \tau - \theta] = FN_{\theta,\tau}^{\eta_2}.$$ ◀

A.3 Proofs from Section 5

Proof of Theorem 12. First, recall that

$$\mathrm{Var}[Q|Y_1, ..., Y_n] = \frac{1}{\frac{1}{\sigma_Q^2} + \frac{n}{\sigma_\eta^2}} = \frac{\sigma_Q^2 \sigma_\eta^2}{\sigma_\eta^2 + n\sigma_Q^2}.$$

Solving for n_2 in the equation $\mathrm{Var}_1[Q|Y_1, ..., Y_{n_1}] = \mathrm{Var}_2[Q|Y_1, ..., Y_{n_2}]$,

$$\frac{\sigma_Q^2 \sigma_{\eta_1}^2}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2} = \frac{\sigma_Q^2 \sigma_{\eta_2}^2}{\sigma_{\eta_2}^2 + n_2\sigma_Q^2}$$

we get

$$\sigma_{\eta_1}^2(\sigma_{\eta_2}^2 + n_2\sigma_Q^2) = \sigma_{\eta_2}^2(\sigma_{\eta_1}^2 + n_1\sigma_Q^2)$$

and hence

$$\sigma_{\eta_1}^2 n_2 = \sigma_{\eta_2}^2 n_1.$$

Extracting n_2, we find that $n_2 = \frac{\sigma_{\eta_2}^2}{\sigma_{\eta_1}^2} n_1$. ◀

Proof of Theorem 13. First, recall that

$$\mathbb{E}[Q|Y_1, ..., Y_n] = \mu_Q + \left[\frac{1}{\frac{\sigma_\eta^2}{\sigma_Q^2} + n}, \ldots\right] \cdot (\mathbf{y} - \mu_y) = \mu_Q + \left[\frac{\sigma_Q^2}{\sigma_\eta^2 + n\sigma_Q^2}, \ldots\right] \cdot (\mathbf{y} - \mu_y)$$

Now,

$$\mathbb{E}_1[Q|Y_1, ..., Y_{n_1}] - \mathbb{E}_2[Q|Y_1, ..., Y_{n_2}] =$$

$$\left[\frac{\sigma_Q^2}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2}, \ldots\right] \cdot (\mathbf{y_1} - \mu_y) - \left[\frac{\sigma_Q^2}{\sigma_{\eta_2}^2 + n_2\sigma_Q^2}, \ldots\right] \cdot (\mathbf{y_2} - \mu_y)$$

$$= \frac{\sigma_Q^2}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2} n_1(\bar{\mathbf{y_1}}) - \frac{\sigma_Q^2}{\sigma_{\eta_2}^2 + n_2\sigma_Q^2} n_2(\bar{\mathbf{y_2}})$$

$$= \frac{\sigma_Q^2 n_1}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2}(\bar{\mathbf{y_1}}) - \frac{\sigma_Q^2 n_2}{\sigma_{\eta_2}^2 + n_2\sigma_Q^2}(\bar{\mathbf{y_2}})$$

$$= \frac{\sigma_Q^2 n_1}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2}(\bar{\mathbf{y_1}}) - \frac{\sigma_Q^2 \frac{\sigma_{\eta_2}^2}{\sigma_{\eta_1}^2} n_1}{\sigma_{\eta_2}^2 + \frac{\sigma_{\eta_2}^2}{\sigma_{\eta_1}^2} n_1\sigma_Q^2}(\bar{\mathbf{y_2}})$$

$$= \frac{\sigma_Q^2 n_1}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2}(\bar{\mathbf{y_1}}) - \frac{\sigma_Q^2 n_1}{\sigma_{\eta_1}^2 + n_1\sigma_Q^2}(\bar{\mathbf{y_2}})$$ ◀

Metric Learning for Individual Fairness

Christina Ilvento
Harvard University, John A. Paulson School of Engineering and Applied Science,
Cambridge, MA, USA
cilvento@g.harvard.edu

Abstract

There has been much discussion concerning how "fairness" should be measured or enforced in classification. Individual Fairness [2], which requires that similar individuals be treated similarly, is a highly appealing definition as it gives strong treatment guarantees for individuals. Unfortunately, the need for a task-specific similarity metric has prevented its use in practice. In this work, we propose a solution to the problem of approximating a metric for Individual Fairness based on human judgments. Our model assumes access to a human fairness arbiter who is free of explicit biases and possesses sufficient domain knowledge to evaluate similarity. Our contributions include definitions for metric approximation relevant for Individual Fairness, constructions for approximations from a limited number of realistic queries to the arbiter on a sample of individuals, and learning procedures to construct hypotheses for metric approximations which generalize to unseen samples under certain assumptions of learnability of distance threshold functions.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases metric learning, individual fairness, fair machine learning

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.2

Related Version `https://arxiv.org/abs/1906.00250`

Funding *Christina Ilvento*: This work was supported by the Smith Family Fellowship, Microsoft Research and the Sloan Foundation.

Acknowledgements The author is grateful for the comments of Cynthia Dwork.

1 Introduction

Determining what it means for an algorithm or classifier to be "fair" and how to enforce any such determination has become a subject of considerable interest as automated decision-making increasingly takes the place of direct human judgment. One attractive definition proposed is Individual Fairness [2], which states that similar individuals should be treated similarly, where similarity is encoded in a task-specific *metric*.

▶ **Definition 1** (Individual Fairness [2]). *Given a universe U, a metric $\mathcal{D} : U \times U \to [0, 1]$ for a classification task with outcome set O, and a distance measure $d : \Delta(O) \times \Delta(O) \to [0, 1]$, a randomized classifier $C : U \to \Delta(O)$ is Individually Fair if and only if for all $u, v \in U$, $\mathcal{D}(u, v) \geq d(C(u), C(v))$.*

Individual Fairness is appealing because each person is assured that her treatment is similar to that of any person similar to her.[1] However, the value of this assurance critically depends on the extent to which the similarity metric ($\mathcal{D}$) faithfully represents society's best

[1] By way of contrast, notions of fairness based on group level statistics can only provide individuals with the guarantee that if they are treated poorly, either someone in a different group is also treated poorly or someone in their group is treated well. Furthermore, many popular notions of statistical group fairness conflict with each other and cannot be satisfied simultaneously [1, 6].

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 2; pp. 2:1–2:11
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

understanding of what constitutes similarity for a given task. Thus, the most significant barrier to implementing Individual Fairness in practice is the need to construct a similarity metric for each classification setting.

In this work we set out a path for constructing metrics for Individual Fairness based on judgments made by a qualified, fair-minded "human fairness arbiter." Our contributions include: (1) a framework for useful approximations to a metric for Individual Fairness; (2) a limited, realistic query model for determining the arbiter's judgments of who is similar to whom; (3) a method for constructing approximations to the true metric with limited queries to the arbiter by using distances from a (set of) representative individual(s); (4) a procedure for generalizing these approximations to unseen samples based on limited learnability assumptions. Throughout this work we make no assumption on the form of the metric or the features included in the learning procedure with the clearly stated exception of Assumption 1 concerning learnability of threshold functions. As our results are built upon a series of sequential steps including new terminology and machinery, we first present an extended introduction to highlight the key concepts, logic and results. These results are expanded and discussed in detail in the full version of the paper [4].

1.1 Model

In this work, we take the viewpoint that fairness is not well described by either accuracy or group statistics alone. Instead, we view fairness as a highly contextual property one can identify but not necessarily describe.[2] Our goal is to produce a metric which results in similarity judgments with which fair-minded people would agree, rather than satisfying any particular statistical properties.[3] The core of our model is the human fairness arbiter, a fair-minded individual who is free from explicit biases or arbitrary preferences, is motivated to engage ethically and honestly in the query protocol, and has sufficient knowledge and contextual understanding of who is similar to whom for a particular task. The arbiter is not expected to provide us a description or specification of the distance metric.

A critical part of learning metrics based on human judgments is determining the type of queries to ask in order to solicit consistent, fast responses. To that end, we assume that we cannot ask the arbiter to consider more than a few individuals at a time, e.g., it is not realistic to ask the arbiter to find the closest pair of elements in the universe.

We ask the arbiter to answer two types of queries in this work: relative distance queries, (e.g., is a closer to b or c), and real-valued distance queries.

▶ **Definition 2** (Real-valued distance query). $\mathsf{O}_{\mathsf{REAL}}(u, v) := \mathcal{D}(u, v)$.

▶ **Definition 3** (Triplet query). $\mathsf{O}_{\mathsf{TRIPLET}}(a, b, c) := \{1 \textit{ if } \mathcal{D}(a, b) < \mathcal{D}(a, c),\ 0 \textit{ if } \mathcal{D}(a, c) \leq \mathcal{D}(a, b)\}$.

Producing a consistent set of real-valued distances is not a natural judgment most people are accustomed to making, so we assume that real-valued queries are very "expensive" for the arbiter to answer. Furthermore, maintaining internal consistency may *increase* the query cost as the number of queries increases. Relative distance queries have been used successfully for human evaluation in image processing and computer vision, e.g. [8, 9], and we anticipate they will be significantly easier for the arbiter to evaluate. Demonstrating how to replace difficult queries with easy queries is a significant part of our contribution.

[2] [3] takes a similar approach in which a judge "knows it when she sees it," but is not required to articulate why a decision is unfair.

[3] We discuss different types of agreement, and the extent to which we fully achieve this goal, in Section 8 of the full paper.

We make several simplifying assumptions about the nature of the human fairness arbiter in the main results of this work. (1) There is either one arbiter or all arbiters agree on all decisions. (2) The arbiter does not change her opinion over the query period. (3) The arbiter's responses are consistent, i.e., if she answers that a is closer to b than it is to c, her responses to real-valued queries will also reflect this relative judgment.[4] For the majority of this work, we focus on the query model specified above, which requires the arbiter to answer with arbitrary precision. We also present a relaxed model which allows the arbiter to answer real-valued queries with bounded noise and does not require arbitrarily small distinctions in relative distances queries. The main results presented are replicated in the relaxed model. As the results are similar, we focus on the more simple exact model in the main presentation of our results.[5]

1.2 Contributions

Approximating the metric by contracting. Our first key observation is that Individual Fairness only requires that we do not *overestimate* distances. This motivates our definition of a *submetric*, which is a contraction of the original metric and can be substituted for the original metric and still maintain Individual Fairness.

Constructing submetrics based on distances from representative elements. Taking the difference in distance to a single reference or "representative" point is one of the simplest ways to produce an underestimate of the distance between two elements. Submetrics based on distances from representative elements form the basis of all of our constructions, and although this may seem simplistic, it has a significant advantage when it comes to deciding which queries to ask the arbiter: *ordering.* An ordering of elements by increasing distance from the representative can be constructed with relative distance (easy) queries used as a comparator. Once this ordering is established, real-valued distances at a given granularity can be layered on top in a *sublinear* number of real-valued (hard) queries.

Choosing representatives. A single representative may not be sufficient to capture all relevant distance information, but combining the information from multiple representative elements can produce a more complete picture of the distances between all pairs of individuals. But which representatives should we choose to maximize distance preservation? We discuss a general, randomized approach and show that given certain properties of the metric, i.e. how tightly packed individuals are, a random set of representatives of reasonable size will have good distance preservation properties.

Generalizing submetrics to unseen samples. Once we have established how to construct a submetric for a fixed sample of elements, our next step is to generalize to unseen samples. Our results are based on an assumption that threshold functions, i.e. binary indicators of whether an element is closer to a representative than a given threshold, are efficiently learnable. We show how to combine threshold functions to simulate rounding distances to a representative and then exhibit appropriate parameters to construct an efficient combined learning procedure.

[4] Please see Section 8 in the full paper for additional details.

[5] Extended discussion of the exact query model and a more general definition of relative queries is included in Section 3 of the full paper. The relaxed query model is discussed in detail in Section 7 of the full paper.

Relaxing arbiter requirements. Finally, we present a relaxation of the arbiter query model in which the arbiter (1) may respond to real-valued queries with arbitrary bounded noise and (2) is not required to make arbitrarily precise distinctions between distances and may instead declare relative comparisons to be "too close to call." This model more closely matches the reality of human arbiters, and our results extend with improvements in query complexity at the cost of increased error magnitude.

1.3 Preliminary terminology and definitions

We refer to the universe of individuals as U, a distribution over the universe of individuals as $\mathcal{U}$, and the size of the universe as $|U| = N$. We write $\mathcal{U}^*$ for the uniform distribution over U. We assume $\mathcal{D} : U \times U \to [0, 1]$ for simplicity. Individual Fairness does not require that distances between individuals be maintained exactly, only that they not be exceeded. This observation motivates our definition of a *submetric* which is a contraction of the true metric, i.e., it does not *overestimate* any distance beyond a small additive error term.[6]

▶ **Definition 4** ($\alpha-$submetric). *Given a metric $\mathcal{D}$, $\mathcal{D}' : U \times U \to [0, 1]$ is an α-submetric of $\mathcal{D}$ if for all $u, v \in U$, $\mathcal{D}'(u, v) \leq \mathcal{D}(u, v) + \alpha$.*

Any classifier which satisfies the distance constraints of the submetric $\mathcal{D}'$ will also satisfy those of $\mathcal{D}$, modulo small additive error.[7] Given an α-submetric it is possible to eliminate the additive error by taking $\max\{0, \mathcal{D}'(x, y) - \alpha\}$. On the other hand, we want to avoid contracting distances to the point of triviality. We say that a submetric is $(\beta, c)-$nontrivial if a β fraction of distances between pairs preserve at least a $c-$fraction of their original distance.[8]

▶ **Definition 5** ($(\beta, c)-$nontrivial). *Given a metric $\mathcal{D}$, a submetric $\mathcal{D}'$ of $\mathcal{D}$ is (β, c)-nontrivial for the distribution $\mathcal{U}$ if $\Pr_{u,v\sim\mathcal{U}\times\mathcal{U}}\left[\frac{\mathcal{D}'(u,v)}{\mathcal{D}(u,v)} \geq c\right] \geq \beta$.*

1.4 Constructing submetrics from arbiter judgments

A core component of this work is constructing submetrics based on distance information (either exact or underestimated) from a single representative element. We define the *representative submetric* $\mathcal{D}_r$ in the following Lemma. (The proof of follows from triangle inequality.)

▶ **Lemma 6.** *Given a representative r, $\mathcal{D}_r(x, y) := |\mathcal{D}(r, x) - \mathcal{D}(r, y)|$ is a 0-submetric of $\mathcal{D}$.*

Given a sample of N individuals, $\mathcal{D}_r$ can be constructed from $O(N)$ queries to $\mathsf{O}_{\mathsf{REAL}}$. Although $O(N)$ may seem good compared with the $O(N^2)$ queries required to reconstruct the whole metric, it can be improved to $O(\log(N))$ by supplementing with relative distance queries. Our general strategy will be to show that (1) an *ordering* of elements by distance from a representative can be constructed using $\mathsf{O}_{\mathsf{TRIPLET}}$ as a comparator, and (2) given this ordering, the real-valued distances between each element and the representative can be closely approximated by labeling the ordering with distances at granularity α, which requires a sublinear number of real-valued queries. Algorithm 1 outlines this process.[9]

[6] This relaxation is very similar to the notion of (d, τ) metric fairness of [5] and approximate metric fairness of [10].

[7] As originally noted in [2], the distance measure need not be a true metric, i.e. it does not strictly need to obey triangle inequality or distinguish unequal elements.

[8] Nontriviality is defined over a product of identical distributions of elements in the universe. There is no general obstacle to extending our results to more complicated scenarios, but definitions of density (presented in the full version in Section 6) would need to be adjusted.

[9] See Section 4 in the full version for the detailed specifications and analysis.

Algorithm 1 (Pseudocode).

```
Inputs: the representative $r$, a set of elements $U$, error parameter $\alpha$, interfaces $\mathsf{O_{TRIPLET}}$ and $\mathsf{O_{REAL}}$.
Output: an $\alpha-$submetric $\mathcal{D}'_r$.
Initialize the submetric $\mathcal{D}'_r(x, y) \leftarrow 0$ for all $x, y \in U \times U$.
Order the elements of $U$ by distance from $r$ using $\mathsf{O_{TRIPLET}}$ as a comparator.
Designate the entire ordered list as the first continuous range.
while there are still ranges left to be labeled do
    Select a range left to be labeled.
    Query $\mathsf{O_{REAL}}(r, \text{first})$ and $\mathsf{O_{REAL}}(r, \text{last})$ for the first and last elements in the range.
        if the difference in distances is $> \alpha$ then
            Split into two continuous ranges, each with half of the elements in the current range.
        else set $\mathcal{D}'_r(r, x)$ to $\mathsf{O_{REAL}}(r, \text{first})$ for each element $x$ in the range.
Set $\mathcal{D}'_r(x, y) = |\mathcal{D}'_r(r, x) - \mathcal{D}'_r(r, y)|$ for all $x, y$ in the ordering.
return $\mathcal{D}'_r$.
```

Theorem 7 states that Algorithm 1 produces an $\alpha-$submetric, which follows from observing that rounding $\mathcal{D}(r, x)$ and $\mathcal{D}(r, y)$ down by at most α results in an increase (or decrease) of at most α in $|\mathcal{D}(r, x) - \mathcal{D}(r, y)|$. The bound of $O(N \log(N))$ relative distance queries follows from a straightforward analysis of sorting. The bound of $O(\max\{\frac{1}{\alpha}, \log(N)\})$ real-valued queries is included in Section 4 in the full paper. Briefly, the analysis considers the maximum number of continuous ranges that, when split, result in one range with difference greater than α and one with less. In the worst case, this results in logarithmic dependency on N or $\frac{1}{\alpha}$.

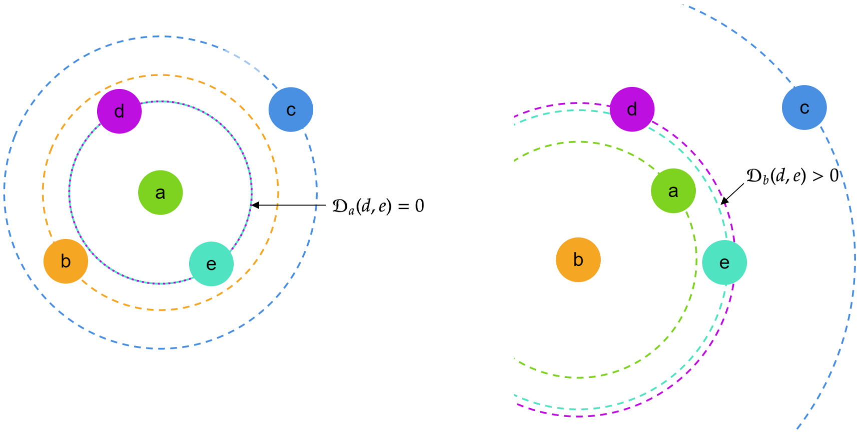
▶ **Theorem 7.** *Algorithm 1 produces an $\alpha-$submetric of $\mathcal{D}$ which preserves $\mathcal{D}(r, u)$ for each $u \in U$ (with additive error $\leq \alpha$) from $O(\max\{\frac{1}{\alpha}, \log(N)\})$ queries to $\mathsf{O_{REAL}}$ and $O(N \log(N))$ queries to $\mathsf{O_{TRIPLET}}$.*

The submetric produced by Algorithm 1 preserves distances between r and other elements well, as $\mathcal{D}'_r(r, x)$ is rounded down by at most α, but we cannot make guarantees on distance preservation between arbitrary pairs without further information. For example, with only the information that u and v are equally distant from r, it is impossible to distinguish whether the distance between u and v is zero or equal to twice their distance from r. (See Figure 1). Submetrics constructed based on different representatives preserve different information about the underlying metric, so we can construct more expressive submetrics by aggregating information from multiple representatives. Taking $\mathsf{maxmerge}(\{\mathcal{D}_i\}, x, y) := \max_i \mathcal{D}_i(x, y)$, it's straightforward to show that if all $\mathcal{D}_i$ are submetrics of $\mathcal{D}$, then the maxmerge of the set is also a submetric of $\mathcal{D}$, and that the merge preserves the "best" distance known for each pair.[10]

1.5 Choosing good representative elements

Although the maxmerge of submetrics based on multiple representatives is an improvement over a single representative, we still cannot make any guarantees about distances between pairs which do not include a representative. There are two approaches one might take to give non-triviality guarantees for arbitrary pairs: (1) develop specialized strategies for combining

[10] Formal analysis of maxmerge and the proof of Lemma 6 appear in Section 3 of the full paper. The proof of Theorem 7 as well as a precise description of Algorithm 1 appear in Section 4 of the full paper.



(a) a chosen as representative.

(b) b chosen as representative.

Figure 1 The impact of representative choice on distance preservation. The distance between each element and the chosen representative is the radius of the shell containing the element. The difference in radii of each pair of shells indicates the distance between the pair of elements under $\mathcal{D}_a$ or $\mathcal{D}_b$. If a is chosen as a representative, notice that d and e are indistinguishable using distance from a alone. Choosing representative b preserve distances better than a, but still does not distinguish d and e very well.

representative submetrics which depend on the structure of the metric, e.g., Euclidean distance, or (2) characterize generic randomized representative selection strategies. In this extended introduction, we focus on the randomized strategies for full generality.

Distance preservation via γ-nets. The crux of the argument for nontriviality with random representatives is (1) a random set of representatives is likely to be "close to" a significant portion of the distribution $\mathcal{U}$, and (2) we can bound the magnitude of underestimates based on the distance from a representative. Below, we formally define a $\gamma-$net to capture the notion of being "close to" or "covering" a set of elements.

▶ **Definition 8.** *A set $R \subseteq U$ is said to form a $\gamma-$net for a subset $V \subseteq U$ under $\mathcal{D}$ if for all balls of radius γ (determined by $\mathcal{D}$) containing at least one element $v \in V$, the ball also contains $r \in R$.*

Intuitively, the distance between r and x will be nearly identical to the distance between a close neighbor of r and x, so we can conclude that if a set of representatives forms a $\gamma-$net for a subset of U, then pairs with at least one element in the net will have their original distance preserved up to a 2γ contraction. (Proofs of Lemmas 9 and 10 follow from triangle inequality.)

▶ **Lemma 9.** *For all $u, v \in U\backslash\{r\}$, $\mathcal{D}(u, v) - \mathcal{D}_r(u, v) \leq \min\{2\mathcal{D}(r, u), 2\mathcal{D}(r, v)\}$, where $\mathcal{D}_r(u, v) := |\mathcal{D}(r, u) - \mathcal{D}(r, v)|$.*

▶ **Lemma 10.** *If a set of representatives $R \subseteq U$ forms a $\gamma-$net for $V \subseteq U$, then for every pair $x, y \in V \times U$ there exists $r \in R$ such that $\mathcal{D}(x, y) - \mathcal{D}_r(x, y) \leq 2\gamma$, where $\mathcal{D}_r(x, y) := |\mathcal{D}(r, x) - \mathcal{D}(r, y)|$.*

Of course, forming a $\gamma-$net for an *arbitrary* γ isn't enough to give a good nontriviality guarantee. To understand how representatives which form a $\gamma-$net will preserve distances, we define *density* and *diffusion* below to characterize the relevant properties of the metric and

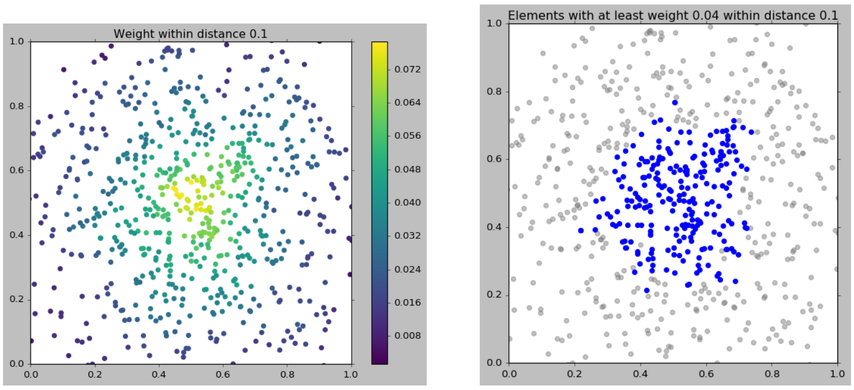


Figure 2 A visualization of the weight of elements, b, within distance $\gamma = .1$ of each element under $\mathcal{U}^*$ for an example universe of points in $[0,1]^2$ where $\mathcal{D}$ is taken as Euclidean distance. The color assigned to each point on the left indicates the weight of elements in the universe which are within distance $\gamma = 0.1$ from the element under the uniform distribution $\mathcal{U}^*$. On the right, the points with at least weight $b = 0.04$ of $\mathcal{U}^*$ within distance $\gamma = 0.1$ are highlighted in blue. This example is $(\gamma = 0.1, a = .31, b = 0.04)-$dense for $\mathcal{U}^*$. That is, 31% of elements in the universe are within distance 0.1 of 4% of the rest of the universe.

distribution. The notion of $(\gamma, a, b)-$dense is intended to capture the weight (a) of elements that have a significant weight (b) on their close neighbors (distance γ) under $\mathcal{U}$ as a way to characterize how likely it is that a randomly chosen representative will be γ-close to a significant fraction of elements.

▶ **Definition 11** ($(\gamma, a, b)-$dense)**.** *Given a distribution $\mathcal{U}$ over U, a metric $\mathcal{D}$ is $(\gamma, a, b)-$dense for $\mathcal{U}$ if there exists a subset $A \subseteq U$ with weight a under $\mathcal{U}$ such that for all $u \in A$ $\Pr_{v \sim \mathcal{U}}[\mathcal{D}(u, v) \leq \gamma] \geq b$.*

$(p, \zeta)-$diffuse, defined below, captures what fraction of distances can tolerate a contraction proportional to ζ without becoming trivial.

▶ **Definition 12** ($(p, \zeta)-$diffuse)**.** *Given a distribution $\mathcal{U}$, a metric $\mathcal{D}$ is $(p, \zeta)-$diffuse if the fraction of distances between pairs of elements in $\mathcal{U} \times \mathcal{U}$ greater than ζ is p, i.e. $\Pr_{u,v \sim \mathcal{U} \times \mathcal{U}}[\mathcal{D}(u, v) \geq \zeta] \geq p$.*

A metric can be described by many combinations of density and diffusion parameters, as illustrated in Figure 2. These parameters are highly related, and we generally consider the combination of $(\gamma, a, b)-$dense and $(p, \frac{2\gamma}{1-c})-$diffuse. Although $\frac{2\gamma}{1-c}$ initially seems an arbitrary quantity, it indicates that a $p-$fraction of pairs will have distances preserved by a factor of c if the maximum contraction for those pairs is no more than 2γ. Thus the values of γ and c, which in turn dictate p, a, and b, (assuming $\zeta = \frac{2\gamma}{1-c}$) can loosely be seen as a tradeoff between how many pairs will have distance preservation guarantees and how significant the guarantees will be.

Nontriviality properties of $\gamma-$nets. Next, we relate the magnitude of γ to the non-triviality properties of the maxmerge of a set of representative submetrics. Lemma 13 states that a submetric based on a set of representatives which form a $\gamma-$net for a subset of U will have nontriviality properties related to the diffusion properties of $\mathcal{D}$ and the weight of the subset in $\mathcal{U}$.

▶ **Lemma 13.** *If a set of representatives $R \subseteq U$ form a $\gamma-$net for weight w of $\mathcal{U}$ and $\mathcal{D}$ is $(p, \frac{2\gamma}{1-c})-$diffuse on $\mathcal{U}$, then the submetric $\mathcal{D}_R(x,y) := \mathsf{maxmerge}(\{\mathcal{D}_r | r \in R\}, x, y)$ is $(p', c)-$nontrivial for $\mathcal{U}$, where $p' \geq p - (1-w)^2$.*

The proof follows from a worst-case analysis of the fraction of pairs with at least one element in the net with distance large enough that a 2γ contraction leaves at least a c-fraction of the original distance. The nontriviality guarantees of Lemma 13 are conservative, and we stress that our goal is to show the possibility of positive results, rather than achieving optimal performance or guarantees.

Representative set size. We now consider how likely it is that a set of random representatives drawn from $\mathcal{U}$ will form a $\gamma-$net for a significant fraction of $\mathcal{U}$. Lemma 14 characterizes the necessary representative set size based on the density and diffusion properties of the metric. The proof follows from characterizing the probability of "hitting" a sufficient weight of the distribution with a sample of a given size, and arguing that no element in our subset of interest can be more than 3γ far from any of the "hitting" elements.

▶ **Lemma 14.** *Given access to unlimited queries to the arbiter, if a metric $\mathcal{D}$ is $(\gamma, a, b)-$dense and $(p, \frac{6\gamma}{1-c})-$diffuse on $\mathcal{U}$, then a random set of representatives R of size at least $\frac{1}{b}\ln(\frac{1}{b\delta})$ will produce a $(p-(1-a)^2, c)$-nontrivial submetric for $\mathcal{U}$ with probability at least $1-\delta$.*

Random sampling is not the only method to construct a $\gamma-$net, and our strategy is motivated by simplicity as much as generality. In practice it may be more efficient to use the distance information from previously selected representatives to inform the selection of the next representative. For example, omitting or down-weighting any candidates that are already very close to existing representatives, or using a greedy strategy.[11]

1.6 Generalizing arbiter judgments

Now that we have shown how to construct a nontrivial submetric with ongoing access to the arbiter, we consider the problem of generalizing the arbiter's responses to unseen samples. Our goal is to construct efficient learners for submetrics as in Valiant's Probably Approximately Correct (PAC) model of learning [7]. However, we do not want to be too prescriptive about the submetric concept class, particularly about the representation of elements. Instead, we will make an assumption about the learnability of *threshold functions* and construct learning procedures for submetrics using threshold functions as building blocks without any additional direct access to labeled or unlabeled samples. More formally, our goal is to produce an efficient submetric learner, defined below.

▶ **Definition 15** (Efficient submetric learner). *A learning procedure is an efficient $\alpha-$submetric learner if for all $\varepsilon, \delta \in (0,1]$, given access to labeled examples, with probability at least $1-\delta$ over the randomness of the sampling and the learning procedure produces a hypothesis h_r such that $\Pr_{x,y \sim \mathcal{U}\times\mathcal{U}}[h_r(x,y) - \mathcal{D}(x,y) \geq \alpha] \leq \varepsilon$ in time $O(poly(\frac{1}{\varepsilon}, \frac{1}{\delta}))$.*

To show how to construct an efficient submetric learner, we first formalize our assumption of learnability of threshold functions. Next, we show how to combine the threshold function hypotheses for each representative to simulate rounding the distance between the representative and each element down to the nearest threshold. Finally, we specify the appropriate parameters for each component to achieve the desired bounds.

[11] Section 6 of the full paper contains proofs for Lemmas 9-14 and extended discussion of specialized strategies for representative selection, in particular strategies taking advantage of known metric structure.

Learnability of threshold functions. Assumption 1 (below) states that for every representative, there exists a set of thresholds and a learner for each threshold which, with high probability, produces an accurate hypothesis for the threshold function which generalizes to unseen samples.[12] ("With high probability" always refers to the probability over the randomness of sampling and the learner.) We first formally define a threshold function, which is a binary indicator of whether a particular element $u \in U$ is within distance $t \in [0, 1]$ of a representative r as $T^r_t(u) := \{1 \text{ if } \mathcal{D}(r, u) \leq t, 0 \text{ otherwise}\}$.

▶ **Assumption 1.** (Informal) *Given a metric $\mathcal{D}$ and a representative r, there exists a set of thresholds $\mathcal{T}$ such that $t \in [0, 1]$ for all $t \in \mathcal{T}$, $0 \in \mathcal{T}$, and $|\mathcal{T}| = O(1)$, and for every $t \in \mathcal{T}$ there exists an efficient learner $L^r_t(\varepsilon_t, \delta_t)$ which for all $\varepsilon_t, \delta_t \in (0, 1]$, with probability at least $1 - \delta_t$, produces a hypothesis h^r_t such that $\Pr_{x\sim\mathcal{U}}[h^r_t(x) \neq T^r_t(x)] \leq \varepsilon_t$ in time $O(poly(\frac{1}{\varepsilon_t}, \frac{1}{\delta_t}))$ with access to labeled samples of $T^r_t(u \sim \mathcal{U})$ for any distribution $\mathcal{U}$.*

Constructing submetric learners from threshold learners. Given Assumption 1, our next step is to determine how to combine the threshold learners into a learner for the representative submetric. (Notice that training data for the threshold function learners can be produced by post-processing the outputs of Algorithm 1.) Our strategy is similar to the rounding strategy used in Algorithm 1, using the threshold functions to identify the largest threshold which underestimates the distance between the representative and the element under consideration. The LinearVote mechanism takes in a set of hypotheses for the thresholds and outputs the threshold with which the most hypotheses agree. When all hypotheses output the correct value of their corresponding threshold function, LinearVote is equivalent to rounding $\mathcal{D}(r, x)$ down to the nearest threshold.

▶ **Definition 16** (LinearVote). *Given an ordered set of thresholds, $\mathcal{T} = \{t_1, t_2, \ldots, t_{|T|}\}$, and a set of hypotheses $H^r_{\mathcal{T}} = \{h^r_{t_1}, h^r_{t_2}, \ldots, h^r_{t_{|T|}}\}$, one corresponding to each threshold function,* $\mathsf{LinearVote}(\mathcal{T}, H^r_{\mathcal{T}}, x) := \arg\max_{t_i} \sum_{t_j<t_i}(1 - h^r_{t_j}(x)) + \sum_{t_j\geq t_i} h^r_{t_j}(x)$.

Algorithm 2 Pseudocode.

```
Inputs: error and failure probability parameters $\varepsilon, \delta$, density parameter $b$, a set of threshold function learners, the threshold set $\mathcal{T}$, and interfaces to the arbiter.
Sample a set of representatives $R \sim \mathcal{U}$ of size $\frac{1}{b}\ln(\frac{2}{b\delta})$ to produce $\gamma$−net with $\Pr \geq 1 - \frac{\delta}{2}$.
Generate labeled training data for the threshold learners via Algorithm 1.
Run each threshold learner $L^r_{t_i}$ with error parameters $\varepsilon_t \leftarrow \frac{\varepsilon}{2|R||\mathcal{T}|}$ and $\delta_t \leftarrow \frac{\delta}{2|R||\mathcal{T}|}$ to produce threshold function hypotheses $h^r_{t_i}$ for $r \in R$ and $t_i \in \mathcal{T}$.
For each representative, produce a hypothesis for distance from the representative by taking $h_r(x, y) := |\mathsf{LinearVote}(\mathcal{T}, \{h^r_{t_i}|t_i \in \mathcal{T}\}, x) - \mathsf{LinearVote}(\mathcal{T}, \{h^r_{t_i}|t_i \in \mathcal{T}\}, y)|$.
Combine the hypotheses for each representative into $h_R(x, y) := \mathsf{maxmerge}(\{h_r|r \in R\}, x, y)$.
return $h_R$.
```

Algorithm 2 combines all of our constructions thus far to create an efficient submetric learner: it chooses a set of representatives, learns threshold functions for each threshold for each representative, and combines the resulting hypotheses using LinearVote and maxmerge

[12] The formal statement of Assumption 1 is included in Section 5.1 of the full paper.

to produce a single submetric hypothesis.[13] Theorem 17 builds on the result of Lemma 14 and concludes that the parametrization of Algorithm 2 results in an efficient submetric learner.

▶ **Theorem 17.** *[Informal] Given a distance metric $\mathcal{D}$, and a distribution $\mathcal{U}$ over the universe, if there exist a set of thresholds $\mathcal{T}$ with maximum gap $\alpha_{\mathcal{T}}$ and efficient learners $\{L^{r}_{t_i \in \mathcal{T}}\}$ as in Assumption 1, and $\mathcal{D}$ is $(\gamma, a, b)-$dense and $(p, \frac{6\gamma+\alpha_{\mathcal{T}}}{1-c})-$diffuse on $\mathcal{U}$, then there exists an efficient $\alpha_{\mathcal{T}}$-submetric learner which produces a hypothesis h_R such that h_R is $(p-(1-a)^2-\varepsilon, c)-$nontrivial for $\mathcal{U}$.*

The proof of Theorem 17 follows from an analysis of the error parameter propagation. We briefly give some intuition for the analysis and implications of the theorem. First, the magnitude $\alpha_{\mathcal{T}}$ error follows from the same single direction rounding argument as for Algorithm 1. The error probability follows from noticing that at least one threshold function must be in error for one of the elements to result in an error in LinearVote. The failure probability "budget" is split evenly between failure to choose a good set of representatives (Line 1) as specified in Lemma 14, and failure of the underlying learning procedures (Line 3) derived by union bound. Compared with Lemma 14, the diffusion and nontriviality parameters are adjusted to take into account the additional rounding error magnitude of $\alpha_{\mathcal{T}}$ introduced by LinearVote and the combined hypothesis error probability ε. In practice, we expect that the set of thresholds which are learnable are unlikely to occur at regular intervals. Post-processing is a valuable tool to reduce the magnitude of $\alpha_{\mathcal{T}}$ (by re-mapping the threshold values in step 4 to reduce the maximum gap), but comes at the cost of reduced nontriviality guarantees.

The desired query complexity to the arbiter follows from basic analysis of the parameters. However, the query complexity bound can be improved significantly by observing that no independence of errors between threshold functions is assumed, allowing a single call to Algorithm 1 for each representative (rather than $|\mathcal{T}|$ calls). The dependence on $|R|$ can also be improved to logarithmic by sorting a single merged list of (representative, element) pairs, but we defer detailed discussion to Sections 5 and 6 of the full paper.

1.7 Relaxing the query model

Our results extend to a relaxed model in which arbiters are not expected to make arbitrarily small distinctions between distances or individuals and may answer real-valued queries with bounded noise. The relaxed model assumes that there are two fixed constants, α_L, the minimum precision with which the arbiter can distinguish elements or distances, and α_H, a bound on the magnitude of the (potentially biased) noise in the arbiter's real-valued responses. For any comparisons with difference smaller than α_L, the arbiter declares the elements indistinguishable or the difference "too close to call." The model allows for a "gray area" between α_L and α_H in which the arbiter may either respond with the true answer or "too close to call." For any differences larger than α_H, the arbiter responds with the true answer.

For the most part, our results translate to the relaxed model with minimal modification to the logic of the proofs to handle two-sided error in real-valued queries. Interestingly, the real-value query complexity improves to constant, as the worst-case behavior in Algorithm 1 is avoided as the arbiter "knows" not to worry about inconsequentially small distances. However, this does result in additional error magnitude, so the improved query complexity

[13] See Sections 5 and 6 in the full version.

does not come for free. Furthermore, unlike the exact model we won't necessarily be able to label a sample with perfect accuracy for every threshold function learner due to the bi-directional error. To handle this labeling problem, we modify the distribution of samples presented to each learner, eliminating samples whose labels are ambiguous, again resulting in increased error. Formal results in the relaxed model are discussed in Section 7 of the full paper.

References

1 Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

2 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.

3 Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2605–2614, 2018.

4 Christina Ilvento. Metric learning for individual fairness. *CoRR*, abs/1906.00250, 2019. arXiv:1906.00250.

5 Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4847–4857, 2018.

6 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017.

7 Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

8 Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

9 Michael J Wilber, Iljung S Kwak, and Serge J Belongie. Cost-effective hits for relative similarity comparisons. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

10 Gal Yona and Guy N. Rothblum. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674, 2018.

Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?

Avrim Blum
Toyota Technological Institute at Chicago, 6045 South Kenwood Avenue, Chicago, IL, 60637, USA
avrim@ttic.edu

Kevin Stangl
Toyota Technological Institute at Chicago, 6045 South Kenwood Avenue, Chicago, IL, 60637, USA
kevin@ttic.edu

Abstract

Multiple fairness constraints have been proposed in the literature, motivated by a range of concerns about how demographic groups might be treated unfairly by machine learning classifiers. In this work we consider a different motivation; learning from biased training data. We posit several ways in which training data may be biased, including having a more noisy or negatively biased labeling process on members of a disadvantaged group, or a decreased prevalence of positive or negative examples from the disadvantaged group, or both. Given such biased training data, Empirical Risk Minimization (ERM) may produce a classifier that not only is biased but also has suboptimal accuracy on the true data distribution. We examine the ability of fairness-constrained ERM to correct this problem. In particular, we find that the Equal Opportunity fairness constraint [14] combined with ERM will provably recover the Bayes optimal classifier under a range of bias models. We also consider other recovery methods including re-weighting the training data, Equalized Odds, and Demographic Parity, and Calibration. These theoretical results provide additional motivation for considering fairness interventions even if an actor cares primarily about accuracy.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases fairness in machine learning, equal opportunity, bias, machine learning

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.3

Funding *Avrim Blum*: Supported in part by the National Science Foundation under grants CCF-1815011 and CCF-1733556.
Kevin Stangl: Supported in part by the National Science Foundation under grant CCF-1815011.

Acknowledgements We would like to thank Jon Kleinberg and Manish Raghavan for their helpful and insightful comments on an earlier draft of this manuscript.

1 Introduction

Machine learning (typically supervised learning) systems are automating decisions that affect individuals in sensitive and high stakes domains such as credit scoring [7] and bail assignment [2, 12]. This trend toward greater automation of decisions has produced concerns that learned models may reflect and amplify existing social bias or disparities in the training data. Examples of possible bias in learning systems include the Pro-Publica investigation of COMPAS (an actuarial risk instrument) [2], accuracy disparities in computer vision systems [5], and gender bias in word vectors [4].

In order to address observed disparities in learning systems, an approach that has developed into a significant body of work is to add demographic constraints to the learning problem that encode criteria that a fair classifier ought to satisfy.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 3; pp. 3:1–3:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Multiple constraints have been proposed in the literature [14, 11], each encoding a different type of unfairness one might be concerned about, and there has been substantial work on understanding their relationships to each other, including incompatibilities between the fairness requirements [8, 6, 16, 20].

In this work, we take a different angle on the question of fairness. Rather than argue whether or not these demographic constraints encode intrinsically desirable properties of a classifier, we instead consider their ability to help a learning algorithm to recover from biased training data and to produce a *more accurate* classifier.

In particular, adding a constraint (such as a fairness constraint) to an optimization problem (such as ERM) would typically result in a lower quality solution. However, if the objective being optimized is skewed (e.g., because training data is corrupted or not drawn from the correct distribution) then such constraints might actually help prevent the optimizer from being led astray, and yield a higher quality solution when accuracy is *measured on the true distribution.*

More specifically, we consider a binary classification setting in which data points correspond to individuals, some of whom are members of an advantaged Group A and the rest of whom are members of a disadvantaged Group B. We want to make a decision such as deciding whether to offer a candidate a loan or admission to college. We have access to labeled training data consisting of (x, y) pairs where x is some set of features corresponding to an individual and y is a label we want to predict for new individuals.

The concern is that the training data is potentially biased against Group B in that *the training data systematically misrepresents the true distribution over features and labels in Group* B, while the training data for Group A is drawn from the true distribution for Group A. We consider several natural ways this might occur. One way is that members of the disadvantaged group might show up in the training data at a lower rate than their true prevalence in the population, and worse, *this rate might depend on their true label.*

For instance, if the positive examples of Group B appear at a much lower rate in the training data than the negative examples of Group B (which might occur for cultural reasons or due to other options available to them), then ERM might learn a rule that classifies all or most members of Group B as negative.

A second form of bias in the training data we consider is bias in the labeling process. Human labelers might have inherent biases causing some positive members of Group B in the training data to be mislabeled as negative, which again could cause unconstrained ERM to be more pessimistic than it should be. Alternatively, both processes might occur together. We examine the ability of fairness constraints to help an ERM learning method recover from these problems.

1.1 Summary of Results

Our main result is that ERM subject to the **Equal Opportunity** fairness constraint [14] recovers the true Bayes optimal hypothesis under a wide range of bias models, making it an attractive choice even for decision makers whose overall concern is purely about accuracy on the true data distribution.

In particular, we assume that under the true data distribution, the Bayes optimal classifiers h_A^* and h_B^* classify the same fraction p of their respective populations as positive[1], h_A^* and h_B^* have the same error rate η on their respective populations, and that these errors are uniformly distributed.

[1] $p = P_{\mathcal{D}_A}(h_A^*(x) = 1) = P_{\mathcal{D}_B}(h_B^*(x) = 1)$. We will allow the classifiers to make decisions based on group membership or alternatively assume we have sufficiently rich data to implicitly infer the group attribute.

However, during the training process we do not have access to the true distribution. We only have access to a biased distribution in a way that implicates the distinct social groups and causes the classifier to be overly pessimistic on individuals from Group B.

We prove that, subject to the above conditions on h^*_A and h^*_B, even with substantially corrupted training data either due to the under-representation of positive examples in Group B or a substantial fraction of positive examples in Group B mislabeled as negative, or both, the Equality of Opportunity fairness constraint will enable ERM to learn the Bayes optimal classifier $h^* = (h^*_A, h^*_B)$, subject to a pair of inequalities ensuring that the labels are not too noisy and Group A has large mass.

Expressed another way, this means that *the lowest error classifier on the biased data satisfying Equality of Opportunity is the Bayes optimal classifier on the un-corrupted data.* These results provide additional motivation for considering fairness interventions, and in particular Equality of Opportunity, even if one cares primarily about accuracy.

Other related fairness notions such as Equalized Odds, Demographic Parity, and Calibration do not succeed in recovering the Bayes optimal classifier under such broad conditions. In fact, we show that given data subject to Under-Representation Bias, Calibration can actually *amplify* the effects of the bias, and so can be worse than doing nothing and instead learning with plain ERM (see Section 3.1).

Our results are in the infinite sample limit and we suppress issues of sample complexity[2] in order to focus on the core phenomenon of the data source being unreliable.

1.2 Related Work

This paper is directly motivated by a model of implicit bias in ranking [17]. In that paper, the training data for a hiring process is systematically corrupted against minority candidates and a method to correct this bias increases both the quality of the accepted candidate and the fraction of hired minority candidates. However, that fairness intervention, the Rooney Rule, does not immediately translate to a general learning setting.

Our results avoid triggering the known impossibility results between high accuracy and satisfying fairness criteria [6, 16] by assuming we have equal base rates across groups. This assumption may not be realistic in all settings, however there are settings where bias concerns arise and there is empirical evidence that base rates are equivalent across the relevant demographic groups, e.g. highly differential arrest rates for some alleged crimes that have similar occurrence rates across groups [19, 21].

Within the fairness literature there are several approaches similar to ours. In particular, our concern with positive examples not appearing in the training data is similar in effect to a selective labels problem [18]. [9] uses data augmentation to experimentally improve generalization under selective label bias.

[13, 22] also consider the training and test data distribution gap we experience in our model and posit differing interpretations of fairness constraints under different worldviews. While we do not explicitly use the terminology in these papers, we believe our view of the gap between the true distribution and the training time distribution is aligned with Friedler et al's concept of the gap between the construct space and the observed space.

[2] Our notion of sample complexity is typical. Let S be the biased training data-set and $ERM_{\mathcal{H}}(S) = \hat{h}$. Given $\epsilon, \delta > 0$, $m(\epsilon, \delta)$ samples ensures with probability greater than $1 - \delta$ that $L_{\mathcal{D}}(\hat{h}) \leq L_{\mathcal{D}}(h^*) + \epsilon$.

Our second bias model, Labeling Bias, is similar to [15]. In that paper, the bias phenomenon is that a biased labeler makes poor decisions on the disadvantaged group and intervenes with a reweighting technique, one that is more complex than our Re-Weighting intervention. However, that paper does not consider the interaction of biased labels with different groups appearing in the data at different rates as a function of their labels.

2 Model

In this section we describe our learning model, how bias enters the data-set, and the fairness interventions we consider.

We assume the data lies in some instance space $\mathcal{X}$, such as $\mathcal{X} = \mathbb{R}^d$. There are two demographic groups in the population, Group A and Group B. Their proportions in the population are given by $P(x \in A) = 1 - r$ and $P(x \in B) = r$ for $r \in (0, 1)$. $x \in A$ can be read as individual x in demographic Group A. Group B is the disadvantaged group that suffers from the effects of the bias model.

Assume there is a special coordinate of the feature vector x that denotes group membership. The data distribution is given by $\mathcal{D}$, and is a pair distributions $(\mathcal{D}_A, \mathcal{D}_B)$, with $\mathcal{D}_A$ determining how $x \in A$ is distributed and $\mathcal{D}_B$ determining how $x \in B$ is distributed.

2.1 True Label Generation

Now we describe how the true labels for individuals are generated. Assume there exists a pair of Bayes optimal classifiers $h^* = (h^*_A, h^*_B)$ with $h^*_A, h^*_B \in \mathcal{H} : \mathcal{X} \to \{0, 1\}$.

We assume that the Bayes optimal classifier h^*_B for Group B may be different from the Bayes optimal classifier h^*_A for Group A. If h^*_A was also optimal for Group B, then we can just learn h^* for both Groups A and B using data only from Group A and biased data concerns fade away. Thus we are learning a pair of classifiers, one for each demographic group.

When generating samples, first we draw a data-point x. With probability $1 - r$, $x \sim \mathcal{D}_A$ (and thus $x \in A$) and with probability r, $x \sim \mathcal{D}_B$ (so $x \in B$).

Once we have drawn a data-point x, we model the true labels as being produced as follows; evaluate $h^*(x)$, using the classifier corresponding to the demographic group of x. If $x \in A$, then $h^*(x) = h^*_A(x)$. If $x \in B$, then $h^*(x) = h^*_B(x)$. However, we assume that h^* is not perfect and independently with probability η, the true label of x does not correspond to the prediction $h^*(x)$.

$$y = y(x) = \begin{cases} \neg\, h^*(x) & \text{with probability} \quad \eta \\ h^*(x) & \text{w.p.} \quad 1 - \eta \end{cases}$$

The labels y after this flipping process are the *true labels* of the training data.[3] We assume that $p = P(h^*_A(x) = 1 | x \in A) = P(h^*_B(x) = 1 | x \in B)$. This combined with the assumption that η is the same for classifiers from both groups implies that the two groups have equal base rates (fraction of positive samples) i.e $p(1 - \eta) + (1 - p)\eta$ (un-normalized).

We denote this label model as $(x, y) \sim P_{\mathcal{D},r}(h^*, \eta)$ for a pair of classifiers $h^* = (h^*_A, h^*_B)$ with $h^*_A, h^*_B \in \mathcal{H}$ where $\mathcal{H} : \mathcal{X} \to \{0, 1\}$ is some hypothesis class with finite VC dimension.

[3] Note this label model is equivalent to the Random Classification Noise model [1]. However the key interpretative difference is that in RCN, $h^*(x)$ is the correct label and those that get flipped are noise, but in our case the y are the true labels and h^* is merely the Bayes optimal classifier given the observed features.

2.2 Biased Training Data

Now we consider how bias enters the data-set. Consider the example of hiring where the main failure mode will be a classifier that is too negative on the disadvantaged group. We explore several different bias models to capture potential ways the data-set could become biased.

The first bias model we call **Under-Representation Bias**. In this model, the positive examples from Group B are under-represented in the training data. Specifically, the biased training data is drawn as follows:

1. m examples are sampled from the distribution $\mathcal{D}$. Thus each $x \sim \mathcal{D}$.
2. The label y for each x is generated according to the label process from Section 2.1 with hypothesis $h^* = (h^*_A, h^*_B)$ and η.
3. For each pair (x, y), if $x \in B$ and $y = 1$, then the data-point (x, y) is discarded from our training set independently with probability $1 - \beta$.

Thus we see fewer positive examples from Group B in our training data. β is the probability a positive example from Group B stays in the training data and $1 > \beta > 0$.

If $\eta = 0$, then the positive and negative regions of h^* are strictly disjoint, so if we draw sufficiently many examples, with high probability, we will see enough positive examples in the positive domain of h^* to find a low empirical error classifier that is equivalent to h^*.[4]

In contrast for non-zero η, our label model interacting with the bias model can induce a problematic phenomenon that fools the ERM classifier. For non-zero η there is error even for the Bayes optimal classifier h^* and thus in the region classified as positive by the Bayes optimal classifier h^* there are positive examples mixed with negative examples. The fraction of negative examples is amplified by the bias process.

If β is sufficiently small, there could in fact be more negative examples of Group B than positive examples in the positive region of h^*_B. If this occurs, then the bias model will snap the unconstrained ERM optimal hypothesis (optimal on the biased data) to classifying all individuals from Group B as negatives. This can be observed in Figure 1.

Under-Representation Bias is related to selective labels in [18] since we are learning on a filtered distribution where the filtering process is correlated with the group label. Our model is functionally equivalent to over-representing the negatives of the in the training data, an empirical phenomenon observed in [21].

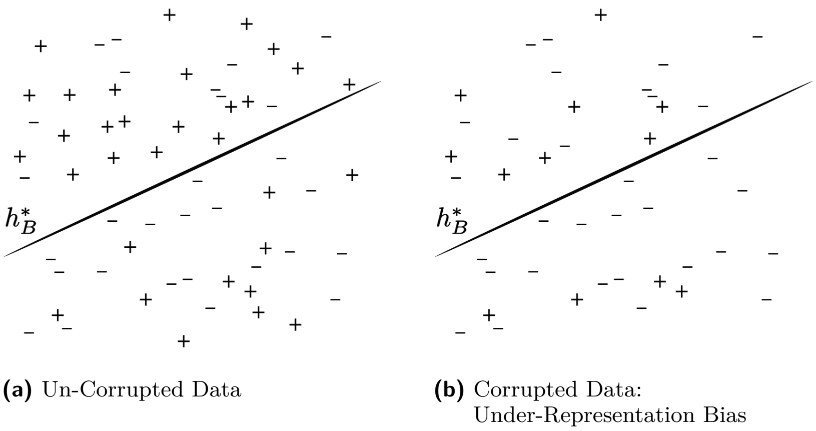
2.3 Alternative Bias Model: Labeling Bias

We now consider a bias model that captures the notion of implicit bias, which we call **Labeling Bias.** In particular, a possible source of bias in machine learning is the label generating process, especially in applications where the sensitive attribute can be inferred by the labeler, consciously or unconsciously. For example, training data for an automated resume scoring system could be based upon the historical scores of resumes created by a biased hiring manager or a committee of experts. This source of labels could then systematically score individuals from Group B as having lower resume scores, an observation noted in randomized real world investigations [3].

Formally, the labeling bias model is:

1. m examples are sampled from the distribution $\mathcal{D}$. Thus each $x \sim \mathcal{D}$.
2. The labels y for each x are generated according to the label process from Section 2.1 with hypothesis $h^* = (h^*_A, h^*_B)$ and η.
3. For each pair (x, y), if $x \in B$ and $y = 1$, then independently with probability ν, the label of this point is flipped to negative.

[4] We would learn with ERM and Uniform Convergence, using the fact that $\mathcal{H}$ has finite VC-dimension.



(a) Un-Corrupted Data

(b) Corrupted Data: Under-Representation Bias

Figure 1 The schematic on the left displays data points with $p = 1/2$, h_B^* as a hyperplane, and $\eta = 1/3$. The schematic on the right displays data drawn from the same distribution subject to the Under-Representation Bias with $\beta_{POS} = 1/3$. Now there are more negative examples than positive examples above the hyperplane so the lowest error hypothesis classifies all examples on the right as negative.

This process is one-sided, so true positives become negatives in the biased training data, so apparent negatives becomes over-represented. We are making a conceptual distinction that the *true* labels (Step 2) are those generated by the original label model and these examples that get flipped by the bias process (Step 3) are not really negative, instead they are just mislabeled.

As ν increases, more and more of the individuals in the minority group appear negative in the training data. Once the number of positive samples is smaller than the number of negative samples above the decision surface h_B^*, then the optimal unconstrained classifier (according to the biased data) is to simply classify all those points as negative.

2.4 Under-Representation Bias and Labeling Bias

We now consider a more general model that combines Under-Representation Bias and Labeling Bias, and moreover we allow either positives or negatives of Group B (or both) to be under-represented. Specifically, we now have *three* parameters: β_{POS}, β_{NEG}, and ν. Given m examples drawn from $P_{\mathfrak{D},r}(h^*, \eta)$, we discard each positive example of Group B with probability $1 - \beta_{POS}$ and discard each negative example of Group B with probability $1 - \beta_{NEG}$ to model the Under-Representation Bias. Next, each positive example of Group B is mislabeled as negative with probability ν to model the Labeling Bias. Note that the under-representation comes first: β_{POS} and β_{NEG} represent the probability of *true* positive and *true* negative examples from Group B staying in the data-set, respectively, regardless of whether they have been mislabeled by the agent's labelers.

2.5 Fairness Interventions

Now we introduce several fairness interventions and define a notion of successful recovery from the biased training distribution.

We consider multiple fairness constraints to examine whether the criteria have different behavior in different bias regimes. The fairness constraints we focus on are **Equal Opportunity**, **Equalized Odds**, **Demographic Parity**, and **Calibration**.

▶ **Definition 1.** *Classifier h satisfies **Equal Opportunity** on data distribution $\mathcal{D}$ [14] if*

$$P_{(x,y)\sim\mathcal{D}}(h(x)=1|y=1,x\in A)=P_{(x,y)\sim\mathcal{D}}(h(x)=1|y=1,x\in B) \qquad (1)$$

This requires that the true positive rate in Group B is the same as the true positive rate in Group A.

Equalized Odds is a similar notion, also introduced in [14]. In addition to requiring Line 1, Equalized Odds also requires that the false positive rates are equal across both groups. Equivalently, we can define **Equalized Odds** as $h\perp A|Y$, meaning that h is independent of the sensitive attribute, conditioned on the true label Y. We also consider **Demographic Parity** $:= P(h(x)=1|x\in A)=P(h(x)=1|x\in B)$ [11]. For each of these criteria, the overall training procedure is solving a constrained ERM problem.[5]

An alternative intervention we study **data Re-Weighting**, where we change the training data distribution to correct for the bias process and then do ERM on the new distribution. The overall gist of how the training data becomes biased in our models is that the positive samples from Group B are under-represented in the training data so we can intervene by up-weighting the observed fraction of positives in the training data from Group B to match the fraction of positives from the Group A training data.

In the training process we only have access to samples from the training distribution and thus when using a fairness criterion to select among models *we check the requirement on the biased training data.*

The last fairness intervention we consider is **Calibration**. Calibration [12, 10, 6, 20] requires that when interpreted as probabilities, the same score communicates the same information for individuals from different demographic groups. Specifically, in the bucket of individuals receiving score s, the same fraction in both demographic groups is in fact truly positive. We focus on Calibration for the case of our binary classifier where there are only two scores, e.g. the scores 0 and 1, so in order for classifier $h=(h_A,h_B)$ to satisfy Calibration, the following equalities must hold. [6]

$$P_{x\sim\mathcal{D}_A}(y=1|h_A(x)=1)=P_{x\sim\mathcal{D}_B}(y=1|h_B(x)=1)$$
$$P_{x\sim\mathcal{D}_A}(y=1|h_A(x)=0)=P_{x\sim\mathcal{D}_B}(y=1|h_B(x)=0)$$

While the other fairness criteria are vigorously debated, Calibration is less contested as an important desiderata of machine learning models. Calibration has been used to defend the epistemic validity of risk prediction instruments [12, 10] and it is claimed that mis-calibrated classifiers may have serious harms and induce undesirable behavior when scores are used by a human actor [20].

Observe that in our model of label generation, the Bayes optimal classifier on the true distribution is the h^* used to generate the labels initially, regardless of the values of η and r. Thus our goal for the learning process is to recover the original optimal classifier h^*, subject to training data from a range of bias models and the true label process with $(x,y)\sim P_{\mathcal{D},r}(h^*,\eta)$. A more effective learning method would recover h^* in a wider range of the model parameters (the parameters that characterize the bias process and the true label process). Accordingly we define **Strong-Recovery**(r,η):

▶ **Definition 2.** *A Fairness Intervention in bias model B satisfies Strong-Recovery(r_0,η_0) if for all $\eta\in[0,\eta_0)$ and all $0<r<r_0$, when given data corrupted by bias model B, the training procedure recovers the Bayes optimal classifier h^*, given sufficient samples, for all $\beta_{POS},\beta_{NEG}\in(0,1]$, $\nu\in[0,1)$, and $p\in(0,1]$.*

[5] We do not consider methods for efficiently solving the constrained ERM problem.

[6] If one of the conditioned on events never occurs, such as a classifier that never classifies anyone from Group B as positive, we treat the associated equality as satisfied.

3 Recovery Behavior Across Bias Models

There are two failure modes for learning a fairness constrained classifier that we will need to be concerned with. First, the Bayes optimal hypothesis may not satisfy the fairness constraint evaluated on the biased data. Second, within the set of hypotheses satisfying the fairness constraint, another hypothesis (with higher error on the true distribution) may have lower error than the Bayes optimal classifier h^* on the biased data. We now describe how the multiple fairness interventions provably avoid or fail to avoid these pitfalls in increasingly complex bias models. We defer formal proofs to Section 4.

3.1 Under-Representation Bias

Equal Opportunity and Equalized Odds both perform well in this bias model and avoid both failure modes, subject to an identical constraint on the bias and demographic parameters.

First, from the definition of the Under-Representation Bias model, observe that h^* satisfies both fairness notions on the biased data, so the first failure mode does not occur.

Second, Equal Opportunity intuitively prevents the failure mode where a hypothesis is produced that appears better than h^* on the biased data, such as classifying all examples from Group B as negative, by forcing the two classifiers to classify the same fraction of positive examples as positive. So, if we classify all the examples from Group B as negative, we have to do the same with Group A, inducing large error on the training data from the majority Group A. In particular, so long as the fraction r of total data from Group B is not too large and η is not too close to $1/2$, this will not be a worthwhile trade-off for ERM (saying negative on all samples will not have lower perceived error on the biased data than h^*) and so it will not produce this outcome.

A formal proof of correctness is given in Section 4.1. Specifically, we prove that Equal Opportunity strongly recovers from Under-Representation Bias so long as

$$(1-r)(1-2\eta) + r((1-\eta)\beta - \eta) > 0 \tag{2}$$

Note that this is true for all $\eta < 1/3$ and $r \in (0, 1/2)$, so we have that Equal Opportunity satisfies Strong-Recovery(1/2,1/3) from Under-Representation Bias. Alternatively, we see that if $r = 1/4$ then the inequality simplifies to at least $3/4(1-2\eta) - \eta/4 = 3/4 - (7/4)\eta$ so we have Strong-Recovery$(1/4, 3/7)$. Equalized Odds also recovers in this bias model with the same conditions as Equal Opportunity.

In contrast, Demographic Parity fails to recover h^* even if $\eta = 0$. If $p = 1/2$, $\eta = 0$, and $\beta = 1/2$ and we originally had n samples, then the Bayes optimal classifier does not satisfy Demographic Parity on the biased data since the fraction of samples that will be labelled positive is $\frac{1}{3} \neq \frac{1}{2}$.

Similarly, if we let $\eta \neq 0, \beta < 1$, then in order to match the fraction of positive classifications made by h_A^*, h_B is forced to classify a larger region of the input spaces as positive than h_B^* would in the absence of biased data and so we do not recover h_B^*.

Another way to intervene in the Under-Representation Bias model would just be to re-weight the training data to account for the under-sampling of positives from Group B. If we really know positives from Group B are under-represented, we can change our objective function $min \sum_{i=1}^{m} I(h(x) \neq y)$ by changing each indicator function such that minimizing the sum of indicators measures the loss on the true distribution and not the loss on the biased training distribution.

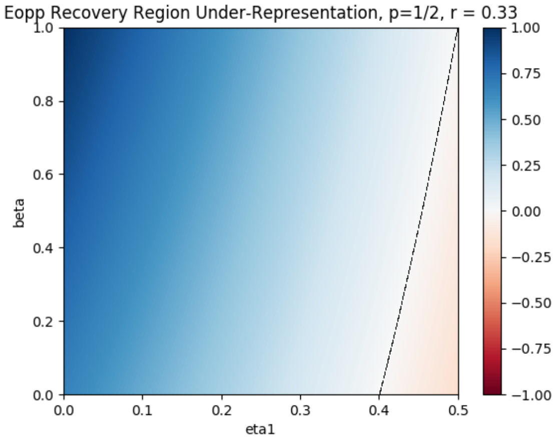


Figure 2 This figure indicates the parameter region such that Equal Opportunity Constrained ERM recovers h^* under the Under-Representation Bias Model and is a visualization of Equation 2. $r = 1/3$ and $p = 1/2$. We label each pair (η, β) with blue if it satisfies the inequality and red otherwise. This plot shows how smaller η means we can recover from lower β. Blue means h^* is recovered. The dashed black line indicates the boundary between recovering h^* and failing to recover h^*.

Define $B^+ = \{x \in B \; s.t. \; y = 1\}$. Then let,

$$I'(h(x), y) = \begin{cases} \frac{1}{\beta} \quad h(x) \neq 1 \quad and \quad x \in B^+ \\ 0 \quad h(x) = 1 \quad and \quad x \in B^+ \\ I(h(x) \neq y) \quad otherwise \end{cases}$$

Then we use this new indicator in the objective function. This new loss function is an unbiased estimator of the true unbiased risk, so uniform convergence on this estimator will suffice to learn h^*. We can infer the value of β from the data for Group A if we know the data from Group B is corrupted by this bias model. One concern with re-weighting in general is that the functional form of the correction is tied to the exact bias model.

As we show in Section 5, Calibration has strange results in this bias model. Specifically, when the bias is such that ERM fails to recover h^* (i.e when $(1 - \eta)\beta < \eta$), then the Calibration constraint can only be satisfied by a trivial classifier that assigns all of Group A to one label and all of Group B to the alternative label. For typical parameters, this will result in Group B being given the negative label and Group A will be assigned as all positive. This will not recover h^* and is in fact substantially worse than merely using ERM. Un-constrained ERM would learn badly on Group B but would recover h^*_A for Group A.

When the bias regime is such that $(1 - \eta)\beta > \eta$, plain ERM recovers h^*, while enforcing Calibration will lead to excess true error on both demographic groups over the true error of h^*. In particular, satisfying Calibration on the biased data requires intentionally classifying some negative input space from Group A as positive and classifying some positive input space from Group B as negative. These results suggest that Calibration is an actively harmful intervention (for both groups) in our model, when compared to plain ERM, across all model parameters.

In summary, for the Under-Representation Bias model, the fairness interventions Equalized Odds, Equal Opportunity, and Re-Weighting recover h^* under a range of parameters. However, Demographic Parity is inadequate even for $\eta = 0$ and will not recover h^* for non-vacuous bias parameters.

3.2 Labeling Bias

In Section 4, we prove that Equal Opportunity constrained ERM on data biased by the Labeling Bias model also finds the Bayes optimal classifier, under similar parameter conditions to the previous bias model.

Interestingly, in contrast to Under-Representation Bias, *Labeling Bias cannot be corrected* by *Equalized Odds*. The problem is the first failure mode. For example, consider $\eta = 0$ but where $\nu \neq 0$. The Bayes optimal classifier h^*_A for Group A has false positive rate of 0 and true positive rate of 1. However, since $\nu > 0$, there is no classifier for Group B that achieves both of these rates simultaneously. In particular, the only way to classify the negative individuals in the positive region as negative is for the classifier to decrease its true positive rate from 1. Therefore, Equalized Odds rules out usage of h^*_A. This violation holds for $\eta \neq 0$ as well.

In contrast, h^* does satisfy Equal Opportunity on the biased data, and given the conditions in Theorem 3, it will be the lowest error such classifier on the biased data.

Demographic Parity experiences similar limitations as in the Under-Representation Bias model.

The Re-Weighting intervention is to change the weighting of observed positives in the training data for Group B so that we have the same fraction of positives in Group B as in Group A. Define $p_{A,1}$:= the fraction of positive individuals in Group A and $p_{B,1}$:= the *observed* fraction of positives in B in the biased data. $p_{A,0}$ and $p_{B,0}$ refer to the observed fraction of negative individuals in Group A and Group B in the biased data.

We need a re-weighting factor Z such that:

$$\begin{aligned} \frac{p_{A,1}}{p_{A,0}} &= \frac{Zp_{B,1}}{p_{B,0}} \\ \frac{p_{A,1}}{1-p_{A,1}} &= \frac{Zp_{A,1}(1-\nu)}{p_{A,0}+p_{A,1}\nu} = \frac{Zp_{A,1}(1-\nu)}{1-p_{A,1}+p_{A,1}\nu} \\ Z &= \frac{1-p_{A,1}(1-\nu)}{(1-\nu)(1-p_{A,1})} \end{aligned}$$

We prove in Section 4.2 that this correction factor will lead to the positive region of h^*_B having a higher weight of positive examples than negative examples and simultaneously the negative region of h^*_B having a higher weight of negative examples than positive examples. This causes ERM to learn the optimal hypothesis h^*. We can infer the value of ν by comparing the fraction of positives in Group A and Group B.

In summary, Equal Opportunity and the Re-Weighting Interventions recover well in this bias model (Labeling Bias) while Equalized Odds and Demographic Parity are inadequate.

3.3 Under-Representation Bias and Labeling Bias

In this most general model that combines the two previous models, Re-Weighting the data is now no longer sufficient to recover the true classifier. For example, consider the case where $\eta = 0$ and $p = 1/4$, $\nu = 1/2$ and $\beta_{NEG} = 1/3$ and $\beta_{POS} = 1$. If there were n points originally from group B, then in expectation $3n/4$ were negative and $n/4$ were positive. After the bias process, in expectation there are $n/4$ negatives on the negative side of h^*, and on the positive side of h^* we have $n/8$ correctly labelled positives and what appear to be $n/8$ negative samples.

The Re-Weighting intervention will not do anything in expectation because the overall fractions are still correct; we have $n/2$ total points with one quarter of them labeled positive. ERM is now indifferent between h^* and labeling all samples from Group B as negative. If we just slightly increase the parameter ν and reduce β_{POS} then in expectation ERM will strictly prefer labeling all the samples negatively.

While the Re-Weighting method fails, we prove that Equal Opportunity-constrained ERM recovers the Bayes optimal classifier h^* as long as we satisfy a condition ensuring that Group A has sufficient mass and the signal is not too noisy. As with the previous model, Demographic Parity and Equalized Odds are not satisfied by h^* on minimally biased data and so they will not recover the Bayes optimal classifier.

4 Main Results

We now present our main theorem formally. Define the biased error of a classifier h as its error rate computed on the biased distribution.

▶ **Theorem 3.** *Assume true labels are generated by $P_{\mathcal{D},r}(h^*, \eta)$ corrupted by both Under-Representation bias and Labeling bias with parameters $\beta_{POS}, \beta_{NEG}, \nu$, and assume that*

$$(1-r)(1-2\eta)+r((1-\eta)\beta_{POS}(1-2\nu)-\eta\beta_{NEG}) > 0 \tag{3}$$

and

$$(1-r)(1-2\eta)+r((1-\eta)\beta_{NEG}-(1-2\nu)\beta_{POS}\eta) > 0 \tag{4}$$

Then $h^* = (h^*_A, h^*_B)$ is the lowest biased error classifier satisfying Equality of Opportunity on the biased training distribution and thus h^* is recovered by Equal Opportunity constrained ERM.

Note $\beta_{POS}, \beta_{NEG} \in (0,1]$, $\nu \in [0,1)$, $\eta \in [0,1/2)$, $r \in (0,1)$ and $p \in (0,1]$. Condition 3 refers to Equation 3 and Equation 4.

This case contains our other results as special cases and in the next section we prove our main theorem in this bias model. Note that if Equation 3 is not satisfied then the all-negative hypothesis will have the lowest biased error among hypotheses satisfying Equal Opportunity on the biased training distribution. Similarly, if Equation 4 is not satisfied then the all-positive hypothesis will have the lowest biased error among hypotheses satisfying Equal Opportunity on the biased training distribution. Thus Theorem 3 is tight. To give a feel for the formula in Theorem 3, note that the case of small r is *good* for our intervention, because the advantaged Group A is large enough to pull the classification of the disadvantaged Group B in the right direction. For example, if $r \leq \frac{1}{3}$ then the bounds are satisfied for all $\eta < \frac{1}{4}$ (and if $r \leq \frac{1}{4}$ then the bounds are satisfied for all $\eta < \frac{1}{3}$) for *any* under-representation biases $\beta_{POS}, \beta_{NEG} > 0$ and *any* labeling bias $\nu < 1$.

Thus, Equal Opportunity Strongly Recovers with $(1/4, 1/3)$ and $(1/3, 1/4)$ in the Under-Representation and Labeling Bias model.

Table 1 Summary of recovery behavior of multiple fairness interventions in bias models.

Intervention	Under-Representation	Labeling Bias	Both
Equal Opportunity-ERM	Yes: $(1-r)(1-2\eta)+r((1-\eta)\beta-\eta)>0$	Yes: $(1-r)(1-2\eta)+r((1-\eta)(1-2\nu)-\eta)>0$	Yes: Using Condition 3
Equalized Odds	Yes: $(1-r)(1-2\eta)+r((1-\eta)\beta-\eta)>0$	No	No
Re-weighting Class B:	Yes	Yes	No

Table 1 summarizes the results in the three core interventions and the three core bias models. Demographic Parity is omitted from the table since it cannot recover under the bias models when $\eta = 0$ and thus is inadequate. The contents of each square indicate if recovery is possible in a bias model with an intervention and what constraints need to be satisfied for recovery.

4.1 Proof of Main Theorem

In this section we present the proof of the main result, **Theorem** 3. We want to show that the lowest biased error classifier satisfying Equal Opportunity on the biased data is h^*, given Condition 3.

The first step of the proof is to show that h^* satisfies Equal Opportunity on the biased training data. Note: the lemmas and claims here are all in the Under-Representation Bias combined with Labeling Bias Model, the most general bias model.

▶ **Lemma 4.** $h^* = (h^*_A, h^*_B)$ *satisfies Equal Opportunity on the biased data distribution.*

Proof. First, let's consider the easiest case with $\eta = 0$, $\beta_{POS} = \beta_{NEG} = 1$, and $\nu = 0$. Recall that h^* is the pair of classifiers used to generate the labels. When $\eta = 0$, h^* is a perfect classifier for both groups so Equal Opportunity is trivially satisfied. Now, let's consider arbitrary $0 \leq \eta < 1/2$. Recall that $p = Pr_{\mathfrak{D}_A}(h^*_A(x) = 1 | x \in A) = Pr_{\mathfrak{D}_B}(h^*_B(x) = 1 | x \in B)$.

By our assumption that Group A and Group B have equal values of p and η we have

$$\Pr(h^*_A(x) = 1 | Y = 1, x \in A) = \frac{p(1-\eta)}{p(1-\eta) + (1-p)\eta} = \Pr(h^*_B(x) = 1 | Y = 1, x \in B)$$

Next consider when we have both Under-Representation Bias and Labeling Bias. Recall that $\beta_{POS}, \beta_{NEG} > 0$ is the probability that a positive or negative sample from Group B is *not filtered* out of the training data while $\nu < 1$ is the probability a positive label is flipped and this flipping occurs after the filtering process. Then,

$$\{\text{True Positive Rate on Group A}\} := \Pr(h^*_A(x) = 1 | Y = 1, x \in A) =$$
$$\frac{p(1-\eta)}{p(1-\eta) + (1-p)\eta} = \frac{p(1-\eta)\beta_{POS}(1-\nu)}{p(1-\eta)\beta_{POS}(1-\nu) + (1-p)\eta\beta_{POS}(1-\nu)}$$
$$= \Pr(h^*_B(x) = 1 | Y = 1, x \in B) := \{\text{True Positive Rate on Group B}\}$$

so Equal Opportunity is still satisfied.

In words, the bias model removes or flips positive points from Group B independent of their location relative to the optimal hypothesis class. Thus positive points throughout the input space are are equally likely to be removed, so the overall probability of true positives being classified as positives is not changed. ◀

Now we describe how a candidate classifier h_B differs from h^*_B. We can describe the difference between the classifiers by noting the regions in the input space that each classifier gives a specific label. This gives rise to four regions of interest with probability mass as follows:

$$p_{1B} = P_{1B}(h_B) := P_{x \in \mathfrak{D}_B}(h^*_B(x) = 1 \wedge h_B(x) = 0)$$
$$p_{2B} = P_{2B}(h_B) := P_{x \in \mathfrak{D}_B}(h^*_B(x) = 0 \wedge h_B(x) = 1)$$
$$p - p_{1B} = P_{x \in \mathfrak{D}_B}(h^*_B(x) = 1 \wedge h_B(x) = 1)$$
$$1 - p - p_{2B} = P_{x \in \mathfrak{D}_B}(h^*_B(x) = 0 \wedge h_B(x) = 0)$$

These probabilities are made with reference to the regions in input space *before* the bias process. p_{1B} and p_{2B} are functions of h_B to make explicit that there may be multiple hypotheses with different functional forms that could allocate the same amount of probability mass to parts of the input space where h^*_B and h_B agree on labeling as positive and negative respectively. The partition of probability mass into these regions is easiest to visualize for hyperplanes but will hold with other hypothesis classes. p_{1A} and p_{2A} are defined similarly with respect to h^*_A and $\mathcal{D}_A$. A schematic with hyper-planes is given in Figure 3. To show that h^* has the lowest error on the true distribution, we first show how given any pair of classifiers h_A and h_B, which jointly satisfy Equal Opportunity (Equal Opportunity) on the biased distribution, we can transform $\{h_A, h_B\}$ into a pair of classifiers still satisfying Equal Opportunity with at most one non-zero parameter from $\{p_{1B}, p_{2B}\}$, and at most one non-zero parameter from $\{p_{1A}, p_{2A}\}$, while also not increasing biased error.

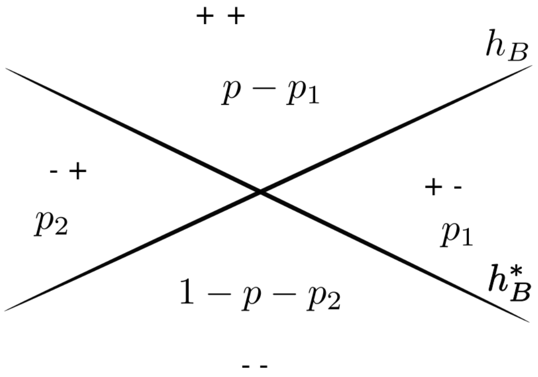


Figure 3 Differences between h_B and h^*_B measured with probabilities in the true data distribution (before the effects of the bias model).

The final step of our proof argues that out of the family of all hypotheses with (1) at most one non-zero parameter for the hypothesis on Group A, (2) at most one non-zero parameter for the hypothesis on Group B, (3) and jointly satisfying Equal Opportunity on the biased data, h^* has the lowest biased error.

These steps combined imply that h^* is the lowest biased error hypothesis that satisfies Equal Opportunity.

▶ **Lemma 5.** *Given classifiers h_A and h_B which satisfy Equal Opportunity on the biased data, there exist classifiers $h^{'}_A$ and $h^{'}_B$ (not necessarily in $\mathcal{H}$) satisfying*

1. *At most one of $\{P_{1A}(h^{'}_A), P_{2A}(h^{'}_A)\}$ is non-zero and at most one of $\{P_{1B}(h^{'}_B), P_{2B}(h^{'}_B)\}$ is non-zero.*
2. *$(h^{'}_A, h^{'}_B)$ has error at most that of (h_A, h_B) on the biased distribution.*
3. *$h^{'}_A$ and $h^{'}_B$ satisfy Equal Opportunity.*

Proof. We want to exhibit a pair of classifiers with lower biased error that zeros out one of the parameters. We do this by modifying each classifier separately, while keeping the true positive rate on the biased data fixed to ensure we satisfy Equal Opportunity.

First, consider Group A and suppose that $P_{1A}(h_A), P_{2A}(h_A) > 0$ since otherwise we do not need to modify h_A. Imagine holding the true positive rate of h_A constant and shrinking p_{2A} towards zero. As we shrink p_{2A}, we must shrink p_{1A} towards zero in order hold the true positive rate fixed (and thus satisfy Equal Opportunity).

The un-normalized[7] True Positive Rate (constrained by Equal Opportunity) is $(p - p_{1A})(1-\eta) + p_{2A}\eta = p(1-\eta) - p_{1A}(1-\eta) + p_{2A}\eta = (p - p_{1B})(1-\eta) + p_{2B}\eta$. Since the $p(1-\eta)$ term is independent of the classifier h_A, keeping the true positive rate constant is equivalent to keeping $C := -p_{1A}(1-\eta) + p_{2A}\eta$ constant.

Define $f(\Delta) = \Delta\frac{\eta}{1-\eta}$. If $C \leq 0$ then we can shrink p_{2A} to 0 and reduce p_{1A} by $f(p_{2A})$, keeping C constant. If $C \geq 0$ we can instead shrink p_{1A} to 0 and reduce p_{2A} by $f^{-1}(p_{1A})$.

Observe for Group A this process will clearly reduce training error since we are decreasing both p_{1A} and p_{2A} and the error on group A is monotone increasing (and linear) with respect to $p_{1A} + p_{2A}$.

We then separately do this same shrinking process for group B. Now we show the biased error decreases for Group B. For a given amount Δ by which we shrink p_{2B}, the overall biased error change for Group B is $\Delta[\eta\beta_{POS}(1-\nu) - (1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu] + f(\Delta)[\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu - (1-\eta)\beta_{POS}(1-\nu)]$, and simplifies to become

$$\begin{aligned} &= \Delta\eta\beta_{POS}(1-\nu) - f(\Delta)(1-\eta)\beta_{POS}(1-\nu) \\ &\qquad + \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + f(\Delta)(\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu) \end{aligned}$$

The first two terms vanish because of $f(\Delta) = \Delta\frac{\eta}{1-\eta}$.

$$\begin{aligned} &= \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + f(\Delta)(\eta\beta_{NEG} + (1-\eta)\beta_{POS}\nu) \\ &= \Delta(-(1-\eta)\beta_{NEG} - \eta\beta_{POS}\nu) + \Delta\frac{\eta^2}{1-\eta}\beta_{NEG} + \Delta\eta\beta_{POS}\nu \\ &= \Delta(\frac{\eta^2}{1-\eta}\beta_{NEG} - (1-\eta)\beta_{NEG}) < 0 \end{aligned}$$

Since this term is negative, we have shown that this modification process decreases error on the biased training data for both Group A and Group B while keeping the true positive rate fixed. $h_A^{'}$ and $h_B^{'}$ are then any functions satisfying these p's (e.g. p_{1A}, p_{2A} etc). ◀

▶ **Lemma 6.** *If h_A and h_B satisfy the Equal Opportunity constraint and each classifier has at most one non-zero parameter, then $p_{1B} = p_{1A}$ and $p_{2B} = p_{2A}$.*

Proof. Recall that the Equal Opportunity constraint requires that these expressions be equal.

$$\begin{aligned} &(p - p_{1A})(1-\eta) + p_{2A}\eta = (p - p_{1B})(1-\eta) + p_{2B} \\ &p_{2A}\eta - p_{1A}(1-\eta) = p_{2B}\eta - p_{1A}(1-\eta) \end{aligned}$$

Then the theorem follows from inspecting the second equality. ◀

This lemma makes explicit that when the classifiers each have only one non-zero parameter and satisfy Equal Opportunity, then the non-zero parameter corresponds to the same region.

▶ **Lemma 7.** *Of hypotheses satisfying ($p_{1A} = p_{1B}$ and $p_{2A} = p_{2B} = 0$) or ($p_{1A} = p_{1B} = 0$ and $p_{2A} = p_{2B}$), if these inequalities hold:*

$$\begin{aligned} (1-r)(1-2\eta) + &r((1-\eta)\beta_{POS}(1-2\nu) - \eta\beta_{NEG}) > 0 \\ &\textit{and} \\ (1-r)(1-2\eta) + &r((1-\eta)\beta_{NEG} - \eta\beta_{POS}(1-2\nu)) > 0 \end{aligned}$$

then the lowest biased error classifier satisfying Equal Opportunity on the biased data is $h^* = (h_A^*, h_B^*)$.

[7] The normalization factor for these rates for Group A and Group B is the same so this term can be cancelled.

Proof. First, we sketch the proof informally. Consider three cases which depend on how the bias process affects the unconstrained optimum for Group B on the biased data. In the first case, in the biased data distribution, the region $X^+ := \{x \ s.t. \ h^*_B(x) = 1\}$ has more positive than negative samples in expectation and the region $X^- := \{x \ s.t. \ h^*_B(x) = 0\}$ has more negative than positive samples in expectation. In the second case, there are more positive than negative samples throughout the entire input space in the biased distribution. In the third and final case, there are more negative than positive samples throughout the input space in the biased distribution.

In these three cases, the optimal hypothesis is exactly one of $\{h^*_B, h^1_B, h^0_B\}$, respectively. The second two hypotheses mean labelling all inputs as positive and labelling all inputs as negative, respectively. These three hypotheses correspond to hypotheses with at most one non-zero parameter.

For instance, h^1_B occurs when $p_{2B} = 1 - p$ and $p_{1B} = 0$. Each of the three hypotheses occur when the one non-zero parameter attains a location on the boundary of its range of values. When p_{2B} is allowed to be non-zero, if instead $p_{2B} = 0$ (and thus it also must be that $p_{1B} = 0$), the hypothesis is equivalent to h^*_B. A similar relationship holds for h^0 and p_1.

In order to show the theorem, we prove that if h^* has lower biased error than $h^1 = (h^1_A, h^1_B)$ and $h^0 = (h^0_A, h^0_B)$ on the biased data distribution, then h^* has the lowest error among all hypotheses with at most one non-zero parameter and satisfying Equal Opportunity.

To see this, consider h_A and h_B with the same non-zero parameter equal to Δ. Then the error of h_A is a linear function of Δ. Similarly, the error of h_B is a linear function of Δ. The overall error of $h = (h_A, h_B)$ is a weighted combination of the error of h^* and the error of h^0 or h^1, so the overall error of h is thus linear in Δ, so the optimal hypothesis parametrized by Δ must occur on the boundaries of the region of Δ, so the optimal hypothesis is one of $\{h^*, h^0, h^1\}$. We then show that the inequalities we assume in the theorem enforce that h^* has strictly lower error than h^0 or h^1. Formally, we enumerate the possible events:

Type	Sign of h^*	Label in Biased Data	Un-Normalized Probability of Event
A	+	+	$R_1 = (1-r)p(1-\eta)$
A	+	-	$R_2 = (1-r)p\eta$
A	-	+	$R_3 = (1-r)(1-p)\eta$
A	-	-	$R_4 = (1-r)(1-p)(1-\eta)$
B	+	+	$R_5 = rp(1-\eta)\beta_{POS}(1-\nu)$
B	+	-	$R_6 = rp[(1-\eta)\beta_{POS}\nu + \eta\beta_{NEG}]$
B	-	+	$R_7 = r(1-p)(\eta\beta_{POS})(1-\nu)$
B	-	-	$R_8 = r(1-p)[(1-\eta)\beta_{NEG} + \eta\beta_{POS}\nu]$

The probabilities on the far right hand side are not normalized. First we show that the $err(h^*) < err(h^1)$. $err(h^*) = R_2 + R_3 + R_6 + R_7$ and $err(h^1) = R_2 + R_4 + R_6 + R_8$, thus $err(h^*) < err(h^1)$ if and only if $R_3 + R_7 < R_4 + R_8$ or thus if

$$(1-r)(1-p)\eta + r(1-p)(\eta\beta_{POS})(1-\nu)$$
$$< (1-r)(1-p)(1-\eta) + r(1-p)[(1-\eta)\beta_{NEG} + \eta\beta_{POS}\nu]$$

Equivalently,

$$0 < (1-r)(1-2\eta) + r[(1-\eta)\beta_{NEG} - \eta\beta_{POS}(1-2\nu)] \tag{5}$$

Now we consider h^* compared to h^0. Then $err(h^0) = R_1 + R_3 + R_5 + R_7$ Then $err(h^*) < err(h^0)$ if and only if $R_2 + R_6 < R_1 + R_5$.

$$(1-r)p\eta + rp[(1-\eta)\beta_{POS}\nu + \eta\beta_{NEG}] < (1-r)p(1-\eta) + rp(1-\eta)\beta_{POS}(1-\nu)$$

Equivalently,

$$0 < (1-r)(1-2\eta) + r((1-\eta)\beta_{POS}(1-2\nu) - \eta\beta_{NEG}) \tag{6}$$

Thus we have shown that the error of h^* is less than the error of of h^1 and h^0 if and only if both Lines 5 and 6 are true, which we assume in our theorem.

Now we show that we error of $h = (h_A, h_B)$ is linear in Δ. There are two cases depending on what parameter of h is non-zero.

Let h be a hypothesis such that $P_{1A}(h_A) = p_{1B} = \Delta$ and $P_{2A}(h_A) = p_{2B} = 0$ and $\Delta \in [0, p]$.

$$\begin{aligned} &err(h) = R_1\frac{\Delta}{p} + R_2\frac{p-\Delta}{p} + R_3 + R_5\frac{\Delta}{p} + R_6\frac{p-\Delta}{p} + R_7 \\ &= \frac{\Delta}{p}err(h^0) + \frac{p-\Delta}{p}err(h^*) \end{aligned}$$

On the other case let $P_{1A}(h_A) = p_{1B} = 0$ and $P_{2A}(h_A) = p_{2B} = \Delta$ and $\Delta \in [0, 1-p]$.

$$\begin{aligned} &err(h) = R_2 + \frac{1-p-\Delta}{1-p}R_3 + \frac{\Delta}{1-p}R_4 + R_6 + \frac{1-p-\Delta}{1-p}R_7 + \frac{\Delta}{1-p}R_8 \\ &= \frac{\Delta}{1-p}err(h^1) + \frac{1-p-\Delta}{1-p}err(h^*) \end{aligned}$$

Thus the error of h is linear in Δ and boundary values for Δ correspond to the hypotheses in $\{h^*, h^0, h^1\}$. These two arguments show that:

1. Any single parameter h is a weighted sum of (h^* and h^0) or is a weighted sum of (h^* and h^1) and so is linear in Δ. The boundary values of Δ correspond to $\{h^*, h^0, h^1\}$.
2. Since the optimal value of a linear function occurs on the boundaries of its range, the optimal Equal Opportunity classifier with at most one non-zero parameter is one of $\{h^*, h^0, h^1\}$.
3. The inequalities in the theorem statement enforce that h^* has lower biased error than either h^0 or h^1, so h^* has the lowest biased error of any single parameter hypothesis satisfying Equal Opportunity. ◀

If the conditions in the Theorem *do not hold*, then h^* will not have lower error than h^0 and h^1.

4.2 Verification Re-Weighting Recovers from Labeling Bias

The way we intervene by Reweighting is we multiply the loss term for mis-classifying positive examples in Group B by a factor Z such that the weighted fraction of positive examples in biased data for Group B is the same as the overall fraction of positive examples in Group A.

The goal of this reweighting is to ensure that the ratio of positive to negative samples in the positive region of h^*_B is greater than 1 while the ratio is less than 1 in the negative region of h^*_B. Thus the re-weighted probabilities need to simultaneously satisfy:

$$\begin{aligned} &\frac{P(y=1|h^*_B(x)=1)}{P(y=0|h^*_B(x)=1)} = \frac{Z[(1-\eta)(1-\nu)]}{(\eta+(1-\eta)\nu)} > 1 \\ &\frac{P(y=1|h^*_B(x)=0)}{P(y=|h^*_B(x)=0)} = \frac{Z[\eta(1-\nu)]}{((1-\eta)+\eta\nu)} < 1 \end{aligned}$$

The two constraints are equivalent to requiring that:

$$\frac{\eta + (1-\eta)\nu}{(1-\eta)(1-\nu)} < Z < \frac{1-\eta+\eta\nu}{\eta(1-\nu)} \tag{7}$$

Recall from Section 3.2 that $Z = \frac{1-P_{A,1}(1-\nu)}{(1-\nu)(1-P_{A,1})}$

First we show the right hand inequality.

$$\begin{aligned}
&\frac{1-p_{A,1}(1-\nu)}{(1-\nu)(1-p_{A,1})} < \frac{1-\eta+\eta\nu}{\eta(1-\nu)}\\
&0 < \frac{1-\eta+\eta\nu}{\eta} - \frac{1-p_{A,1}(1-\nu)}{(1-p_{A,1})}
\end{aligned}$$

Observe that both terms are linear in ν. When $\nu = 0$, the inequality becomes $\frac{1-\eta}{\eta} - \frac{1-p_{A,1}}{1-p_{A,1}} = \frac{1-\eta}{\eta} - 1 > 0$. In our bias model $\nu \in [0,1)$, but if $\nu = 1$, the inequality becomes $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$. Thus Equation 7 holds if both $\frac{1-\eta}{\eta} - 1 > 0$ and $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$.

$\frac{1-\eta}{\eta} - 1 > 0$ is clearly true because $0 < \eta < 1/2$.

To see that $\frac{1}{\eta} - \frac{1}{1-p_{A,1}} > 0$, note that this is equivalent to $\eta < 1 - p_{A,1}$, where the right-hand-side is the overall fraction of negative examples in A. This is clearly true because the positive region of h^*_A has exactly an η fraction of negatives, and the negative region of h^*_A has a $1-\eta > \eta$ fraction of negatives.

Now we show the left hand inequality in Equation 7.

$$\begin{aligned}
&\frac{\eta + (1-\eta)\nu}{(1-\eta)(1-\nu)} < \frac{1-P_{A,1}(1-\nu)}{(1-\nu)(1-P_{A,1})}\\
&\frac{\eta + (1-\eta)\nu}{(1-\eta)} < \frac{1-P_{A,1}(1-\nu)}{1-P_{A,1}}\\
&0 < \frac{1-P_{A,1}(1-\nu)}{(1-P_{A,1})} - \frac{\eta+(1-\eta)\nu}{(1-\eta)} \qquad (8)
\end{aligned}$$

We follow a similar linearity argument to above. For $\nu = 1$, Equation 8 becomes $\frac{1}{1-p_{A,1}} - \frac{1}{1-\eta} > 0$. This holds if $1 - p_{A,1} < 1 - \eta \iff \eta < p_{A,1}$. This is clearly true because the negative region of h^*_A has exactly an η fraction of positives, and the positive region of h^*_A has a $1-\eta > \eta$ fraction of positives. For $\nu = 0$, Equation 8 becomes $1 - \frac{\eta}{1-\eta} > 0$ which holds since $0 < \eta < 1/2$.

5 Calibration Results

▶ **Theorem 8.** *Assume the training data is corrupted by Under-Representation Bias with parameter $\beta < 1$. For any such β, h^* does not satisfy Calibration on the biased data and thus Calibration constrained ERM will return a hypothesis that has strictly worse true error than the true error of h^*. This occurs even when $(1-\eta)\beta > \eta$, i.e. in the bias regime such that plain ERM on the biased data would recover h^*.*

Moreover, if bias is such that $(1-\eta)\beta < \eta$ and thus ERM on the biased data will not recover h^*, then the unique ERM solution that satisfies Calibration on the biased data is a trivial classifier, meaning that all individuals from Group A receive one label (the positive label) and all individuals from Group B receive the opposite label.

Proof. Recall that Calibration of hypothesis $h = (h_A, h_B)$ requires that both Eq. 9 and 10 hold simultaneously.

$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 1) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 1) \tag{9}$$
$$P_{x\sim\mathcal{D}_A}(y = 1|h_A(x) = 0) = P_{x\sim\mathcal{D}_B}(y = 1|h_B(x) = 0) \tag{10}$$

We assume that if one of the terms is vacuous in the Calibration constraints , then that constraint is still satisfied. In other words, if one bin is non-empty for one group while the corresponding bin for the other group is empty, we assume that bin satisfies Calibration. Due to the effects of the bias model positive samples from Group B appear in the training data with lowered frequency and so the equalities in Equations 9 and 10 become:

$$P_{x\sim A}(y = 1|h^*_A(x) = 1) > P_{x\sim B}(y = 1|h^*_B(x) = 1) \tag{11}$$
$$P_{x\sim A}(y = 1|h^*_A(x) = 0) > P_{x\sim B}(y = 1|h^*_B(x) = 0) \tag{12}$$

Thus $h^* = (h^*_A, h^*_B)$ violates calibration for any $\beta < 1$ and any other hypothesis satisfying calibration will have strictly greater error on the true data distribution. Intuitively, for h to be Calibrated it will need to reduce the left-hand side of Equation 11 because it cannot increase the right-hand side and will have to increase the right-hand side of Equation 12 because it cannot decrease the left-hand side. As a result, its true error will be strictly larger than that of h^*.

Now, consider $(1 - \eta)\beta < \eta$. In this case, plain ERM will not recover h^*. With this amount of bias, then:

$$\begin{aligned} P_{x\sim A}(y = 1|h^*_A(x) = 1) &> P_{x\sim A}(y = 1|h^*_A(x) = 0) \\ &> P_{x\sim B}(y = 1|h^*_B(x) = 1) > P_{x\sim B}(y = 1|h^*_B(x) = 10) \end{aligned}$$

Satisfying Calibration with non-trivial classifiers requires achieving an equality with one side being a non-negative combination of the first two probabilities, and the other side being a non-negative combination of the second two probabilities. Since these inequalities are all strict, this is clearly not possible, so the only way to satisfy calibration is to use a trivial classifier that assigns all of Group A to one label, and all of Group B to the other label.[8] ◀

6 Conclusion

In this paper we have shown that Equal Opportunity constrained ERM will recover from several forms of training data bias, including Under-Representation Bias (where positive and/or negative examples of the disadvantaged group show up in the training data at a lower rate than their true prevalence in the population) and Labeling Bias (where each positive example from the disadvantaged group is mislabeled as negative with probability $\nu \in (0, 1)$), in a clean model where the Bayes optimal classifiers h^*_A, h^*_B satisfy most fairness constraints on the *true* distribution and the errors of h^*_A, h^*_B are uniformly distributed. The high-level message of this paper is that fairness interventions need not be in competition with accuracy and may improve classification accuracy if training data is unrepresentative or biased; however these results will be connected to the true data distributions and features of

[8] Which trivial classifier is selected by ERM will depend on p and r. If $1 - r > r$ and $p > 1/2$, then Group A will be all positive and Group B all negative. While if $1 - r > r$ and $p < 1/2$, then then Group A will be all positive and Group B all negative.

the biased data-generation process. It would be interesting to consider other ways in which training data could be biased, and other assumptions on the optimal classifiers, to determine what kinds of interventions might be most appropriate for different biased-data scenarios.

References

1 Dana Angluin and Philip Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, April 1988. doi:10.1007/BF00116829.

2 Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

3 Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004.

4 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

5 Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

6 Alexandra Chouldechova. Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.

7 Danielle Keats Citron and Frank Pasquale. The Scored Society: Due Process for Automated Predictions. *Wash. L. Rev.*, 89:1, 2014.

8 Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

9 Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint*, 2018. arXiv:1807.00905.

10 William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

11 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness Through Awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012. doi:10.1145/2090236.2090255.

12 Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80:38, 2016.

13 Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016. arXiv:1609.07236.

14 Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

15 Heinrich Jiang and Ofir Nachum. Identifying and Correcting Label Bias in Machine Learning. *CoRR*, abs/1901.04966, 2019. arXiv:1901.04966.

16 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017. doi:10.4230/LIPIcs.ITCS.2017.43.

17 Jon M. Kleinberg and Manish Raghavan. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 33:1–33:17, 2018. doi:10.4230/LIPIcs.ITCS.2018.33.

18 Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284. ACM, 2017.

19 Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

20 Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

21 Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online, Forthcoming*, 2019.

22 Samuel Yeom and Michael Carl Tschantz. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. *arXiv preprint*, 2018. arXiv:1808.08619.

Can Two Walk Together: Privacy Enhancing Methods and Preventing Tracking of Users

Moni Naor[1]
Department of Computer Science and Applied Mathematics,
Weizmann Institute of Science, Rehovot, 76100, Israel
http://www.wisdom.weizmann.ac.il/~naor
moni.naor@weizmann.ac.il

Neil Vexler
Department of Computer Science and Applied Mathematics,
Weizmann Institute of Science, Rehovot, 76100, Israel
neil.vexler@weizmann.ac.il

Abstract

We present a new concern when collecting data from individuals that arises from the attempt to mitigate privacy leakage in multiple reporting: tracking of users participating in the data collection via the mechanisms added to provide privacy. We present several definitions for untrackable mechanisms, inspired by the differential privacy framework.

Specifically, we define the trackable parameter as the log of the maximum ratio between the probability that a set of reports originated from a single user and the probability that the same set of reports originated from two users (with the same private value). We explore the implications of this new definition. We show how differentially private and untrackable mechanisms can be combined to achieve a bound for the problem of detecting when a certain user changed their private value.

Examining Google's deployed solution for everlasting privacy, we show that RAPPOR (Erlingsson et al. ACM CCS, 2014) is trackable in our framework for the parameters presented in their paper.

We analyze a variant of randomized response for collecting statistics of single bits, Bitwise Everlasting Privacy, that achieves good accuracy and everlasting privacy, while only being reasonably untrackable, specifically grows linearly in the number of reports. For collecting statistics about data from larger domains (for histograms and heavy hitters) we present a mechanism that prevents tracking for a limited number of responses.

We also present the concept of Mechanism Chaining, using the output of one mechanism as the input of another, in the scope of Differential Privacy, and show that the chaining of an ε_1-LDP mechanism with an ε_2-LDP mechanism is $\ln \frac{e^{\varepsilon_1+\varepsilon_2}+1}{e^{\varepsilon_1}+e^{\varepsilon_2}}$-LDP and that this bound is tight.

2012 ACM Subject Classification Security and privacy → Privacy-preserving protocols

Keywords and phrases Differential Privacy, Surveillance

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.4

Related Version A full version of the paper is available at [17], https://arxiv.org/abs/2004.03002.

Funding *Moni Naor*: Israel Science Foundation (no. 950/16)

Acknowledgements We thank Alexandra Korolova for many discussions and suggestions regarding the RAPPOR project as well as Guy Rothblum and Kobbi Nissim for much appreciated comments for their invaluable insights into various parts of this work. Part of this work was done while the second author was visiting the Simons Institute as part of the Data Privacy: Foundations and Applications program.

[1] Incumbent of the Judith Kleeman Professorial Chair.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 4; pp. 4:1–4:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

The cure should not be worse than the disease. In this paper we raise the issue that mechanisms for Differentially Private data collection enable the tracking of users. This wouldn't be the first time an innocent solution for an important problem is exploited for the purposes of tracking. Web cookies, designed to let users maintain a session between different web pages, is now the basis of many user tracking implementations. In the Differential Privacy world, we examine how various solutions meant to protect the privacy of users over long periods of time actually enable the tracking of participants.

To better understand this, consider the following scenario: A browser developer might wish to learn what which are the most common homepages, for caching purposes, or perhaps to identify suspiciously popular homepages that might be an evidence for the spreading of a new virus. They develop a mechanism for collecting the URLs of users' homepages. Being very privacy aware, they also make sure that the data sent back to them is Differentially Private. They want to ensure they can collect this data twice a day without allowing someone with access to the reports to figure out the homepage of any individual user.

If fresh randomness is used to generate each differentially private report, then the danger is that information about the users homepage would be revealed eventually to someone who follows the user's reports. We strive to what we call "Everlasting Privacy", the property of maintaining privacy no matter how many collections were made. In our example, the users achieve everlasting privacy by correlating the answers given at each collection time: e.g. a simple way is that each user fixes the randomness they use, and so sends the same report at each collection.

Now consider Alice, a user who reports from her work place during the day and from her home during the evening. At every collection, Alice always reports regarding the same homepage[2], and therefore (since the randomness was fixed) sends identical reports at home and at work. An eavesdropper examining a report from the work IP address and a report from Alice's home IP address would notice that they are the same, while if they examined a report generated by Alice and one generated by Bob (with the same homepage) they will very likely be different. This allows the adversary to find out where Alice lives.

To elaborate, correlation based solutions open the door to the new kind of issue, tracking users. The correlation between reports can be used as an instrument of identifying individuals, in particular it makes the decision problem of whether or not two sets of reports originated from the same user much easier. This concern has been suggested by the RAPPOR project [13] but without a formal definition, or analysis in the framework where their solution was provided.

The problem of tracking users is related to the problem of point change detection, i.e. identifying when a stream of samples switched from one distribution to another. While this problem has been researched in the past under the lens of privacy by Cummings et al. [5, 4], these works focused on private release of point change detection, i.e. how to enable researchers to detect changes in the sampled distribution while not being too reliant on any specific sample. Our goal is different. We wish to *prevent* change point detection as much as we can; as in our case, a change in distribution correlates to a change in private value. Detecting a change in private value jeopardizes the privacy of the user (think of a case where the gender is changed).

[2] The reader may be wondering why bother reporting about the same value if it does not change. For instance it may for purposes of aggregating information about the currently online population.

1.1 Our Contributions

The main conceptual contribution of this work is the definition of reports being untrackable, presented in Section 3. Roughly, the definition states that the distribution on outputs generated by a single user needs to be sufficiently close to that generated by two users. For the discussion on motivation and possible variants see Section 3.4

▶ **Definition 1** (informal). *A mechanism M is (γ, δ)-Untrackable for k reports, if for any k reports*

$$Pr\left[\textit{Reports were generated by one user}\right] \leq e^{\gamma} Pr\left[\textit{Reports were generated by two users}\right] + \delta.$$

We present a formal definition to Everlasting Privacy. Roughly speaking, a mechanism is (γ, δ)-Everlasting Privacy if executing it any number of times is (γ, δ)-DP. Our main goal is to simultaneously achieve both tracking prevention and everlasting privacy, while maintaining a reasonable accuracy for the global statistics. We explore the implications of this new definition, specifically how it composes and what a fixed state that is reported in a noisy manner can achieve.

We describe how our tracking definitions can be extended to the change point detection framework, namely to bound the probability that a change in the user's private value is ever detected. In that section we also discuss the necessity of correlating answers between data collections to ensure Differential Privacy, and define various general constructions for mechanisms that can achieve this Everlasting Differential Privacy.

As a tool for analyzing such constructions, in Section 4 we prove a theorem about running a Local Differential Privacy mechanism on the output of another such mechanism.

▶ **Theorem 2** (informal). *A mechanism that consists of running an ε_2-LDP mechanism on the result of an ε_1-LDP mechanism results in $\frac{1}{2}\varepsilon_1 \cdot \varepsilon_2$-LDP for small ε_1 and ε_2.*

Theorem 20 and Corollary 21 provide the formal statement and proof.

We then continue to analyze Google RAPPOR's [13] performance under the framework of tracking. We show the pure tracking bound RAPPOR achieves as well as estimate its "average" case performance. We cocnslude that according to our definition of untrackable, RAPPOR achieves poor protection guarantees. This is presented in Section 5.

As a warm up, in Section 6 we present a mechanism that deals with data collection of a single private bit from each participant. One can view it as the extension of randomized response in this setting. Each user generates a bit at random and remembers it. At each collection, the user generates an new bit and sends the XOR of the private bit, the remembered bit and the new bit. The remembered bit is generated by flipping one biased coin, parameterized by ε_1. The new bits are generated from fresh coin flips from another biased coin, parameterized by ε_2. The aggregator collects all the reports and outputs estimated frequencies for both 0 and 1. We prove that for a choice of privacy parameters $\varepsilon_1, \varepsilon_2 < 1$, and for n participating users, the mechanism has the properties:

(i) It is ε_1-Everlasting Differentially Private.

(ii) Accuracy: the frequency estimation of 0 and 1 is no further than $\tilde{O}\left(\frac{1}{\varepsilon_1 \cdot \varepsilon_2 \cdot \sqrt{n}}\right)$ from the actual values.

(iii) It is $\lfloor \frac{k}{2} \rfloor \varepsilon_2$-untrackable for k reports.

In Section 7 we present a mechanism that allows the collection of statistics of users private values when their data is d bits. This mechanism is particularly relevant for the problems of heavy hitters estimations and histograms. The mechanism's state consists of the results of the inner product of the private value with multiple vectors in a way that is Differentially Private, reporting one such vector and the private result of the inner product at each data collection. The aggregator collects all the reports and produces an estimate for the frequencies of all possible values, such that the sum of frequencies is 1. We prove that for a choice of privacy parameters $\varepsilon < 1$, setting the state to consist of L reports, and for n participating users, the mechanism has the properties:

(i) It is (ε, δ)-Approximate Everlasting Differentially Private.

(ii) The estimation of the frequency of all values is no further than $\tilde{O}\left(\frac{1}{\varepsilon'}\sqrt{\frac{d}{n}}\right)$ from the actual frequency, for $\varepsilon' = \frac{\varepsilon}{2\sqrt{2L\ln\left(\frac{1}{\delta}\right)}}$.

(iii) It is $\left(0, \frac{k^2}{L}\right)$-untrackable.

Concretely, to obtain (ε, δ)-Everlasting Privacy and α accuracy, then for k reports the guarantee on the mechanism is $\left(0, \widetilde{O}\left(\frac{k^2}{\alpha^2\varepsilon^2 n}\right)\right)$-Untrackable.

Coming up with better bounds or showing the inherent limitations is the main open direction we propose (see Section 8).

2 Preliminaries

2.1 Differential Privacy

For background on Differential Privacy see Dwork and Roth [10] or Vadhan [18].

Throughout most of this paper we consider a variant of Differential Privacy, called *Local Differential Privacy*. Local Differential Privacy regards mechanisms where each individual user runs on their own data to create a report, which is then sent to the server and aggregated there to produce a population level result. The setting we consider is one where the aggregator access the users' data only through a randomized mapping, a mechanism, that has the following property:

▶ **Definition 3** ([15]). *Let $\varepsilon, \delta > 0$. A mechanism $M : U \mapsto O$ is (ε, δ)-Local Differentially Private if for every two possible inputs, $u, u' \in U$, and $\forall S \subseteq O$, $Pr\left[M\left(u\right) \in S\right] \leq e^{\varepsilon} \cdot Pr\left[M\left(u'\right) \in S\right] + \delta$.*

One of the significant properties of Differential Privacy is the way it composes. Composing two mechanisms that are $(\varepsilon_1, \delta_1)$ and $(\varepsilon_2, \delta_2)$-Differentially Private respectively is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-Differentially Private. A small deterioration in the δ parameter achieves a great improvement in the ε parameter of the composition.

▶ **Theorem 4** (Advanced composition for Differential Privacy [11]). *Let $\delta' > 0$. The k fold composition of (ε, δ)-Differentially Private mechanisms is $(\varepsilon', k\delta + \delta')$-Differentially Private for $\varepsilon' = \sqrt{2k\ln\left(1/\delta'\right)}\varepsilon + k\varepsilon\left(e^{\varepsilon} - 1\right)$.*

Another useful property of Differential Privacy is that running any function on the output of an (ε, δ)-Differentially Private mechanism is (ε, δ)-Differentially Private. That is, Differential Privacy is *closed under post-processing.*

When using the same mechanism to collect reports multiple times, if not done carefully, the privacy guarantee might deteriorate as the number of collections periods grows. We define *Everlasting Differential Privacy* as an upper bound on the privacy parameter of a mechanism, no matter how many times it is executed, as long as the private data had not changed. Definition 10 formalizes this idea.

2.2 Background

The need for everlasting privacy became apparent since the early stages of the Differential Privacy research. As mentioned in Section 2, independent repetitive executions of Differential Privacy mechanisms inevitably deteriorate the privacy guarantee. While Theorem 4 teaches us that the privacy guarantee can grow as low as only the square root of the number of reports, practical implementations might require users to participate in as many as thousands of data collections (e.g. anything requiring daily reports).

This led researchers to suggest data collection mechanisms that allow numerous data collections, while maintaining individuals' privacy. Certain solutions, such as Google's RAPPOR [13] and Microsoft's dBitFLip [6], use the concept of statefulness, maintaining some data between executions. This enables them to correlate outputs between executions, which allows for a manageable upper bound of the privacy leakage that does not rely on the number of collections made. This effectively allows for a privacy guarantee that holds forever, namely Everlasting Privacy.

Heavy Hitter Mechanisms

Two problems that have been very interesting for data collectors are the histogram and heavy hitters problems. In the histogram problem the goal is to accurately estimate the frequencies of all possible values the population might hold. The heavy hitters problem is about identifying the most common values amongst the population. Both histograms and heavy hitters in the local model has been researched before by Bassily et al. [2, 1], who used Hadamard transformations on the users private data that allow users to send succinct reports to the curator while allowing the required statistics to be generated very efficiently. These works do not fit our framework, as they intrinsically allow for trackability. In their solution, each user is associated with a specific piece of some shared randomness. The aggregator must know to which piece of randomness a specific report belongs to, essentially forcing their solution to be highly trackable. The techniques used in their paper are similar to the ones used by Naor et al. [16]. In that work the authors use an inner product mechanism to identify and ban the most common passwords. This enables the increase in the effective time an adversary will need to invest in order to guess a user's password. Their mechanism maintains Differential Privacy to prevent the leakage of each individual's password, but it does not maintain Everlasting Privacy. They also mention a modification to their scheme achieves Everlasting Privacy, by reusing the same random vector for all future inner products, but such a solution is highly trackable. The inner product mechanisms used in [2, 1, 16] were the inspiration of our Noisy Inner Product mechanism presented in 7.

Continual Observation and Pan Privacy

Other models and solutions to long-lasting privacy have been developed as well, such as the Continual Observation model in [7, 3]. The goal is to maintains differential privacy for values that change over time, e.g. a counter that updates over time, or streams of data, like traffic conditions and so on. This solution is in the central, or streaming, model and not in the local model. Another model is that of Pan Privacy, where the goal is to maintain privacy even if the internal representation of the secret state is leaked from time to time (Dwork et al. [8]). In Erlignsson et al. [12] this idea was extended, transforming the mechanism in [7] to the local model, in order to solve the 1-bit histogram problem, and thus achieving privacy over extended periods of time. The transformation means that every user reports genuinely only once throughout all data collections, thus resulting in accuracy that relies linearly in the number of times their value changed. This suggests that accuracy will drop as collection times increase.

Joseph et al. [14] suggested an approach where at the beginning, a global update occurs, where each individual participates in a private histogram estimation. At each subsequent potential collection time, each user compares their current contribution to the histogram compared to the last time a global update occurs. Depending on how different it is, they are more likely to suggest that another global update occurs. If enough users vote in favor, the curator initiates another round of global update, creating a more accurate histogram. This solution allows for collections to be made from users only when it is likely that the previously computed output is no longer accurate, greatly increasing the privacy guarantee of individuals. On the other hand, their accuracy analysis relies on the existence of a small number of user types, where all users of the same type behave identically.

3 Stateful Mechanisms and Tracking

Consider a mechanism for users to report their values to a center.Such mechanisms may be *stateless*, i.e. ones that receive an input and (probabilistically) produce an output, or *stateful* mechanisms, ones that receive in addition to the input a *state* and produce in addition to an output a state for the next execution. The power of stateful mechanisms is that they enable the correlation of outputs between different executions through the states passed from one execution to the next.

3.1 Definitions of Mechanisms and Report Stream Generators

Stateless mechanisms are randomized mappings for which each execution is independent of the others. Stateless mechanisms receive the user's data and publicly available information, namely auxiliary information, and output a report. The publicly available information can be anything known to all parties, like time of day, value of some publicly accessible counter, etc.

▶ **Definition 5** (Stateless Mechanism). *A stateless mechanism M is a randomized mapping from a user's data and auxiliary information to the domain of reports, $M : U \times A \times \{0,1\}^\star \mapsto R$. In our setting it is used to generate a stream of reports, $r_1, r_2, \ldots$, where each report is generated independently.*

Stateless mechanisms might provide very poor everlasting privacy, as each iteration reveals more information about the user's data.

Therefore, to achieve everlasting privacy one must *correlate* the reports sent by the user(s) (see for instance [9] where this is proved for counting queries). For this we define *Stateful Mechanisms*, where the mechanism maintains a state that is updated with each call to the mechanism.

▶ **Definition 6** (Fully Stateful Mechanism). *A fully stateful mechanism M is a randomized mapping from a user's data, current state and auxiliary information to the domain of reports and to a new state, $M : U \times S \times A \times \{0,1\}^\star \mapsto R \times S$. In our setting it is used to generate a stream of reports $r_1, r_2, \ldots$ and a stream of states $s_0 = \perp, s_1, \ldots$, such that each pair of state and report are generated by the previous state, auxiliary information and the user's data, $r_i, s_i = M(u, s_{i-1}, a_{i-1})$.*

Notice that the execution number and all previous outputs can be encoded into the state. A fully stateful mechanism can achieve everlasting privacy by correlating answers using the data stored in the state. For example, it can execute a DP mechanism on the user's data and remember the result, reporting the same result whenever queried.

One shortcoming of correlating the reports in such a manner is that it might be used as an identifier by an adversary, potentially allowing the adversary to identify that a group of reports all originated from the same user, thus allowing tracking other activities of the user (see Section 3.2).

We define *Permanent State Mechanisms* as mechanisms that maintain the same state once set, i.e. $s_1 = s_2 = s_3 \ldots$. As we shall see, such mechanisms are very convenient to work with and have good properties wrt composition.

Report stream generators (RSG) are mappings that use mechanisms to generate a stream of reports. The responsibility of the RSG is to get the user's data and iteratively call the mechanism.

▶ **Definition 7** (Stateless Report Stream Generator). *For a domain of user data U, a range of reports R, and a report stream size n, a Stateless Report Stream Generator using a stateless mechanism M is a mapping $G_n^M : U \mapsto R^n$, that acquires the auxiliary information required at each step and calls M to generate the reports $r_1, \ldots, r_n$.*

Similarly, stateful RGSs use fully stateful mechanisms to generate the stream of reports.

▶ **Definition 8** (Stateful Report Stream Generator). *For a domain of user data U, a range of reports R, and a report stream size n, a Stateful Report Stream Generator using a fully stateful mechanism M is a mapping $G_n^M : U \mapsto R^n$, that acquires the auxiliary information required at each step and calls M with the state of the current step to generate report r_i and the next step's state s_i.*

3.2 Everlasting Privacy and Tracking

The problem we focus on is the ability of an adversary to distinguish whether or not a set of reports originated from a single user or by two users (or more, see Section 3.4). For example, If an adversary had two sets of reports belonging to two different IP addresses, the adversary could learn if those IP addresses belong to the same user or not (potentially identifying the user's work place or home address). The definition of untrackable we propose is inspired by definition of Differential Privacy.

▶ **Definition 9.** *For a domain of user data U, a range of reports R and a report stream size k, a report stream generator G_k^M is (γ, δ)-**untrackable** if for all user data $u \in U$, for all subsets of indices $J \subseteq [k], J^{\complement} = [k] \setminus J$ and $\forall T \subseteq R^k$ we have:*

$$Pr\left[G_k^M(u) \in T\right] \leq e^{\gamma} \cdot Pr\left[G_{|J|}^M(u) \in T_J\right] \cdot Pr\left[G_{k-|J|}^M(u) \in T_{J^{\complement}}\right] + \delta$$

and

$$Pr\left[G_{|J|}^M(u) \in T_J\right] \cdot Pr\left[G_{k-|J|}^M(u) \in T_{J^{\complement}}\right] \leq e^{\gamma} \cdot Pr\left[G_k^M(u) \in T\right] + \delta.$$

For report stream generators that are (γ, δ)-untrackable, an adversary has only a small advantage in distinguishing between the following two cases: the reports originated from a single user or two users. A discussion for the idea behind this definition and its benefits can be found in Section 3.4. If we want this property to hold for any possible output (i.e. always have the ambiguity), then we can demand that the mechanism be $(\gamma, 0)$-untrackable. We call such mechanisms γ-untrackable. We leverage the similarity to DP show composition theorems on untrackable mechanisms.

Everlasting Privacy is meant to limit the leakage of information users suffer, no matter how many executions a mechanism had. For the following definitions let T be a collection of report streams. For a set of indices J let T_J be the collection of partial report stream, where the reports taken are those in indices J.

▶ **Definition 10** (Everlasting Privacy). *For a domain of user data U, a range of reports R, a report stream generator G_k^M is (ε, δ)-**Everlasting Privacy** if for all user data $u, u' \in U$, for all report stream size k and for all sets of output streams $T \subseteq R^k$, $Pr\left[G_k^M(u) \in T\right] \leq e^{\varepsilon} Pr\left[G_k^M(u') \in T\right] + \delta$.*

If a mechanism is $(\varepsilon, 0)$-Everlasting Privacy we say it is ε-Everlasting Privacy.

These definitions are tightly related to the problem of change-point detection. We define undetectability similarly to untrackability, only we do not assume both report sets originated from the same private data:

▶ **Definition 11.** *For a domain of user data U, a range of reports R and a report stream size k, a report stream generator G_k^M is (γ, δ)-**undetectable** if for all pairs of user data $u, u' \in U$, for all subsets of indices $J \subseteq [k], J^{\complement} = [k] \setminus J$ and $\forall T \subseteq R^k$ we have:*

$$Pr\left[G_k^M(u) \in T\right] \leq e^{\gamma} \cdot Pr\left[G_{|J|}^M(u) \in T_J\right] \cdot Pr\left[G_{k-|J|}^M(u') \in T_{J^{\complement}}\right] + \delta$$

and

$$Pr\left[G_{|J|}^M(u) \in T_J\right] \cdot Pr\left[G_{k-|J|}^M(u') \in T_{J^{\complement}}\right] \leq e^{\gamma} \cdot Pr\left[G_k^M(u) \in T\right] + \delta.$$

We can now connect being untrackable and everlasting privacy with being undetectable.

▶ **Theorem 12.** *A mechanism that is (γ, δ)-untrackable and (ε, δ')-everlasting differentially private is also $(\gamma + \varepsilon, \delta_{\max})$-undetectable, for $\delta_{\max} = \max\left\{e^{\varepsilon}\delta + \delta', \delta + e^{\gamma}\delta'\right\}$.*

The proof for this theorem can be found in the full version of the paper.

3.3 Tracking Bounds, Composition Theorems and Generalizations

For the special case of Permanent State Mechanisms, we can show an upper bound on the untrackable parameter. If the mechanism is ε-Differentially Private in its state, i.e. the mechanism protects the privacy of the state, then the untrackable parameter grows linearly in ε:

▶ **Theorem 13.** *A Permanent state mechanism whose reports are generated by an ε-Differentially Private mechanism receiving the state as its input is $\lfloor \frac{k}{2} \rfloor \varepsilon$-untrackable for k reports.*

The proof for this theorem can be found in the full version of the paper.

An important question is how tracking composes, i.e. how does a user's participation in multiple Report Stream Generators affect his untrackable guarantees. The similarity between the definition of untrackability and differential privacy allows us to apply results regarding the latter to obtain results on the former. We show an advanced composition for untrackable mechanisms that is analogous to advanced composition for differential privacy [11] and Theorem 4.

▶ **Theorem 14** (Advanced composition for untrackability). *Let m be a positive integer. Let $\{M_i\}_{i\in[m]}$ be m mechanisms that are (γ,δ)-untrackable for k_i reports respectively. The composition of these mechanisms, $\widehat{M}$, is $(\gamma',m\delta+\delta')$-untrackable for*

$$\gamma' = \sqrt{2m\ln\left(1/\delta'\right)}\cdot\gamma + m\cdot\gamma\left(e^{\gamma}-1\right).$$

The proof for this theorem, as well as the formal definition of composition, can be found in A.1

Another important question is what can be said about the untrackable guarantees in the settings where the reports are split into more than two sets, i.e. when we want to answer the question whether some reports were generated by a single user or any number of users. For this we define untrackable for n users for k reports.

▶ **Definition 15** (Multiple User Untrackable). *For a domain of user data U, a range of reports R and a report stream size k, and n users, a report stream generator G_k^M is γ-multiple user untrackable if for all user data $u\in U$, all partitions $P=\{P_i\}_{i\in[n]}$ of $[k]$ into n parts, and all output stream sets $T\subseteq R^k$:*

$$e^{-\gamma}\leq\frac{\prod_{j\in[n]}Pr\left[G_{|P_j|}^M\left(u\right)\in T_{P_j}\right]}{Pr\left[G_k^M\left(u\right)\in T\right]}\leq e^{\gamma}.$$

We show two connections between Definitions 9 and 15: the first is a general bound, essentially saying that the untrackable parameter increases linearly in the number of users.

▶ **Theorem 16.** *A mechanism that is γ-untrackable for k reports, is $(n-1)\,\gamma$-multiple user untrackable for n users for k reports.*

We can significantly improve this bound for *permanent state mechanisms* by leveraging the fact that their untrackable parameter is linear in the number of reports used.

▶ **Theorem 17.** *A permanent state mechanism M, whose reports are generated by an ε-Differentially Private mechanism receiving the state as its input, is $\lceil\log n\rceil\left\lfloor\frac{k}{2}\right\rfloor\varepsilon$-multiple user untrackable for n users for k reports.*

The proofs of these theorems can be found in Section A.2 and A.3

3.4 Discussion

The way we defined untrackable is not the only one possible. The "typical" attack we wish to prevent is against an adversary that sees many sets of reports and tries to identify two that belong to the same user. However, making this the basis of a definition might result in weak guarantees, as it disregards any prior information that an adversary might have. The adversary might know that Alice only lives in one of two houses, and only tries to identify where she lives. Our definition is designed to protect against exactly this kind of attacker, who only tries to distinguish whether a stream of reports was generated by Alice, or partly by Alice and partly by Bob.

Another natural definition is to prevent distinguishing whether a stream of reports was generated by any combination of users vs. any other combination of users. Our definition, though appearing weaker than this one, actually implies it, with some deterioration to the parameter; Theorem 16 suggests that the parameter deteriorates linearly in the number of users, while Theorem 15 suggests that in some cases it can deteriorate logarithmically.

Our definition also implies that it would be hard to decide whether any two reports were both generated by Alice, or one by Alice and one by Bob. This property might seem tempting as a basis of an alternative untrackable definition, but it is too weak on its own. A mechanism that has this property might have very poor protection against adversaries with access to more than two reports.

Finally, Theorem 12 teaches us that our definition, when combined with everlasting privacy, naturally extends to the problem of change point (un)detection. That is, a mechanism that adheres both to the everlasting privacy requirement and our untrackable definition also protects the fact that a user changed their private value.

In conclusion, This definition is strong enough to protect users against reasonable adversaries, i.e. ones who have some prior knowledge about the locations of users. On the other hand, while seeming weaker than other definitions it actually implies them. Additionally, as can be seen in Sections 6 and 7, it is achievable while also allowing for reasonable everlasting privacy guarantees and accuracy.

4 Mechanism Chaining

In this section we generalize the idea presented in Theorem 13 of using a Differential Privacy mechanism on the output of another such mechanism. We first provide a formal definition for this mechanism chaining, and then state and prove two theorems about the Differential Privacy guarantee achieved by doing such chaining. The first weak, but intuitive, the second much more powerful and also optimal.

4.1 Definitions

We now present mechanism chaining in three different settings: In the first setting we simply define the chaining of two mechanisms as taking the output of the first and using it as the input of the second.

▶ **Definition 18** (2 Local Mechanism Chaining). *Given two mechanisms $A : U \to V$ and $B : V \to O$, the chaining of these two mechanisms $\mathcal{M}_{B \circ A} : U \to O$ is defined as $\mathcal{M}_{B \circ A}(u) = B(A(u))$.*

The second setting we examine is the chaining of k mechanism, and the third and final setting is the chaining of k families of mechanisms that are not necessarily local. They are not relevant for the rest of this paper, but for completeness we present them in the full version of the paper.

4.2 Differential Privacy Guarantees for Two Mechanism Chaining

We now present a tight bound on the Differential Privacy guarantee of the chaining of two mechanisms. We begin by presenting the "Basic Chaining Upper Bound", which is not tight, but is perhaps more intuitive. We then present a better upper bound called the "Advanced Chaining Upper Bound". Basic Chaining simply says that the resulting Differential Privacy is no worse than the Differential Privacy of either mechanisms.

▶ **Theorem 19** (Basic Chaining). *Given two mechanisms $A : U \to V$ and $B : V \to O$ that are ε_1-LDP and ε_2-LDP respectively, $\mathcal{M}_{B \circ A} : U \to O$ is $\min\{\varepsilon_1, \varepsilon_2\}$-LDP.*

The advanced chaining bound is always better:

▶ **Theorem 20** (Advanced Chaining)**.** *Given two mechanisms $A : U \to V$ and $B : V \to O$ that are ε_1-LDP and ε_2-LDP respectively, $\mathcal{M}_{B \circ A} : U \to O$ is $\ln \frac{e^{\varepsilon_1+\varepsilon_2}+1}{e^{\varepsilon_1}+e^{\varepsilon_2}}$-LDP.*

The proof of these theorem can be found in the full version of the paper. The privacy parameter can be upper bounded by a more simple bound that is meaningful for small ε_1 and ε_2:

▶ **Corollary 21.** *Given two mechanisms $A : U \to V$ and $B : V \to O$ that are ε_1-LDP and ε_2-LDP respectively, $\mathcal{M}_{B \circ A} : U \to O$ is $\frac{1}{2}\varepsilon_1 \cdot \varepsilon_2$-LDP.*

When ε_1 or ε_2 are greater than 2 this upper bound is worse than the bound in Theorem 19, let alone the optimal one in Theorem 20, but otherwise this bound has little error compared to the optimal bound and is easier to work with.

5 (Un)Trackability in RAPPOR

Equipped with a new framework to analyze tracking, we first consider one of the most significant deployments of a differentially private mechanism, used in all Chrome copies, and analyze its trackability. Introduced in [13], RAPPOR is a DP mechanism designed to allow repeated collection of telemetry data from users in Chrome. This mechanism was the starting point of this work, since some of the goals stated in the original paper indicate the desirability of being untrackable.

Roughly speaking, RAPPOR reports a value (e.g. the homepage of a user) from a large set. It does so with the help of a Bloom filter that initially encodes a set that contains a single element, the desired value. A Bloom filter's output is an all 0 array that is set to 1 at locations corresponding to hashes of the value. The mechanism proceeds to randomly flip bits in the Bloom filter, generating what we call the Permanent Randomization. At each point in time when data is to be collected, the mechanism generates a report by taking a copy of the Permanent Randomization and, again, randomly flipping bits and reporting the resulting array. The details of the mechanism can be found in the original paper, but for completeness we also present them in the full version of the paper.

In the paper introducing RAPPOR, the authors mention that preventing tracking of users is an issue with their construction: *"RAPPOR responses can even affect client anonymity, when they are collected on immutable client values that are the same across all clients: if the responses contain too many bits (e.g. the Bloom filters are too large), this can facilitate tracking clients, since the bits of the Permanent randomized responses are correlated"*. On the other hand, when talking about the reason behind the second phase of the mechanism execution, generating a report from the permanent randomization, they mention that *"Instead of directly reporting B' [The Permanent Randomization] on every request, the client reports a randomized version of B'. This modification significantly increases the difficulty of tracking a client based on B', which could otherwise be viewed as a unique identifier in longitudinal reporting scenarios"*. We wish to show that in our framework, using the same parameters they used in the RAPPOR data collections, RAPPOR is more aligned with the first statement than with the second. We analyzed RAPPOR's untrackable parameter in the worst case setting, which can be found in the full version of the paper. We present an analysis of the "average case" behavior of RAPPOR.

Estimated Percentile of the Trackability Random Varaible

We estimate the statistics of the trackabiltiy random variable for RAPPOR. In essence, the trackability random variable is the distribution of trackability leaks that happen when participating in the mechanism. The pure version of the untrackable bound in Definition 9 is an upper bound on the possible values of the trackability random variable.

Formally, denote the RAPPOR mechanism by R. For k reports we define a vector of partitions $\vec{J} = \{J_i\}_{i\in[k]}$, where $J_i = [i]$. We also define two report vectors $\vec{T} = \{T_i\}_{i\in[k]}$ and $\vec{T'} = \{T'_i\}_{i\in[k]}$, where T_i is drawn from the product distribution $\left(G_i^R(u), G_{n-i}^R(u)\right)$ and all of the T'_i are drawn from $G_n^R(u)$. The trackability random variable for k reports is the value: $\tau := \max\left\{\max_{i\in\left[\lfloor\frac{k}{2}\rfloor\right]} C_{T_i,J_i}, \max_{i\in\left[\lfloor\frac{k}{2}\rfloor\right]} C_{T'_i,J_i}\right\}$, where $C_{T,J}$ is:

$$C_{T,J} := \left|\ln\frac{Pr\left[G_n^R(u) = T\right]}{Pr\left[G_{|J|}^R(u) = T_J\right]\cdot Pr\left[G_{n-|J|}^R(u) = T_{J^\complement}\right]}\right|.$$

The random variable τ is the maximum measured tracking for the $\lfloor\frac{k}{2}\rfloor$ cases where the reports are generated by two users and the $\lfloor\frac{k}{2}\rfloor$ cases where the reports are generated by one user. In our setting a mechanism should protect against both types of cases.

The measures of interest are percentiles of the trackability random variable distribution. We estimate the median and the 90^{th} percentile of the trackability random variable. The full version of the paper presents the details of the estimation process.

Figure 1, in Appendix B, shows the estimated median and 90^{th} percentile of the Trackability random variable for between 2 and 15 reports, and their respective 95% confidence interval. Our estimation shows that RAPPOR's trackability random variable's median is better than the worst case trackability, but reaches high values, around 5 after as few as 10 reports. The 90^{th} percentile is worse, reaching trackability of 5 after as little as 7 reports.

6 Bitwise Everlasting Privacy Mechanism

We present a mechanism for collecting statistics about the distribution of a single bit in the population, in such a way that everlasting privacy is maintained. Our mechanism is a permanent state one, using a state that consists of a noisy copy of the private bit. At each report, the user sends a noisy version of the state, effectively sending a doubly noisy version of their private bit. We show the mechanism achieves good accuracy, and reasonable everlasting privacy. Since this mechanism is a permanent state mechanism, we can use Theorem 13 to give a less than reasonable upper bound on the untrackable parameter of this mechanism. We show, however a lower bound of the untrackable parameter of this mechanism that is not far off from the upper bound in Theorem 13.

Consider the mechanism where each user holds one bit, b. First they generate a permanent randomization, $b' = b \oplus x$, where $x \sim \text{Ber}\left(\frac{1}{e^{\varepsilon_1}+1}\right)$. Then at each report they generate a report bit, $r = b' \oplus y$, where $y \sim \text{Ber}\left(\frac{1}{e^{\varepsilon_2}+1}\right)$. The aggregator receives these reports from all users and invokes the frequency oracle to output an estimate:

$$\tilde{p}_0 = \frac{e^{\varepsilon_1+\varepsilon_2} + 1 - \left(e^{\varepsilon_1}+1\right)\left(e^{\varepsilon_2}+1\right)\sum_{i\in[n]} r_i}{\left(e^{\varepsilon_1}-1\right)\left(e^{\varepsilon_2}-1\right)}$$

and $\tilde{p}_1 = 1 - \tilde{p}_0$. Let $\tilde{p}$ be the vector whose coordinates are $\tilde{p}_0$ and $\tilde{p}_1$. Let p be the vector of true frequencies.

Privacy, Accuracy and Trackability

Bitwise Everlasting Privacy is ε_1-EDP, outputs $\tilde{p}$ such that with probability $1-\beta$:

$$\left\|\tilde{p}-p\right\|_\infty \leq \frac{(\varepsilon_1+2)(\varepsilon_2+2)}{\varepsilon_1 \cdot \varepsilon_2}\sqrt{\frac{32\ln(2/\beta)}{n}}$$

and is $\left\lfloor \frac{k}{2} \right\rfloor \varepsilon_2$-Untrackable, but no better than $\frac{k}{2}\varepsilon_2 - \varepsilon_1 - \ln 2$-Untrackable. The proof of these claims can be found in the full version of the paper.

7 Report Noisy Inner Product

In this section we present a method for collecting statistics about users' data when it is encoded in a vector of d bits. This mechanism allows us to solve the heavy hitters or histograms problems, while maintaining everlasting privacy. This solution achieves good accuracy with high probability and is effectively untrackable with high probability, but only for a "not so large" number of reports (where "not so large" is approximately the square root of the number of vectors in the state).

The "delta" part of the untrackable bound of this solution can be small, but most likely not *cryptographically* small. While in Differential Privacy one should make sure the "delta" part is cryptographically small, it is not clear whether or not the same requirement applies to the framework of tracking.

The construction of this mechanism follows a general transformation from a Locally Differential Privacy mechanism to an Everlasting Privacy mechanism with certain trackability parameters: memorize a fixed number (L) of executions of a local privacy preserving computation. At each collection the mechanism mimics one of these stored executions, choosing one of them at random. Everlasting Privacy is maintained by the finite access to a user's data: only L total different executions are ever available to the adversary. On the other hand, in terms of trackability, as long as no two different stored execution are played, there is no difference between one user and two users. No guarantees are given if the same stored execution is chosen twice.

In our instantiation of this idea, Report Noisy Inner Product is based on creating a state that contains random d-bit vectors as well as their noisy inner product with the user's private value.

In this setting there are n users. Let:

$$\varepsilon' := \frac{\varepsilon}{2\sqrt{2L\ln\left(\frac{1}{\delta}\right)}}.$$

At initialization, every user i, with private value $u_i \in \{0,1\}^d$ chooses L random vectors $\{v_{i,j}\}_{j\in[L]}$, $v_{i,j} \in \{0,1\}^d \setminus \{\vec{\mathbf{0}}\}$, and L noisy bits $\{x_{i,j}\}_{j\in[L]}$ such that $x_{i,j} \sim \text{Ber}\left(\frac{1}{e^{\varepsilon'}+1}\right)$ and calculates $b_{i,j} = \langle v_{i,j}, u_i\rangle \oplus x_{i,j}$.

At each time of collection, every user j picks at random a vector from the state generated in the previous step. That is, they choose one of the v_i's generated before and the corresponding result of the inner product s_i. They then send it to the server. We refer to the report user i sends at a given collection time as (V_i, B_i) (i.e. the vector and the noisy inner product). The aggregator receives these reports from all users and invokes the frequency oracle to output an estimate:

$$\tilde{p}_u := \frac{2^d-1}{2^d n}\frac{e^{\varepsilon'}+1}{e^{\varepsilon'}-1}\sum_{i\in[n]}(-1)^{\langle V_i, u\rangle \oplus B_i} + \frac{1}{2^d}.$$

Since we never choose the vector $\vec{0}$ as one of the vectors of the state, we introduce a small bias to the probability that a report will agree with any other value than the one used to generate it. This bias is corrected by the multiplicative $\frac{2^d-1}{2^d}$ factor and the additive $\frac{1}{2^d}$ factor, resulting in an unbiased estimator.

Let p be the entire true frequency vector and $\tilde{p}$ as the entire estimated frequency vector.

7.1 Privacy, Accuracy and Trackability

The mechanism Report Noisy Inner Product (RNIP) maintains (ε, δ)-Approximate Everlasting Privacy, outputs $\tilde{p}$ such that with probability $1-\beta$:

$$\begin{aligned}\left\|\tilde{p}-p\right\|_\infty &\le \frac{\varepsilon'+2}{\varepsilon'}\sqrt{\frac{8\ln(2^{d+1}/\beta)}{n}}\\ &= O\left(\sqrt{\frac{\ln(2^{d+1}/\beta)\ln(1/\delta)L}{n\varepsilon^2}}\right).\end{aligned}$$

And it is $\left(0, \frac{k^2}{L}+\frac{L^2}{2^d}\right)$-untrackable for k reports.

The proofs of these claims can be found in the full version of the paper and are similar to the analysis in [16].

7.2 Parameter Selection

When deploying this mechanism, the significant parameters considered are the everlasting privacy and desired accuracy. In our setting we have n users and our data consists of values that can be encoded into d bits. Assume we wish to have everlasting privacy (ε, δ) and accuracy α with probability $1-\beta$. By the results of the accuracy analysis, the required value of the Differential Privacy parameter of every report, which we denoted ε', needs to be at least

$$\frac{2\sqrt{2\ln(2^{d+1}/\beta)}}{\alpha\cdot\sqrt{n}-\sqrt{2\ln(2^{d+1}/\beta)}}.$$

For most interesting settings we can assume that $\alpha > 2\sqrt{\frac{2\ln(2^{d+1}/\beta)}{n}}$, which allows us to choose $\varepsilon' = \frac{4}{\alpha}\sqrt{\frac{2\ln(2^{d+1}/\beta)}{n}}$. Once we have ε' we can say that the mechanism needs to have a state of size at most $L = \left\lfloor\frac{\varepsilon^2}{8\varepsilon'^2\ln(1/\delta)}\right\rfloor$. This means that the mechanism is $\left(0, \frac{k^2}{L}\right)$-Untrackable for k reports.

To summarize, if we were to require (ε, δ)-Everlasting Privacy and α accuracy with probability at least $1-\beta$, then for k reports we can guarantee:

$$\left(0, \widetilde{O}\left(\frac{k^2}{\alpha^2\varepsilon^2 n}\right)\right)\text{-Untrackable.}$$

where the $\widetilde{O}$ hides the logarithmic factors in the relaxation parameter for differential privacy δ, the failure probability β and the size of the vectors d.

8 Conclusions and Open Problems

The issue of using differentially private mechanisms in order to track users is a newly formulated problem. While avoiding tracking is very natural, it has not been investigated before in a formal manner. The notion of Everlasting Privacy is very tempting, and indeed, some companies implemented and deployed it. But Everlasting Privacy should be handled with caution; We have shown that one such deployment of Everlasting Privacy left much to be desired in terms of the untrackable parameter. The risks of tracking are real, and as such every mechanism deployed to a user base must try to prevent it as much as it can.

Many questions concerning tracking are open and the results presented here should be treated as a preliminary investigation. The most important one is how do you combine the constraints on accuracy and on everlasting differential privacy to produce a lower bound on the untrackable parameter. In particular, are the schemes of Sections 6 and 7 the best one can hope for, or are there better mechanisms? One downside of the scheme of Section 7 is the rapid deterioration in the untrackable parameter once k reaches $\sqrt{L}$. Is there a scheme with a more graceful degradation of the untrackable parameter?

The mechanisms we presented are permanent state mechanisms. Perhaps mechanisms which transform the state between executions can achieve better untrackable parameter bounds? Doing such a construction is delicate, since if not done correctly one of two things might happen:

1. The Differential Privacy guarantee will decline the more the state alters.
2. The accuracy will decline, as many different inputs might converge to the same states over time.

But perhaps a clever construction of a mechanism that transforms its state can achieve a much better untrackable parameter bound for given Differential Privacy and accuracy requirements.

Also, perhaps everlasting privacy is an unreasonable demand. A mechanism that achieves privacy for many executions, but not for infinite executions, can be very suitable for practical purposes as well. If so, how can we extend these results to these "long-lasting" privacy mechanisms?

References

1 Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2288–2296, 2017. URL: `http://papers.nips.cc/paper/6823-practical-locally-private-heavy-hitters`.

2 Raef Bassily and Adam D. Smith. Local, private, efficient protocols for succinct histograms. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 127–135. ACM, 2015. `doi:10.1145/2746539.2746632`.

3 T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3):26:1–26:24, 2011. `doi:10.1145/2043621.2043626`.

4 Rachel Cummings, Sara Krehbiel, Yuliia Lut, and Wanrong Zhang. Privately detecting changes in unknown distributions. *CoRR*, abs/1910.01327, 2019. `arXiv:1910.01327`.

5 Rachel Cummings, Sara Krehbiel, Yajun Mei, Rui Tuo, and Wanrong Zhang. Differentially private change-point detection. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 10848–10857, 2018. URL: `http://papers.nips.cc/paper/8280-differentially-private-change-point-detection`.

6 Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3574–3583, 2017. URL: `http://papers.nips.cc/paper/6948-collecting-telemetry-data-privately`.

7 Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In Leonard J. Schulman, editor, *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 715–724. ACM, 2010. `doi:10.1145/1806689.1806787`.

8 Cynthia Dwork, Moni Naor, Toniann Pitassi, Guy N. Rothblum, and Sergey Yekhanin. Pan-private streaming algorithms. In *Proceedings of Innovation in Computer Science, ICS 2010*, pages 66–80. Tsinghua University Press, 2010. URL: `http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/6.html`.

9 Cynthia Dwork, Moni Naor, and Salil P. Vadhan. The privacy of the analyst and the power of the state. In *FOCS*, pages 400–409, 2012. `doi:10.1109/FOCS.2012.87`.

10 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. `doi:10.1561/0400000042`.

11 Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 51–60. IEEE Computer Society, 2010. `doi:10.1109/FOCS.2010.12`.

12 Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2468–2479. SIAM, 2019. `doi:10.1137/1.9781611975482.151`.

13 Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 1054–1067. ACM, 2014. `doi:10.1145/2660267.2660348`.

14 Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2381–2390, 2018. URL: `http://papers.nips.cc/paper/7505-local-differential-privacy-for-evolving-data`.

15 Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 531–540. IEEE Computer Society, 2008. `doi:10.1109/FOCS.2008.27`.

16 Moni Naor, Benny Pinkas, and Eyal Ronen. How to (not) share a password: Privacy preserving protocols for finding heavy hitters with adversarial behavior. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, 2019*, pages 1369–1386. ACM, 2019. `doi:10.1145/3319535.3363204`.

17 Moni Naor and Neil Vexler. Can two walk together: Privacy enhancing methods and preventing tracking of users, 2020. `arXiv:2004.03002`.

18 Salil Vadhan. *The Complexity of Differential Privacy*, pages 347–450. Springer International Publishing, Cham, 2017. `doi:10.1007/978-3-319-57048-8_7`.

A Proofs for Section 3

A.1 Proof of Theorem 14

First, we properly define the composition of m mechanisms, $\{M_i\}_{i\in[m]}$. In our setting a user generate m report streams using the m mechanisms. Namely, each M_i was used to generate k_i reports. Let the sum of all k_i's be k. Let $\widehat{M}$ be the composition of all the M_i's. Formally, $\widehat{M}$ will first generate k_1 reports from the report stream generator of M_1, it will then continue to generate k_2 reports from the report stream generator of M_2, and so on, until all k reports were generated. Let the indices of the reports generated by M_i be J_i, i.e. $J_1 = \{1, ..., k_1\}$, $J_2 = \{k_1 + 1, ..., k_1 + k_2\}$, and so on. Notice that the probability that $G_k^{\widehat{M}}$, on input u, will generate a report stream t of k reports is exactly:

$$Pr\left[G_k^{\widehat{M}}(u) = t\right] = \prod_{i\in[m]} Pr\left[G_{k_i}^{M_i}(u) = t_{J_i}\right]$$

since all mechanisms M_i are executed independently.

We are now ready to prove Theorem A.1.

Proof. In the definition of untrackable, we consider whether a set of reports were generated by a single user or two, according to any partition. To prove that the composition mechanism is $(\gamma', m\delta + \delta')$-untrackable we will prove that the bound in the definition holds for every partition possible.

Consider the partition $P = \{P_1, P_2\}$ of all the reports generated by $\widehat{M}$, where P_1 are the reports associated with the first user and P_2 are the reports associated with the second. We will split this partition into partitions for each mechanism separately. Namely, for every M_i we define a partition $P^i = \left\{P_1^i, P_2^i\right\}$, such that each $P_1^i = P_1 \cap J_i$ and similarly for P_2^i. The partition P^i is exactly the partition on reports generated by M_i induced by P. Notice that we allow P_1^i (or P_2^i) to be empty.

Consider new mechanisms $\{F_i\}_{i\in[m]}$. that each receives as input a bit b. For every F_i, If $b = 0$ the mechanism outputs a stream generated by one copy of M_i, and if $b = 1$ the mechanism outputs a stream generated by two independent copies of M_i according to partition P_i. If either P_1^i or P_2^i are empty, the output F_i will not depend on its input b. If the M_i's are (γ, δ)-untrackable then the F_i's are (γ, δ)-differentially private. This allows us to use Advanced Composition for differential privacy (Theorem 4) to say that the m-fold composition of all of the F_i's is $(\gamma', m\delta + \delta')$-differentially private for γ' as is in the theorem statement. Notice that conditioned on all mechanisms receiving input 1, the output product distribution over reports is identical to the case where all reports, for each mechanism, were generated by two users. Similarly if all inputs are 0, the output product distribution over reports is identical to the case where all reports, for each mechanism, were generated by one user. This implies that the composition of the original M_i's, $\widehat{M}$, is $(\gamma', m\delta + \delta')$-untrackable. ◀

A.2 Proof of Theorem 16

Proof. The proof for this theorem is very intuitive. Since the mechanism is γ-untrackable for k reports, it is also γ-untrackable for fewer reports. This teaches us that by paying no more than γ we can reduce the question of being untrackable for n users to the question of being untrackable for $n-1$ users. Continuing this until we have 2 users costs us $(n-2)\gamma$ to the untrackable parameter, resulting in a total of $(n-1)\gamma$ untrackable.

Formally, we prove this by induction. Assume that a mechanism M is $(t-1)\,\gamma$-untrackable for t users for k reports. We wish to prove that the mechanism M is $t\gamma$-untrackable for $t+1$ users for k reports. The base case, $t=2$, follows directly from the fact that the mechanism is γ-untrackable for k reports.

If the mechanism M is γ-untrackable for k reports, then it is γ-untrackable for fewer reports as well. We denote $Pr\left[A\right] := Pr\left[G^M_{|A|}\left(u\right) \in T_A\right]$. Notice that for all user data $u \in U$, all partitions $P = \{P_i\}_{i\in[t+1]}$ of $[k]$ into $t+1$ parts and all output stream sets $T \subseteq R^k$:

$$\begin{aligned}\prod_{j\in[t+1]} Pr\left[P_j\right] &= Pr\left[P_1\right]\cdot Pr\left[P_2\right]\cdot \prod_{j\in[t+1]\setminus\{1,2\}} Pr\left[P_j\right]\\ &\le e^{\gamma} Pr\left[P_1\cup P_2\right] \prod_{j\in[t+1]\setminus\{1,2\}} Pr\left[G^M_{|P_j|}\left(u\right)\in T_{P_j}\right]\\ &\le e^{t\gamma} Pr\left[G^M_k\left(u\right)\in T\right]\end{aligned}$$

Where the first inequality is due to the mechanism being γ-untrackable for $|P_1|+|P_2|$ reports and the second inequality is due to the induction hypothesis.

Similarly, in the other direction:

$$\begin{aligned}Pr\left[G^M_k\left(u\right)\in T\right] &\le e^{(t-1)\gamma} Pr\left[P_1\cup P_2\right] \prod_{j\in[t+1]\setminus\{1,2\}} Pr\left[P_j\right]\\ &\le e^{t\gamma} Pr\left[P_1\right]\cdot Pr\left[P_2\right]\cdot \prod_{j\in[t+1]\setminus\{1,2\}} Pr\left[P_j\right]\\ &= e^{t\gamma}\prod_{j\in[t+1]} Pr\left[P_j\right]\end{aligned}$$

Where the first inequality is due to the induction hypothesis and the second inequality is due to the mechanism being γ-untrackable for $|P_1|+|P_2|$ reports. ◀

A.3 Proof of Theorem 17

Proof. The proof for this is also rather intuitive. Theorem 13 teaches us that we can exchange the probability that two sets of reports, of size totaling k', originated from two users to the probability they originated from one user by paying no more than $\left\lfloor\frac{k'}{2}\right\rfloor\varepsilon$ in the untrackable parameter. By combining pairs of users, we can use this to reduce the question of being untrackable for n users to the question of untrackable for $\left\lceil\frac{n}{2}\right\rceil$ users, by paying no more than $\left\lfloor\frac{k}{2}\right\rfloor\varepsilon$. By repeating this process $\lceil\log n\rceil - 1$ times we can reduce the question of untrackable for n users to the question of untrackable for 2 users, by paying no more than $\lceil\log n\rceil\left\lfloor\frac{k}{2}\right\rfloor\varepsilon$.

Formally, we prove this by induction. Assume that the mechanism is $\lceil\log t\rceil\left\lfloor\frac{k}{2}\right\rfloor\gamma$-untrackable for t users for k reports. We wish to prove that the mechanism is $\lceil\log\left(t+1\right)\rceil\left\lfloor\frac{k}{2}\right\rfloor\gamma$-untrackable for $t+1$ users for k reports. The base case $t=2$ follows directly from the fact that the mechanism is γ-untrackable for k reports.

Assume t is odd. The proof is very similar when it is even, but for simplicity we will only show it for odd values of t. We denote $\ell := \lceil\log\left(t+1\right)\rceil$ and use the same notation for $Pr\left[A\right]$ as before. Notice that for all user data $u\in U$, all partitions $P=\{P_i\}_{i\in[t+1]}$ of $[k]$ into $t+1$ parts and all output stream sets $T\subseteq R^k$:

$$\begin{aligned}\prod_{j\in[t+1]} Pr\left[P_j\right] = &\prod_{j\in\left[\frac{t+1}{2}\right]} Pr\left[P_{2j}\right]\cdot Pr\left[P_{2j+1}\right]\\ \leq &\prod_{j\in\left[\frac{t+1}{2}\right]} e^{\left\lfloor\frac{\left|P_{2j}\cup P_{2j+1}\right|}{2}\right\rfloor\gamma} Pr\left[P_{2j}\cup P_{2j+1}\right]\\ \leq &\, e^{\left\lfloor\frac{k}{2}\right\rfloor\gamma}\prod_{j\in\left[\frac{t+1}{2}\right]} Pr\left[p_{2j}\cup P_{2j+1}\right]\\ \leq &\, e^{\ell\left\lfloor\frac{k}{2}\right\rfloor\gamma} Pr\left[G_k^M\left(u\right)\in T\right]\end{aligned}$$

Where the first inequality is due to Theorem 13 and the third inequality is due to the induction hypothesis.

Similarly, for the other direction:

$$\begin{aligned}Pr\left[G_k^M\left(u\right)\in T\right] \leq &\, e^{(\ell-1)\left\lfloor\frac{k}{2}\right\rfloor\gamma}\prod_{j\in\left[\frac{t+1}{2}\right]} Pr\left[P_{2j}\cup P_{2j+1}\right]\\ \leq &\, e^{\ell\left\lfloor\frac{k}{2}\right\rfloor\gamma}\prod_{j\in\left[\frac{t+1}{2}\right]} Pr\left[P_{2j}\right]\cdot Pr\left[P_{2j+1}\right]\\ = &\, e^{\ell\left\lfloor\frac{k}{2}\right\rfloor\gamma}\prod_{j\in[t+1]} Pr\left[P_j\right]\end{aligned}$$

Where the first inequality is due to the induction hypothesis and the second inequality is due to Theorem 13. ◀

B Estimated Median and 90th Percentile Of RAPPOR Figure

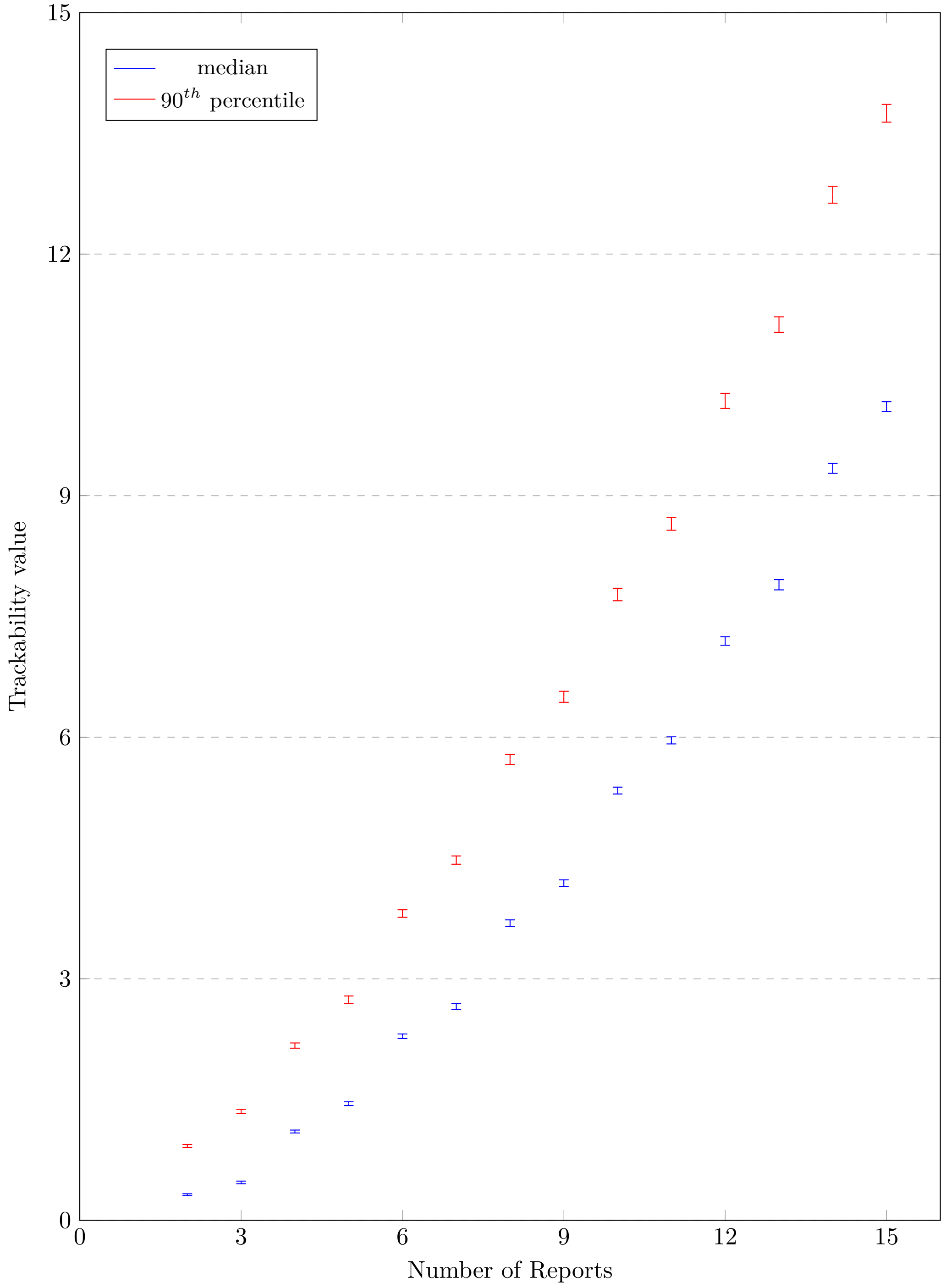


Figure 1 Growth of the estimated median and 90^{th} percentile of the trackability random variable of RAPPOR as a function of the number of reports.

Service in Your Neighborhood: Fairness in Center Location

Christopher Jung
University of Pennsylvania, Philadelphia, PA, USA
chrjung@seas.upenn.edu

Sampath Kannan
University of Pennsylvania, Philadelphia, PA, USA
kannan@cis.upenn.edu

Neil Lutz
Iowa State University, Ames, IA, USA
nlutz@istate.edu

Abstract

When selecting locations for a set of centers, standard clustering algorithms may place unfair burden on some individuals and neighborhoods. We formulate a fairness concept that takes local population densities into account. In particular, given k centers to locate and a population of size n, we define the "neighborhood radius" of an individual i as the minimum radius of a ball centered at i that contains at least n/k individuals. Our objective is to ensure that each individual has a center that is *within at most a small constant factor of her neighborhood radius.*

We present several theoretical results: We show that optimizing this factor is NP-hard; we give an approximation algorithm that guarantees a factor of at most 2 in all metric spaces; and we prove matching lower bounds in some metric spaces. We apply a variant of this algorithm to real-world address data, showing that it is quite different from standard clustering algorithms and outperforms them on our objective function and balances the load between centers more evenly.

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains

Keywords and phrases Fairness, Clustering, Facility Location

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.5

Related Version A full version of the paper is available at `https://arxiv.org/abs/1908.09041`.

Funding *Christopher Jung*: Research supported by a grant from the Quattrone Center, University of Pennsylvania.
Sampath Kannan: Research supported in part by NSF Grants CCF-1733794 and CCF-1763307.

Acknowledgements We thank Moni Naor and Omer Reingold for helpful discussions, and we thank Anupam Gupta for alerting us to the similarities between our Theorem 2 and Lemma 1 of Chan, Dinitz, and Gupta [7].

1 Introduction

Fairness in decision making has become an important research topic as more and more classification decisions, such as college admissions, bank loans, parole and sentencing, are made with the assistance of machine learning algorithms [10]. Such decisions are made on individuals at a particular point in time, and although they have long-term consequences on the individuals affected, there is at least the prospect of these decisions being revisited with new data about these individuals. In contrast, certain infrastructural decisions, for example, about where to locate hospitals, schools, library branches, police stations, or fire stations have long-term consequences on all the residents of a town, district, or county. Stories about

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 5; pp. 5:1–5:15
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

neighborhoods not being adequately served make frequent headlines. Such stories range from food deserts in inner cities because of the absence of supermarkets that sell fresh fruit and vegetables to lack of access to medical services in rural areas [14, 26].

In many situations, equal treatment of all individuals requires clustering individuals into roughly equal-sized groups and allocating the same amounts of resources (such as schools or hospitals) to each group. This means that these resources will naturally be located farther away on average from residents of a sparse district. This is generally accepted by society, and one could argue that it is in fact the just way to allocate resources. Thus, even where all individuals are entitled to equal treatment, it is admissible, even desirable, to discriminate based on geographic location. However, even in situations where there is great geographic variation in density, and concomitant variation in how resources are allocated, we would like to ensure some form of fairness to each individual.

We ask what might be a fair way to locate k hospitals, say, in an area with varying population densities. A standard formulation such as the k-center problem is problematic for at least two reasons:

First, in a good k-center solution, a hospital located in an urban area would be overcrowded. Thus, one kind of fairness we want is *load balance*; the numbers of people served by each center should be as close to equal as possible. Intuitively the definition we give seems tailored to provide such balance, and we confirm this empirically.

Second, people living in areas with different population densities have different expectations for a reasonable distance to travel to a hospital. In rural areas, it would be unreasonable for a resident to expect to find a hospital within a mile, say, of her residence, but in an urban area this might be an entirely reasonable expectation. This is reinforced by the fact that individuals in dense urban areas – especially dense, low-income urban neighborhoods – are more likely to rely on bicycles or public transit and less likely to have access to a car [24].

Taking this perspective, consider the problem of serving a population P of n people using k centers, for a given k. On average, we expect each center to serve n/k people. An individual i might reasonably hope that the center that serves i is no farther than the $(\lceil \frac{n}{k} \rceil)^{\text{th}}$ nearest individual from i, including i itself. Thus, for a given P and k, we define the *neighborhood radius* $NR(i)$ to be the distance from i to its $(\lceil \frac{n}{k} \rceil - 1)^{\text{th}}$ nearest neighbor.

Unfortunately, it is not always possible to find a solution with k centers where each individual finds a center within her neighborhood radius. Hence our goal is to optimize how far we deviate from this ideal. Given a solution S that specifies the placement of the k centers, let $d(i, S)$ denote the distance from individual i to the closest center in S. Let

$$\alpha(S) = \max_i \frac{d(i, S)}{NR(i)}$$

denote the maximum factor by which an individual's distance to the center nearest to her, exceeds her neighborhood radius. We say that an algorithm achieves α-fairness if the solution S it produces has $\alpha(S) \leq \alpha$. The goal of this paper is to design an efficient algorithm to locate k centers that achieves a small value of α, the maximum factor by which any individual's fair expectations are not met.

Fair k-Center: For as small a value of α as feasible, given n points in a metric space, and a number k, find a solution S^* consisting of a subset of at most k of the given points so that $\alpha(S^*) \leq \alpha$.

One could formulate a Steiner version of this problem, where centers are allowed at arbitrary points in the metric space, but we do not consider this variant in this paper. We also formulate an extremal version of the problem: For a given metric space, what is the worst-case value, over all possible configurations of points in the metric space, of $\alpha(S^*)$? We perform empirical comparisons between our fair k-center formulation and the standard k-center, k-means, and k-medians formulations. Using algorithms designed for each of these optimization problems, we select sets of center locations based on two geographical data sets from Fairfax County, Virginia and Allegheny County, Pennsylvania.

Our results are as follows:

- There is an efficient algorithm that achieves $\alpha = 2$ for any set of points and any parameter k, in any metric space (Theorem 2). We have come to learn that the same algorithm was discovered earlier in a different context by [7].
- Finding the optimal α for a given set of points and parameter k is NP-hard (Theorem 8).
- There are metric spaces and configurations of points for which $\alpha = 2$ is the best possible (Proposition 6). For Euclidean spaces there are configurations that require $\alpha = \sqrt{2}$ (Proposition 7).
- On real data standard clustering algorithms achieve worse α than is achieved by an algorithm we describe (Table 1).
- Associating with any algorithm a vector of at most k values giving the number of points assigned to each center, and viewing load balance as the variance of this vector, our algorithm empirically achieves much better load balance than the other clustering algorithms (Table 2).

1.1 Related Work

There is a rapidly growing body of literature on fair clustering [9, 17, 18, 3, 8, 6, 5]. Most of this work has attempted to optimize standard k-center, k-means, and k-medians objective functions, but under some fairness constraints. In particular, there has been a focus on group fairness: requiring that each group must have approximately equal representation across clusters. Two of the motivations for our fairness notion are that outliers should not disproportionately affect clustering outcomes, and that cluster sizes should be roughly equal. These motivations are shared, respectively, by density-based clustering [27, 1], in which data points in sparse regions are treated as noise, and by load-balanced clustering [19].

2 Defining α-fairness

We consider a nonempty collection P of (not necessarily distinct) points in a metric space (X, d) and some positive integer parameter $k \leq |P| = n$. A *centers algorithm* takes an instance (P, k) as input and returns a set $S \subseteq P$ with $|S| \leq k$ of designated *centers*. The travel distance from a point $x \in X$ to S is $d(x, S)$, the minimum distance from x to a center in S:

$$d(x, S) = \min\{d(x, s) : s \in S\}\,.$$

The goal of a centers algorithm is to select a "good" set of centers according to some criterion. For example, the following are well-studied optimization problems based on natural objective functions for assessing the quality of a solution set.

- **k-Center:** minimize the maximum travel distance among individuals in P, $\max_{i \in P} d(i, S)$ [2].
- **k-Medians:** minimize the average travel distance, or equivalently, $\sum_{i \in P} d(i, S)$ [16].
- **k-Means:** minimize the sum of the squares of the travel distances, $\sum_{i \in P} d(i, S)^2$ [21].

The objective function we introduce, unlike those above, is based on the *neighborhood radius* at a point $x \in X$, $NR(x)$. That is, the minimum radius r such that at least $|P|/k$ of the points in P are within distance r of x:

$$NR_{P,k}(x) = \min\{r : |B_r(x) \cap P| \geq n/k\} \ ,$$

where $B_r(x)$ is the closed ball of radius r around x. When P and k are clear from context, we omit these subscripts and simply denote the neighborhood radius at x by $NR(x)$.

We quantify the fairness of a set of centers S on a set of points according to the worst ratio between travel distance and neighborhood radius for any point in P:

$$\alpha_{P,k}(S) = \sup_{i \in P} \frac{d(i,S)}{NR_{P,k}(i)} \ ,$$

adopting the conventions that $0/0 = \infty/\infty = 1$ and $c/0 = \infty$ for any $c > 0$.

Given a centers algorithm A and a constant α, we say that A achieves α*-fairness* on an instance (P,k) if

$$\alpha_{P,k}(A(P,k)) \leq \alpha \, .$$

We say that A is α*-fair* in the given metric space (X,d) if it achieves α-fairness on every instance. That is, if $\alpha_{P,k}(A(P,k)) \leq \alpha$ for all $P \subseteq X$ and all $1 \leq k \leq n$.

Solutions that are optimal for other standard objective functions can be infinitely unfair with respect to α, as shown by the following example on the real line.

▶ **Example 1.** Let $k = 3$ and consider $P = \{-x, 0, 0, 1, 1, x\}$, where x is some large number, as pictured in Figure 1. The optimal solution with respect to the k-center, k-medians, and k-means objective functions is to place one center at either 0 or 1 and the other two centers at $-x$ and x. But the neighborhood radius is 0 at 0 and 1, and whichever of these is not chosen as a center will have to travel a distance of 1 to the nearest center, meaning that

$$\alpha(\{-x, 0, x\}) = \alpha(\{-x, 1, x\}) \geq \frac{1}{0} = \infty \, .$$

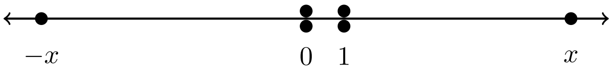


Figure 1 For the above population with $k = 3$, optimizing for $\max_{i \in P} d(i,S)$, $\sum_{i \in P} d(i,S)$, or $\sum_{i \in P} d(i,S)$ will yield a solution that is not α-fair for any finite α.

Although we do not consider allowing Steiner points as centers in this paper, it is clear that even optimal Steiner solutions to the three classical problems – all of which place only one center in the interval $[0,1]$ – do no better in terms of our fairness objective. This example demonstrates that a different approach is needed to achieve even the weakest of fairness guarantees. In the appendix, we include two more examples of metric spaces in which strong fairness guarantees are easy to achieve.

3 Theoretical Results

Given an instance (P,k), let $\alpha^*_{P,k}$ be the minimum value such that α-fairness can be achieved. In this section we prove that

$$1/2 \leq \alpha^*_{P,k} \leq 2$$

always holds, that equality is possible at each end of that bounding interval, and that $\alpha^*_{P,k}$ is NP-hard to compute.

3.1 A 2-Fair Algorithm

We now give an algorithm, 2FairKCenter, that achieves 2-fairness on every instance and in every metric space. In each iteration, 2FairKCenter chooses a center s with minimum neighborhood radius among the set Z of candidate centers. Then, it removes from Z all points i that are sufficiently close to s.

We have recently become aware that achieving 2-fairness is equivalent to finding a (k/n)-density net, as defined by Chan, Dinitz, and Gupta in the context of constructing slack spanners [7]. In proving that ϵ-density nets can be found in polynomial time for all $\epsilon \in (0,1)$, that work describes an algorithm that is essentially identical to 2FairKCenter. In order to keep this paper self-contained, we include the algorithm description and proof of 2-fairness here.

Algorithm 1 2FairKCenter(P, k).

```
$Z = P$
$S = \emptyset$
while $S \neq \emptyset$ do
    choose $s \in \arg\min_{i \in Z} NR_{P,k}(i)$
    $S = S \cup \{s\}$
    $Z = \{i \in Z : d(i,s) > NR_{P,k}(i) + NR_{P,k}(s)\}$
end
return $S$
```

▶ **Theorem 2.** *2FairKCenter is 2-fair in every metric space.*

Proof. Fix a metric space (X, d), let $P \subseteq X$, and let $k \leq n$. Let s_j be the j^{th} center added to S. For any $j' > j$, the definition of the set S guarantees that $B_{NR(s_j)}(s_j)$ and $B_{NR(s_{j'})}(s_{j'})$ are disjoint. Thus, the balls

$$B_{NR(s_1)}(s_1), B_{NR(s_2)}(s_2), \ldots$$

are all pairwise disjoint, and by the definition of neighborhood radius, each includes at least n/k points in P. It follows that there can be at most k centers, so this algorithm will output a valid solution.

Now, a point $i \in P$ is excluded from Z only when there is some $s \in S$ such that $d(i,s) \leq NR(i) + NR(s)$, which is at most $2 \cdot NR(i)$ by our choice of s. So when our algorithm terminates, $d(i,S)/NR(i) \leq 2$ holds for all $i \in P$. Thus, 2-fairness is achieved on (P, k), and the algorithm is 2-fair. ◀

3.2 Lower Bounds

We now give four lower bounds on fairness: We prove that it is never possible to achieve better than 1/2-fairness on any instance; that no algorithm can be better than 1-fair, regardless of the metric space; that there exist metric spaces in which no algorithm can be better than 2-fair; and that no algorithm can be better than $\sqrt{2}$-fair in Euclidean spaces of dimension greater than 1. The first three of these results demonstrate the tightness of Example 9, Example 10 (See appendix for these examples.), and Theorem 2, respectively. We defer the proofs for the following propositions to the appendix

▶ **Proposition 3.** *In every metric space (X, d), for all $S \subseteq P \subseteq X$ and $1 \leq k \leq |S|$, we have $\alpha_{P,k}(S) \geq 1/2$.*

Combining Theorem 2 with Proposition 3 immediately yields the following.

▶ **Corollary 4.** *2FairKCenter is a 4-approximation algorithm to the best α achievable for any configuration of points in any metric space.*

▶ **Proposition 5.** *For all metric spaces (X, d) and all $\alpha < 1$, there is no centers algorithm that is α-fair in (X, d).*

▶ **Proposition 6.** *There exists a metric space (X, d) such that for all $\alpha < 2$, there is no centers algorithm that is α-fair in (X, d).*

▶ **Proposition 7.** *For all $m \geq 2$ and all $\alpha < \sqrt{2}$, there is no centers algorithm that α-fair in m-dimensional Euclidean space.*

3.3 NP-Completeness

The k-center, k-medians, and k-means problems are all known to be NP-hard, and the problem of checking, for a given instance, whether a given value of the objective function is achievable, is NP-complete [11, 22, 15]. We now show in the following theorem that the same is true for our objective function α.

▶ **Theorem 8.** *The problem of determining whether* 1*-fairness can be achieved on a given instance is NP-complete.*

Proof. This problem is a special case of the hitting set problem, where the sets are

$$S_i = \{j \in P : d(i, j) \leq \alpha \cdot NR_{P,k}(j)\}$$

for each $i \in P$. So it belongs to NP.

We prove NP-hardness by reduction from the dominating set problem. Let G be a graph on a set U of n vertices, and let $1 \leq k \leq n$; without loss of generality, we assume that $n - k$ is even. We construct a new graph G' that contains G as a subgraph and also has the following:

- a set V of $2n$ vertices with degree 1 such that each vertex $u \in U$ is adjacent to two vertices $u_1, u_2 \in V$, and
- a set W of $6n - 6k$ vertices arranged as $\frac{3}{2}(n - k)$ disjoint 4-cycles.

Letting $P = U \cup V \cup W$ and $k' = 3n - 2k$, we will show that G has a dominating set of size k if and only if there is a set $S \subseteq P$ with $|S| = k'$ and $\alpha_{P,k'}(S) \leq 1$. Since (G', k') can be efficiently computed from (G, k), this will suffice to prove the theorem.

Suppose that G has a dominating set D of size k, and consider the set of centers $S = D \cup T$, where T is a set consisting of two vertices from each of the squares in W, so that $d(w, S) = d(w, T) \leq 1$ for all $w \in W$. Now,

$$\frac{n}{k'} = \frac{n + 2n + 6n - 6k}{3n - 2k} = 3\,,$$

so $NR_{P,k'}(w) = 1$ for each $w \in W$. Each $u \in U$ has at least two neighbors – namely, u_1 and u_2 – so we also have $NR_{P,k'}(u) = 1$ for all $u \in U$, and it follows immediately that $NR_{P,k'}(v) = 2$ for each $v \in V$. The fact that $D \subseteq S$ is a dominating set means that $d(u, S) \leq 1$ for each $u \in U$ and therefore that $d(v, S) \leq 2$ for each $v \in V$. Thus, $\alpha_{P,k'}(S) = 1$.

Conversely, suppose that there is some set $S \subseteq P$ with $|S| = k'$ and $\alpha_{P,k'}(S) \leq 1$. Then S must contain at least two vertices from each square in W, so letting $Y = S \cap (U \cup V)$, we have

$$|Y| \leq k' - (3n - 3k) = k\,.$$

For each $u \in U$, we must have

$$d(u, Y) = d(u, S) \leq NR_{P,k'}(u) = 1\,.$$

We construct a set D by taking each vertex in $Y \cap V$ and replacing it with the adjacent vertex in U. Then $|D| \leq |Y| \leq k$, and for each $u \in U$, $d(u, D) \leq d(u, Y)$, meaning that D is a dominating set for G. We conclude that this problem is NP-complete. ◀

4 Experiments

In this section we measure the fairness of three standard clustering algorithms on two geographic data sets, and we compare their performance to that of a modified version of 2FairKCenter that still guarantees 2-fairness but attempts to be even fairer.

4.1 A Heuristic Refinement of 2FairKCenter

Our algorithm 2FairKCenter always yields a 2-fair solution, but this solution might be less than optimal and use fewer than k centers. To avoid this situation, we introduce AlphaFairKCenter, a version of 2FairKCenter that is parameterized by a fairness guarantee parameter, α. This algorithm achieves α-fairness on every instance by essentially the same argument we used to prove that Algorithm 1 is 2-fair. The catch is that the output will not necessarily be a valid solution: For $\alpha < 2$, Algorithm 2 may select more than k centers on a given instance (P, k).

Algorithm 2 AlphaFairKCenter(α, P, k).

```
$Z = P$
$S = \emptyset$
while $S \neq \emptyset$ do
    choose $s \in \arg\min_{i \in Z} NR_{P,k}(i)$
    $S = S \cup \{s\}$
    $Z = \{i \in Z : d(i, s) > \alpha \cdot NR_{P,k}(i)\}$
return $S$
```

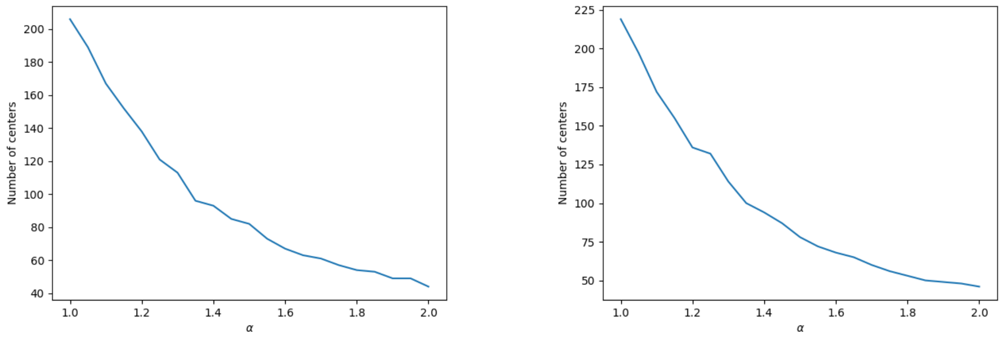
For each instance (P, k), we define a function $f_{P,k} : [1, 2] \to \mathbb{N}$ by

$$f_{P,k}(\alpha) = |\text{AlphaFairKCenter}(\alpha, P, k)|\ ,$$

the number of centers chosen by AlphaFairKCenter with parameter α on instance (P, k). Our goal is to find a small α such that $f_{P,k}(\alpha) \leq k$. In order to do this, we will perform a binary search on the interval $[1, 2]$, recursively searching the lower half of the interval when $f_{P,k}(\alpha) > k$ and the upper half of the interval otherwise.

If $f_{P,k}$ is a monotonic function, then this search will find

$$\inf\{\alpha \in [1, 2] : f_{P,k}(\alpha) \leq k\}$$



(a) Fairfax County, Virginia.

(b) Allegheny County, Pennsylvania.

Figure 2 The number of centers chosen by AlphaFairKCenter for different values of α with $k = 100$ given address points in two counties.

up to arbitrary precision. Intuitively, $f_{P,k}$ has a general tendency to be decreasing – a weaker fairness guarantee requires fewer centers – but in fact $f_{P,k}$ is not necessarily monotonic, and local extrema may cause our search to select a larger α than is necessary.

Fortunately, as shown in Figure 2, $f_{P,k}$ seems to behave monotonically at coarse scales on real data. Furthermore, deviations from monotonicity cannot affect the validity of the solution we find, only its optimality. Hence, this binary search appears to be a useful heuristic, and we employ it in our algorithm FairKCenter. In addition to an instance (P, k), this algorithm takes as input a precision parameter t that determines the depth of the binary search.

Algorithm 3 FairKCenter(t, P, k).

```
$low = 1$
$high = 2$
for $i = 1, 2, \ldots, t$ do
    $mid = (low + high)/2$
    if $|\text{AlphaFairKCenter}(mid, P, k)| \leq k$ then
        $high = mid$
    else
        $low = mid$
return AlphaFairKCenter$(high, P, k)$
```

4.2 Experimental Setup

We applied our algorithm FairKCenter to select 100 center locations in two American counties: Fairfax County, Virginia, and Allegheny County, Pennsylvania. Fairfax County is located near Washington, D.C., and is primarily suburban. According to the 2010 United States Census [25], its population density is 1068 people per square kilometer, with census tracts ranging in density from 56 to 23,397 people per square kilometer. Allegheny County contains the city of Pittsburgh as well as many of its suburbs and exurbs; in the 2010 Census, the county's population density was 647 people per square kilometer, with census tracts ranging in density from 48 to 12,474 people per square kilometer.[1]

[1] The stated ranges of population density exclude the few census tracts with fewer than 100 people. Some census tracts are uninhabited.

The “populations” for our experiment were the sets of all address points in each county, not the locations of individual people. The Fairfax data set contains 537,514 address points, and the Allegheny data set contains 370,776. The data sets were published by Fairfax County GIS and the Allegheny County / City of Pittsburgh / Western PA Regional Data Center, respectively [13, 23]. We measured Euclidean distance after projecting (latitude, longitude) pairs onto the plane using the Universal Transverse Mercator coordinate system.

The bottleneck for our algorithm in terms of running time is calculating the neighborhood radius for each point. In order to accelerate this process, we used the Python library KD-tree [4]. The KD-tree data structure allows us to quickly query the distance to any point's $(\lceil n/k \rceil - 1)^{\text{th}}$ nearest neighbor, which is exactly the definition of the neighborhood radius at that point.

We compared the performance of FairKCenter to standard algorithms for the k-means, k-medians, and k-center problems:

- For k-means, we used the Python library sklearn, which employs either Lloyd's algorithm [20] or Elkan's algorithm [12], depending on the problem size and parameters.
- For k-medians, we used the Python library pyclustering to execute a variant of Lloyd's algorithm that calculates a median instead of a centroid in each iteration.
- For k-center, we implemented the standard greedy approximation algorithm [28].

For each county, we assessed the performance of each algorithm according to our fairness objective function α as well as the k-means, k-medians, and k-center objective functions. We also measured how well each algorithm balanced the load by finding the standard deviation in the number of addresses served by each center.

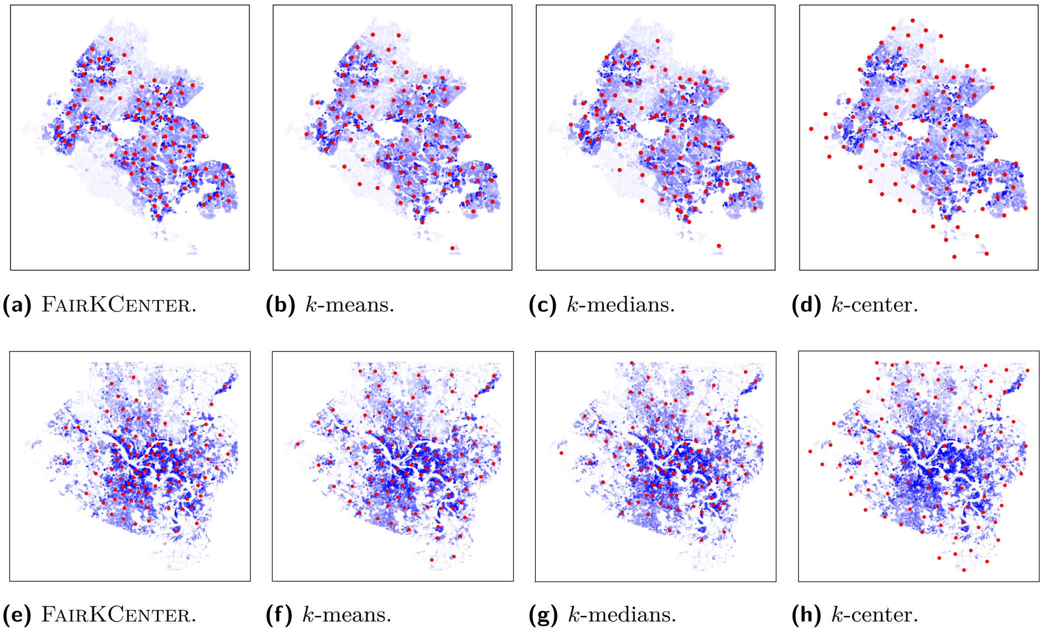
4.3 Experimental Results

In Figure 3, we show the population density map of Fairfax County and Alllegheny County along with 100 centers whose locations were determined by FairKCenter, k-means, k-medians, and k-center. Each tiny blue point, whose transparency has been slightly lowered in order to better show the population density of each region, corresponds to an address point, and each red point is a center.

In Table 1, we show how each algorithm performs in terms of each problem's objective function for each dataset. The values are in units of meters for the k-medians and k-center objective functions, and square meters for the k-means objective function.

Table 1 Performance of each algorithm on Fairfax County and Allegheny County with respect to various objective functions.

		Objective function			
	Algorithm	α	k-means	k-medians	k-center
Fairfax	FairKCenter	1.34306	1811007	1373.79	10002.64
	k-means	1.45643	1137163	1217.07	5662.06
	k-medians	1.80263	1613910	1393.39	6446.81
	k-center	2.57986	2027176	1675.85	2925.80
Allegheny	FairKCenter	1.33721	3600461	1902.78	11615.59
	k-means	1.57453	2082104	1632.00	6183.67
	k-medians	1.90726	3020040	1841.34	7835.79
	k-center	2.67804	3763269	2272.11	3815.03



(a) FairKCenter. **(b)** k-means. **(c)** k-medians. **(d)** k-center.

(e) FairKCenter. **(f)** k-means. **(g)** k-medians. **(h)** k-center.

Figure 3 Placing 100 centers in Fairfax County (a–d) and Allegheny County (e–h), using FairKCenter and algorithms for the k-means, k-medians, and k-center problems.

It is immediately apparent that FairKCenter tends to place more centers in denser regions, compared to other algorithms. This is consistent with the intuition behind the algorithm, as address points in dense regions have relatively smaller neighborhood radius. Although the maximum travel distance is increased significantly relative to the other algorithms, these large travel distances are experienced only by few residents of particularly sparse areas. The increase in average travel distance is more modest, and in Fairfax County our algorithm does even better than the k-medians algorithm with respect to the k-medians objective function. In exchange for these compromises, our algorithm does significantly better with respect to α, ensuring that no individual will needs to venture too far from their density-dependent neighborhood.

Furthermore, FairKCenter balances the load more evenly across centers than other algorithms. Table 2 shows the standard deviation in the number of address points served by each center, i.e., the number of points for which that center is the nearest. For both counties, this value is significantly lower for FairKCenter than for the other algorithms. Our algorithm balances load particularly well compared to the k-center algorithm, which essentially ignores population density.

Table 2 Standard deviation in cluster sizes.

	County	
Algorithm	Fairfax	Allegheny
FairKCenter	1032.49	1696.95
k-means	1344.17	2273.53
k-medians	1630.06	1922.53
k-center	2758.44	5691.10

5 Conclusion

We have formulated a simple geometric concept that captures an intuitive notion of fairness: To whatever extent possible, an individual should have access to resources within her own neighborhood. We have proved basic properties of this fairness concept, given a general approximation algorithm for its optimization, and shown that this algorithm performs well on real data.

One potential future direction for this work is to refine the notion of what constitutes an individual's "neighborhood" for a given purpose. We used the inverse of local population density as a proxy for the size of a neighborhood, and there are good reasons to believe that these two are correlated. But a more sophisticated approach to defining neighborhood size – and possibly shape – might incorporate data on transit times and availability of different modes of transportation. More ambitiously, cellular location data might be used to establish the extent of the common "orbits" of residents of a given small area.

On the theoretical side, an obvious next direction is to find algorithms that yield stronger approximation ratios. Our algorithm FairKCenter improved on 2FairKCenter in our experiments by less aggressively eliminating candidate centers, but we do not have any theoretical characterization of the instances on which FairKCenter will achieve fairness that is strictly better 2FairKCenter. The monotonicity property that is necessary to ensure a fully successful binary search does not hold in general, but one might still be able to identify critical values of α in this range, solve the problem at each of these critical values, and use the smallest one that results in a number of centers that is at most k.

Another interesting extension would be to allow Steiner points, removing the restriction that the solution set S is a subset of the population P. While *prima facie* it looks like we might have to consider infinitely many such possible centers, one can show that the only points we need to consider as centers are points x such that for some r, the boundary of $B_r(x)$ contains at least 3 of the input points. This reduces the number of Steiner points to consider to at most n^3.

Finally, while we have tight lower and upper bounds on α for arbitrary metric spaces, for Euclidean spaces we have a lower bound of $\sqrt{2}$ and an upper bound of 2. It would be interesting to close the gap, perhaps by improving the upper bound for this case.

References

1 Amineh Amini, Ying Wah Teh, and Hadi Saboohi. On density-based data streams clustering algorithms: A survey. *J. Comput. Sci. Technol.*, 29(1):116–141, 2014. doi:10.1007/s11390-014-1416-y.

2 T. Asano, B. Bhattacharya, M. Keil, and F. Yao. Clustering algorithms based on minimum and maximum spanning trees. In *Proceedings of the Fourth Annual Symposium on Computational Geometry*, SCG '88, pages 252–257, New York, NY, USA, 1988. ACM. doi:10.1145/73393.73419.

3 Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 405–413, 2019. URL: http://proceedings.mlr.press/v97/backurs19a.html.

4 Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975. doi:10.1145/361002.361007.

5 Suman Kalyan Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4955–4966, 2019. URL: http://papers.nips.cc/paper/8741-fair-algorithms-for-clustering.

6 Ioana Oriana Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*, pages 18:1–18:22, 2019. doi:10.4230/LIPIcs.APPROX-RANDOM.2019.18.

7 T. H. Hubert Chan, Michael Dinitz, and Anupam Gupta. Spanners with slack. In Yossi Azar and Thomas Erlebach, editors, *Algorithms – ESA 2006*, pages 196–207, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

8 Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 1032–1041, 2019. URL: http://proceedings.mlr.press/v97/chen19d.html.

9 Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5029–5037, 2017. URL: http://papers.nips.cc/paper/7088-fair-clustering-through-fairlets.

10 A. Chouldechova and A. Roth. The Frontiers of Fairness in Machine Learning. *arXiv e-prints*, October 2018. arXiv:1810.08810.

11 Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004. doi:10.1023/B:MACH.0000033113.59016.96.

12 Charles Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 147–153. AAAI Press, 2003. URL: http://dl.acm.org/citation.cfm?id=3041838.3041857.

13 Fairfax County GIS. Address points, 2019. URL: https://catalog.data.gov/dataset/address-points-b4b16.

14 Heather Haddon and Annie Gasparro. Companies and government seek new answers for food deserts. *The Wall Street Journal*, October 2016. URL: https://www.wsj.com/articles/companies-and-government-seek-new-answers-for-food-deserts-1476670262.

15 Dorit S. Hochbaum. When are NP-hard location problems easy? *Annals OR*, 1(3):201–214, 1984. doi:10.1007/BF01874389.

16 Anil K. Jain and Richard C. Dubes. Algorithms for clustering data, 1988.

17 Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3448–3457, 2019. URL: http://proceedings.mlr.press/v97/kleindessner19a.html.

18 Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3458–3467, 2019. URL: http://proceedings.mlr.press/v97/kleindessner19b.html.

19 Ying Liao, Huan Qi, and Weiqun Li. Load-balanced clustering algorithm with distributed self-organization for wireless sensor networks. *IEEE sensors journal*, 13(5):1498–1506, 2012.

20 S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 2006. doi:10.1109/TIT.1982.1056489.

21 J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. URL: https://projecteuclid.org/euclid.bsmsp/1200512992.

22 Nimrod Megiddo and Kenneth J. Supowit. On the complexity of some common geometric location problems. *SIAM J. Comput.*, 13(1):182–196, 1984. doi:10.1137/0213014.

23 Allegheny County / City of Pittsburgh / Western PA Regional Data Center. Allegheny county address points, 2018. URL: `https://catalog.data.gov/dataset/allegheny-county-address-points-07dff`.

24 Issi Romem. Getting around, or just getting by? Where people live with fewer cars, 2019. URL: `https://www.trulia.com/research/people-per-vehicle-map/`.

25 U.S. Census Bureau. Population, housing units, area, and density: 2010 – county – census tract, 2010 census summary file 1, 2010 Census. URL: `https://factfinder.census.gov`.

26 Kelly Virella. Doctors and health workers reflect on rural america's limited access to care. *The New York Times*, 2018. URL: `https://www.nytimes.com/2018/07/19/reader-center/rural-health-care.html`.

27 Wei-Tung Wang, Yi-Leh Wu, Cheng-Yuan Tang, and Maw-Kae Hor. Adaptive density-based spatial clustering of applications with noise (dbscan) according to data. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 445–451. IEEE, 2015.

28 David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2011.

A Additional Examples from Section 2

▶ **Example 9** (A star graph). Consider the graph metric on a star graph: a graph with vertex set $X = \{z, x_1, \dots, x_m\}$ and edge set $\{\{z, x_1\}, \dots, \{z, x_m\}\}$, as in Figure 4. If $(m+1)/2 < k < m+1$, then $NR_{X,k}(z) = 1$ and

$$NR_{X,k}(x_1) = \cdots = NR_{X,k}(x_m) = 2\,.$$

As long as $z \in S$, we have $d(z, S) = 0$ and

$$d(x_1, S) = \cdots = d(x_m, S) = 1\,,$$

so an algorithm that selects z as a center achieves 1/2-fairness for any such instance.

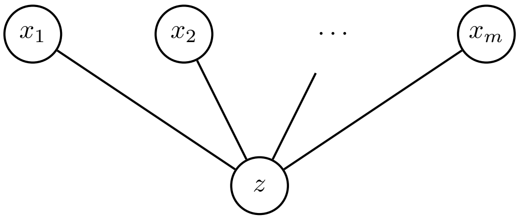


Figure 4 1/2-fairness can be achieved in this instance by selecting z as a center.

▶ **Example 10** (The discrete metric). Consider any nonempty set X under the discrete metric, where $d(x, y) = 0$ if $x = y$ and 1 otherwise. Let $P \subseteq X$ be any nonempty set, let $1 \leq k \leq n$, let $S \subseteq P$ be any nonempty set of size k, and let $i \in P$. If $k = n$, then we have $NR(i) = d(i, S) = 0$. Otherwise, $NR(i) = 1$ and $d(i, S) \in \{0, 1\}$. Hence, there is a 1-fair algorithm for discrete metric spaces: Select any nonempty set of centers.

Most metric spaces do not share the essential property of Examples 9 and 10: the existence of an extremely central point that is close to all other points. In general, achieving fairness requires more care about how one distributes multiple centers.

An ideal situation for the goal of 1-fairness would if the population were arranged in well-separated “villages” containing n/k individuals each, where the diameter of each village is less than the space between villages. In this case, placing a single center anywhere in each village would achieve 1-fairness.

But consider what happens if two villages are brought closer together: Some of an individual’s $\lceil n/k \rceil - 1$ closest neighbors might then reside in the other village, meaning that her neighborhood radius no longer encompasses the entirety of her own village, possibly including that village’s center. This general situation is more difficult, but in the one-dimensional case, 1-fairness can still be achieved, as the following example shows.

Algorithm 4 RealLineFairKCenter(P, k).

```
$S = \emptyset$
while $P \neq \emptyset$ do
    $s = \min P$
    $S = S \cup \{s\}$
    $P = P \setminus B_{NR_{P,k}}(s)$
return $S$
```

▶ **Example 11** (The real line). Given any finite set $P \subseteq \mathbb{R}$, Algorithm 4 starts from the left and takes every $\lceil n/k \rceil^{\text{th}}$ point. Notice that the population P in which the neighborhood radius is determined changes with each iteration.

Each iteration removes at least n/k points from P, so the algorithm will terminate with at most k centers. The $\lceil n/k \rceil$ closest points to any point on the line must include the $j\lceil n/k \rceil^{\text{th}}$ smallest point for some $1 \leq j \leq k$, so this algorithm is 1-fair.

Unfortunately, this approach cannot be extended to higher dimensions. As we show in Section 3, 1-fairness is not always achievable, even in the Euclidean plane.

B Proofs of Propositions in Subsection 3.2

▶ **Proposition 3.** *In every metric space (X, d), for all $S \subseteq P \subseteq X$ and $1 \leq k \leq |S|$, we have $\alpha_{P,k}(S) \geq 1/2$.*

Proof. For each center $s \in S$, define the set

$$J(s) = \{i \in P : d(i, s) = d(i, S)\},$$

the set of points in P for which s is a closest center. There are at most k centers in S, and $\bigcup_{s \in S} J(s) = P$, so there must be some center $\hat{s}$ with $J(\hat{s}) \geq n/k$. Now, let

$$\hat{\imath} \in \arg\max_{i \in J(\hat{s})} d(i, \hat{s}).$$

For each $j \in J(\hat{s})$, we know that $d(j, \hat{s}) \leq d(\hat{\imath}, \hat{s})$, so by the triangle inequality, $d(\hat{\imath}, j) \leq 2d(\hat{\imath}, \hat{s})$. Thus, $B_{2d(\hat{\imath},\hat{s})}(\hat{\imath})$ contains all of $J(\hat{s})$, meaning that

$$\left|B_{2d(\hat{\imath},\hat{s})} \cap P\right| \geq |J(\hat{s})| \geq n/k,$$

so $NR_{P,k}(\hat{\imath}) \leq 2d(\hat{\imath}, \hat{s}) = 2d(\hat{\imath}, S)$. It follows that $\alpha_{P,k}(S) \geq 1/2$. ◀

▶ **Proposition 5.** *For all metric spaces (X, d) and all $\alpha < 1$, there is no centers algorithm that is α-fair in (X, d).*

Proof. Suppose the number of available centers is the same as the size of the population: $k = n$. Then $NR_{P,k}(x) = 0$ for all x, so $\alpha_{P,k}(A(P,k))$ is 1 if A places a center at every point in P and ∞ otherwise. ◀

▶ **Proposition 6.** *There exists a metric space (X, d) such that for all $\alpha < 2$, there is no centers algorithm that is α-fair in (X, d).*

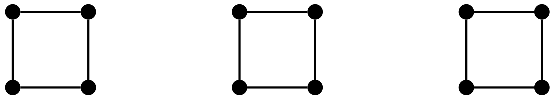


Figure 5 Under a graph metric, it is impossible to do better than 2-fairness when choosing four centers from among this set of points.

Proof. Consider the example in Figure 5, with 12 points under a graph metric with $k = 4$ and $P = X$. For every point $i \in P$, we have $NR(i) = 1$. But for any choice of three centers, some square will have at most one center, and one point in that square will therefore have travel distance at least 2. Thus, no centers algorithm in this metric space can be α-fair for any $\alpha < 2$. ◀

▶ **Proposition 7.** *For all $m \geq 2$ and all $\alpha < \sqrt{2}$, there is no centers algorithm that α-fair in m-dimensional Euclidean space.*

Proof. This holds by essentially the same example used to prove Proposition 6: P consists of 12 points arranged in three squares of unit side, where the distance between the squares is greater than the diameter of the squares, and $k = 4$. Once again, every point has neighborhood radius 1, and for any solution S, some square will have at most one center. Some other point i in that square will have travel distance $d(i, S) \geq \sqrt{2}$, and it follows immediately that $\alpha(S) \geq \sqrt{2}$. ◀

Bias In, Bias Out? Evaluating the Folk Wisdom

Ashesh Rambachan
Department of Economics, Harvard University, Cambridge, MA, USA
asheshr@g.harvard.edu

Jonathan Roth
Microsoft Research, Redmond, WA, USA
Jonathan.Roth@microsoft.com

Abstract

We evaluate the folk wisdom that algorithmic decision rules trained on data produced by biased human decision-makers necessarily reflect this bias. We consider a setting where training labels are only generated if a biased decision-maker takes a particular action, and so "biased" training data arise due to discriminatory selection into the training data. In our baseline model, the more biased the decision-maker is against a group, the more the algorithmic decision rule favors that group. We refer to this phenomenon as *bias reversal*. We then clarify the conditions that give rise to bias reversal. Whether a prediction algorithm reverses or inherits bias depends critically on how the decision-maker affects the training data as well as the label used in training. We illustrate our main theoretical results in a simulation study applied to the New York City Stop, Question and Frisk dataset.

2012 ACM Subject Classification Applied computing → Economics; Computing methodologies → Machine learning

Keywords and phrases fairness, selective labels, discrimination, training data

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.6

Funding *Ashesh Rambachan*: Supported by the NSF Graduate Research Fellowship under Grant DGE1745303.
Jonathan Roth: Supported by the NSF Graduate Research Fellowship under Grant DGE1144152.

Acknowledgements We are grateful to Isaiah Andrews, Talia Gillis, Ed Glaeser, Nir Hak, Nathan Hendren, Larry Katz, Jens Ludwig, Sendhil Mullainathan, Paul Novosad, Aaron Roth, Ben Roth, Hannah Shaffer, Jann Spiess, and seminar participants at Harvard, the Federal Reserve Bank of Philadelphia, and Quantco for valuable comments and feedback.

1 Introduction

Algorithms have the promise to improve upon human decision-making in a variety of settings, but concerns abound that algorithms may produce decision rules that are biased against particular groups. A particular fear is that if the training data is generated by human decision-makers that discriminate against a particular group, then the algorithm will reflect this bias. This concern is captured by the common refrain "bias in, bias out" [4, 28].

In this paper, we evaluate the folk wisdom that algorithms trained on data produced by biased human decision-makers will necessarily inherit bias. Through the lens of a classic model of discrimination in economics, we consider the case where "biased" training data arise due to discriminatory selection into the training data and illustrate that algorithms trained over such biased training data do not necessarily inherit bias. In fact, for a common class of prediction exercises, we show that the opposite is true: The more biased the decision-maker is against a group in the training data, the more favorable the algorithm is toward that group. We refer to this phenomenon as *bias reversal*. We clarify the conditions that give rise to bias reversal and discuss how alternative biases in the training data affect resulting algorithms.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 6; pp. 6:1–6:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

We consider a baseline model with three elements that together produce bias reversal. First, we consider a setting in which labels in the training data are created only if a decision-maker chooses to take a particular action. This is commonly known as the *selective labels problem* [24, 19]. For instance, we may only obtain data on whether a pedestrian is carrying contraband if a police officer chooses to search the pedestrian. Likewise, a college may only obtain data on a student's academic performance in college if an admissions officer chooses to accept the student, a bank may only obtain data on a borrower's creditworthiness if a loan officer chooses to grant the borrower a loan, and a firm may only obtain data on a job applicant's productivity if that applicant is hired. Second, we follow a classic literature in the economics of discrimination and assume that the decision-maker is a *taste-based discriminator* against the disadvantaged group [5, 1, 23, 2, 3]. This means that the decision-maker acts as if they receive a different payoff (or face a different cost) for taking the action of interest against a particular group. This may arise due to preferences, costs, or misperceptions. As a result, bias in our model manifests itself through selection into the training data. Finally, we assume that the decision-maker has access to *unobservables*, which are features that are informative about the label of interest but are unavailable in the observed training data. Each of these three elements – selective labels, taste-based discrimination and unobservables – are critical to bias reversal.

In this baseline model, we then show that the *more biased* the decision-maker is against the disadvantaged group, the *more favorable* the resulting algorithmic decision rule is toward the disadvantaged group. For example, in settings where police officers are biased in their decision to search pedestrians for contraband, an algorithmic decision rule trained to predict whether a pedestrian is carrying contraband using previously conducted searches would search fewer African American pedestrians than if police officers were unbiased in their search decisions. Similarly, in settings where managers are biased against African-Americans in hiring decisions, an algorithmic decision rule trained to predict employee performance using data on previously hired employees would hire more African American applicants than if the managers were unbiased in their hiring decisions.

To illustrate the intuition for this result, consider the example of police searches. Suppose that police assess the probability that an individual is carrying contraband, and search people with high assessed probabilities. Police base their search decision on a number of factors that are recorded in the data (the time of stop, location, demographics of the individual), as well as subjective information that is not recorded in the data (their evaluation of the individual's behavior). Because police choose to search individuals with risky behavior that is unobservable to the data scientist, an algorithm trained to predict whether contraband was found using a sample of conducted searches will tend to make predictions that are too high for the general population. However, this selection issue will be mitigated for African Americans if police officers are racially biased. Indeed, in the extreme case where police officers are so biased that they search *all* African Americans, regardless of underlying risk, then there will be no selection on unobservable behavior for African Americans in the training data. Thus, the more biased are police officers, the more favorable is the training data for African Americans, and hence the more the algorithm learns to favor African Americans.

We emphasize that our results do not imply that biased data can never produce biased algorithms. Rather, our results highlight that whether an algorithm does or does not inherit bias depends crucially on the form of the bias and the training of the algorithm. To illustrate this, we consider modifications to our baseline model that can produce effects in line with the usual "bias in, bias out" intuition. First, bias reversal crucially depends on the fact that the algorithm is trained to predict the outcome of interest (carrying contraband in the policing example) in the sample where the outcome is available. The typical "bias in, bias out" result can be obtained if either i) the algorithm is instead trained to predict the human decision, or ii) the outcome of interest is assumed to be zero for those not selected by the human decision-maker. Second, while we assume that selection into the training data is determined by a biased decision-making process, we assume that the label of interest is measured without bias. This rules out "label bias," an additional source of bias in training data mentioned in the literature on algorithmic fairness – see [9] for a discussion.[1]

This paper relates to several recent works that study fairness and discrimination across computer science and the social sciences. First, several papers consider properties of algorithms that are trained on selectively-labelled data. [19] and [24] define the selective labels problem and discuss its implications for evaluating the predictive performance of algorithms. [17] studies how the selective labels problem impacts fairness-adjusted predictors. [13] illustrates that the selective labels problem cannot be addressed via standard sample selection procedures and propose new methodology to deal with it. [11] shows that when there are selective labels, an algorithm can improve upon human decisionmaking if the human decisions are sufficiently noisy. [27] proposes a causal modeling approach to estimating fair prediction functions in the presence of unobserved features. [18] studies the related problem of how a fairness-minded decision-maker (e.g. college admissions officer) should select a screening rule if the selected data from that screening decision are used downstream by a Bayesian decision-maker (e.g. employer). Our work is also related to a series of legal papers that have argued that automating decisions will magnify discrimination due to historical biases in existing training data – see [4], [7], [28]. In contrast, our results suggest that for certain prediction exercises, historical biases in training data can produce automated decision rules that may reverse discrimination. Conversely, our results also imply that if an algorithm is trained on data that is produced by a decision-maker that exhibits explicit affirmative action towards a group, the algorithm could, in fact, inherit bias.

Our analysis abstracts away from several potentially important considerations that could be considered in future work. First, we assume that the outcome Y itself is measured without bias. This is often a significant concern in many empirical settings of interest. Second, our main theoretical results focus on properties of the optimal, population prediction function under squared loss – i.e., the conditional expectation of the outcome given the features – and abstracts away from finite-sample considerations. Although our simulation evidence indicates that our results still hold in finite-sample, this deserves further attention. Extending these results to more general loss functions may also be of interest. Third, our results focus on an algorithmic decision rule that is trained "naively" by the data-scientist, meaning that they do not adjust for selection into the data nor impose any additional fairness criteria. Finally, we focus attention on a taste-based model for discrimination. Other models of discriminating behavior may yield different conclusions. For example, discrimination may arise due to stereotypes (e.g. [6]) or differential noise in the decision-maker's predictions across groups (e.g. [25]).

The remainder of this paper is structured as follows. Section 2 presents our baseline model. Section 3 states our main results and Section 4 discusses extensions. Section 5 illustrates our results in simulations based on New York Stop, Question and Frisk data. We place all proofs in the Appendix.

[1] In the policing example, label bias would arise if police officers discriminated against African-Americans by fabricating evidence against them or ignoring evidence against whites.

2 A Model of Biased Decisions

In this section, we develop a model wherein the training data given to a predictive algorithm is generated by a biased human decision-making process. For the sake of exposition, we discuss the model in the context of police bias in pedestrian searches and refer to the decision-maker as the police throughout. This will more clearly connect our theoretical results with our empirical application to New York Stop, Question and Frisk. However, this model is broadly applicable to other settings with selective labels such as college admissions, loan decisions, and hiring decisions, among many others. We discuss the connection to these other settings in Section 4.1.

Police officers wish to search individuals that have a high probability of carrying contraband. Following [5] and a large literature in economics, police officers are taste-based discriminators against African Americans.[2] Based on the search decisions of police officers, data are then revealed to the data scientist. If a police officer searches an individual, the data scientist observes the result of that search (was the individual carrying contraband?), some characteristics of the individual and the stop (age, gender, location of stop, time of stop, etc.) as well the race of the individual. The data scientist then uses this training data to construct an algorithm to predict which individuals are most likely to be carrying contraband. We focus on analyzing properties of the predictive algorithm produced by the data scientist.

2.1 The population

Individuals in the population are characterized by the random vector (X, U, R, Y). Let $X \in \mathcal{X}$ denote some characteristics about the individual that are typically recorded after a police search such as age, gender, location of stop, time of stop, etc. Let $U \in \mathcal{U}$ denote characteristics of an individual that are observed by a police officer prior to a search but are typically not recorded. For example, this may consist of the police officer's evaluation of the individual's behavior prior to the stop or the individual's behavior during the stop. Importantly, U is observed by the police officer but is unobserved to the data scientist. Finally, $R \in \{0, 1\}$ denotes the race of the individual with $R = 1$ for African Americans, and $Y \in \{0, 1\}$ denotes whether the individual is carrying contraband. The population is described by the joint distribution $\mathbb{P}$ of the random vector (X, U, R, Y).

2.2 Police decisions

Police officers observe the characteristics (X, U, R) of each individual and decide whether to search that individual. Police officers receives a positive payoff $b > 0$ if they find contraband after searching an individual and without loss of generality, we normalize this payoff to one, $b = 1$. Police officers receive a payoff of zero if the individual is not searched and incur a cost $c > 0$ for every search.

In addition, police officers are taste-based discriminators against African Americans and receive an additional payoff $\tau > 0$ from searching African Americans. The parameter τ parametrizes the degree to which the police are biased against African Americans. The larger the magnitude of τ, the more biased the police are against African Americans. In sum, the police's payoffs from conducting a search are $Y + \tau R - c$. In order to maximize their expected payoff, the police decide whether to search according to a threshold rule:

$$S^*(X, U, R) = 1\left(\mathbb{E}[Y|X, U, R] \geq c - \tau \cdot R\right).$$

[2] Unlike [2] and [23], we do not assume that the individual's decision to carry contraband responds to police search decisions. As a result, we do not introduce an equilibrium concept such as Nash equilibrium.

The bias of the police implies that a lower threshold for search is applied to African Americans. In this sense, the police are biased against African Americans. We assume for now that the police make their decisions based on an optimal prediction of Y given (X, U, R) under squared loss (i.e., $\mathbb{E}[Y|X, U, R]$), although in Section 4.2 we show that our main results extend to the case where the police use noisy estimates of the conditional expectation under certain regularity conditions.

2.3 The prediction problem

The data scientist then observes data consisting of individuals that are stopped by the police. There are "selective labels" – the data scientist only observes whether an individual was carrying contraband (Y) if the police searched the individual $(S^* = 1)$. The data-scientist thus observes the pair (Y, X, R, S^*) for those with $S^* = 1$. In some of our results, we will also consider what happens if the data scientist is able to observe (X, R, S^*) but not Y for those who are not searched by the police. Let $\hat{\mathbb{P}}_\tau$ denote the joint distribution of the data that is revealed to the data scientist. We index the probability distribution of the observed data by the police's discrimination parameter τ as our results focus on comparative statics over τ.

Using the observed data, the data scientist constructs a predictive algorithm of whether an individual is carrying contraband Y using the observed features (X, R). In our baseline model, we suppose that the data scientist trains the algorithm using only the data where the outcome is available $(S^* = 1)$. We abstract from the estimation problem and consider properties of the optimal predictor under squared loss, $\mathbb{E}_{\hat{P}_\tau}[Y|X, R, S^* = 1]$ (i.e., the conditional expectation over the distribution of observed data). For now, we suppose that race is included as a feature; we discuss relaxing this assumption in Section 4.3.

3 Baseline results

3.1 Bias reversal

We now present our bias reversal result, which examines how an algorithm trained to predict Y in the searched sample $(S^* = 1)$ can reverse bias. We first sketch the intuition and then formally state the result.

Since the police incorporate the unobservable U into their search decision, the training data of conducted searches will tend to be composed of individuals that have values of U associated with higher probability of $Y = 1$. As a result, the predictive algorithm trained on the selected training data will tend to over-predict the label Y for the whole population. However, as the police officers become more biased, this selection problem becomes less severe for African Americans. Intuitively, the more biased are the police officers against African Americans, the more likely they are to search any given African American, and so there is less selection on the unobservable U. In the extreme case where $\tau \geq c$, police officers search all African Americans, and there is no selection on the unobservable U for African Americans. The predictive algorithm thus becomes more favorable to African Americans as the police officers become more biased.

▶ **Theorem 1.** $\mathbb{E}_{\hat{P}_\tau}[Y|X = x, R = 1, S^* = 1]$ *is weakly decreasing in* τ *for all* $x \in \mathcal{X}$ *and* τ *such that* $\hat{P}_\tau(S^* = 1 \,|\, X = x, R = 1) > 0$. *Likewise,* $\mathbb{E}_{\hat{P}_\tau}[Y|X = x, R = 0, S^* = 1]$ *is constant in* τ *for all* $x \in \mathcal{X}$ *and* τ *such that* $\hat{P}_\tau(S^* = 1 \,|\, X = x, R = 0) > 0$.

Theorem 1 shows that as the police become more biased against African Americans, the predictions of the algorithm trained on the selected data become more favorable to African Americans, in the sense that African Americans are predicted to have lower risk of carrying contraband. This implies the more biased the police are against African Americans, the fewer African Americans will be searched by an automated search rule that uses these predictions.

► **Corollary 2.** *Consider the automated search rule:*

$$S_\tau^{automated}(x,r) = 1\left(\mathbb{E}_{\hat{\mathbb{P}}_\tau}[Y|X=x, R=r, S^*=1] \geq c_{min}\right)$$

for some $c_{min} \in [0,1]$.[3] *Then* $S_\tau^{automated}(x,1) \leq S_{\tau'}^{automated}(x,1)$ *for any* $\tau' < \tau$*, so any African-American searched under* τ *is also searched under* τ'*, whereas* $S_\tau^{automated}(x,0)$ *does not depend on* τ*. It follows that the fraction of African Americans searched under* $S_\tau^{automated}$ *(i.e.,* $\mathbb{E}[S_\tau^{automated}(X,R) \,|\, R=1]$*) is decreasing in* τ*, whereas the fraction of whites searched under* $S_\tau^{automated}$ *is constant in* τ*.*

Corollary 2 states that an automated search rule based on a threshold rule using $\mathbb{E}_{\hat{\mathbb{P}}_\tau}[Y|X=x, R=r, S^*=1]$ searches fewer African-Americans the larger is the bias τ in the training data.

These results clarify the manner in which the bias of police officers influences the algorithmic treatment of African Americans under an automated search rule. We do not take a stance directly on whether the algorithm's treatment of African Americans for any given τ is "fair" in a formal sense.[4] However, any sensible notion of fairness would suggest that if a given decision rule is unfair to African Americans, then any decision rule that is "harsher" to African Americans (i.e. more likely to search any given African American) and treats whites the same is at least as unfair. Therefore, Theorem 1 and Corollary 2 suggest that if a decision rule based upon a prediction function trained on data produced by police officers that discriminate against African Americans ($\tau > 0$) is unfair, then a decision rule which is based upon a prediction function trained on data produced by police officers that are unbiased against African Americans ($\tau = 0$) would be even *more* unfair to African Americans.

We refer to the phenomenon in which the more biased is the human-decisionmaker, the more favorable is the algorithmic decision rule to the minority group as "*bias reversal.*" Although presented in the context of police searches, we show this phenomenon extends to other settings such as loan applications, hiring decisions, and college admissions in Section 4.1.

3.2 Bias inheritance for alternative prediction exercises

In Theorem 1 and Corollary 2, we assumed that the data scientist constructs an algorithm to predict the observed label Y using the training data for the searched sample ($S^* = 1$). We now consider what happens if a different label and sample is used. First, the data scientist may instead predict the human decision S^* itself over the full population. This is a common type of prediction problem in some contexts. For example, a series of papers note that using the human decision as the label is common in training algorithms to automate hiring decisions [10, 11, 30]. For this prediction exercise, bias reversal no longer holds. Instead, the prediction function inherits bias – as the police become more biased against African-Americans, the predictions of the algorithm trained in this way become less favorable to African-Americans.

[3] We implicitly assume that $P(S^* = 1 \,|\, X = x, R = r) > 0$ for almost every (x, r), so that the search rule is well-defined.

[4] Results in [22] highlight that an algorithm cannot simultaneously satisfy several common definitions of fairness if the base rates of risk differ across groups.

▶ **Theorem 3.** $\mathbb{E}\left[S^*|X=x,R=1\right]$ *is weakly increasing in* τ *for all* $x \in \mathcal{X}$*.*

A second alternative prediction exercise that the data scientist may consider is to predict the compound outcome that the individual was searched by the police and that the individual was carrying contraband - that is, predict the label $Y \cdot S^*$ over the full sample. Put otherwise, the data scientist imputes the missing label Y to be zero if $S^* = 0$. This type of prediction exercise is common in certain medical applications (e.g., see [29]). In this case, we again find that the prediction function inherits bias from the police officers' discriminatory search decisions.

▶ **Theorem 4.** $\mathbb{E}\left[YS^*|X=x,R=1\right]$ *is weakly increasing in* τ *for all* $x \in \mathcal{X}$*.*

Theorem 3 and Theorem 4 immediately imply that an automated decision rule that is based upon predictions of S^* or YS^* will inherit bias – that is, search more African Americans as police officers become more biased.

▶ **Corollary 5.** *Consider the automated search rule* $\check{S}_{\tau}^{automated}(x,r) = 1\left(\hat{Y}(x,r) \geq c_{min}\right)$ *for some* $c_{min} \in [0,1]$*, where* $\hat{Y}(x,r) = \mathbb{E}\left[S^* \mid X=x, R=r\right]$ *or* $\hat{Y}(x,r) = \mathbb{E}\left[YS^* \mid X=x, R=r\right]$*. Then,* $\check{S}_{\tau}^{automated}(x,1) \leq \check{S}_{\tau'}^{automated}(x,1)$ *for* $\tau < \tau'$*, so any African American that is searched under* τ *is also searched under* τ'*. It follows that the fraction of African Americans searched under* $\check{S}_{\tau}^{automated}$ *(i.e.* $\mathbb{E}\left[\check{S}_{\tau}^{automated}(X,R) \mid R=1\right]$*) is increasing in* τ*.*

The key distinction between these alternative prediction exercises and our earlier result is that bias now drives a wedge between the true outcome of interest and the label that the algorithm is trained on (S^* or $Y \cdot S^*$), but the human bias does not affect sample composition. By contrast, in the original setting that predicts Y over the selected sample with $S^* = 1$, the bias affects the prediction exercise only through sample composition. This is a crucial yet subtle difference. Taken together, these results show that the choices of label (Y vs. S^* vs. $Y \cdot S^*$) and training sample ($S^* = 1$ vs. full sample) play a key role in determining whether human biases propagate into algorithmic predictions and automated decisions, formalizing an argument made heuristically in [21]. Table 1 summarizes our results across the three prediction exercises considered.

Table 1 Summary of prediction exercises.

Outcome	Training sample	Comparative static
Y	$S^* = 1$	Bias reversal
S^*	Full sample	Bias inheritance
$Y \cdot S^*$	Full sample	Bias inheritance

4 Extensions

4.1 When discrimination yields fewer labels for the disadvantaged group

In other settings of interest with selective labels such as loan applications, hiring decisions, and college admissions, *fewer* labels are generated when the decision-maker is biased. For example, if a hiring manager is biased against African Americans, fewer African American applicants are hired. A simple extension shows that an analogous comparative static still holds: the more biased the decision-maker is against a group, the more the resulting algorithmic decision rule favors that group.

As an example, consider a hiring manager that predicts the productivity Y of job applicants using features X that are observable to the data scientist and features U that are unobservable to the data scientist. A biased hiring manager applies a higher predicted-productivity threshold for African Americans than for whites. This means the more biased is the hiring manager, the fewer African-Americans will enter the training data. However, the African-Americans who do enter the training data will be more positively selected on U (i.e., on unobservables positively correlated with productivity). Thus, the more biased is the hiring manager against African-Americans, the higher will be the algorithm's predicted productivity for African-Americans and the more African Americans will be hired by an algorithmic hiring rule.

Formally, consider a modified selection rule $\tilde{S}(X, U, R) = 1\left(\mathbb{E}[Y|X, U, R] \geq c + \tilde{\tau}R\right)$, with $\tilde{\tau} > 0$. Define $\hat{P}_{\tilde{\tau}}$ to be the joint distribution of the data revealed to the data scientist under the selection rule $\tilde{S}(X, U, R)$. This model is equivalent to the model considered earlier with $\tau = -\tilde{\tau}$. We thus obtain the immediate corollary to Theorem 1.

▶ **Corollary 6.** $E_{\hat{P}_{\tilde{\tau}}}[Y|X = x, R = 1, \tilde{S} = 1]$ *is weakly increasing in* $\tilde{\tau}$ *for all* x,$\tilde{\tau}$ *such that* $\hat{P}_{\tilde{\tau}}(\tilde{S} = 1|X = x, R = 1) > 0$*, while* $E_{\hat{P}_{\tilde{\tau}}}[Y|X = x, R = 0, \tilde{S} = 0]$ *is constant in* $\tilde{\tau}$*, for all* $x, \tilde{\tau}$ *such that* $\hat{P}_{\tilde{\tau}}(\tilde{S} = 1|X = x, R = 0) > 0$.

Moreover, consider the automated hiring rule:[5]

$$\tilde{S}^{automated}_{\tilde{\tau}}(x, r) = 1\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\tau}}}[Y|X = x, R = r, \tilde{S} = 1] \geq c_{min}\right)$$

for $c_{min} \in [0, 1]$*. Then,* $\tilde{S}^{automated}_{\tilde{\tau}}(x, 1) \leq \tilde{S}^{automated}_{\tilde{\tau}'}(x, 1)$ *for any* $\tilde{\tau} < \tilde{\tau}'$*. It follows that the fraction of African Americans hired under* $\tilde{S}^{automated}_{\tilde{\tau}}$ *(i.e,* $E[\tilde{S}^{automated}_{\tilde{\tau}}(X, R) \,|\, R = 1]$*) is increasing in* $\tilde{\tau}$*.*

In unpacking this result, it is useful to distinguish between statistical bias and "favorability" of the algorithm. As the decision-maker becomes more biased, the predictions of the algorithm become more biased in a statistical sense, meaning that the magnitude of $E[Y \mid X, R = 1, \tilde{S} = 1] - E[Y \mid X, R = 1]$ becomes larger. However, this statistical bias works in a way that makes algorithmic decision rules more likely to select members of the discriminated against group.

These results highlight that the phenomenon of bias reversal is not dependent on the selective labels problem leading to more labels to be collected for the disadvantaged group, and is thus applicable to a range of settings with selective labels.

4.2 Noisy decision-making

We next show that the results in Section 3 are robust to allowing for random noise in the officers' decisions. In the baseline model, we assumed that the police officers are able to correctly combine the available information to construct accurate predictions about risk, $\mathbb{E}[Y|X, R, U]$ and thereby rank order individuals correctly. Extensive work in the social sciences suggest that this does not hold in many applications of interest. For example, [19] suggest that even experienced judges are unable to accurately predict recidivism in bail decisions. We now show that the comparative static in Theorem 1 still holds if police officers have independent random noise in their risk assessments.

[5] We implicitly assume that $P(\tilde{S} = 1 \,|\, X = x, R = r) > 0$ for almost every (x, r), so that the hiring rule is well-defined.

▶ **Proposition 7.** *Suppose police search according to*

$$S^{noise}(X, U, R, \epsilon) = 1\Big(\mathbb{E}[Y | X, U, R] + \epsilon \geq c - \tau \cdot R\Big),$$

for a random prediction error ϵ, where the distribution $\epsilon \,|\, X, R$ has strictly increasing hazard and $\epsilon \perp\!\!\!\perp (Y, U) \,|\, (X, R)$. Then, $\mathbb{E}_{\hat{P}_\tau}[Y | X = x, R = 1, S^{noise} = 1]$ is weakly decreasing in τ for all $x \in \mathcal{X}$ and τ such that $\hat{P}_\tau(S^{noise} = 1 \,|\, X = x, R = r) > 0$.

Similarly, the comparative statics derived for the alternative prediction exercises are also robust to noisy decision-making.

▶ **Proposition 8.** *The conclusions of Theorem 3 and Theorem 4 hold replacing S^* with S^{noise}.*

4.3 Excluding group membership from the predictive algorithm

We now consider what happens if the data scientist is forbidden from using group membership in the predictive algorithm. For example, it may be illegal for a predictive algorithm to explicitly use race as a feature [21, 15]. In this case, the prediction function in the baseline model now takes the form $\mathbb{E}_{\hat{\mathbb{P}}_\tau}[Y | X, S^* = 1]$.

Whether the comparative static in bias still holds now depends on whether group membership R is "reconstructable" from the observed features X. That is, it depends on whether group membership is predictable from the observed features. If group membership is perfectly reconstructable, then these results trivially hold for a prediction function that does not use group membership as $\mathbb{E}_{\hat{\mathbb{P}}_\tau}[Y | X, S^* = 1] = \mathbb{E}_{\hat{\mathbb{P}}_\tau}[Y | X, R, S^* = 1]$.

If group membership is not perfectly reconstructable, then one can construct examples in which the gap in average predictions across groups for a group-blind algorithm moves in the opposite direction as the gap in average predictions across groups for an algorithm that includes race. The direction of the effect will depend on whether the marginally searched individual in the $R = 1$ group is more "similar" to the average person with $R = 0$ or $R = 1$. As a simple example, suppose there is only one observed, binary feature X. Suppose that among whites, $X = 1$ with probability $1 - \epsilon$ for some small $\epsilon > 0$. Among African Americans, $X = 0$ with probability $1 - \epsilon$. Then, if the marginally searched African American has feature $X = 1$, then an increase in the bias of police officers will have a larger effect on the average prediction for whites than African Americans, as there are relatively more whites among the group with $X = 1$ in the observed data. Conversely, if the marginally searched African American has feature $X = 0$, then it will have a larger effect on the average prediction for African Americans than whites. The same intuition holds for the alternative prediction exercises that we considered earlier.

The reconstruction problem has been discussed at length elsewhere – see, among many others, [20, 26, 8, 12]. Typically, it is thought that if race is reconstructable from other features, then algorithms will exhibit bias or discriminate against minority groups. Our results illustrate that this is not true generally. If group membership is reconstructable, then an algorithm that is blind to group membership may exhibit bias reversal (Theorem 1), in line with results in [14, 20, 15].

5 Application: New York City Stop, Question and Frisk

We now apply these results to the New York Stop, Question and Frisk (SQF) data. We synthetically create a training data set that is produced by biased search decisions and illustrate the key comparative statics described in Section 3.

5.1 Data description

SQF was a program in New York City that allowed the police to temporarily stop, question, and search individuals on the street. We use publicly available, stop-level data that contains information on all stops conducted as part of the SQF program from 2008-2013, totalling over 4 million stops of pedestrians and over 350,000 searches [16]. For each recorded stop, we observe whether the stopped individual was searched for contraband and if so, an indicator for whether contraband was found. The data also contains several detailed characteristics of the stopped individual and the circumstances of the stop. The features in the data include the stopped individual's age, gender, and build, and the time and location of the stop, which we treat as the observable features X. Importantly, we also observe the race R of the stopped individual. We restrict attention to stops of non-Hispanic whites and African Americans. The data also include the officer's stated reason for conducting the stop, e.g. "carrying a suspicious object" or "displayed behavior indicative of a drug transaction." We treat these responses as the unobservable features U that are available to the officer at the time of the search decision but are unavailable to the data scientist. This is analogous to "soft information" about the individual that may be available to the officer at the time of the stop but may be unavailable in certain data sets.

5.2 Simulation design

We conduct a simulation exercise that trains an algorithm to predict whether a stopped individual is carrying contraband on synthetic training datasets that are generated from the original SQF data. Across synthetic training datasets, we vary the degree of bias against African Americans in search decisions by selectively "undoing" observed searches. We then examine how changing the degree of bias against African Americans affects the resulting algorithm's predictions.

More concretely, we first subset the data to only include stops in which searches were conducted ($S^* = 1$). We then randomly split the searched SQF stops into two partitions. In the first partition, we construct a predictor for carrying contraband among stops with searches. The predictor estimates $\mathbb{E}[Y|X, R, U, S^* = 1]$, where X is a feature vector that includes demographic information about the stopped individual such as age, gender and build as well as the location and time of the stop, and U is the officer's stated reason for the stop. We construct the predictor using logistic regression, matching the approach of previous research using this data [16, 17]. In the held-out partition, we then use the estimated prediction function to construct a synthetic search flag $\hat{S}$. For individuals with $\hat{Y} = \hat{\mathbb{E}}[Y|X, R, U, S^* = 1] \leq c_R$, we set $\hat{S} = 0$ and treat them as if they had not been searched. For individuals with $\hat{Y} > c_R$ for $R \in \{0, 1\}$, we set $\hat{S} = 1$. This produces a synthetic dataset at the search thresholds (c_0, c_1) in which we observe $(Y, X, R, \hat{S})$ for each observation. Finally, we re-estimate the prediction function over the synthetically searched observations. We estimate the functions $\mathbb{E}[Y|X, R, \hat{S} = 1]$, $\mathbb{E}[\hat{S}|X, Y]$ and $\mathbb{E}[Y\hat{S}|X, R]$ using logistic regression and examine properties of the estimated prediction functions. We repeat this simulation for a variety of different thresholds c_0, c_1 to construct a series of synthetically searched observations at different levels of bias, defined as $\tau = c_0 - c_1$, against African Americans. We vary c_0, c_1 so that 50 percent of the synthetic dataset is always searched and only the composition of searches between African Americans and whites vary. We vary the fraction of searches that are conducted on African Americans from 80 to 95%.

5.3 Simulation results

Figure 1 plots the results from our simulation exercise. The X-axis plots the discrimination parameter $\tau = c_0 - c_1$ across synthetic datasets. Larger values of τ correspond with a search rule that is more biased against African Americans. The Y-axis plots the fraction of African Americans that fall in the top 50 percent of predicted risk using the prediction function estimated over the synthetic dataset. The predictions from our results in Section 3 hold sharply. First, as the police become more biased against African Americans, the prediction function $\hat{\mathbb{E}}[Y|X, R, \hat{S} = 1]$ becomes more favorable to them. In particular, fewer African Americans fall in the top half of predicted risk as τ increases. This illustrates our result of bias reversal in a concrete application of interest. Second, as the police become more biased against African Americans, the prediction functions $\hat{\mathbb{E}}[\hat{S}|X, R]$ and $\hat{\mathbb{E}}[Y\hat{S}|X, R]$ become less favorable to African Americans. As τ increases, more African Americans fall in the top half of predicted risk. For these prediction functions, "bias in" implies "bias out."

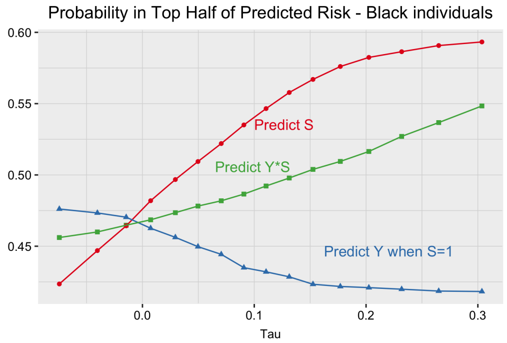


Figure 1 NYC SQF Simulation Results.

6 Conclusion

In this paper, we evaluated the folk wisdom that algorithmic decision rules trained on data that are produced by biased human decision-makers will necessarily inherit this bias. We showed that in an important class of prediction exercises, the opposite holds: The more biased the decision-maker towards a group, the more favorable is the algorithm towards that group. We refer to this phenomenon as "*bias reversal*." We then showed that an important determinant of whether one obtains bias reversal or "bias in, bias out" is whether the human bias affects sample selection or the measured label. When we consider whether algorithms will inherit human biases, it is therefore important to think carefully about the form of the human bias, how it affects the training sample, as well as how the labels and features are selected for the predictive algorithm.

References

1 Joseph Altonji and Rebecca Blank. Race and gender in the labor market. In Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics, Volume 3C*, pages 3143–3259. North Holland, 1999.

2 Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, 96(1), 2006.

3 David Arnold, Will Dobbie, and Crystal Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

4 Solon Barocas and Andrew Selbst. Big data's disparate impact. *California Law Review*, 104, 2016.

5 Gary Becker. *The Economics of Discrimination.* University of Chicago Press, 1957.

6 Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016.

7 Anupam Chander. The racist algorithm. *Michigan Law Review*, (6):1023–1046, 2017.

8 Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency '19*, pages 339–348, 2019.

9 Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. Technical report, Stanford University Working Paper, 2018.

10 Bo Cowgill. Bias and productivity in humans and machines: Theory and evidence. Technical report, Columbia business School Working Paper, 2018.

11 Bo Cowgill. Bias and productivity in humans and machines. Technical report, Columbia Business School Working Paper, 2019.

12 Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. Technical report, arXiv preprint, 2017. `arXiv:1707.08120`.

13 Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. Technical report, arXiv preprint, 2018. `arXiv:1807.00905`.

14 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

15 Talia Gillis and Jann Spiess. Big data and discrimination. *The University of Chicago Law Review*, 86:459–487, 2019.

16 Sharad Goel, Justin Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1), 2016.

17 Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2444–2453. PMLR, 2018.

18 Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. Technical report, arXiv preprint, 2018. `arXiv:1808.09004`.

19 Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 2018.

20 Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, 2018.

21 Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 80:1–62, 2018.

22 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. Technical report, arXiv preprint, 2016. `arXiv:1609.05807`.

23 John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *The Journal of Political Economy*, 109(1), 2001.

24 Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

25 Danielle Li. Expertise vs. bias in evaluation: Evidence from the nih. *American Economic Journal: Applied Economics*, 9(2), 2017.

26 Zachary Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8125–8135, 2018.

27 David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 349–358, 2019.

28 Sandra Mayson. Bias in, bias out. *The Yale Law Journal*, 128(8):2122–2473, 2018.

29 Sendhil Mullainathan and Ziad Obermeyer. Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–80, 2017.

30 Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic employment screening: Evaluating claims and practices. Technical report, arXiv preprint, 2019. `arXiv:1906.09208`.

A Proofs

Proof of Theorem 1

Proof. Define $\mu_{X,R,U} := \mathbb{E}[Y|X,R,U]$ and $\mu_{X,R} := \mathbb{E}[Y|X,R]$. Let $U^* = \mu_{X,R,U} - \mu_{X,R}$, so that $\mu_{X,R,U} = \mu_{X,R} + U^*$. Note that $S^* = 1$ if and only if $U^* \geq T(X,R,\tau)$ for the threshold $T(X,R,\tau) = (c - \tau \cdot R) - \mu_{X,R}$. Applying the law of iterated expectations,

$$\begin{aligned}\mathbb{E}\left[Y|X=x,R=r,S^*=1\right] &= \mathbb{E}\left[Y|X=x,R=r,U^* \geq T(x,r,\tau)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[Y|X=x,R=r,U\right]|X=x,R=r,U^* \geq T(x,r,\tau)\right] \\ &= \mu_{x,r} + \mathbb{E}\left[U^*|X=x,R=r,U^* \geq T(x,r,\tau)\right].\end{aligned}$$

Note that for $r = 1$, $T(x,r,\tau)$ is decreasing in τ. It follows immediately that $E[U^*|X = x, R = r, U^* \geq T(x,r,\tau)]$ is weakly decreasing in τ, which gives the first desired result. Likewise, when $r = 0$, $T(x,r,\tau)$ does not depend on τ, which gives the second result. ◀

Proof of Theorem 3

Proof. By the law of iterated expectations, $\mathbb{E}[S^*|X = x, R = 1] = \mathbb{E}[\mathbb{E}[S^* \mid X = x, U, R = 1]]$. Then,

$$\begin{aligned}\mathbb{E}[S^*|X=x,R=1] &= \int_{u\in\mathcal{U}} \mathbb{E}[S^*|X=x,U=u,R=1]\,dF(u) \\ &= \int_{\{u\in\mathcal{U}:S^*(x,u,1)=1\}} dF(u) \\ &= \int_{\{u\in\mathcal{U}:\mathbb{E}[Y|X=x,U=u,R=1]\geq c-\tau\}} dF(u).\end{aligned}$$

It follows that for $\tau_1 < \tau_2$,

$$\mathbb{E}[S^*|X=x,R=1,\tau=\tau_2] - \mathbb{E}[S^*|X=x,R=1,\tau=\tau_1] = \int_{u\in\mathcal{U}_{12}} dF(u),$$

for $\mathcal{U}_{12} = \{u \in \mathcal{U} : c - \tau_2 \leq \mathbb{E}[Y|X = x, U = u, R = 1] \leq c - \tau_1\}$, which gives the desired result. ◀

Proof of Theorem 4

Proof. The proof of this result is analogous to Theorem 3, replacing S^* with YS^*. ◀

Proof of Proposition 7

The proof of Proposition 7 uses the following lemma.

▶ **Lemma 9.** *Suppose the police search individuals according to*

$$S^{noise}(X, U, R, \epsilon) = 1\Big(\mathbb{E}[Y|X, U, R] + \epsilon \geq c - \tau \cdot R\Big),$$

for a random prediction error ϵ. Suppose that $\epsilon \perp\!\!\!\perp U|X, R$ and the distribution of $\epsilon \mid X, R$ has an increasing hazard, i.e. $\frac{f(\epsilon|X,R)}{1-F(\epsilon|X,R)}$ is increasing in ϵ for $f(\cdot \mid X, R)$ the conditional density function of ϵ. Then, $\mu_{X,R,U}|\{S^{noise} = 1, X, R = 1\}$ has the monotone likelihood ratio property in $-\tau$, where $\mu_{X,R,U} = \mathbb{E}[Y|X, U, R]$ as before.

Proof. The police choose $S^{noise} = 1$ if and only if $\mu_{X,R,U} + \epsilon \geq c - \tau \cdot R$, or equivalently, if and only if $\epsilon \geq c - \tau \cdot R - \mu_{X,R,U}$. Consider $\mu_1 < \mu_2$ in the support of $\mu_{X,R,U}$. Then,

$$\begin{aligned}
&\frac{\mathbb{P}\left(\mu_{X,R,U} = \mu_1 | S^{noise} = 1, X, R\right)}{\mathbb{P}\left(\mu_{X,R,U} = \mu_2 | S^{noise} = 1, X, R\right)} \\
&= \frac{\mathbb{P}\left(S^{noise} = 1 | \mu_{X,R,U} = \mu_1, X, R\right)}{\mathbb{P}\left(S^{noise} = 1 | \mu_{X,R,U} = \mu_2, X, R\right)} \times \frac{\mathbb{P}\left(\mu_{X,R,U} = \mu_1 | X, R\right) / \mathbb{P}\left(S^{noise} = 1 | X, R\right)}{\mathbb{P}\left(\mu_{X,R,U} = \mu_2 | X, R\right) / \mathbb{P}\left(S^{noise} = 1 | X, R\right)} \\
&= \frac{\mathbb{P}\left(\epsilon \geq c - \tau \cdot R - \mu_1 | X, R\right) \cdot \mathbb{P}\left(\mu_{X,R,U} = \mu_1 | X, R\right)}{\mathbb{P}\left(\epsilon \geq c - \tau \cdot R - \mu_2 | X, R\right) \cdot \mathbb{P}\left(\mu_{X,R,U} = \mu_2 | X, R\right)} \\
&= \frac{\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]\right) \cdot \mathbb{P}\left(\mu_{X,R,U} = \mu_1 | X, R\right)}{\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\right) \cdot \mathbb{P}\left(\mu_{X,R,U} = \mu_2 | X, R\right)}
\end{aligned}$$

where the first equality follows from Bayes' Rule, the second equality uses the definition of S^{noise} and the conditional independence of ϵ and U, and the third applies the definition of the CDF. Now, differentiating with respect to $-\tau$:

$$\begin{aligned}
&\frac{\partial}{\partial(-\tau)}\left(\frac{\mathbb{P}\left(\mu_{X,R,U} = \mu_1 | S^{noise} = 1, X, R\right)}{\mathbb{P}\left(\mu_{X,R,U} = \mu_2 | S^{noise} = 1, X, R\right)}\right) = \\
&R \cdot \left(\frac{f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\right)}{\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\right)^2} - \right. \\
&\left. \frac{f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]\right)}{\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\right)^2}\right) \times \frac{\mathbb{P}\left(\mu_{X,R,U} = \mu_1 | X, R\right)}{\mathbb{P}\left(\mu_{X,R,U} = \mu_2 | X, R\right)},
\end{aligned}$$

Clearly, this derivative is zero if $R = 0$. If $R = 1$, the derivative is greater than or equal to zero if and only if

$$\begin{aligned}
&f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\right) \\
&- f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]\left(1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]\right) \geq 0
\end{aligned}$$

or equivalently,

$$\frac{f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]}{1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_1\right]} \geq \frac{f_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]}{1 - F_{\epsilon|X,R}\left[c - \tau \cdot R - \mu_2\right]}. \tag{1}$$

However, since $\mu_1 < \mu_2$, we have $c - \tau \cdot R - \mu_1 > c - \tau \cdot R - \mu_2$, and so (1) holds if $\epsilon|X, R$ has increasing hazard. ◀

Proof of Proposition 7. Returning to Proposition 7, we follow an argument that is analogous to the proof of Theorem 1. As before, define $\mu_{X,R,U} := \mathbb{E}[Y|X,R,U]$. Note that $S^{noise} = 1$ if and only if $\mu_{X,R,U} + \epsilon \geq c - \tau \cdot R$. Applying the law of iterated expectations,

$$\begin{aligned}\mathbb{E}\left[Y \mid X = x, R = r, S^{noise} = 1\right] &\overset{(1)}{=} \mathbb{E}\left[\mathbb{E}\left[Y \mid X, R, U, \epsilon\right] \mid X = x, R = r, S^{noise} = 1\right] \\ &\overset{(2)}{=} \mathbb{E}\left[\mu_{X,R,U} \mid X = x, R = r, S^{noise} = 1\right],\end{aligned}$$

where (1) uses the law of iterated expectations and that S^{noise} is simply a function of X, U, R, ϵ and (2) uses $\epsilon \perp\!\!\!\perp Y \,|X, U, R$. The result then follows from Lemma 9. ◀

Proof of Proposition 8

Proof. Analogous to the proofs of Theorem 3 and Theorem 4, replacing expectations over U with expectations over the joint distribution of (U, ϵ). ◀

Individual Fairness in Pipelines

Cynthia Dwork
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
Radcliffe Institute for Advanced Study, Cambridge, MA, USA
Microsoft Research, Mountain View, CA, USA
dwork@seas.harvard.edu

Christina Ilvento
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
cilvento@g.harvard.edu

Meena Jagadeesan
Harvard University, Cambridge, MA, USA
mjagadeesan@college.harvard.edu

Abstract

It is well understood that a system built from individually fair components may not itself be individually fair. In this work, we investigate individual fairness under pipeline composition. Pipelines differ from ordinary sequential or repeated composition in that individuals may drop out at any stage, and classification in subsequent stages may depend on the remaining "cohort" of individuals. As an example, a company might hire a team for a new project and at a later point promote the highest performer on the team. Unlike other repeated classification settings, where the degree of unfairness degrades gracefully over multiple fair steps, the degree of unfairness in pipelines can be arbitrary, even in a pipeline with just two stages.

Guided by a panoply of real-world examples, we provide a rigorous framework for evaluating different types of fairness guarantees for pipelines. We show that naïve auditing is unable to uncover systematic unfairness and that, in order to ensure fairness, some form of dependence must exist between the design of algorithms at different stages in the pipeline. Finally, we provide constructions that permit flexibility at later stages, meaning that there is no need to lock in the entire pipeline at the time that the early stage is constructed.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases algorithmic fairness, fairness under composition, pipelines

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.7

Related Version `https://arxiv.org/abs/2004.05167`

Funding *Cynthia Dwork*: Research supported in part by NSF grant CCF-1763665 and Microsoft Research.
Christina Ilvento: This work was funded by the Sloan Foundation.

1 Introduction

As algorithms reach ever more deeply into our daily lives, there is increasing concern that they be *fair*. The study of the theory of algorithmic fairness was initiated by Dwork et al. [5], who introduced the solution concept of *individual fairness*. Roughly speaking, individual fairness requires that similar individuals receive similar distributions on outcomes. Dwork and Ilvento [6] examined the behavior of individual fairness (and various group notions of fairness) under composition. They showed that although competitive composition, i.e., when two different tasks "compete" for individuals, can result in arbitrarily bad behavior under composition, fairness under simple repeated or sequential classifications (for the same task) degrades gracefully, similar to degradation of differential privacy loss under

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 7; pp. 7:1–7:22
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

multiple computations.[1] In this work we expand the investigation of individual fairness under sequential composition to the case of *cohort pipelines*. Cohort pipelines differ from ordinary sequential composition in that each stage of the pipeline considers only the remaining cohort of individuals and may change its classification strategy conditioned on the set of individuals remaining.

Cohort pipelines are common: many data-driven systems consist of a sequence of cohort selection or filtering steps, followed by decision or scoring steps. A running exemplary scenario in this work will be a two-stage cohort pipeline: a company hires a team (cohort) of individuals to work on a project and subsequently promotes the highest performer on the team to a leadership position. Although the team selection may be fair in the sense that similarly qualified candidates have similar chances of being chosen for the team, the selection of the highest performer critically depends on the *other members of the team.* As we will see, being compared fairly to other members of the cohort in each stage doesn't imply fairness of the entire pipeline, as the competitive landscape can vary between similar individuals.

Indeed, a fair cohort selection mechanism [6] can exploit the “myopic” nature of the promotion stage to skew overall fairness. This can happen either through good intentions (e.g., choosing teams so that members of a minority group always have a mentor on the team) or malice (e.g., ensuring that minority candidates are almost always paired with a more qualified majority candidate): in both these cases minorities suffer significantly reduced chances of promotion.[2] Unlike other repeated classification settings in which the degree of unfairness of multiple fair steps degrades gracefully, the degree of unfairness in cohort pipelines can be arbitrary, even in a pipeline with just two stages. Furthermore, we demonstrate that construction of malicious pipelines under naïve auditing of fairness is straightforward and both computationally and practically feasible.

In this work we examine the subtle issues that arise in cohort-based pipelines, focusing on short pipelines consisting of a single cohort selection step followed by a scoring step. We formalize fairness desiderata capturing the issues unique to pipelines (not shared by ordinary sequential composition), and give constructions for *robust* cohort selection mechanisms that behave well under (i.e., are robust to) pipeline composition with a variety of future scoring policies. In particular, we demonstrate that it is possible to design cohort selection mechanisms that are robust to a rich family of subsequent scoring functions given a simple description of a *policy* governing the behavior of the family.[3] This provides, for example, a means for enabling a company to choose an individually fair hiring procedure that will be robust to many possible compensation functions (all adhering to the policy) chosen at a later date. Guided by a panoply of real-world examples, this work provides a rigorous framework for evaluating and ensuring different types of fairness guarantees for pipelines.

We now summarize our contributions. First, we formalize what it means for the outcomes of a pipeline, which include both the outcome of the initial cohort selection step and the score conditioned on being chosen, to be fair.[4] We then extend this fairness notion to describe how a cohort selection mechanism can be *robust* to a scoring policy, i.e., to compose fairly with any cohort scoring function chosen from a permissible set. Although the choice of scoring function may not depend on the cohort, the scores assigned to any individual may be highly

[1] Note that although fairness degrades gracefully in these scenarios, it does not rule out the existence of feedback loops which arbitrarily amplify unfairness, see e.g., [12, 20].

[2] See Appendix A for additional examples.

[3] Formally, we can think of a policy as a description of a set of permitted scoring functions.

[4] Bower et al. consider fairness in pipelines for a group-based definition of fairness, and primarily consider the accuracy of the final pipeline decision [1].

dependent on their cohort "context." Second, we determine how the scoring policy imposes conditions on the cohort selection mechanism. In particular, we show that there is a natural way to describe the set of cohort *contexts* in which similar individuals are treated similarly by all functions permitted by the policy, and we demonstrate that assigning similar individuals to similar *distributions* over cohort contexts is sufficient (and sometimes necessary) to ensure pipeline robustness. Third, we provide constructions for cohort selection mechanisms which are both robust to a rich set of practical scoring policies and permit flexibility in selection of the original cohort.

2 Model and Definitions

2.1 Preliminaries

We base our model on individual fairness, as proposed in [5]. The intuition behind individual fairness is that "similar individuals should be treated similarly." What constitutes similarity for a particular classification task is provided by a metric which captures society's best understanding of who is similar to whom. Below we formally define individual fairness as in [5] with a natural Lipschitz relaxation.

▶ **Definition 1** (α-Individual Fairness [5]). *Given a universe of individuals U, and a metric $\mathcal{D} : U \times U \to [0,1]$ for a classification task with outcome set O, and a distance metric $d : \Delta(O) \times \Delta(O) \to [0,1]$ over distributions over outcomes, a randomized classifier $C : U \to \Delta(O)$ is $\alpha-$individually fair if and only if for all $u, v \in U$, $d(C(u), C(v)) \leq \alpha\mathcal{D}(u,v)$.*

We use the phrase "similar individuals are treated similarly" as a shorthand for the individual fairness Lipschitz condition. Individual fairness was originally proposed in the context of independent classification, i.e., each individual is classified exactly once, independently of all others. However, in many practical settings individuals cannot be classified independently, particularly when there are a limited number of positive classifications available (e.g., a university which can only accept a limited number of students each year, an advertiser with a limited budget). Dwork and Ilvento formalized this problem as the "cohort selection problem," in which a set of exactly n individuals must be selected such that the probabilities of selection conform to individual fairness constraints [6].

▶ **Definition 2** (Cohort Selection Problem [6]). *Given a universe U of individuals, an integer n, and a task with metric $\mathcal{D}$, select a cohort $C \subseteq U$ of exactly n individuals such that $|\Pr[u \in C] - \Pr[v \in C]| \leq \mathcal{D}(u,v)$. We call such a mechanism an individually fair cohort selection mechanism.*

Our work extends the investigation into fair composition by considering composition within a *pipeline* of cohort selection and scoring steps. We focus on the case of a two-step pipeline, and we assume for simplicity that the metric for the cohort selection and scoring functions are the same.

▶ **Definition 3** (Two-stage Cohort pipeline). *Given a universe of individuals U, a two-stage cohort pipeline consists of: a set of permissible cohorts $\mathcal{C} \subseteq \mathsf{Pow}(U)\backslash\emptyset$ (where $\mathsf{Pow}(U)$ indicates the power set of U), a single (randomized) cohort selection mechanism A which outputs a single cohort $C \subseteq \mathcal{C}$, a set of scoring functions $\mathcal{F} : \mathcal{C} \times U \to [0,1]$, and a scoring function $f \in \mathcal{F}$. The two-stage cohort pipeline procedure is $A \circ f$.*

Table 1 Terminology.

Term	Definition
U	The universe of individuals
$\mathcal{D} : U \times U \to [0,1]$	The individual fairness metric
$\mathcal{C} \subseteq \mathsf{Pow}(U) \backslash \emptyset$	The set of permissible cohorts
$\mathcal{F}$	The family of permitted scoring functions.
$f : \mathcal{C} \times U \to [0,1]$	A scoring function. $f(C,x)$ is *undefined* whenever $x \notin C$, and throughout this work, whenever we write $f(C,x)$, where x is any element in U, it is the case that $x \in C$.
$A : U \to \mathcal{C}$	An individually fair cohort selection mechanism.
$\mathbb{A}(C) \in [0,1]$	The probability that A outputs the cohort C.
$\mathcal{C}_u \subseteq \mathcal{C}$	The subset of permissible cohorts containing u.
$p(u) \in [0,1]$	The probability A outputs a cohort containing U.

We now briefly introduce supporting terminology (summarized in Table 1). For $C \in \mathcal{C}$, let $\mathbb{A}(C)$ denote the probability that A outputs C, where the probability is over the randomness in the cohort selection mechanism A operating on the universe U. We denote the set of cohorts containing u as $\mathcal{C}_u$, and the probability that A selects u can be expressed $p(u) = \sum_{C \in \mathcal{C}_u} \mathbb{A}(C)$. As initial constraints on A and $\mathcal{F}$, we assume that A is an individually fair cohort selection mechanism and that each $f \in \mathcal{F}$ is individually fair within the cohort it observes, i.e., it is *intra-cohort individually fair*:

▶ **Definition 4** (Intra-cohort individual fairness). *Given a cohort C, a scoring function $f : \mathcal{C} \times U \to [0,1]$ is intra-cohort individually fair if for all $C \in \mathcal{C}$, $\mathcal{D}(u,v) \geq |f(C,u) - f(C,v)|$ for all $u, v \in C$.*

Although intra-cohort fairness constrains f to be individually fair *within* a particular cohort, $f(C_1, u)$ can differ arbitrarily from $f(C_2, u)$ if $C_1 \neq C_2$. For ease of exposition we sometimes refer to C as the "cohort context" or simply the "context" of u for $u \in C$.

▶ Remark 5 (Intra-cohort individual fairness is insufficient). A pipeline consisting of an individually fair cohort selection mechanism and intra-cohort individually fair scoring function may result in arbitrarily unfair treatment. For example, suppose $\mathcal{X} = \{X_1, X_2, \ldots\}$ is a partition of U, and A chooses a cohort X_i uniformly at random. Suppose f assigns score 1 to all members of the cohort corresponding to X_*, and otherwise assigns score 0. A is not only individually fair, it selects each element with an equal probability; f is not only intra-cohort individually fair, it treats all members of a given cohort equally; yet the pipeline can result in arbitrarily large differences in scores for similar individuals. Furthermore, this observation holds for *any* partition including adversarially chosen partitions. Although this abstract example suffices to prove the point, we include an extensive set of realistic pipeline examples, analogous to the "Catalog of Evils" of [5], in Appendix A. We also include a practical method for *malicious* pipeline construction in Appendix B of the full version.

An important part of the pipeline definition is the contextual behavior of f, i.e., the behavior of the second stage of the pipeline may depend on the selected cohort C. The simplistic solution to this problem is to design and evaluate the whole pipeline for fairness as a single unit, i.e., requiring that similar individuals have similar distributions over $\Delta(O_{pipeline})$. Although such evaluation would catch unfairness, it (1) doesn't provide explicit guidance for designing any given component, (2) may miss certain pipeline-specific fairness issues (see Examples 7 and 9), and (3) "locks" the pipeline into a single monolithic strategy, which is highly impractical. For example, employers frequently need to change compensation policies due to changing market conditions. However, changing compensation policies due to disliking a particular member of a cohort, e.g., switching to equal bonuses for all team members if the company does not like the individual who would have received the largest bonus, is not permitted in our model. Indeed, later stages in the pipeline may be completely ignorant of the existence of prior stages, e.g., a manager deciding on employee compensation may be unaware of automated resume screening.

This motivates our design goal of *robustness*: designing the cohort selection mechanism A which composes well with *every* function in $\mathcal{F}$, rather than expecting the scoring function to properly analyze and respond to the choices made in the original cohort selection mechanism design. As a result, the only communication necessary between the steps is the description of $\mathcal{F}$. With this in mind, a deceptively(!) simple extension of Definition 1 gives our fairness desideratum for pipelines.

▶ **Definition 6** (α-Individual Fairness and Robustness for Pipelines (informal)). *Consider the pipeline consisting of $(\mathcal{C}, A, \mathcal{F})$, with outcome space $O_{pipeline}$. For $f \in \mathcal{F}$, the pipeline instantiated with f satisfies $\alpha-$individual fairness with respect to the similarity metric $\mathcal{D}$ and a distance measure $d : \Delta(O_{pipeline}) \times \Delta(O_{pipeline}) \to [0,1]$ if $\forall u, v \in U$, $d([f \circ A](u), [f \circ A](v)) \leq \alpha\mathcal{D}(u,v)$.*

*If the pipeline satisfies $\alpha-$individual fairness with respect to every $f \in \mathcal{F}$, i.e., if $\forall f \in \mathcal{F}$ and $\forall u, v \in U$, $d([f \circ A](u), [f \circ A](v)) \leq \alpha\mathcal{D}(u,v)$, we say that A is $\alpha-$*robust *to $\mathcal{F}$ with respect to $d, \mathcal{D}$.*

We model the contextual nature of the problem by allowing the behavior of each $f \in \mathcal{F}$ to depend on the cohort, rather than allowing f to be chosen adaptively in response to the selected cohort. This modeling choice still allows us to capture the contextual nature of scoring policies, while keeping our abstractions clean.[5]

2.2 Fairness of pipelines

Lurking in this informal definition are two subtle choices critical to pipeline fairness: (1) how should distributions over $O_{pipeline}$ be interpreted, and (2) what distance measure d is appropriate for measuring differences in distributions over $O_{pipeline}$. In the remainder of this section, we consider these two questions and frame the notion of robustness parametrized by the two axes: distribution and distance measure over distributions.

2.2.1 Choosing the interpretation of the distribution

To account for the fact that individuals not selected by A never receive a score from f the relevant outcome space is the union of possible scores and "not selected," i.e., $O_{pipeline} := [0,1] \cup \{\perp\}$. Thus conditioning on whether an individual was selected or not changes the interpretation of the distribution over the outcome space and, more importantly, changes the *perception* of fairness.

▶ **Example 7** (Perception of conditional probability). Suppose Alice (a) and Bob (b) are similar but not equal job candidates, i.e., $\mathcal{D}(a,b) \in (0, 0.1]$. Consider an individually fair cohort selection mechanism, A which either selects a cohort containing one of Alice or Bob or neither and satisfies $p(a) = p(b) = p^*$. Consider the fairness constraint on the scoring function f for the unconditional distribution over $O_{pipeline}$: $|p(a)f(a) - p(b)f(b))| \leq \mathcal{D}(a,b)$, which

[5] See Appendix A for explicit examples of modeling adaptation to changing market conditions.

simplifies to $p^*|f(a) - f(b))| \leq \mathcal{D}(a, b)$. (Note: as Alice and Bob never appeared together in a cohort, there is no intra-cohort fairness condition.) The constraint on the difference in treatment by f is essentially diluted by a factor of p^*.

Enforcing fairness on the unconditional distribution essentially allows the company to hand out job offers of the following form: "Congratulations you are being offered a position at Acme Corp., you can expect a promotion after one year with probability $x\%$." Alice and Bob may *receive* offers will equal probability, but the values of x printed on the offer may be wildly different, and as such they will perceive the value of the job offer differently.

The choice of conditional or unconditional distribution boils down to what perception of fairness is important. In the case of bonuses or promotions awarded long after hiring, the conditional perception may be particularly important. However, on shorter time frames or if the only consequential outcome is the final score, the unconditional distribution may be more appropriate (e.g., resume screening immediately followed by interviews).[6] We consider two approaches which capture these different perspectives: the **unconditional distribution** $S_u^{N,A,f}$, treats the $\perp$ outcome as a score of 0 and the **conditional distribution** $S_u^{C,A,f}$ conditions on u being selected in the cohort. More formally:

▶ **Definition 8** (Conditional and unconditional distributions). *Let $S_u^{A,f} \in \Delta(O_{pipeline})$ be the distribution over outcomes arising from the pipeline, i.e., $f \circ A$. $S_u^{A,f}$ places a probability of $1 - p(u)$ on $\perp$, and for $s \in [0, 1]$, $S_u^{A,f}$ places a probability of $\sum_{C \in \mathcal{C}} \Pr[f(C, u) = s]A(C)$ on s.*

- *The **unconditional distribution** $S_u^{N,A,f}$ is identical to $S_u^{A,f}$ with the exception that it treats the $\perp$ outcome as if it had score 0. That is, for $0 < s \leq 1$, $S_u^{N,A,f}$ places a probability of $\sum_{C \in \mathcal{C}} \Pr[f(C, u) = s]\mathbb{A}(C)$ on s; at $s = 0$, $S_u^{N,A,f}$ has a probability of $1 - p(u) + \sum_{C \in \mathcal{C}} \Pr[f(C, u) = 0]\mathbb{A}(C)$.*
- *The **conditional distribution** $S_u^{C,A,f}$ has probability $\frac{\sum_{C \in \mathcal{C}} \Pr[f(C,u)=s]\mathbb{A}(c)}{p(u)}$ for each score $s \in [0, 1]$, i.e., it is $S_u^{A,f}$ conditioned on the positive outcome of $A(C)$.*[7]

Each of these approaches can be viewed as a method for converting a distribution $S_u^{A,f}$ over $O_{pipeline}$ to distributions $S_u^{C,A,f}$ and $S_u^{N,A,f}$ over $[0, 1]$.

2.2.2 Distance measures over distributions

The natural approach for measuring distances between distributions would be to use expectation: that is, $d^{uncond,\mathbb{E}}(S_u^{A,f}, S_v^{A,f}) := |\mathbb{E}[S_u^{N,A,f}] - \mathbb{E}[S_v^{N,A,f}]|$ and $d^{cond,\mathbb{E}}(S_u^{A,f}, S_v^{A,f}) := |\mathbb{E}[S_u^{C,A,f}] - \mathbb{E}[[S_v^{C,A,f}]|$. Difference in expectation generally captures the unfairness in the examples discussed thus far. However, a subtle issue can arise from the *certainty* of outcomes, which requires greater insight into the distribution of scores.

[6] Although in this work we consider pipelines with a single relevant metric, the conditional versus unconditional question is critically important when metrics differ between stages of the pipeline. For example, the metric for selecting qualified members of a team may be different than the metric for choosing an individual from the team to be promoted to a management role, as the two stages in the pipeline require different skillsets.

[7] This definition is not defined if $p(u) = 0$, since it does not make sense to consider a "conditional distribution" if u is never selected to be in the cohort (and thus never receives a score). In defining robustness of a cohort selection mechanism, we should thus restrict to considering $u \in U$ where $p(u) > 0$ (and individual fairness of the cohort selection mechanism on its own would provide fairness guarantees over the probabilities $p(u)$). For simplicity, we do not explicitly mention this modification.

▶ **Example 9** (Certainty of outcomes)**.** Consider two equally qualified job candidates, Charlie and Danielle. As these two candidates are equally qualified, they should clearly be offered jobs and promotions with equal probability. Recall the company's pleasant form letter for job offers from Example 7, "Congratulations you are being offered a position at Acme Corp., you can expect a promotion after one year with probability $x\%$." Danielle receives an offer with $x = 70\%$ (with probability p^*), but Charlie receives either an offer with $x = 100\%$ (with probability $0.7p^*$) or an offer with $x = 0\%$ (with probability $0.3p^*$). Although both are offered jobs with equal probability and their expectations of promotion are equal, Charlie's offers have *certainty* of promotion (or no promotion) whereas Danielle's promotion fate is uncertain.

As Example 9 illustrates, expected score does not entirely capture problems related to the *distribution* of scores rather than the average score. Although total-variation distance is a natural choice for evaluating such distributional differences, it is too strong for this setting. For example, if Charlie receives a score of 0.7 with probability 1 (over randomness of the entire pipeline), while Danielle receives a score of $0.7 - \epsilon$ with probability 0.5 and a score of $0.7 + \epsilon$ with probability 0.5, then the total variation distance would be 1, though these outcomes are intuitively very similar. We therefore introduce the notion of *mass-moving distance* over probability measures. Mass-moving distance combines total variation distance with earthmover distance to reflect that similar individuals should receive similar distributions over close (rather than identical) sets of scores.

▶ **Definition 10** (Mass-moving distance)**.** *Let γ_1 and γ_2 be probability mass functions over finite sets $\Omega_1 \subseteq [0,1]$ and $\Omega_2 \subseteq [0,1]$, respectively. Let $V \subseteq [0,1]$ be the set of real values $v \in [0,1]$ such that there exist probability mass functions $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ over $[0,1]$ with finite supports $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$, respectively, where:*

1. *Nothing moves far and mass is conserved. For $i = 1, 2$, there is a function $Z_i : [0,1] \to \Delta(\tilde{\Omega}_i)$ such that:*
   - **a.** *Nothing moves far. For all $x \in [0,1]$ and $y \in Supp(Z_i(x))$, it holds that $|x - y| \leq 0.5v$.*
   - **b.** *Mass is conserved. For all $y \in \tilde{\Omega}_i$, it holds that $\tilde{\gamma}_i(y) = \sum_{x \in \Omega_i} z_i^x(y)\gamma_i(x)$, where z_i^x is the probability mass function of the distribution $Z_i(x)$.*
2. *Total variation distance is small. It holds that $0.5v \geq TV(\tilde{\gamma}_1, \tilde{\gamma}_2) := \frac{1}{2}\sum_{w \in \tilde{\Omega}_1 \cup \tilde{\Omega}_2} |\tilde{\gamma}_1(w) - \tilde{\gamma}_2(w)|$.*

Then we let $MMD(\gamma_1, \gamma_2) = \inf(V)$.

A simple way to think about mass-moving distance is to break the definition down into two steps: (1) transforming the original distributions over scores into distributions over a single shared set of *adjusted scores* and (2) moving mass between the distributions over adjusted scores.

Since there is a natural association between probability distributions over $[0,1]$ and probability mass functions over $[0,1]$, Definition 10 also gives a notion of distance between probability distributions.[8] In the example of Charlie and Danielle receiving scores of 0.7 or $0.7 \pm \varepsilon$ described above, the mass-moving distance is at most 2ϵ since $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ can both be taken to be the probability measure that places the full mass of 1 on 0.7.

[8] We slightly abuse notation and use $MMD(\mathcal{X}_1, \mathcal{X}_2)$ for probability distributions $\mathcal{X}_1$ and $\mathcal{X}_2$, to denote $MMD(\gamma_1, \gamma_2)$ where γ_1 is the probability mass function associated to $\mathcal{X}_1$ and γ_2 is the probability mass function associated to $\mathcal{X}_2$.

Using mass-moving distance, we specify two additional complementary distance measures $d^{cond,MMD}(S_u^{A,f}, S_v^{A,f}) := MMD(S_u^{C,A,f}, S_v^{C,A,f})$ and $d^{uncond,MMD}(S_u^{A,f}, S_v^{A,f}) := MMD(S_u^{N,A,f}, S_v^{N,A,f})$.

2.3 Robustly fair pipelines

Recall our informal notion that a cohort selection mechanism A is robust to a family of scoring functions $\mathcal{F}$ if the composition of A and any $f \in \mathcal{F}$ is individually fair. We can now formalize robustness as either **conditional** or **unconditional** with respect to either **expected score** or **mass moving distance** over score distributions. By evaluating the properties of each combination of distribution and distance measure, we can capture a range of subtle fairness desiderata in pipelines.[9]

▶ **Definition 11** (Robust pipeline fairness). *Given a universe U, a metric $\mathcal{D}$, let A be an individually fair cohort-selection mechanism and let $\mathcal{F}$ be a collection of intra-cohort individually fair scoring functions $\mathcal{C} \times U \to [0,1]$. Choose $d \in \{d^{cond,\mathbb{E}}, d^{uncond,\mathbb{E}}, d^{cond,MMD}, d^{uncond,MMD}\}$, a distance measure over $S_u^{A,f}$. We say A is α-**robust w.r.t $\mathcal{F}$** for d if $d(S_u^{A,f}, S_v^{A,f}) \leq \alpha\mathcal{D}(u,v)$ for all $u, v \in U$ and for all $f \in \mathcal{F}$.*

Throughout the rest of this work, we will examine robustness properties in terms of particular settings of d. As one might expect, mass moving distance over score distributions is a stronger condition than expected score, and conditional robustness implies unconditional robustness up to a Lipschitz relaxation.[10]

3 Conditions for Success

In this section, we describe conditions on A that will result in our desired robustness properties with respect to a class of scoring functions $\mathcal{F}$. We first consider the description of $\mathcal{F}$ available to A, i.e., the policy. The simplest method of specifying the policy by describing all $f \in \mathcal{F}$ prohibits adding f with similar or identical fairness properties to $\mathcal{F}$ at a later point and is highly unrealistic (and potentially intractable). In practice, we expect policies to govern how differently f can treat individuals within different contexts, rather than enumerating the permitted functions. To that end, we propose policies in the form of a *distance function over (cohort, individual) pairs*, $\delta^{\mathcal{F}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$. This distance function specifies the maximum difference in score between two (cohort, individual) pairs $\delta^{\mathcal{F}}((C_1, u), (C_2, v)) := \sup_{f \in \mathcal{F}} |f(C_1, u) - f(C_2, v)|$. $\delta^{\mathcal{F}}$ captures the salient fairness behavior of the family of scoring functions, while being succinct in comparison to maintaining a list of all supported f directly. In fact, as we will show in Lemma 13, a partial description or an overestimate of $\delta^{\mathcal{F}}$ will also suffice. To illustrate our policy descriptions, consider the following two families:

1. $\mathcal{F}_1$ ignores the cohort context entirely, and treats each $u \in U$ the same regardless of the cohort, i.e., $\mathcal{F}_1 = \{f \mid \exists g : U \to [0,1] \text{ s.t. } f(C,u) = g(u) \text{ for all } (C,u) \in \mathcal{C} \times U\}$.
2. $\mathcal{F}_2$ treats u and v similarly within the same context, but has no constraint on treatment in different contexts, i.e., $\mathcal{F}_2 = \{f \mid f((C \backslash \{u\}) \cup \{v\}, v) - f(C,u)| \leq \mathcal{D}(u,v) \text{ for all } u, v \in U \text{ and } \forall C \in \mathcal{C} \text{ s.t. } u \in C, v \notin C\}$.

[9] Note that these choices for d are not the only possible choices, and the framework can be extended to different choices of distribution and distance measure to address other fairness concerns.

[10] See Propositions E.2 and E.1 in the full version. Interestingly, we show in the full version that for some classes of score functions, guaranteeing individual fairness w.r.t mass-mover distance fairness is "equivalent" to guaranteeing individual fairness w.r.t expected score.

Recall that intra-cohort individual fairness requires that the scoring functions in both families must treat u and v similarly if they appear in the same cohort, i.e., $\mathcal{D}(u,v) \geq |f(C,u) - f(C,v)|$.

For the family $\mathcal{F}_1$, we observe that $\delta^{\mathcal{F}_1}((C_1,u),(C_2,v)) = \mathcal{D}(u,v)$, and, intuitively, the designers of A will not need to consider the behavior of $\mathcal{F}$ in their design of A. On the other hand, for $\mathcal{F}_2$, we observe that $\delta^{\mathcal{F}_2}((C,u),(C,v)) = \mathcal{D}(u,v)$ for any cohort C, but $\delta^{\mathcal{F}_2}((C,u),(C',v))$ can be much greater than $\mathcal{D}(u,v)$ for $C' \neq C$. For this reason, composition planning for A is non-trivial. As one would expect, $\delta^{\mathcal{F}}$ heavily influences the strength of conditions on A.

3.1 A's Task: Designing Mechanisms Compatible with $\delta^{\mathcal{F}}$

We now describe how to design A to guarantee robustness with respect to $\mathcal{F}$, given (possibly overestimates of) the distance function $\delta^{\mathcal{F}}$ over (cohort, individual) pairs describing $\mathcal{F}$. The conditions on A will roughly consist of making sure that A assigns *similar individuals to similar distributions over cohort contexts*, where similarity of (cohort, individual) pairs is defined with respect to $\delta^{\mathcal{F}}$.

Although $\delta^{\mathcal{F}}$ is a succinct description of a policy, it is more intuitive when designing with composition in mind to translate $\delta^{\mathcal{F}}$ into a set of “mappings” specifying which (cohort, individual) pairs will be treated similarly by $f \in \mathcal{F}$. That is, for each pair $u, v \in U$, we can describe $\delta^{\mathcal{F}}$ as a partitioning $\mathcal{P}_{u,v}$ of $(\mathcal{C}_u \times u) \cup (\mathcal{C}_v \times v)$ such that each partition or “cluster” has small diameter with respect $\delta^{\mathcal{F}}$, i.e., within a cluster $\delta^{\mathcal{F}}((C_1,u),(C_2,v)) \leq \mathcal{D}(u,v)$. The collection of partitions over all pairs of individuals then defines the mapping.

▶ **Definition 12** (Mapping based on δ). *For each pair of distinct individuals u and v, consider the subset $\mathcal{P}_{u,v} := (\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$ of (cohort, individual) pairs. Consider a partition of $\mathcal{P}_{u,v}$ into clusters that respects δ, i.e., that satisfies the following condition: if $(C_1,x),(C_2,y)$ are in the same cluster*[11]*, then $\delta((C_1,x),(C_2,y)) \leq \mathcal{D}(u,v)$. Let $n_{u,v}$ (and $n_{v,u}$) be the number of clusters of the partition. We call a collection of such partitions for each pair $u, v \neq U$ a **mapping** of $\mathcal{C}$ that **respects δ**.*

Mappings interact well with distance functions δ' that overestimate $\delta^{\mathcal{F}}$, as larger distances between (cohort, individual) pairs imposes more strict conditions on cluster membership. Lemma 13 states that a mapping that respects δ' will also respect $\delta^{\mathcal{F}}$, although the resulting conditions on the mapping might be more restrictive.

▶ **Lemma 13.** *Let $\delta' : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ be a distance function. Suppose that δ' has the property that for all pairs of cohort contexts $(C_1,x),(C_2,y) \in \mathcal{C} \times U$, it holds that $\delta'((C_1,x),(C_2,y)) \geq \delta^{\mathcal{F}}((C_1,x),(C_2,y))$. If a mapping respects δ', then the mapping also respects $\delta^{\mathcal{F}}$.*

Proof. Consider any pair of individuals u and v, and consider any mapping that respects δ'. In the partition corresponding to u and v, if (C_1,x) and (C_2,y) are in the same cluster, then it holds that $\delta^{\mathcal{F}}((C_1,x),(C_2,y)) \leq \delta'((C_1,x),(C_2,y)) \leq \mathcal{D}(u,v)$. Thus, the mapping respects $\delta^{\mathcal{F}}$, as desired. ◀

[11] Note that $x, y \in \{u,v\}$. Recall that (C_1,u) and (C_2,u) may appear in the same cluster, and thus it is possible that $x = y$.

Table 2 Policy and mapping terminology.

Term	Definition
$\delta^{\mathcal{F}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0, 1]$.	distance function specifying the maximum difference in treatment between (cohort,individual) pairs by any $f \in \mathcal{F}$. $\delta^{\mathcal{F}}((C_1, u), (C_2, v))$ is undefined if $u \notin C_1$ or $v \notin C_2$.
$M_{u,v} : \mathcal{C}_u \to \mathbb{N}$	a mapping of the cohorts containing u to clusters containing (C, u).
$n_{u,v}$	The number of clusters in a mapping
$\mathcal{M}_\delta$	the set of all mappings which respect δ.

We now briefly introduce supporting terminology for policies and mappings (summarized in Table 2). To succinctly refer to the clusters in a mapping, we define label functions $M_{u,v} : \mathcal{C}_u \to \mathbb{N}$ and $M_{v,u} : \mathcal{C}_v \to \mathbb{N}$ such that $M_{u,v}(C)$ is the label of the cluster containing (C, u) and $M_{v,u}(C)$ is the label of the cluster containing (C, v). We use $n_{u,v}$ (or $n_{v,u}$) to denote the number of clusters in a mapping. We also refer to the set of functions $(M_{u,v})_{u \neq v \in U}$, which entirely specify the partitions, as a mapping. Valid mappings for δ are not necessarily unique, as there may be more than one way to partition $\mathcal{P}_{u,v}$ into clusters with diameter bounded by $\mathcal{D}(u, v)$. We let $\mathcal{M}_\delta$ be the set of mappings that respect δ.

Given a mapping of $\delta^{\mathcal{F}}$ (or of an overestimate δ'), we can now interpret "distributions over cohorts" induced by A as "distributions over clusters" induced by A. Formally, we convert the distributions over cohorts into measures over $[n_{u,v}]$ for each pair $(u, v) \in U \times U$. As a result, "similar distributions over cohorts" will turn out to mean "similar measures over $[n_{u,v}]$."

▶ **Definition 14.** *Let $(M_{u,v})_{u \neq v \in U}$ be a mapping of $\mathcal{C}$. For $u, v \in U$, we define measures $q^1_{u,v}$ and $q^2_{u,v}$ over the sample space $[n_{u,v}]$ as follows:*

1. *The **unconditional measure over cohorts** $\mathbf{q^1_{u,v}}$ on the sample space $[n_{u,v}]$ for each (u, v) ordered pair is defined as follows. For $i \in [n_{u,v}]$, we let $q^1_{u,v}(i) = \sum_{C \in \mathcal{C}_u | M_{u,v}(C) = i} \mathbb{A}(C)$.*[12]
2. *The **conditional measure over cohorts** $\mathbf{q^2_{u,v}}$ on the sample space $[n_{u,v}]$ for each (u, v) ordered pair is defined as follows. For $i \in [n_{u,v}]$, we let $q^2_{u,v}(i) = \frac{\sum_{C \in \mathcal{C}_u | M_{u,v}(C) = i} \mathbb{A}(C)}{p(u)}$.*[13][14]

We now specify sufficient conditions for robustness in terms of distances between these measures over $[n_{u,v}]$. The conditions require that for each pair $u, v \in U$, A assigns similar probabilities to cohorts containing u and cohorts containing v within each cluster.

▶ **Definition 15** (α-Notions 1 and 2)**.** *Let $(M_{u,v})_{u \neq v \in U}$ be a mapping of $\mathcal{C}$. For $u, v \in U$, let $q^1_{u,v}$ and $q^2_{u,v}$ be defined as in Definition 14. We define α-Notions 1 and 2 as follows:*

1. *For $\alpha \geq 0.5$, we say that A satisfies α-**Notion 1** if for all $u, v \in U$ such that $\mathcal{D}(u, v) < 1$, $TV(q^1_{u,v}, q^1_{v,u}) \leq (\alpha - 0.5)\mathcal{D}(u, v)$. (The 0.5 arising in Notion 1 comes from having to "smooth out" $q^1_{u,v}$ to a probability measure in a later step.)*
2. *For $\alpha \geq 0$, we say that A satisfies α-**Notion 2** if for all $u, v \in U$ such that $\mathcal{D}(u, v) < 1$, $TV(q^2_{u,v}, q^2_{v,u}) \leq \alpha\mathcal{D}(u, v)$.*

[12] This is not necessarily a probability measure, since the total sum on the sample space is $p(u) \leq 1$, but it is finite.

[13] Observe that this is in fact a probability measure since $p(u) = \sum_{C \in \mathcal{C}_u} \mathbb{A}(C) = \sum_{i=1}^{M_{u,v}} \sum_{C \in \mathcal{C}_u | M_{u,v}(C) = i} \mathbb{A}(C)$.

[14] Like in Definition 8, this definition is not defined if $p(u) = 0$, since it does not make sense to consider a "conditional distribution" if u is never selected to be in the cohort (and thus never receives a score). We should thus restrict to considering $u \in U$ where $p(u) > 0$ (and individual fairness of the cohort selection mechanism on its own would provide fairness guarantees over the probabilities $p(u)$). For simplicity, in this extended abstract, we do not explicitly mention this modification.

Our main result is that these conditions guarantee pipeline robustness for composition with any $f \in \mathcal{F}$ with respect to mass-moving distance (and thus expected score).[15] Theorem 16 states that so long as A satisfies Notion 1 (resp. 2) for the mappings associated with $\mathcal{F}$, then A will be robust with respect to $\mathcal{F}$.

▶ **Theorem 16** (Robustness to Post-Processing). *Let $\mathcal{F}$ be a class of scoring functions, let $\alpha \geq 0.5$ be a constant. Suppose that $(M_{u,v})_{u \neq v \in U}$ is in $\mathcal{M}_{\frac{1}{2\alpha}\delta^{\mathcal{F}}}$. If A is individually fair and satisfies α-Notion 1 (resp. α-Notion 2) for $(M_{u,v})_{u \neq v \in U}$, then A is 2α-robust w.r.t. $\mathcal{F}$ for $d^{uncond,MMD}$ (resp. $d^{cond,MMD}$).*

The proof of Theorem 16 appears in Appendix B.1 of the full version.

Furthermore, these conditions are necessary both for mass-moving distance and the weaker condition of expected score for certain rich classes of scoring functions.

▶ **Theorem 17** (Informal). *Let d be any metric in $\left\{d^{uncond,MMD}, d^{cond,MMD}, d^{cond,\mathbb{E}}, d^{uncond,\mathbb{E}}\right\}$. Loosely speaking, given $\mathcal{F}$ described by mappings such that inter-cluster distances are much larger than intra-cluster distances, the requirements on A in Theorem 16 are **necessary** for achieving robustness w.r.t. d.*

We formalize Theorem 17 in Appendix B of the full version.[16]

4 Robust Mechanisms

Although the conditions specified in the previous section are quite strict, and indeed some pathological scoring function families admit no robust solutions, we can nonetheless construct robust cohort selection mechanisms for rich classes of scoring policies.[17] We exhibit mechanisms robust to two broad classes of policies:

1. **Individual interchangeability:** replacing a single individual in the cohort does not change treatment of the cohort too much, i.e., policies like $\delta^{\mathcal{F}_2}$.
2. **Quality-based treatment:** cohorts with similar quality "profiles" are treated similarly. That is, the scoring function only considers the set of qualifications represented within a cohort and is agnostic to the specific individual(s) exhibiting a given qualification.

These policies cover a wide range of realistic scenarios and allow for significant flexibility and adaptability in the choice of f. In this section, we demonstrate that these policies also admit a variety of efficient and *expressive* constructions for A, i.e., A that may assign a wide range of probabilities $p(u)$ to individuals.

▶ Remark 18. As previously noted, robustness is trivial for the class of scoring functions which ignore the cohort context ($\mathcal{F}_1$). We formalize this observation in the following proposition:

▶ **Proposition 19.** *Consider the mapping that, for each pair of individuals u and v, places all of the cohort contexts in $(\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$ into the same cluster. If A is individually fair, then A satisfies 0.5-Notion 1 and 0.5-Notion 2 w.r.t. this mapping.*

[15] See Corollary B.1.1 in the full version for a formal statement of the relationship between MMD and expected score.

[16] See Theorem B.5 and Theorem B.6.

[17] See Appendix D.1 of the full version for an example of $\mathcal{F}$ which admits no robust A.

4.1 Individual interchangeability

To describe the interchangeability policy, we specify a distance function $\delta^{\mathsf{int}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ that requires that "swapping" any individual in a cohort does not result in significantly different treatment. More formally:

▶ **Definition 20** (Individual interchangeability policy).

$$\delta^{\mathsf{int}}((C,u),(C',v)) = \begin{cases} \mathcal{D}(u,v) & \textit{if } C = C' \\ \mathcal{D}(u,v) & \textit{if } C' = (C \setminus \{u\}) \cup \{v\}\,. \\ 1 & \textit{otherwise.} \end{cases}$$

δ^{int} can be viewed as an overestimate of $\delta^{\mathcal{F}_2}$, or as a partial specification of the distance function on a subset of $(\mathcal{C} \times U) \times (\mathcal{C} \times U)$, trivially completed to 1 on other pairs of cohort context pairs. δ^{int} is naturally translated into a simple mapping: for any pair of individuals u and v, the partition corresponding to u and v in the mapping consists of clusters of size 2 consisting of "corresponding" (cohort, individual) pairs. This follows from observing that if an individual u receives some score $f(C,u)$ in a cohort C, if u were replaced by $v \notin C$, then v would receive a score in $[f(C,u) - \mathcal{D}(u,v), f(C,u) + \mathcal{D}(u,v)]$. More formally:

▶ **Definition 21** (Swapping Mapping). *Let $\mathcal{C}$ be the set of all subsets of U with exactly k individuals. The* ***swapping mapping*** *is defined as follows. For each pair of individuals $u, v \in C$:*

1. *For $C \in \mathcal{C}$ such that $u, v \in C$, the partition includes the cluster $\{(C,u),(C,v)\}$.*
2. *For $C \in \mathcal{C}$ such that $u \in C$ and $v \notin C$, the partition includes the cluster $\{(C,u),((C \setminus \{u\}) \cup \{v\}, v)\}$.*

It is straightforward to verify that the swapping mapping respects δ^{int}.

For the swapping mapping, there is a simple condition under which cohort selection mechanisms satisfy unconditional robustness (Notion 1): monotonicity.

▶ **Definition 22** (Monotonic cohort selection). *Suppose that $\mathcal{C}$ is the set of cohorts of size k. A cohort selection mechanism A is* ***monotonic*** *if for all pairs of individuals $u, v \in U$, for any $C' \subseteq U$ such that $|C'| = k-1$ and $u, v \notin C'$, if $p(u) \leq p(v)$ then $\mathbb{A}(C' \cup \{u\}) \leq \mathbb{A}(C' \cup \{v\})$.*

The intuition for the link between the monotonicity property and the swapping mapping is that the probability masses on a cohort containing u and a cohort containing v that are paired in the swapping mapping are directionally aligned and cannot diverge by more than $\mathcal{D}(u,v)$.

▶ **Lemma 23.** *Suppose that $\mathcal{C}$ is the set of cohorts of size k. If A is monotonic, then A satisfies* 0.5*-Notion 1 for the swapping mapping.*

Both PermuteThenClassify and WeightedSampling, cohort selection mechanisms proposed in [6], are monotonic, efficient and have a high degree of expressivity.[18]

However, monotonicity alone is not sufficient to guarantee conditional robustness (Notion 2) for the swapping mapping (see Appendix B of the full version). Borrowing intuition from PermuteThenClassify, we give a novel, efficient, individually fair cohort selection mechanism that achieves conditional robustness (Notion 2) for the swapping mapping:

[18] See Appendix B of the full version for detailed descriptions of these mechanisms and formal proofs of the monotonicity property.

▶ **Mechanism 24** (Conditioning Mechanism). *Given a target cohort size k, a universe U and a distance metric $\mathcal{D}$, initialize an empty set S. For each individual $u \in U$:*

1. *Assign a weight $w(u)$ such that $|w(u) - w(v)| \leq \mathcal{D}(u, v)$, i.e., the weights are individually fair.*
2. *Draw from $\mathbb{1}_u \sim Bern(w(u))$, (i.e., flip a biased coin with weight $w(u)$). If $\mathbb{1}_u$, add u to S.*

If $|S| \geq k$, return a uniformly random subset of S of size k.[19] *Otherwise, repeat the mechanism.*

We show that under mild conditions, the Conditioning Mechanism is satisfies Notion 2, concludes in a small number of rounds, and allows for a high degree of expressivity. (See Appendix D of the full version for a formal statement and proof details.)

4.2 Quality-based treatment

One downside of the monotonic mechanisms proposed for δ^{int} is that they require that any cohort with a single individual swapped is considered with nearly the same probability as the original cohort. In practice, this is problematic when A needs to ensure that each cohort has a certain structure. For example, when hiring a team of software engineers, designers and product managers, the proportion of each type of team member is important, and arbitrary swaps are not desirable from the perspective of team structure. By restricting to scoring functions that only consider the quality profile of a cohort, i.e., how many individuals from each quality group are represented in a cohort, A can construct highly *structured* cohorts, so long as the structure of the cohort is valid with respect to the fairness metric $\mathcal{D}$.

We now consider robust mechanisms for policies predicated on additional structure within the metric over U. In particular, we assume the existence of a partition of the universe U into one or more "quality groups" $q_1, \ldots, q_n$. These quality groups satisfy the property that the distances within a quality group are smaller than distances between quality groups. *How much* smaller is determined by a parameter β. More formally,

▶ **Definition 25.** *Let $\beta \leq 1$ be a constant and $n \geq 1$ be an integer. Consider a partitioning of a U into subsets $q_1, \ldots, q_n$, i.e., "quality groups", and let $\mathcal{D}^*$ be a metric on U. Now, we define metrics D on $\{1, \ldots, n\}$ and $\mathcal{D}^i$ for $1 \leq i \leq n$ on q_i as follows: we let $D(i, j) = \inf_{u \in q_i, v \in q_j} \mathcal{D}^*(u, v)$ and $\mathcal{D}^i$ be the restriction of $\mathcal{D}^*$ to q_i. We call the metric $\mathcal{D}^*$ endowed with quality groups $q_1, \ldots, q_n$* ***β-quality-clustered*** *if for all $1 \leq i \leq n$, we have that*

$$\max_{u,v \in q_i} \mathcal{D}^i(u, v) \leq \beta \min_{j \neq i} D(i, j).$$

Notice that any metric $\mathcal{D}^*$ is trivially 1-clustered with respect to the trivial quality group $q_1 = U$. The benefit of endowing $\mathcal{D}^*$ with a greater number of quality groups is to exploit additional structure of the metric, when any exists.

For simplicity in the specification of the relevant policy and family of scoring functions we introduce a **quality profile function** P to count the number of individuals in each quality group in a cohort: that is, $P : 2^U \to \{(x_1, \ldots, x_n) \mid x_i \in \mathbb{Z}^{\geq 0}\}$, and the ith coordinate of

[19] One might imagine a mechanism that conditions on exactly k individuals being chosen, but this mechanism can be arbitrarily far from individually fair. Consider $k - 1$ individuals with weight 1 and $|U| - k - 1$ individuals with weight 0.9. Conditioning exactly k individuals would cause $|p(u) - p(v)|$ to diverge arbitrarily for $w(u) = .9$ and $w(v) = 1$.

$P(C)$ is $|C \cap q_i|$. Loosely speaking, the quality-based treatment policy requires that the only information about a cohort utilized by the scoring functions is its quality profile. We now formally define $\mathcal{F}_3$ and an associated policy δ^{quality} :

▶ **Definition 26.** *Let $\beta \leq 1$ be a constant. Suppose that $\mathcal{D}$ is endowed with quality groups $q_1, \ldots, q_n$ and $\mathcal{D}$ is β-quality-clustered. We define $\mathcal{F}_3$ to be the set of intra-cohort individually fair score functions $f : \mathcal{C} \times U \to [0,1]$ satisfying the following conditions:*

1. *For $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, if u and v that are in the same quality group, then $f(C,u) = f(C',v)$.*
2. *For integers $1 \leq i \neq j \leq n$, $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, and any individuals $u \in q_j$ and $v \in q_j$, it holds that $|f(C,u) - f(C',v)| \leq D(i,j)$.*

When each quality group is homogeneous in terms of individual "quality", this corresponds to score functions that are determined purely by "quality".[20] As in Section 4.1, we specify a distance function $\delta^{\text{quality}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ that overestimates $\delta^{\mathcal{F}_3}$, but still preserves enough of the fairness structure to construct the desired mapping.

▶ **Definition 27** (Quality-based treatment policy). *Given a universe U, a set of permissible cohorts $\mathcal{C}$ and distance metrics and quality groups as in Definition 26,*

1. *For $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, if $u \in q_j$ and $v \in q_j$, then $\delta^{\text{quality}}((C,u),(C',v)) = 0$.*
2. *For integers $1 \leq i \neq j \leq n$, $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, and any individuals $u \in q_j$ and $v \in q_j$, we set $\delta^{\text{quality}}((C,u),(C',v)) = D(i,j)$.*

The core intuition is that a nice mapping exists when $\mathcal{C}$ is "symmetric with respect to individuals in each quality group." It is helpful here to consider a bipartite graph $G = (A, B, E)$, where A has one vertex for each subset of the universe U, B has one vertex for each possible profile of a subset of U, and there is an edge $(a,b) \in E$ precisely when b is the profile of a, that is $b = P(a)$.

Fix any $\mathcal{C}$, and consider the subgraph $G' = (A', B', E')$ of G induced by the vertices in A corresponding to members of $\mathcal{C}$, the edges incident on these vertices, and the subset of B induced by these edges. We say that $\mathcal{C}$ is **quality-symmetric** if for all $b' \in B'$ it is the case that E' contains all the edges in E (in the original graph) incident on b'.

That is, $\mathcal{C}$ contains all cohorts obtained by swapping out individuals from the same quality group. If $\mathcal{C}$ is quality-symmetric, then consider the following mapping.

▶ **Definition 28** (Quality-Based Mapping). *Let $\beta \leq 1$ be a constant. Suppose that $\mathcal{D}$ is endowed with quality groups $q_1, \ldots, q_n$ and $\mathcal{D}$ is β-quality-clustered. Suppose $\mathcal{C}$ is quality-symmetric. The **quality-based mapping** is defined as follows. For each pair of individuals $u, v \in C$, let $\mathcal{P}_{u,v} = (\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$. For each $(x_1, \ldots, x_n) \in P(\mathcal{C}_u \cup \mathcal{C}_v)$, the partitioning of $\mathcal{P}_{u,v}$ contains a cluster of the form $\{(C,x) \in \mathcal{P}_{u,v} \mid P(C) = (x_1, \ldots, x_n)\}$.*

We verify that the quality-based mapping indeed respects δ^{quality} (and thus respects $\delta^{\mathcal{F}_3}$ by Lemma 13). If u and v are in the same quality group, then the diameter of each cluster under δ^{quality} is 0, which is trivially upper bounded by $\mathcal{D}(u,v)$. On the other hand, if u and v are in different quality groups q_i and q_j respectively, then the diameter of each cluster is no more than $D(i,j) \leq \mathcal{D}(u,v)$. Thus, the properties of a mapping are satisfied by the quality-based mapping.

[20] In this case, $\mathcal{F}_3$ includes Equal Treatment, Promotion, Stack Rank, and Fixed Bonus (discussed in Appendix A) when scores are based on the "quality" of "performance" of individuals.

In this scenario, the quality-based *mapping* captures the intuition for the fairness structure of $\mathcal{F}_3$ much better than $\delta^{\mathsf{quality}}$. The mapping groups together all cohorts with the same quality profile (i.e., the same number of individuals in each quality group), capturing the intuition that the only information that a score function in $\mathcal{F}_3$ utilizes about a cohort is the quality profile.

As the score function behavior does not depend on the specific individuals in a quality group, A should have significant freedom to choose individuals within each quality group while still satisfying robustness w.r.t. $\mathcal{F}_3$. We will show that once the number of members of each quality group in the cohort is decided, utilizing any individually fair cohort selection mechanism within each quality group will satisfy our conditions. Moreover, our mechanisms have some flexibility in deciding the quality profile as well.

▶ **Mechanism 29** (Quality Compositional Mechanisms)**.** *Let $\beta \leq 1$ be a constant, and suppose that $\mathcal{D}$ endowed with quality groups $q_1, \ldots, q_n$ is β-quality-clustered. Suppose also that $\mathcal{C}$ is quality-symmetric. For each $1 \leq i \leq n$ and each $1 \leq x_i \leq |q_i|$, let A_{i,x_i} be a $\mathcal{D}^i$-individually fair mechanism selecting x_i individuals in q_i. We define the **quality compositional mechanism** for $\{A_{i,x_i}\}$ as follows. Let $\mathcal{X}$ be any distribution over n-tuples of nonnegative integers $(x_1, \ldots, x_n) \in P(\mathcal{C})$.*

1. *Draw $(x_1, \ldots, x_n) \sim \mathcal{X}$.*
2. *Independently run A_{i,x_i} for each $1 \leq i \leq n$, and return the union of the outputs of all of these mechanisms.*

In the next lemma, we show that when a quality composition mechanism only selects cohorts whose quality projection vectors $(x_1, \ldots, x_n)$ are "close" to an inter-quality group distance multiple of $(|q_1|, \ldots, |q_n|)$, Notion 1 is achieved. (This requirement essentially says that the relative proportion of selected individuals in each quality group needs to be approximately reflective of the relative proportion of individuals in each quality group in the universe, scaled by the difference between the quality groups in the original metric. This type of requirement turns out be necessary for basic individual fairness guarantees, by the constrained cohort impossibility result in [6].) Moreover, under stronger conditions, we show that Notion 2 is also achieved.

▶ **Lemma 30.** *Let $\beta \leq 0.5$ be a constant, and suppose that $\mathcal{D}$ endowed with quality groups $q_1, \ldots, q_n$ is β-quality-clustered. Suppose also that $\mathcal{C}$ is quality-symmetric, and let $\mathcal{X}$ be any distribution over $(x_1, \ldots, x_n) \in P(\mathcal{C})$ such that $|\frac{x_i}{|q_i|} - \frac{x_j}{|q_j|}| \leq (1 - 2\beta) D(i, j)$. If A is a quality compositional mechanism, then:*

1. *A is always individually fair.*
2. *A always satisfies 0.5-Notion 1.*
3. *A satisfies 0.5-Notion 2 for $\mathcal{D}$ and $\delta^{\mathcal{F}}$ if **either** of the following conditions hold:*
   a. *(One set) $|Supp(\mathcal{X})| = 1$ (i.e., one "canonical" $(x_1, \ldots, x_n)$), or*
   b. *(0-1 metric) $D(i, j) = 1$ for $1 \leq i \neq j \leq n$ and $\mathcal{D}^i(u, v) = 0$ for $1 \leq i \leq n$.*

The quality compositional mechanisms provide a greater degree of structure in cohort selection than the monotone mechanisms giving in Section 4.1. The Conditioning Mechanism and similar monotone mechanisms are forced to select individuals essentially independently, with the only dependence stemming from the cohort size constraint. However, structured cohorts are necessary in a number of practical applications, as previously noted. Although $\delta^{\mathsf{quality}}$ imposes more constraints on the permitted $\mathcal{F}$ than δ^{int}, the basis for these constraints is likely to be tolerated well in legitimate use cases in which structure is important.

Moreover, the company has flexibility in selecting individuals within each experience group, as any individually fair mechanism can be utilized. This offers significantly more flexibility than selecting members in each quality group uniformly at random. Such flexibility is particularly crucial, for example, if a company further wants to ensure that tech company teams have a mixture of software engineers and product managers. The individually fair mechanisms within each quality group can help achieve this balance through selecting balanced subsets of engineers and product managers. In essence, the quality compositional mechanisms allow flexibility in cohort selection while still satisfying robustness for $\mathcal{F}_3$, due to restrictions on the behavior of scoring functions in $\mathcal{F}_3$.

5 Discussion and Future Work

We have presented a framework for evaluating the robustness of cohort selection as part of a pipeline. We've demonstrated that naive auditing strategies concerning average cohort quality or score are unable to uncover significant fairness problems. We've also shown that many reasonable policies for cohort selection and subsequent scoring can conflict with each other resulting in very poor fairness outcomes. Furthermore, we've demonstrated that a malicious pipeline designer can easily use composition problems to disguise bad behavior. Despite these hurdles, we've shown that it is possible to construct pipelines that are fair. In particular we've shown that constructing cohort selection mechanisms that are robust to composition with a family of scoring functions is possible. By framing the problem in terms of robustness, we address the concern that placing requirements on future designs is nearly unenforceable, whereas designing the current stage to be robust to a large class of potential future policies can give much better practical guarantees. Finally, we've shown robust cohort selection mechanisms that compose well with reasonable scoring function families.

In the process of exploring robustness and fairness in pipelines, we uncovered a number of interesting questions for future work. **Policy complexity**: we have considered a set of concise and practical policies in this work, but the trade-off between policy complexity and the expressiveness of cohort selection has not been fully characterized. **Fair Matching**: choosing a cohort is very similar to the problem of assigning an individual to an existing cohort. However, in the traditional matching literature, significant emphasis is placed on individuals' and teams' preferences over placements, rather than external fairness criteria. Is it possible to simultaneously achieve a good matching, in the sense of satisfying preferences or stability, and individual fairness? **Quantifying the tradeoff**: There are significant differences in the difficulty for constructing mechanisms which satisfy the conditional, versus unconditional, notion of robustness. Is it possible to more directly quantify the tradeoff in mechanism expressivity between these two settings? **Different metrics**: Handling different metrics in the pipeline: we considered just one metric throughout the entire pipeline, but using different metrics for different stages of the pipeline may be valid. For example, in the case of promoting an individual contributor to a management position, the metric for "manager" may be different. **Ranking instead of scoring**: although ranking with hard cutoffs does not satisfy individual fairness, it is frequently used in practice. Can the model we have outlined with respect to scoring be translated to ranking, e.g., incorporating the results of [7]?

6 Related Work

There is a wide variety of work concerning fairness in machine learning [9, 14, 24, 17, 3, 18, 25, 22, 10, 16, 15, 11, 12, 20, 19, 4, 23, 5]. Individual fairness was introduced by Dwork et al. [5]. Dwork and Ilvento studied composition of combination of individually fair and group fair classifiers [6]. Two other recent lines of work have also considered composition problems and fair systems. First, several works have studied the problem of feedback loops, in which decisions that previous time steps, such as where to send law enforcement officers, influence outcomes at later time steps potentially unrelated to the original decision [12, 8, 21]. Bower et al. study fairness in a pipeline of decisions under a group-based notion of fairness [1]. They primarily consider the combination of multiple non-adaptive sequential decisions, evaluating fairness at the end of the pipeline. Second, several works have considered competitive scenarios, such as advertising, in which many (potentially fair or unfair) classifiers compete for individuals [2, 13]. Although not explicitly addressing composition, recent work considering fairness in rankings, e.g., [7], also address fairness in a setting in which outcomes, in this case rankings, naturally depend on the outcomes of others.

References

1 Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines. *CoRR*, abs/1707.00391, 2017. arXiv:1707.00391.

2 L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Toward controlling discrimination in online ad auctions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 4456–4465, 2019.

3 Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

4 Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

5 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012.

6 Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 33:1–33:20, 2019.

7 Cynthia Dwork, Michael Kim, Omer Reingold, Guy Rothblum, and Gal Yona. Learning from outcomes: Evidence-consistent rankings. In *60th Annual IEEE Symposium on Foundations of Computer Science November 9-12, 2019, Baltimore, Maryland*, pages 106–125, 2019.

8 Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 160–171, 2018.

9 Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2605–2614, 2018.

10 Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

11 Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

12 Lily Hu and Yiling Chen. Fairness at equilibrium in the labor market. *CoRR*, abs/1707.01590, 2017. `arXiv:1707.01590`.

13 Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 348–358, 2020.

14 Christopher Jung, Michael J. Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *CoRR*, abs/1905.10660, 2019. `arXiv:1905.10660`.

15 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2569–2577, 2018.

16 Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 656–666, 2017.

17 Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4847–4857, 2018.

18 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017.

19 Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981, 2019.

20 Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6196–6200, 2019.

21 Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

22 David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3381–3390, 2018.

23 Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint*, 2017. `arXiv:1706.08519`.

24 Gal Yona and Guy N. Rothblum. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674, 2018.

25 Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 325–333, 2013.

A Extended Motivating Examples

In each example, we consider a universe U comprised of individuals belonging to two groups, a majority group S and a minority group T, such that the majority group is k times as large as the minority group (i.e., $k|T| = |S|$). For the particular employment task in question, there is a known metric $\mathcal{D}$ which specifies who is similar to whom for the purposes of this task. For simplicity, we assume that $\mathcal{D}$ is one-dimensional, i.e., each individual u has a qualification $q_u \in [0,1]$, and $\mathcal{D}(u,v) := |q_u - q_v|$. We assume that S and T have an equal distribution of talents: more specifically, for every qualification level q, there are exactly k times as many individuals with qualification q in S as there are in T. We assume that there is a nontrivial range of qualifications in $[0,1]$, and we will generally assume that the company prefers to hire the most highly qualified candidates, but in order to fill the number of positions open cannot hire only maximally qualified candidates. We use Q_H to refer to the subset of individuals who are highly qualified.

Our examples are based on a set of facially neutral company compensation policies. We now give precise descriptions of these policies in the form of a scoring function, and indicate where the scoring policies must be adjusted to give intra-cohort individual fairness. (As we will see later, even adjusting the policies to be intra-cohort individually fair won't be enough to prevent bad behavior under composition.)

1. **Fixed Bonus Pool**: A fixed pool of bonus money B is assigned to each team and is split between the members of each team, with the highest achieving members receiving larger portions of the pool. More formally, given a cohort of individuals $C = \{x_1, \dots, x_c\}$ of size c with qualifications $\{q_{x_1}, \dots, q_{x_c}\}$, the scoring function f_B assigns a bonus share b_i to each individual x_i such that $\sum_{u \in C} b_u = 1$, optimized to ensure that individuals with higher qualification receive larger bonuses.
   In particular, f_B can either be a simple proportional mechanism, e.g., $f_B(u) \propto q_u$, or it can be optimized for specific goals, e.g., maximizing the difference in compensation between the most and least qualified individuals, creating an even spread of compensations, etc. For example, the company could choose f_B using the following optimization to choose the largest "weighted spread" to maximize the objective of increasing the difference in compensation based on difference in qualification: $argmax_{\{b_u \in [0,1]\}}\{\sum_{u,v \in C}(b_u - b_v)(q_u - q_v)\}$ subject to $|b_u - b_v| \leq |q_u - q_v|$ for all $u, v \in C$ and $\sum_{u \in C} b_u = 1$.
   This optimization will tend to choose bonus shares that maximize the differences in bonuses between individuals with significantly different qualifications within the cohort. Notice that the scoring function has no way of knowing what other cohorts may or may not appear and with what probabilities, and so it only optimizes within the particular cohort C.
2. **Stack Rank**: The bottom 10% of each team may be fired or put on "performance plans".
   Formally, $f(C,u) := \begin{cases} 1 \text{ if } \frac{|\{v | q_u > q_v\}|}{|C|} \leq 0.1, \\ 0 \text{ otherwise} \end{cases}$
   However, this strict cut off violates intra-cohort individual fairness, as two nearly equally qualified individuals might find themselves on opposite sides of the cutoff. Alternatively, we can construct a scoring function which closely approximates the desired policy but still satisfies intra-cohort individual fairness, by optimizing subject to the intra-cohort fairness constraints. For example, taking $\mathbb{O}_u$ to be the indicator that u is in the bottom 10% of the cohort, one could use the following optimization to maximize the probability that

only the bottom 10% are placed on performance plans: $argmax_f \sum_{u \in C} f(C,u)\mathbb{O}_u + (1 - f(C,u))(1 - \mathbb{O}_u)$subject to $|f(C,u) - f(C,v)| \leq |q_u - q_v|$ for all $u, v \in C$. Alternatively, if exactly 10% of the cohort should be put on performance plans, Permute-Then-Classify can be applied or an additional constraint on the expected number of employees placed on performance plans could be added to the optimization above in order to satisfy *intra-cohort* individual fairness.

3. **Equal Treatment**: Each team's bonus is determined by average performance of the team (assumed to be proportional to average quality) and awarded equally to each member. Formally, the scoring function f first chooses the total bonus amount $B_C \propto B \sum_{u \in C} q_u$, and then assigns $b_u = \frac{B_C}{|C|}$ for all $u \in C$. Intra-cohort individual fairness for f is trivial, as every individual is treated equally.
4. **Promotion**: Choose the single most qualified person on the team to promote, based on performance. As in the case of stack ranking, strictly implementing this policy will violate intra-cohort individual fairness, as nearly equal individuals may be treated very differently. As above we can satisfy *intra-cohort* individual fairness by posing the relevant optimization question, and Permute-then-Classify (see Appendix B of the full version) can be used to select exactly one individual for promotion.

We now show that these compensation policies can cause significant unfairness for T when combined with simple hiring protocols. In each case, we state the set of cohorts the company intends to select from, and we assume that the company uses a method similar to the one described in Appendix C.3 of the full version to derive a fair set of weights to use to sample a single cohort in an individually fair way.[21]First, we consider the "packing" hiring protocol.

▶ **Example 31** (Packing)**.** Suppose that in the past, the company had a particular problem retaining employees from the minority group T and in order to address this problem, the company ensures that individuals with high potential from T are always hired together into the same team for mutual support. On the other hand, talented members of S are spread out between the other teams, to make sure that there is at least one highly talented individual on each team. Formally, the company specifies the set of cohorts $\mathcal{C}_{packing} = \{C \in \mathcal{C} \mid (|C \cap T \cap Q_H| > 1 \wedge |C \cap S \cap Q_H| = 0) \oplus (|C \cap T \cap Q_H| = 0 \wedge |C \cap S \cap Q_H| = 1)\}$, where Q_H is the set of highly qualified candidates, and samples a single cohort from the set such that individual fairness is satisfied.

"Packing" results in lower compensation for T for Fixed Bonus Pool, Stack Rank, and Promotion compensation policies

"Packing" causes talented members of T to be on teams of higher average quality than those with talented members of S. As a result, members of T will receive lower bonuses and promoted less often than members of S. Thus, this seemingly beneficial practice can backfire when composed with certain compensation policies.

One may imagine that utilizing a "splitting" strategy, where qualified members of T are separated from other qualified members to increase their chance of "standing out" on teams, would solve this issue.

[21] We omit the details of the method and the particulars of the conditions on the set of cohorts specified as they are easy to fulfill in these settings.

▶ **Example 32** (Splitting)**.** The company chooses teams where highly qualified members of T are always the only highly qualified member of their team, giving them the opportunity to stand out and be recognized for their talent. More formally, the company chooses from the set of cohorts $\mathcal{C}_{splitting} = \{C \in \mathcal{C} \mid (|C \cap T \cap Q_H| = 1 \wedge |C \cap S \cap Q_H| = 0) \oplus (|C \cap T \cap Q_H| = 0 \wedge |C \cap S \cap Q_H| \geq 1)\}$. In each cohort containing a highly qualified member of T, there are no other highly qualified individuals (from either T or S).

Though this policy no longer leads to lower compensation for T for Stack Rank, Fixed Bonus Pool, and Promotion, "Splitting" results in lower compensation for T for Equal Treatment, because the practice causes talented members of T to be on teams of lower average quality than talented members of S. As a result, with Equal Treatment, qualified S will receive greater compensation than qualified T. Splitting can also occur when members of T are primarily hired via outreach. For example, suppose that a company has been trying to form a team to work on a difficult or low prestige task (e.g., Fortran code maintenance). All of the talented candidates in S pass on the job offer because they are confident they can do better, so HR reaches out more aggressively to candidates in T. These candidates may be more willing to take the job because they are less confident about their other options. Thus, even without an explicit policy in place to choose minority candidates to be the singular most qualified member on a less qualified team, these situations can still arise from the interactions between the hiring procedure and the job market.

▶ Remark 33. The motivation for both of these policies could be malicious, and determining whether the stated goals or justifications were legitimate aims of the policy would be difficult.

One may imagine that these issues could be addressed by ensuring that qualified members of T and qualified members of S appearing on teams with similar average quality. However, a malicious company can still cause members of T to receive lower compensation.

▶ **Example 34** (Adversarial ranking)**.** Suppose that the company did not want any member of the T to be chosen for promotion or wished to depress their compensation relative to the members of S. The company decides to choose teams such that, for each team, there is a correspondence between the members of T and S included in the team, such that the members of S are almost always more talented than their counterparts in T. (Given the equal distribution of talents of T and S, there may be an excess member of T that is allowed to be the most qualified, but this is a singular case.) More formally, the company chooses from $\mathcal{C}_{adv.ranking} = \{C \in \mathcal{C} \mid \exists G : C \cap T \to C \cap S \text{ s.t. } \forall u \in C \cap T,\, q_u < q_{G(u)}\}$.

"Adversarial Ranking" is particularly catastrophic for T for Promotion or Stack Ranking if the hard cutoff (not intra-cohort individually fair) versions are used. Although ensuring intra-cohort individual fairness helps, members of T will always be seeing depressed levels of promotion, higher levels of firing, and lower levels of compensation except in the case of Equal Treatment. Thus "Adversarial Ranking" keenly illustrates that average team quality is not sufficient to ensure that individuals are truly being treated fairly in cohort-based pipelines. We stress that Adversarial Ranking can also be efficiently achieved using the procedure described in Appendix C.3 of the full version.

Sample Cohorts

To illustrate these issues, we include Figures 1a and 1b to compare the example scoring functions for a pair of cohorts, demonstrating the issues outlined above.

	Qualification	Fixed Pool Bonus	Equal Bonus
Cohort 1			
Alice	0.8	35	60
Bob	0.7	25	60
Charlie	0.5	5	60
Dan	0.2	0	60
Eve	0.8	35	60
Cohort 2			
Frank	0.8	57	40
George	0.6	36	40
Harriet	0.1	0	40
Ivan	0.2	0	40
Julia	0.3	7	40

(a) Bonus score function comparisons for two cohorts, each containing five individuals of varying qualifications. Cohort 1 has an average qualification of 0.6, and Cohort 2 has an average qualification of 0.4. In the fixed pool bonus, a total pool of 100 is split between the members of the cohorts. The same optimization is used for both cohorts, that is according the maximum possible bonus to the most qualified individual(s). Notice that in Cohort 1, Alice and Eve have to share the top bonus (35 each), but in Cohort 2, Frank doesn't have to split the top bonus (57). Notice also that George and Julia receive higher bonuses than Bob and Charlie, even though they are (much) less qualified. On the other hand, in the equal bonus setting Frank receives a lower bonus than both Alice and Eve, even though he's equally qualified.

	Qualification	Promotion	Stack Rank (IF)	Stack Rank (exact, not IF)
Cohort 1				
Alice	0.8	35%	0	0
Bob	0.7	25%	10%	0
Charlie	0.5	5%	30%	0
Dan	0.2	0	60%	1
Eve	0.8	35%	0	0
Cohort 2				
Frank	0.8	57%	0	0
George	0.6	36%	0	0
Harriet	0.1	0	43%	1
Ivan	0.2	0	33%	0
Julia	0.3	7%	24%	0

(b) Promotion score function comparison of the cohorts from Figure 1a. The promotion policy attempts to maximize the probability of promotion for the most qualified individuals, subject to the individual fairness constraints and that the expected number of promotions is 1. In this case, essentially the same observations apply as in the fixed pool bonus setting. In the case of Stack rank, both cohorts are optimized to maximize the probability of placing the least qualified person on a performance plan. Notice that Dan is much more likely to be placed on a performance plan than the equally qualified Ivan, due to the larger number of less qualified individuals in Cohort 2. Although it might seem that the exact stack rank policy, rather than the individually fair version, would be less likely to have this problem, in fact in this case Dan is still treated differently than Ivan.

Abstracting Fairness: Oracles, Metrics, and Interpretability

Cynthia Dwork
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
Radcliffe Institute for Advanced Study, Cambridge, MA, USA
Microsoft Research, Mountain View, CA, USA
dwork@seas.harvard.edu

Christina Ilvento
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
cilvento@g.harvard.edu

Guy N. Rothblum
Department of Computer Science and Applied Mathematics,
Weizmann Institute of Science, Rehovot, Israel
rothblum@alum.mit.edu

Pragya Sur
Harvard University, Center for Research on Computation and Society, Cambridge, MA, USA
pragya@seas.harvard.edu

Abstract

It is well understood that classification algorithms, for example, for deciding on loan applications, cannot be evaluated for fairness without taking context into account. We examine what can be learned from a *fairness oracle* equipped with an underlying understanding of "true" fairness. The oracle takes as input a (context, classifier) pair satisfying an arbitrary fairness definition, and accepts or rejects the pair according to whether the classifier satisfies the underlying fairness truth. Our principal conceptual result is an extraction procedure that learns the underlying truth; moreover, the procedure can learn an approximation to this truth given access to a *weak* form of the oracle. Since every "truly fair" classifier induces a coarse metric, in which those receiving the same decision are at distance zero from one another and those receiving different decisions are at distance one, this extraction process provides the basis for ensuring a rough form of *metric fairness*, also known as *individual fairness*.

Our principal technical result is a higher fidelity extractor under a mild technical constraint on the weak oracle's conception of fairness. Our framework permits the scenario in which many classifiers, with differing outcomes, may all be considered fair.

Our results have implications for *interpretablity* – a highly desired but poorly defined property of classification systems that endeavors to permit a human arbiter to reject classifiers deemed to be "unfair" or illegitimately derived.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Algorithmic fairness, fairness definitions, causality-based fairness, interpretability, individual fairness, metric fairness

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.8

Related Version A full version of the paper is available at `https://arxiv.org/pdf/2004.01840.pdf`.

Funding *Cynthia Dwork*: This research was supported in part by NSF grant CCF-1763665 and Microsoft Research.
Christina Ilvento: This work was supported by Microsoft Research and the Sloan Foundation.
Guy N. Rothblum: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), the Israel Science Foundation (grant number 5219/17), and Microsoft Research.

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 8; pp. 8:1–8:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Pragya Sur: This research was supported by the Center for Research on Computation and Society, Harvard John A. Paulson School of Engineering and Applied Sciences, and in part by Microsoft Research.

Acknowledgements This research was conducted, in part, while the authors were at Microsoft Research, Silicon Valley.

1 Introduction

Definitions of fairness, for example, in the context of accept/reject classification algorithms, mostly fall into two main categories: group fairness definitions are requirements on various forms of statistical equality in the treatment of disjoint demographic groups; *individual* (or *metric*) fairness requires that individuals that are similar with respect to the classification task at hand should be treated similarly by the classifier. Although intuitively appealing, group fairness definitions suffer from internal inconsistency and incompatibility [4, 3, 21, 1]. (See also [27].)

On the other hand, individual fairness requires a task-specific similarity metric, which may be difficult to find.[1] *Counterfactual Fairness*, proposed by Kusner et al. in 2017 [22], is an approach to capturing individual fairness via *counterfactual reasoning* as put forth by Pearl [28]. Counterfactual fairness seeks to prevent discrimination based on protected attributes, such as race or sexual preference, by requiring that individuals' outcomes "would have been" the same in a counterfactual world in which these attributes have different values. To make such an assertion, the definition relies on a causal model that captures the ways in which these attributes influence other attributes relevant to classification. Thus, to evaluate whether a predictor for loan default is counterfactually fair for sexual orientation, one would construct a causal model reflecting the relationships between sexual orientation and the features weighed by the predictor, and then determine whether the predictions are inappropriately dependent on orientation. In this definition, the causal model replaces the metric as the specification of fairness. The classifier is evaluated for fairness in the context of the causal model just as in metric fairness the classifier is evaluated for fairness in the context of the metric.

As widely noted, and partially addressed in later work (Kilbertus et al., 2019 [18]), this approach suffers from the fact that different data generation models can give rise to the same distribution on outcomes. In particular, a blatantly unfair classifier can satisfy the definition when paired with a suitably contrived model, showing that the choice of model is itself a vector for unfairness.

More generally, the maxim "All models are wrong but some models are useful" highlights the dangers of a fairness definition that evaluates a classifier in the context of a stated model: What are the semantics of having the (model, classifier) pair satisfy the definition when the model is wrong (which is always the case!)?

To complete the counterfactual fairness approach (among others), one might assume the existence of an expert that can judge whether or not a classifier is "truly fair" in a given context. For example, a domain expert may reject a (causal model, classifier) pair for home loan decisions that satisfies the technical definition of counterfactual fairness but in which the model has been contrived to use zip code instead of race in order to obfuscate racial bias. In this work we investigate what can be learned by interacting with such an expert.

[1] See, however, the recent proposal of [11].

We abstract the problem by instantiating "true fairness" (and the expert who knows what this is) via an oracle that holds a collection $\mathcal{T} \subseteq \{0,1\}^{|\mathcal{X}|}$ of vectors specifying the classification outcome for each individual in the universe $\mathcal{X}$ of possible individuals[2]. Each $t \in \mathcal{T}$ corresponds to a classifier that the oracle considers to be fair, at least in some context. As an example, one might imagine that the oracle has access to the true data generation model, and it evaluates classifiers in this single context.

Our principal conceptual result is an (inefficient) *extraction procedure* that learns the underlying truth (collection $\mathcal{T}$) held by the oracle under the assumption that the contexts of interest are of bounded size. Once the assumption is cleanly stated it is not surprising that $\mathcal{T}$ can be extracted by brute force, so this first contribution is the conceptual framing of the problem (Sections 2 and 3). This result makes no assumptions about the set of fair classifiers accepted by the oracle, nor about the particular context(s) that make the oracle accept a classifier. We extract the full set of classifiers for which there exists *some* context that makes the oracle accept.

Under the assumption that counterfactual fairness (or any other causality-based definition, such as path-specific effects [26]), combined with the true causal model (or an appropriate approximation), genuinely captures fairness, our results imply that one can extract, from an oracle with access to the true model, a coarse metric for individual fairness. This holds because every classifier induces a coarse metric in which those receiving positive decisions are at distance zero from one another, and similarly for those receiving negative decisions. (See Remark 1.)

We then turn to *weak* oracles, which solve a more relaxed promise problem. Each weak oracle $\tilde{\mathcal{O}}$ is a relaxation, based on a given notion of closeness of classifiers, of a strong oracle, $\mathcal{O}$. Weak oracles always accept the (context, classifier) pairs accepted by their strong counterparts, but only reject (context, classifier) pairs where the classifier is "far" from an accepted classifier for the given distance notion.

We consider two types of closeness in defining weak oracles: Hamming distance, where the reconstruction problem is straightforward, provided the members of $\mathcal{T}$ are sufficiently separated[3], and an asymmetric *transportation cost* $\mathcal{C}$ that does not satisfy the triangle inequality. Our transportation cost is closely related to individual fairness: $\mathcal{C}(t \to c)$ captures the number of pairs of individuals that are treated similarly in t but differently in c. In essence, the transportation cost notion requires *less* of the oracle: δ-weak oracles[4] with this notion of distance may not know *how* individuals should be treated for the task at hand, but may have a sense of who should be treated similarly to whom. This lack of decisiveness on the part of the oracle makes extraction much more difficult. Not only does it lead to a transportation cost that is not even a distance function, but it also limits what can possibly be extracted even if $\delta = 0$: under this notion, the distance between a classifier and its complement is 0! [5] In consequence, rather than aiming to extract the set of fair classifiers, we extract a set of fair partitions, where each partition specifies which individuals are similar to each other. The partition can also be viewed as a coarse metric.

[2] We focus on the case of binary, deterministic classifiers. Such a classifier can only satisfy individual fairness if for all individuals u, v, $d(u,v) \in \{0,1\}$, where d is the task-specific metric. In Remark 1 we discuss amplification of this technique to a richer class.

[3] Much as it is possible to learn a mixture of Gaussians provided the means are sufficiently far apart.

[4] Oracles only guaranteed to reject classifiers at distance greater than δ from all $t \in \mathcal{T}$.

[5] Our techniques apply to a symmetrized version of $\mathcal{C}(t \to c)$, defined by the fraction of pairs of individuals that disagree between t and c (Section 3). This case is, in fact, easier than the transportation cost.

Our principal technical result is a high fidelity extractor in the transportation cost model, under a mild technical constraint on the weak oracle's conception of fairness. For $t \in \mathcal{T}$, define $t^{\text{flip}}(x) = 1 - t(x)$ for all $x \in \mathcal{X}$. The assumption is: for $t \in \mathcal{T}$ for which the weak oracle rejects t^{flip}, it also rejects all classifiers very close (in Hamming distance) to t^{flip}.

Interpretability

Our results have implications for *interpretablity* – a highly desired but poorly defined property of classification systems that endeavors to permit a human arbiter to reject classifiers deemed to be "unfair" or illegitimately derived. If "interpretability" permits a knowledgeable human to distinguish truly fair from truly unfair classifiers, then there is a procedure to extract from the human information a measure of similarity for pairs of individuals. Roughly speaking, we can get our hands on a metric, even when the closeness notion for classifiers is the Hamming distance on their vector representation, which is unrelated to metric fairness!

▶ Remark 1. In this work, the "metric" we extract from the oracle is crude: all distances are either 0 or 1. Metrics of this type can be amplified to yield a richer class of metrics by considering a collection of oracles with varying tolerance for unfairness. For example, given a metric $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$, we can instantiate k approximations $\{d'_1, \ldots, d'_k\}$ of d such that $d'_i(u, v) := 1$ if $d(u, v) > \frac{k}{i}$ and 0 otherwise. Given access to an oracle for each threshold, we can apply the extraction procedure multiple times to (approximately) recover this set of $\{0, 1\}-$metrics. The recovered collection can then be combined to form an approximation of d, using the threshold combination procedure developed in [11]. See also [8, 12, 14] for demonstrations of the usefulness of coarse metrics.

Related Work

There is a vast literature on algorithmic fairness. The theory of algorithmic fairness was first studied by Dwork et al. in 2012 [4]. In addition to defining individual fairness, this work noted that sensitive attributes may be holographically embedded in the data, showed the benefits of utilizing, rather than trying to suppress, the sensitive information; showed the power of Individual Fairness when given a metric; examined the group fairness property of demographic parity and gave examples motivating its dismissal as a fairness solution concept, and provided a metric-based approach to Fair Affirmative Action. Earlier work suggested concrete approaches based on training on a modified dataset in which the proportion of positive labels is equal in disjoint demographic groups, in the hopes that a classifier trained on these new labels will imbibe the group fairness properties of the training data [29, 15]. A second approach added a regularization term to the classification training objective to quantify the degree of bias or discrimination [16, 2]. Subsequent work saw heavy investment in algorithms satisfying group-based criteria, even in the face of the negative results about the compatibility of natural group fairness objectives [27, 3, 21, 17, 5]. Individual fairness, predicated on access to a similarity metric, proceeded more slowly, although the literature contains several works extending the theory [5, 30, 8, 20]. Recent work [11] combines insights from HCI and computational learning theory to learn an approximation to a metric known to a human arbiter with surprisingly few queries. An intriguing "middle ground" enforces *calibration* (in the case of scoring functions [10]) *simultaneously* for large numbers of intersecting subpopulations (see [6] for a treatment of fair *rankings* in this setting). A variant of the multiple intersecting groups approach [17] enforces *Equalized Odds* [9] among all pairs of groups simultaneously. An economics justification for Equalized Odds is put forth in [13]. Equalized Odds and related candidate fairness criteria are criticized through the lens of graphical models [1].

Still other work employs deep learning to build *fair representations* of individuals that, speaking intuitively, retain much useful information for classification or even transfer learning, but "screen out" sensitive demographic information [31, 7, 25]. Finally, there is also a vast literature on *interpretability.* See [24] for a discussion of what this might mean (and hurdles to be overcome); the course notes of Lakkaraju [23] contain a wealth of examples and references for this literature.

Our work was inspired by the elegant proposal of Counterfactual Fairness by Kusner, Loftus, Russell, and Silva [22]. A related definition of fairness concentrates on path-specific effects [26] (see also [19]). Kilbertus et al. design tools to assess the sensitivity of fairness measures to unmeasured confounding for a popular class of noise models [18].

Organization

The rest of this paper is organized as follows. Section 2 introduces the definitions used in this work. Section 3 states the main contributions and motivates our use of oracles. Section 4 describes our algorithms, and Section 5 concludes with a discussion of necessary assumptions.

2 Definitions

We consider a universe $\mathcal{X}$ of individuals, each represented by a vector of p attributes. We will assume each vector of attributes represents a unique individual. Determining whether or not the representation of the individuals is sufficient to permit fair classification is a fascinating topic beyond the reach of this paper; here we assume an affirmative answer. Since our work may be viewed as negative results, this assumption only strengthens the contribution.

A *classifier* maps individuals to $\{0,1\}$, $C : \mathcal{X} \to \{0,1\}$. It is often convenient to think of classifiers as vectors $c \in \{0,1\}^{|\mathcal{X}|}$, with $c_i \in \{0,1\}$ being the classification of the ith individual in some canonical ordering. We completely identify an individual and its index i, so we will often write $i \in \mathcal{X}$ to denote the ith individual in this ordering.

It is sometimes convenient to think of a classifier as partitioning $\mathcal{X}$ into two groups according to their classification outcomes. Unless otherwise specified, we use lower case letters to denote classifiers and the corresponding upper case letter to denote the partition. For a classifier c, we let $c^0 = \{i \in \mathcal{X} | c_i = 0\}$ and $c^1 = \{i \in \mathcal{X} | c_i = 1\}$. We sometimes refer to c^0 as the *Left Hand Side* of the partition c, denoted $\mathcal{LHS}(c)$, and c^1 as the *Right Hand Side*, denoted $\mathcal{RHS}(c)$. The *flip* of a partition is a swap of its left and right sides; in vector form, $c^{\text{flip}} = 1 - c$, *i.e.*, $\forall i \in \mathcal{X}, c_i^{\text{flip}} = 1 - c_i$. Constant classifiers have the property that for some $v \in \{0,1\}$, $c_i = v$, $\forall i \in \mathcal{X}$.

It is also sometimes convenient to think of individuals in $\mathcal{X}$ as vertices, and to think of the classifier as a two-coloring of the complete graph on $\mathcal{X}$ (see Figure 1). Monochromatic edges indicate pairs of individuals who are treated the same by the classifier.

Contexts and Valid Pairs

Many fairness notions require that classifiers be considered in some form of *context.* For example, in the case of counterfactual fairness the context is given by a causal model [6]. We therefore abstract the notion of a fairness definition $\mathcal{W}$ as a set of (context, classifier) pairs.

[6] See [1] for a general discussion of the need for context.

▶ **Definition 2** (validity). *If $(\mathcal{A}_{\mathcal{W}}, c) \in \mathcal{W}$ then $(\mathcal{A}_{\mathcal{W}}, c)$ is said to be a valid pair* under $\mathcal{W}$. *In our work $\mathcal{W}$ is typically fixed, in which case we may simply refer to* valid pairs.

Boundedness

We assume there is a procedure for enumerating all contexts, whose running time is a fixed function of $|\mathcal{X}|$. For example, we might consider the case in which the context is given by a causal graph constrained to have a number of vertices linear in p (the number of attributes) and the functions computed at each vertex can be described by circuits of size polynomial in $|\mathcal{X}|$. We note that without this assumption it is not even clear how to represent a context $\mathcal{A}_{\mathcal{W}}$ for the purposes of determining whether or not some $(\mathcal{A}_{\mathcal{W}}, c) \in \mathcal{W}$.

Oracles

We view the fairness definition $\mathcal{W}$ as a filter, and hypothesize the existence of an *oracle* to rule on the acceptability of valid pairs. A useful intuition, for example, with counterfactual fairness in mind, is that the oracle knows the true data generation model $\mathcal{A}_{\mathcal{W}}{}^*$, and is willing to accept exactly valid pairs $(\mathcal{A}_{\mathcal{W}}{}^*, c) \in \mathcal{W}$; alternatively, the oracle may be willing to accept valid pairs $(\mathcal{A}_{\mathcal{W}}, c)$ whenever $\mathcal{A}_{\mathcal{W}}$ enjoys certain properties. However, we make no explicit assumptions: Formally, the oracle is specified by a subset of $\mathcal{W}$. It takes as input a valid pair $(\mathcal{A}_{\mathcal{W}}, t) \in \mathcal{W}$ and either accepts $(\mathcal{O}(\mathcal{A}_{\mathcal{W}}, t) = 1)$ or rejects $(\mathcal{O}(\mathcal{A}_{\mathcal{W}}, t) = 0)$.

▶ **Definition 3** (Strong Oracle). *A strong oracle is completely specified by the valid pairs that it accepts.*

It is convenient to name the collection of classifiers associated with acceptance by the strong oracle, that is, to define $\mathcal{T} = \{t \in \{0,1\}^{|X|} \mid \exists \mathcal{A}_{\mathcal{W}} : \ \mathcal{O}(\mathcal{A}_{\mathcal{W}}, t) = 1\}$.

Weak Oracles

Every weak oracle is a relaxation of a strong oracle. Weak oracles differ from their corresponding strong oracles by relaxation of the conditions for acceptance: weak oracles will accept whatever the associated strong oracles accept, but may also accept valid pairs.

▶ **Definition 4** (δ-Weak Oracle for Hamming distance). *Fix an arbitrary strong oracle $\mathcal{O}$ with associated classifiers $\mathcal{T}$. For $\delta > 0$ we say that $\mathcal{O}_{\mathcal{H}}$ is a* δ-weak oracle relaxation of $\mathcal{O}$, *based on the Hamming distance, if*

1. *$\mathcal{O}_{\mathcal{H}}$ accepts all valid pairs accepted by $\mathcal{O}$;*
2. *$\mathcal{O}_{\mathcal{H}}$ rejects valid pairs whose classifiers are far (in Hamming distance) from all classifiers in $\mathcal{T}$: Let $(\mathcal{A}_{\mathcal{W}}, c) \in \mathcal{W}$. If $\forall t \in \mathcal{T}, d_{\mathrm{H}}(c,t) > \delta$, then $\tilde{\mathcal{O}}(\mathcal{A}_{\mathcal{W}}, c) = 0$. Here, for $u, v \in \{0,1\}^{|\mathcal{X}|}$, $d_{\mathrm{H}}(u,v) := \{i \in \mathcal{X} \mid u_i \neq v_i\}$.*

On the remaining valid pairs, $\mathcal{O}_{\mathcal{H}}$ may behave arbitrarily.

The definition of a weak oracle *based on transportation cost* requires one additional concept.

▶ **Definition 5** (δ-faithfulness). *For $c, t \in \{0,1\}^{|\mathcal{X}|}$, we say that c is* δ-faithful *to t if*

$$\frac{1}{\binom{n}{2}} \sum_{i \neq j} \mathbf{1}\{t_i = t_j, c_i \neq c_j\} \leq \delta. \tag{1}$$

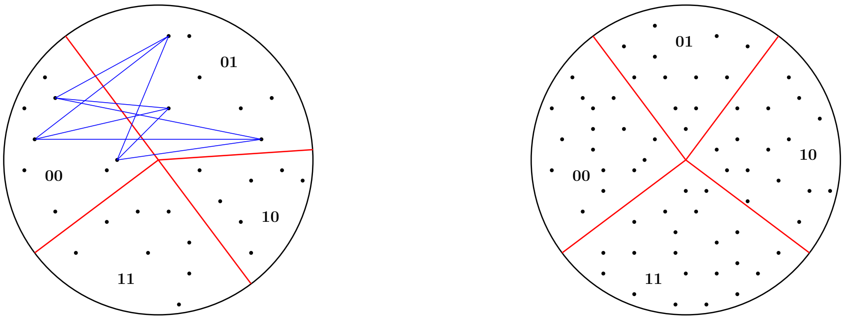


Figure 1 (Left) The (complete) labeled graph $G^{u\rightarrow v}$ for an ordered pair of classifiers (u, v). There is a vertex for each $i \in \mathcal{X}$. Vertices are labeled with two-bit strings indicating their classifications under u (first bit) and v (second bit). So vertices i in the quadrant $G_{00}^{u\rightarrow v}$ have $u_i = v_i = 0$, vertices in $G_{01}^{u\rightarrow v}$ have $u_i = 0$ and $v_i = 1$, and so on. The transportation cost $\mathcal{C}(u \rightarrow v)$ is captured by the number $|G_{00}^{u\rightarrow v}| \cdot |G_{01}^{u\rightarrow v}|$ of edges between the 00 and 01 quadrants (a few drawn in blue on left) plus the number of edges between the 11 and 10 quadrants (none drawn). For example, if $i \in G_{00}^{u\rightarrow v}$ and $j \in G_{01}^{u\rightarrow v}$ this says that the edge (i, j) was monochromatic (both vertices colored zero) in u but is polychromatic in v (because $v_i = 0 \neq v_j = 1$). (Right) Illustration of Assumption 14(2): all quadrants are substantial.

Note that faithfulness is not symmetric. Typically, we will consider faithfulness when c is a candidate classifier and t is an element of the set $\mathcal{T}$ associated with an oracle. δ-faithfulness suggests a natural transportation cost capturing the answer to the question, "Starting from t, how many monochromatic edges in t do we need to "break" when we transition to c?" We let $\mathcal{C}(t \rightarrow c)$ denote this transportation cost (Figure 1). This transportation cost is asymmetric and does not satisfy the triangle inequality.

▶ **Definition 6** (δ-neighborhood). *The* δ-neighborhood *of a classifier* $t \in \{0,1\}^{|\mathcal{X}|}$*, denoted* $\Gamma_\delta(t)$*, is the set of all* $c \in \{0,1\}^{|\mathcal{X}|}$ *such that* c *is* δ*-faithful to* t*.*

▶ **Definition 7** (δ-Weak Oracle for transportation cost). *Fix an arbitrary strong oracle* $\mathcal{O}$ *with associated classifiers* $\mathcal{T}$*. For* $\delta > 0$ *we say that* $\tilde{\mathcal{O}}$ *is a* δ-weak oracle relaxation of $\mathcal{O}$*, based on the transportation cost* $\mathcal{C}$*, if*

1. $\tilde{\mathcal{O}}$ *accepts all valid pairs accepted by* $\mathcal{O}$*;* $\forall(\mathcal{A}_\mathcal{W}, c)$ *such that* $\mathcal{O}(\mathcal{A}_\mathcal{W}, c) = 1$, $\tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 1$*;*
2. $\tilde{\mathcal{O}}$ *rejects valid pairs whose classifiers are far (in transportation cost) from all classifiers in* $\mathcal{T}$*: Let* $(\mathcal{A}_\mathcal{W}, c) \in \mathcal{W}$*. If* $\forall t \in \mathcal{T}, c \notin \Gamma_\delta(t)$*, then* $\tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 0$*.*

There are no further constraints on oracles other than being deterministic. Note that there may be many weak oracle relaxations of a given strong oracle $\mathcal{O}$.

3 Main Contributions

Our principle contributions are extraction procedures that recover the underlying truth held by oracles. Recall that a strong oracle is associated with a set $\mathcal{T}$ of classifiers. Formally, an *extraction procedure* is a program that, using only access to an oracle $\mathcal{O}$, outputs the list $\mathcal{T}$ associated with $\mathcal{O}$. Intuitively, one may imagine that this set arises from some ground truth provided by the concept $\mathcal{W}$. To illustrate, let $\mathcal{W}$ be the notion of counterfactual fairness and $\mathcal{M}$ the true causal model explaining actual functional relationships between all the relevant variables for the task. An oracle may believe that all classifiers that satisfy the counterfactual fairness definition with respect to this "true" causal model are indeed truly fair. Then, $\mathcal{T}$ equals the set of all counterfactually fair classifiers with respect to $\mathcal{M}$. The data analysts have no knowledge of $\mathcal{M}$ whatsoever, but hope to learn about fair classifiers by interacting with the oracle. We provide algorithms that achieve this goal – starting with the simpler case of the strong oracle, subsequently moving on to weak oracles.

Recall that every strong oracle $\mathcal{O}$ has an associated set $\mathcal{T}$ of classifiers, such that $\forall t \in \mathcal{T}$, $\exists \mathcal{A}_{\mathcal{W}}$ with $(\mathcal{A}_{\mathcal{W}}, t) \in \mathcal{W}$ and $\mathcal{O}(\mathcal{A}_{\mathcal{W}}, t) = 1$; $\mathcal{O}$ rejects all valid pairs $(\mathcal{A}_{\mathcal{W}}, c)$ with $c \notin \mathcal{T}$. To begin with, we establish the following.

▶ **Theorem 8.** *For any fairness notion $\mathcal{W}$ satisfying the boundedness condition, and for any strong oracle $\mathcal{O}$ accepting a subset of $\mathcal{W}$, there exists an extraction procedure interacting with $\mathcal{O}$ whose running time is bounded by a function of $|\mathcal{X}|$. The output of the extraction procedure is the set $\mathcal{T}$ associated with $\mathcal{O}$.*

Under the assumption of bounded length contexts, Theorem 8 can be achieved simply via exhaustive search, since our extraction procedures are allowed to be inefficient. The primary contribution of this result is thus conceptual – that it is feasible to extract the ground truth from $\mathcal{O}$ under our framing of the problem. In the spirit of prior examples, if $\mathcal{W}$ is counterfactual fairness and $\mathcal{T}$ the set of all counterfactually fair classifiers with respect to the true causal model (which we do not have any access to, but let's say the oracle has complete knowledge about), then in principle one can learn all of these classifiers. Note that each of these is equivalent to a partitioning of the universe, and can therefore be viewed as a metric (albeit a simple one). Intuitively, one can interpret the oracle as a highly knowledgeable human expert with a deep understanding of the true underlying relationships between the variables relevant for the task, but who is unable to enunciate them – however, the expert is able to tell whether a classifier is fair or not by "looking" at it. Our result demonstrates that, given access to such an expert, a systematic strategy can successfully learn all the fair classifiers. Thus, in settings where learning the true causal model is extremely hard (if not impossible), and hence, reliably implementing counterfactual fairness (or any other causality-based notion) may be out of scope, our results suggest that developing efficient query models to interact with human experts suffices for fair classification, since these directly learn metric information from the expert (recall Remark 1).

While it is helpful to think of an all-knowing expert, who can accurately identify fair classifiers and task-appropriate contexts, our framework can be applied to *any* expert. We can also extract from an imperfect expert, *e.g.* one who can only reason about simple contexts, and accepts a subset of the fair classifiers (or even accepts some unfair ones!). The better the expert, the better the classifiers (or metrics) we extract will be.

Weak Oracles

A weak oracle accepts every valid pair accepted by a strong oracle; in addition, it rejects valid pairs whose classifiers are far from all classifiers in $\mathcal{T}$. However, a weak oracle may behave arbitrarily on the remaining pairs. Nonetheless, we are able to extract even when we do not know how the oracle will behave on these remaining pairs, and this is a strength of our framework. Since the oracle only provides fuzzy information, in the sense that it may behave at will on several pairs, we can only hope to recover $\mathcal{T}$ up to some error. Different notions of distance that determine what should be judged "far" lead to different instantiations of the weak oracle. The conversation around the right notion of distance lies beyond the scope of this paper. Here, we consider two notions, Hamming distance and transportation cost, and provide algorithms in each case that approximately recover $\mathcal{T}$.

Extraction from a Hamming distance based weak oracle

To begin with, we consider a natural distance measure – the Hamming distance. Recall that any weak oracle is a relaxation of a strong oracle $\mathcal{O}$ with an associated set $\mathcal{T}$ of classifiers. A weak oracle $\tilde{\mathcal{O}}$ based on the Hamming distance accepts any valid pair accepted by $\mathcal{O}$, and rejects a valid pair $(\mathcal{A}_{\mathcal{W}}, c)$ for which $d_{\mathrm{H}}(c,t) > \delta$ for every $t \in \mathcal{T}$, and otherwise behaves arbitrarily. Recovery of an approximation to the elements in $\mathcal{T}$ is then trivial (via exhaustive search) as long as any two elements $t, u \in \mathcal{T}$ satisfy $d_{\mathrm{H}}(t,u) > 4\delta$. (Details omitted.)

Extraction from a transportation cost based weak oracle

Our primary technical contribution is an extraction algorithm that approximately recovers elements of $\mathcal{T}$ from a weak oracle based on the transportation cost $\mathcal{C}(t \to c)$ (Definition 7). Theorem 9, stated next, says that the Sharp Extraction Algorithm (Algorithm 3, Section 4) produces a list of classifiers, each of which corresponds to a unique member of $\mathcal{T}$. Recall that for any classifier c, $c^0 = \{i \in \mathcal{X} | c_i = 0\}$ and $c^1 = \{i \in \mathcal{X} | c_i = 1\}$.

▶ **Theorem 9.** *Suppose $\mathcal{O}$ is a strong oracle with an associated set $\mathcal{T}$ of classifiers, and $\tilde{\mathcal{O}}$ is a δ-weak relaxation of $\mathcal{O}$ under the transportation cost $\mathcal{C}(t \to c)$ (Definition 7). Then under Assumption 14, the list of classifiers $(P_1, \ldots, P_m, Q_1, \ldots, Q_m)$ obtained from Sharp Extraction (Algorithm 3, Section 4) satisfies the following: Fix any index j. There exists $t \in \mathcal{T}$ such that*

$$P_j{}^0 \subset t^0 \quad (\mathit{or} \;\; t^1) \qquad \mathit{and} \qquad Q_j{}^1 \subset t^1 \quad (\mathit{resp.}\; t^0), \tag{2}$$

simultaneously,

$$\begin{aligned} |P_j{}^1 \cap t^0| \quad (\mathit{resp.} \;\; t^1) &\leq \left(\frac{\tau_j - \sqrt{\tau_j^2 - 2\delta}}{2} \right) n, \quad \mathit{and} \\ |Q_j{}^0 \cap t^1| \quad (\mathit{resp.} \;\; t^0) &\leq \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 - 2\delta}}{2} \right) n. \end{aligned} \tag{3}$$

Above, $\tau_j = |t^0|/n$ and $\tilde{\tau}_j = 1 - \tau_j$. For classifiers P_j, Q_j and P_k, Q_k with different indices, the corresponding elements of $\mathcal{T}$ that satisfy the aforementioned property are also different.

Sharp Extraction precisely pins down the elements of $\mathcal{T}$ up to a small error margin as specified by (3). The smaller the value of δ, the lower the overall error. In general, for every $t \in \mathcal{T}$, Sharp Extraction recovers nearly as many members of t^0 and t^1 as possible (without recovering the exact classification outcomes). To see this, observe that the fraction of pairs of individuals that P_j erroneously splits in two groups, when they belong to the same group in the underlying element of $\mathcal{T}$, can be bounded by

$$\left(\frac{\tau_j - \sqrt{\tau_j^2 - 2\delta}}{2} \right) \left(\frac{\tau_j + \sqrt{\tau_j^2 - 2\delta}}{2} \right) + \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 - 2\delta}}{2} \right) \left(\frac{\tilde{\tau}_j - \sqrt{\tilde{\tau}_j^2 + 2\delta}}{2} \right) = \delta.$$

Since $\tilde{\mathcal{O}}$ is a δ-weak relaxation of $\mathcal{O}$, intuitively, one cannot hope to accurately cluster additional pairs of individuals from this weak oracle model, suggesting that Sharp Extraction achieves the best we may hope for in such a setting.

Extraction from a weak oracle based on a symmetrized transportation cost

Extraction algorithms can also be developed for weak oracles based on the symmetrized version of the transportation cost

$$\mathcal{C}^s(u \leftrightarrow v) := \frac{1}{\binom{n}{2}} \sum_{i \neq j} \left[\mathbf{1}\{u_i = u_j, v_i \neq v_j\} + \mathbf{1}\{u_i \neq u_j, v_i = v_j\}\right]. \tag{4}$$

A weak oracle $\tilde{\mathcal{O}}$ based on this notion accepts any valid pair accepted by its associated strong oracle $\mathcal{O}$, and rejects any valid pair $(\mathcal{A}_{\mathcal{W}}, c)$ for which $\mathcal{C}^s(t \leftrightarrow c) > \delta$ for every $t \in \mathcal{T}$. This case is, in fact, easier to handle than the asymmetric version. Thus, algorithms that work for weak oracles based on the asymmetric transportation cost can be simplified to suit the needs of a weak oracle based on the symmetrized version (4).

Conclusion

Classification algorithms cannot be evaluated for fairness without taking context into account. Several works in the fairness literature posit the existence of fair and wise human judges, and the ability of humans to recognize unfairness when they see it seems to be a linchpin of interpretability. We have explored what can be learned from a fairness oracle that evaluates (context, classifier) pairs satisfying a definition of fairness, accepting or rejecting according to a hypothesized fairness "truth". The oracle abstraction captures any human judge, or algorithm, or benchmark test; the extraction procedures described here do not need to "understand" the oracle's decisions. Even so, the procedures produce rudimentary metrics for the classification task at hand. The procedure can be amplified to improve the expressive power of the metric. These existence proofs are evidence for the conjecture that a metric is *always* at the heart of fairness.

Metrics can be combined with arbitrary loss functions to obtain individually fair classifiers satisfying a wide range of objectives [4]. Metrics learned on a sample of the population can sometimes be generalized to unseen examples [11]. An *efficient* metric extraction procedure would mean that it is essentially no harder to find a metric than to build good causal models and accompanying classifiers. This is an exciting direction for future research.

4 Extraction Algorithms

This section summarizes our extraction algorithms and key ingredients used therein.

▶ **Definition 10** (δ-Balanced classifier). *A classifier c is said to be* δ-balanced *if both $|c^0| > \sqrt{2\delta}|\mathcal{X}|$ and $|c^1| > \sqrt{2\delta}|\mathcal{X}|$.*

We let $\mathcal{B}$ denote the set of all δ-balanced classifiers of $\mathcal{X}$; henceforth, we simply call these the balanced classifiers.

Two classifiers are said to be aligned if they have relatively few disagreements. Recall that, for a classifier $c \in \{0,1\}^{|\mathcal{X}|}$, $c^0 = \{i \,|\, c_i = 0\}$ and c^1 is defined analogously.

▶ **Definition 11** (Close alignment). *Classifiers p and q in $\{0,1\}^{|\mathcal{X}|}$ are* in close alignment *if $|p^0 \cap q^1|, |q^0 \cap p^1|| \leq \sqrt{\frac{\delta}{2}}|\mathcal{X}|$.*

Furthermore, define the following sets (Figure 1) for any $v, c \in \{0,1\}^{|\mathcal{X}|}$,

$$G_{00}^{v \to c} = \{i \,|\, v_i = 0, c_i = 0\}, \qquad G_{01}^{v \to c} = \{i \,|\, v_i = 0, c_i = 1\}, \tag{5}$$

$$G_{10}^{v \to c} = \{i \,|\, v_i = 1, c_i = 0\}, \qquad G_{11}^{v \to c} = \{i \,|\, v_i = 1, c_i = 1\}. \tag{6}$$

The proof of Theorem 9 and the description of the algorithms require the following lemma.

▶ **Lemma 12.** *Let u, v be arbitrary balanced classifiers accepted by the weak oracle $\tilde{O}$ in Theorem 9. Then exactly one of the following must hold:*

1. *There exists $t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. In this case,*

$$\min\{|G_{00}^{u\to v}|, |G_{01}^{u\to v}|\} \leq \sqrt{2\delta}n \leq \min\{|G_{10}^{u\to v}|, G_{11}^{u\to v}\}. \qquad \text{Situation A}$$

2. *There does not exist any $t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. In this case, there must exist $t_1, t_2 \in \mathcal{T}$ such that $t_1 \neq t_2^{\text{flip}}$, $u \in \Gamma_\delta(t_1), v \in \Gamma_\delta(t_2)$, and that at least one of the following holds*

$$\min\{|G_{00}^{u\to v}|, |G_{01}^{u\to v}|\} > \sqrt{2\delta}n, \qquad \min\{|G_{10}^{u\to v}|, G_{11}^{u\to v}\} > \sqrt{2\delta}n. \qquad \text{Situation B}$$

Our main algorithm, Sharp Extraction, builds on Fuzzy Extraction, presented in Algorithm 1.

Informal Description of Fuzzy Extraction Algorithm

The fuzzy extraction algorithm seeks to associate candidate classifiers c with elements of $\mathcal{T}$, roughly, guided by the transportation cost $\mathcal{C}(t \to c)$. In particular, the algorithm aims to recover $\Gamma_\delta(t)$ for each $t \in \mathcal{T}$. As with the reconstruction from a strong oracle, by the boundedness requirement for contexts, we can find $\boldsymbol{V} = \{c \in \{0,1\}^{|\mathcal{X}|} \mid \exists \mathcal{A}_\mathcal{W} : \tilde{\mathcal{O}}(\mathcal{A}_\mathcal{W}, c) = 1\}$ by enumeration. The algorithm starts by finding $\boldsymbol{V}$. The algorithm then prunes out all unbalanced classifiers, setting $\boldsymbol{V}_\mathcal{B} = \boldsymbol{V} \cap \mathcal{B}$. This is the starting point for recovering $\mathcal{T}$, the classifiers associated with the strong oracle $\mathcal{O}$ of which $\tilde{\mathcal{O}}$ is a relaxation.

At a high level, Fuzzy Extraction works as follow: for an arbitrary pair $u, v \in \boldsymbol{V}_\mathcal{B}$, the algorithm checks whether u and v are both in $\Gamma_\delta(t)$ for some $t \in \mathcal{T}$. From Lemma 12, this can be detected simply by inspection, that Situation A holds. If so, the algorithm clusters u and v into the same group, and otherwise, to different groups. In this manner, the algorithm builds up a collection of sets, each of which contains classifiers that are all in $\Gamma_\delta(t)$ for some $t \in \mathcal{T}$.

In addition, the algorithm takes special care to track when $u, v \in \boldsymbol{V}_\mathcal{B}$ are in close alignment (Definition 11; roughly speaking, they are closely aligned if they are close in Hamming distance). Using this information, the algorithm builds a collection of sets, which we call *orbits*, such that each set contains classifiers in close alignment with some $t \in \mathcal{T}$ (or with t^{flip}, where $t \in \mathcal{T}$). Thus, for each $t \in \mathcal{T}$, Fuzzy Extraction produces (1) an orbit containing t and elements in close alignment with t, and, (2) an orbit consisting of elements in close alignment with t^{flip} (but not necessarily containing t^{flip}).

Implications of Fuzzy Extraction Algorithm

The Fuzzy Extraction algorithm teases apart whether any two balanced accepted classifiers belong to the δ-neighborhood of the same or different elements of $\mathcal{T}$. Note that, $\tilde{\mathcal{O}}$ provides relatively vague information – there could be a large number of valid pairs on which $\tilde{\mathcal{O}}$ behaves arbitrarily. In the full paper, we establish that despite such imprecise information, Fuzzy Extraction distinguishes δ-neighborhoods of different elements of $\mathcal{T}$ successfully, and recovers all balanced accepted members of the neighborhoods.

Algorithm 1 Fuzzy Extraction.

```
input  : The universe $\mathcal{X}$ and an oracle $\tilde{\mathcal{O}}$ .
output : A collection of $2|\mathcal{T}|$ disjoint subsets of $\{0,1\}^{|\mathcal{X}|}$.
Find $\boldsymbol{V} = \{c \in \{0,1\}^{|\mathcal{X}|} \mid \exists \mathcal{A}_{\mathcal{W}} : \tilde{\mathcal{O}}(\mathcal{A}_{\mathcal{W}}, c) = 1\}$. Construct $\boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V} \cap \mathcal{B}$. Set $\ell = 1$;
while $\boldsymbol{V}_{\mathcal{B}} \neq \phi$ do
    Choose a classifier $c_\ell$ from $\boldsymbol{V}_{\mathcal{B}}$ and set $\mathbf{Orb}_\ell := \{c_\ell\}$. If $c_\ell^{\text{flip}} \in \boldsymbol{V}_{\mathcal{B}}$, set $\mathbf{Orb}'_\ell = \{c_\ell^{\text{flip}}\}$, else set $\mathbf{Orb}'_\ell = \phi$;
    for $c \in \boldsymbol{V}_{\mathcal{B}} \backslash \{c_\ell\}$ do
        if CheckSituationA( $c_\ell, c$ ) =“Situation A holds” then
            if $G_{00}^{c_\ell \to c} > \sqrt{2\delta}\mathcal{X} \geq G_{01}^{c_\ell \to c}$ then
                update
                    $$\mathbf{Orb}_\ell = \mathbf{Orb}_\ell \cup \{c\},\ \ \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c\},$$
                and if $c^{\text{flip}} \in \boldsymbol{V}_{\mathcal{B}}$, update
                    $$\mathbf{Orb}'_\ell = \mathbf{Orb}'_\ell \cup \{c^{\text{flip}}\}, \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c^{\text{flip}}\};$$
            else
                update
                    $$\mathbf{Orb}'_\ell = \mathbf{Orb}'_\ell \cup \{c\},\ \ \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c\},$$
                and if $c^{\text{flip}} \in \boldsymbol{V}_{\mathcal{B}}$, update
                    $$\mathbf{Orb}_\ell = \mathbf{Orb}_\ell \cup \{c^{\text{flip}}\}, \boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c^{\text{flip}}\}.$$
            end
        end
    end
    Set $\boldsymbol{V}_{\mathcal{B}} = \boldsymbol{V}_{\mathcal{B}} \backslash \{c_\ell\}$, $\ell = \ell + 1$;
end
Return $\mathbf{Orb}_1, \ldots, \mathbf{Orb}_{\ell-1}, \mathbf{Orb}_1, \ldots, \mathbf{Orb}'_{\ell-1}$, and additional sets
 $\mathbf{Orb}_\ell = \{(c_i = 1, \forall i \in \mathcal{X})\}$ $\mathbf{Orb}'_\ell = \{(c_i = 0, \forall i \in \mathcal{X})\}$, whenever a constant
 classifier is in $\boldsymbol{V}$.
```

Algorithm 2 CheckSituationA: Checks whether Situation A from Lemma 12 holds.

```
input  : An ordered pair of classifiers $(c, u) \in \{0,1\}^{|\mathcal{X}|}$.
output : Either “Situation A holds” or “Situation A does not hold”.
if  $\min\{|G_{00}^{c\to u}|, |G_{01}^{c\to u}|\} \leq \sqrt{2\delta}\mathcal{X} < \max\{|G_{00}^{c\to u}|, |G_{01}^{c\to u}|\}$
and
$\min\{|G_{10}^{c\to u}|, |G_{11}^{c\to u}|\} \leq \sqrt{2\delta}\mathcal{X} < \max\{|G_{10}^{c\to u}|, |G_{11}^{c\to u}|\}$  then
    return “Situation A holds”
else
    “Situation A does not hold”
end
```

Intuition for the Sharp Extraction Algorithm

We now focus on the Sharp Extraction algorithm. Fix a strong oracle $\mathcal{O}$ with associated set $\mathcal{T}$ of classifiers, and let $\tilde{\mathcal{O}}$ be a δ-weak relaxation of $\mathcal{O}$. Run Algorithm 1 with weak oracle $\tilde{\mathcal{O}}$ to obtain a collection of orbits.

For this informal discussion, fix $t \in \mathcal{T}$ and let **Orb** be the orbit from Algorithm 1 that contains t, possibly together with some classifiers in close alignment with t. Algorithm 3 applies a screening procedure to the elements of each orbit. Applied to the members of **Orb**, the procedure may screen out some $u \in$ **Orb**, meaning that it determines definitively that $u \notin \mathcal{T}$, but other vectors $v \in$ **Orb** may remain; in particular, t will remain.

The screening procedure invokes a primitive MergeOnes (Definition 13) with the key property that $\forall v \in$ **Orb**, MergeOnes$(t, v) \in$ **Orb**. Thus, to test if $v \in$ **Orb** is a valid candidate for t, the algorithm tests whether MergeOnes$(u, v) \in$ **Orb** for all $u \in$ **Orb**.

Algorithm 3 Sharp Extraction.

```
input  : The universe $\mathcal{X}$ and the $\tilde{\mathcal{O}}$ considered in Theorem 9.
output : A list $P_1, \ldots, P_m, Q_1, \ldots, Q_m$ of classifiers of the universe $\mathcal{X}$.
Run Fuzzy Extraction with the inputs $\mathcal{X}$ and $\tilde{\mathcal{O}}$. Let
  $\boldsymbol{Orb}_1, \ldots, \boldsymbol{Orb}_m, \boldsymbol{Orb}'_1, \ldots, \boldsymbol{Orb}'_m$ denote the output. For $j = 1, \ldots, m$, define
  $\Pi_j, \Pi'_j, \Gamma_j, \Gamma'_j = \phi$. ;
for $j = 1, \ldots, m$ do
    for $c \in \boldsymbol{Orb}_j$ do
        if for all $c' \in \mathbf{Orb}_j$, MergeOnes$(c, c') \in \mathbf{Orb}_j$, update $\Pi_j = \Pi_j \cup c$ ;
        if for all $c' \in \mathbf{Orb}_j$, MergeZeros$(c, c') \in \mathbf{Orb}_j$, update $\Gamma_j = \Gamma_j \cup c$ ;
    end
    for $c' \in \boldsymbol{Orb}'_j$ do
        if for all $c'' \in \mathbf{Orb}'_j$, MergeZeros$(c', c'') \in \mathbf{Orb}'_j$, update $\Pi'_j = \Pi'_j \cup c'$ ;
        if for all $c'' \in \mathbf{Orb}'_j$, MergeOnes$(c', c'') \in \mathbf{Orb}'_j$, update $\Gamma'_j = \Gamma'_j \cup c'$;
    end
end
for $j = 1, \ldots, m$ do
    set $u_j =$ MergeOnes$(\Pi_j)$, $v_j =$ MergeZeros$(\Pi'_j)$, $w_j =$ MergeZeros$(\Gamma_j)$, $x_j =$
      MergeOnes$(\Gamma'_j)$, and define

      $P_j = \begin{cases} u_j & \text{if } |{u_j}^0| \geq |{v_j}^1| \\ v_j^{\text{flip}} & \text{if } |{v_j}^1| > |{u_j}^0| \end{cases}$ ,   $Q_j = \begin{cases} w_j & \text{if } |w_j^1| \geq |x_j^0| \\ x_j^{\text{flip}} & \text{if } |x_j^0| > |{w_j}^1| \end{cases}$ .
end
Return $P_1, \ldots, P_m, Q_1, \ldots, Q_m$;
// The classifier $c$ with ${P_j}^0 \subseteq c^0$ and ${Q_j}^1 \subseteq c^1$ accurately recovers $t^0, t^1$
   (or $[t^{\text{flip}}]^0, [t^{\text{flip}}]^1$) upto a small error margin (Theorem 9).
```

Let **In** $\subseteq$ **Orb** be the subset of **Orb** screened in. Let us arbitrarily name these elements $u_1, u_2, \ldots, u_k$. If there is exactly one element in **In**, then $u_1 = t$. This is excellent: the algorithm found an element of $\mathcal{T}$. Consider now the more general case of multiple elements. Choose any ordering of **In**, say, $(u_1, u_2, \ldots, u_k)$. Then since $u_1 \in$ **In**, we have $w =$ MergeOnes$(u_1, u_2) \in$ **Orb**. Now, since $u_3 \in$ **In**, MergeOnes$(w, u_3) \in$ **Orb**. We can continue in this way until we have merged in u_k, and we see by induction that at every step the merge remains in **Orb**.

MergeOnes is commutative and associative, so we can assume without loss of generality that $u_1 = t$, which will simplify the description of the properties of the merging of all the classifiers in **In**. We first define the operation.

▶ **Definition 13** (MergeOnes). *For a set of classifiers $\mathcal{C} = \{c_1, \dots, c_k\}$, define the* MergeOnes *operation applied to $\mathcal{C}$ by* MergeOnes$(\{c_1, \dots, c_k\})$ *to be a new classifier w such that*

$$w^1 = \cup_{j \in [k]} {c_j}^1, \qquad w^0 = \mathcal{X} \backslash w^1 \tag{7}$$

MergeZeros *is defined analogously and denoted* MergeZeros$(\{c_1, \dots, c_k\})$*.*

Let $w = \text{MergeOnes}(t, u)$ for an arbitrary $u \in \{0,1\}^{|\mathcal{X}|}$. Then w^1 contains all the elements with positive classification under t (and other elements that are positive under u). Since w^0 contains none of these, we have that $w^0 \subseteq t^0$.

Let $w = \text{MergeOnes}(\{u \in \textbf{In}\})$. Then by the above reasoning also for this w, we have $w^0 \subseteq t^0$. To argue that w^0 recovers most of t^0, we show that $t^0 \cap w^1$ must be small. This follows from the fact that $w \in \textbf{Orb}$, as argued above. This ensures that $t^0 \cap w^0$, is in fact, large. A similar reasoning applies to the MergeZeros procedure, which is used to create another subset $\textbf{In}' \subseteq \textbf{Orb}$ with the property that $\text{MergeZeros}(\textbf{In}')^1$ has a large intersection with t^1.

In its final phase, still focusing on a single t and its corresponding orbit **Orb**, the algorithm combines the left side of the output of the MergeOnes procedure and the right side of the output of the MergeZeros procedure to create a classifier that substantially agrees with t (for elements $i \notin (\text{MergeOnes}(\textbf{In}))^0 \cup (\text{MergeZeros}(\textbf{In}'))^1$ it assigns an arbitrary value). This completes the high level intuition for the Sharp Extraction algorithm. Theorem 9 provides the formal guarantees.

5 Discussion on Assumptions

Our main theorem relies on the following crucial assumption regarding the structure of $\mathcal{T}$.

▶ **Assumption 14.** *Assume that every $t \in \mathcal{T}$ obeys the following structure.*

1. *Every $t \in \mathcal{T}$ that is not a constant classifer, must be 4δ-balanced.*
2. *For any $t, u \in \mathcal{T}$, if $t^{\text{flip}} \neq u$, then at most one of $G_{00}^{t \to u}, G_{01}^{t \to u}, G_{10}^{t \to u}$ and $G_{11}^{t \to u}$ can be empty. Furthermore, whenever one of these sets is non-empty, it must contain strictly more than $2\sqrt{2\delta}\mathcal{X}$ elements (Figure 1).*
3. *For every $t \in \mathcal{T}$ such that $t^{\text{flip}} \notin \mathcal{T}$, let $\tilde{\mathcal{O}}(\mathcal{A}_{\mathcal{W}}, c) = 0$ whenever c is in close alignment with t^{flip}.*

We provide some intuition for the assumptions, deferring details to the full paper. Note that, if $t \in \mathcal{T}$ is imbalanced, then most $c \in \Gamma_\delta(t)$ will also be imbalanced. However, imbalanced classifiers are problematic: they belong to δ-neighborhoods of more than one $t \in \mathcal{T}$, even when these are far apart. The presence of several imbalanced classifiers confuses our algorithms. By requiring that all $t \in \mathcal{T}$ be well balanced (Assumption 14(1)), we ensure that sufficiently many $c \in \Gamma_\delta(t)$ are also balanced.

To recover $\mathcal{T}$, our algorithms need to tease apart the following situations for any two classifiers $u, v \in \{0,1\}^{|\mathcal{X}|}$: (a) $\exists t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$, and (b) $\nexists t \in \mathcal{T}$ such that $u, v \in \Gamma_\delta(t)$. Our intuition is that if elements of $\mathcal{T}$ are well-separated as defined via transportation costs, then this should be possible; however, our proof requires Assumption 14(2) that implies separation but is not equivalent. We do not know if this stronger condition can be relaxed.

Finally, for any $t \in \mathcal{T}$, the flipped classifier t^{flip} may or may not belong to $\mathcal{T}$. But, the neighborhoods of t and t^{flip} are identical – this creates additional challenges when $t \in \mathcal{T}$ but $t^{\text{flip}} \notin \mathcal{T}$. Assumption 14(3) protects from complications arising in this case.

References

1 Benjamin R Baer, Daniel E Gilbert, and Martin T Wells. Fairness criteria through the lens of directed acyclic graphical models. *arXiv preprint*, 2019. `arXiv:1906.11333`.

2 Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

3 Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

4 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.

5 Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 33:1–33:20, 2019.

6 Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

7 Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint*, 2015. `arXiv:1511.05897`.

8 Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

9 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

10 Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1944–1953, 2018.

11 Christina Ilvento. Metric learning for individual fairness. *arXiv preprint*, 2019. `arXiv:1906.00250`.

12 Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

13 Christopher Jung, Sampath Kannan, Changwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. Manuscript shared with authors.

14 Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint*, 2019. `arXiv:1905.10660`.

15 Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

16 Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

17 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.

18 Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. *arXiv preprint*, 2019. `arXiv:1907.01040`.

19 Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

20 Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852, 2018.

21 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

22 Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

23 Himabindu Lakkaraju. Course notes for compsci 282br, harvard university: Interpretability and explainability in machine learning, 2019.

24 Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

25 David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint*, 2018. arXiv:1802.06309.

26 Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

27 Roland Neil and Christopher Winship. Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, 2:73–98, 2019.

28 Judea Pearl. *Causality*. Cambridge university press, 2009.

29 Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

30 Gal Yona and Guy N. Rothblum. Probably approximately metric-fair learning. In *ICML*, 2018.

31 Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

The Role of Randomness and Noise in Strategic Classification

Mark Braverman
Department of Computer Science, Princeton University, NJ, USA
mbraverm@cs.princeton.edu

Sumegha Garg
Department of Computer Science, Princeton University, NJ, USA
sumeghag@cs.princeton.edu

Abstract

We investigate the problem of designing optimal classifiers in the "strategic classification" setting, where the classification is part of a game in which players can modify their features to attain a favorable classification outcome (while incurring some cost). Previously, the problem has been considered from a learning-theoretic perspective and from the algorithmic fairness perspective.

Our main contributions include

- Showing that if the objective is to maximize the efficiency of the classification process (defined as the accuracy of the outcome minus the sunk cost of the qualified players manipulating their features to gain a better outcome), then using randomized classifiers (that is, ones where the probability of a given feature vector to be accepted by the classifier is strictly between 0 and 1) is necessary.
- Showing that in many natural cases, the imposed optimal solution (in terms of efficiency) has the structure where players never change their feature vectors (and the randomized classifier is structured in a way, such that the gain in the probability of being classified as a "1" does not justify the expense of changing one's features).
- Observing that the randomized classification is not a *stable* best-response from the classifier's viewpoint, and that the classifier doesn't benefit from randomized classifiers without creating instability in the system.
- Showing that in some cases, a *noisier signal* leads to better equilibria outcomes – improving both accuracy and fairness when more than one subpopulation with different feature adjustment costs are involved. This is particularly interesting from a policy perspective, since it is hard to force institutions to stick to a particular randomized classification strategy (especially in a context of a market with multiple classifiers), but it is possible to alter the information environment to make the feature signals inherently noisier.

2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Machine learning theory

Keywords and phrases Strategic classification, noisy features, randomized classification, fairness

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.9

Funding *Mark Braverman*: Research supported in part by the NSF Alan T. Waterman Award, Grant No. 1933331, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

1 Introduction

Machine learning algorithms are increasingly being used to make decisions about the individuals in various areas such as university admissions, employment, health, etc. As the individuals gain information about the algorithms being used, they have an incentive to adapt their data so as to be classified desirably. For example, if a student is aware that a university heavily weighs SAT score in their admission process, she will be motivated to achieve a higher SAT score either through extensive test preparation or multiple tries. Such efforts by the students might not change their probability of being successful at the university, but are enough to fool the admissions' process. Therefore, under such "strategic manipulation" of one's data, the predictive power of the decisions are bound to decrease. One way to prevent such manipulation is by keeping the classification algorithms a secret, but this is not a practical solution to the problem, as some information is bound to leak over time and the transparency of these algorithms is a growing social concern. Thus, this motivates the study of algorithms that are optimal under "strategic manipulation". The problem of gaming in the context of classification algorithms is a well known problem and is increasingly gaining researchers' attention, for example, [8, 1, 9, 16, 4].

[2] and [8] modeled strategic classification as a Stackelberg competition– the algorithm (Jury) goes first and publishes the classifier, and then the individuals get to transform their data, after knowing the classifier, incurring certain costs to manipulate. The individuals would manipulate their features as long as the cost to manipulate is less than the advantage gained in getting the desirable classification. We assume that such manipulations don't change the actual qualifications of an individual. A natural question is: what classifier achieves optimal classification accuracy under the Stackelberg competition? These papers considered the task of strategic classification when the published classifier is deterministic. We study the role of randomness (and addition of noise to the features) in strategic classification and define the Stackelberg equilibrium for probabilistic classifiers, that assigns a real number in $[0, 1]$, to each individual and a classification outcome o, representing the probability of being classified as o.

As higher SAT scores are preferred by a university, the students would put an effort in increasing their SAT score, thereby, forcing the university to raise the score bar to optimize its accuracy (under the Stackelberg equilibrium). Due to this increased bar of acceptance, even the students who were above the true cutoff would have to put an extra effort to achieve a SAT score above this raised bar. And this effort is entirely the result of gaming in the classification system. We define the *cost of strategy* for a published classifier to be the total extra effort, it induced, amongst the qualified individuals of the population. Then, we define the *efficiency* of a published classifier to be its classification accuracy minus the cost of strategy under the Stackelberg equilibrium. A natural question here is: what classifier achieves the optimal efficiency? The efficiency of a published classifier represents the total impact of the classifier on all the agents in the Stackelberg equilibrium.

In normal classification problems it is never a good idea to use randomness, since one should always adhere to the best/utility maximizing action based on the prediction. Just as in games, randomness may lead to better solution in strategic classification, the paper aims to start understanding tradeoffs between efficiency losses due to randomness and efficiency gains through better equilibria induced by the randomized classifier.

Gaming in classification adds to the plethora of fairness concerns associated with classification algorithms, when the costs of manipulation are different across subpopulations. For example, a high weightage of SAT scores (for university admissions) favors the subgroups of the society that have the resources to enroll in test preparation or attempt the test multiple times. Further, varying costs across the subpopulations can lead to varied efforts put by identically qualified individuals, belonging to different subpopulations, to achieve the same outcome. [16] and [9] study the disparate effects of strategic classification on subpopulations (we will discuss these papers more in the related work section). [9] observes that a single classifier might have different classification errors on subpopulations due to the varying cost of manipulations. We also study the effect of strategic manipulation on the classification errors across subpopulations and how randomized classifiers or noisy features may reduce the disparate effects.

Strategic classification is a well known problem and there has been research in many other aspects of strategic classification, for example, learning the optimal classifier efficiently when the samples might also be strategic [8, 4], mechanism design under strategic manipulation [3, 5, 12], and studying the manipulation costs that actually change the inherent qualifications [14, 15]. The focus of this paper is theoretically demonstrating the role of randomness and noise in the strategic setting.

1.1 Our contributions

Above, we talked about how strategic manipulation can deteriorate the classification accuracy and lead to unfair classification. We investigate the different scenarios of the classification task that help in regaining the lost accuracy and fairness guarantees. Our entire work is based on *one-dimensional feature space.*

1.1.1 Randomized classifiers

Firstly, we formulate the strategic classification task, when the published classifier is randomized. Instead of publishing a single binary classifier (for 2 classification outcomes, 0 and 1), the Jury publishes a distribution of classifiers and promises to pick the final classifier from that distribution. Another interpretation is that the Jury assigns a value in $[0, 1]$ to each feature value, which represents the probability of an individual with this feature being classified as 1. The individuals manipulate their features, after knowing the set of classifiers but not the final classifier, incurring certain costs according to the *cost function.*

Not surprisingly, we show through examples that a probabilistic classifier can achieve strictly higher expected accuracy and efficiency than any binary classifier under strategic setting. Note that, without any strategic manipulation, a randomized classifier has no advantage over deterministic classifiers in terms of classification accuracy. The intuition is as follows: using randomness, the Jury can discourage the individuals from manipulating their features by making the advantage gained by any such a manipulation small enough.

For *simple* cost functions, we then characterize the randomized classifier that achieves optimal efficiency. We prove that such a classifier sets the probabilities (of being classified as 1) such that none of the individuals have an incentive to manipulate their feature. Given two features x and x' in the feature space, let $c(x, x')$ denote the cost of manipulating one's feature from x to x'. Informally, we say a cost function c is *simple* when all the costs are non-negative, the cost to manipulate to a "less" qualified feature is 0, and the costs are sub-additive, that is, manipulating your feature x directly to x'' is at least easier than first manipulating it to x' and then to x''. The characterization theorem, stated informally, is as follows:

▶ **Theorem 1** (Informal statement of Theorem 3). *For simple cost functions, the most efficient randomized classifier is such that the best response of all the individuals is to reveal their true features.*

This characterization, in addition to being mathematically clean, allows us to infer the following: let A and B be two subpopulations (identical in terms of qualifications) such that the costs to manipulation are *higher* for individuals in A than in B, then the optimal efficiency obtained for the subpopulation A is greater than that in B.

1.1.2 Obstacles to using a randomized classifier

Till now, we have argued the benefits of using a probabilistic classifier. However, the degree to which it is possible to use or commit to a randomized strategy varies depending on the setting. There are two main drivers impeding the implementation of the most efficient Stacklberg equilibrium. Firstly, in many real-life classification settings, it might be unacceptable to use a probabilistic classifier, for example, due to legal restrictions (applicants with identical features must obtain identical outcomes). Secondly, for the more complicated scenario with multiple classifiers (such as college admissions), the effect of each Jury on the overall market is small, hence, diminishing the incentive to stick to a randomized strategy "for the benefit of the market as a whole". Informally, the best response of a single Jury, when the other classifiers commit to using a randomized classifier, is not a randomized classifier. And even if we got the Juries to commit to randomization, the final probabilities of classification depends on the number of classifiers (k) and hence, the implementation of the most efficient randomized classifier needs coordination between the multiple classifiers. Analyzing the equilibria for multiple classifiers is beyond the scope of this paper but we illustrate the instability of randomized classifier as follows. We show that unless Jury is able to commit to the published randomized classifier, such a classifier is not a stable solution to strategic classification. As mentioned above, randomization helps because of the following observation: if the difference between the probabilities, of being classified as 1 at *adjacent* features is small, the individuals have no incentive to manipulate their features. But, once the Jury knows that no one changed their feature, her best response, then, is to use the classifier that achieves best accuracy given the *true* features.

Formally, we show (Theorem 5) that for any published randomized classifier that achieves strictly higher accuracy compared to any deterministic classifier under Stackelberg equilibrium, Jury has an opportunity to improve its utility and get strictly better accuracy using a classifier different from the published.

The shortcomings of a randomized classifier can be redeemed by addition of noise to the features.

1.1.3 Addition of noise to the features

This brings us to our second scenario that uses noisy features for classification. Every individual has an associated private signal that identifies their qualification. The Jury sees a feature that is a noisy representation of this private signal. The individuals, after incurring certain cost, can effectively manipulate their private signal such that the features are a noisy representation of this updated private signal. Again, the assumption is that such a manipulation didn't change the true qualifications of an individual. We show, through an example of a cost function and a noise distribution, that in the strategic setting, using a deterministic classifier, the Jury achieves better accuracy when the features are noisy than any deterministic classifier in the noiseless case, that is, when Jury gets to see the private signal. This is counter-intuitive at first glance because under no strategic manipulation, noise can only decrease Jury's accuracy.

We also show examples where noisy features can help in achieving fairer outcomes across subpopulations. Let A and B be two subpopulations *identical* in qualifications but having different (but not extremely different) costs of manipulation (and $|A| \leq |B|$; A is a minority). We show, through an example, that no matter whether the minority has higher or lower costs of manipulation than the majority, it is at a disadvantage when Jury publishes a single deterministic classifier to optimize its overall accuracy (noiseless strategic setting). Here, by disadvantage, we mean that the minority has lower classification accuracy than the majority. Next, we show that the addition of appropriate noise to the private signals, in the same example, can ensure that Jury's best response classifier is fair across subpopulations. This is not that surprising as making the features completely noisy also lead to same outcomes for the subpopulations. However, such an addition of noise can also sometimes increase Jury's overall accuracy (improving both accuracy and fairness). We consider the case where the Jury would publish a single classifier for both the subpopulations (for e.g., either because A is a protected group and the Jury is not allowed to discriminate based on the subgroup membership or because the Jury has not yet identified these subpopulations and the differences in their cost functions). Informally, our results, can be stated as follows:

▶ **Theorem 2** (Informal statement of Theorems 6,7,8)**.** *Let A and B be two subpopulations that are* identical *in qualifications. Let $c_A \neq c_B$ be the cost functions for subpopulations A and B respectively. In Case 1, Jury gets to see the private signals and publishes a single deterministic classifier that achieves optimal overall accuracy (sum over the two subpopulations) under the Stackelberg equilibrium (for the cost functions c_A and c_B). In Case 2, the features are noisy representations of the private signal; Jury publishes a single deterministic classifier that achieves optimal overall accuracy under the Stackelberg equilibrium (knowing that the features are noisy). There exists an instantiation of the "identical qualifications" such that*

1. *If $|A| < |B|$, that is, A is a minority, for a wide set of costs functions c_A, c_B, A is always at a disadvantage when in Case 1.*
2. *There exists a setting of the "noise" (η) for each of the above cost functions, such that, Jury's best response in Case 2, is always fair, that is, achieves equal classification accuracy on the subpopulations.*
3. *There exists cost functions c_A, c_B from this wide set of cost functions, and corresponding noise η, such that Jury's accuracy in Case 2 is strictly better than in Case 1.*

This result has potentially interesting policy implications, since it is easier, both practically and legally, to commit to using noisier signals (for example by restricting the types of information available to the Jury) than to commit to disregarding pertinent information ex-post (as in randomized classification). Therefore, future mechanism design efforts involving strategic classification should carefully consider the mechanisms of information disclosure to the Jury.

1.2 Related Work

[8, 2] initiated the study of strategic classification through the lens of Stackelberg competition. [9, 16, 10] study the effects of strategic classification on different subpopulations and how it can exacerbate the social inequity in the world. [9] also made the observation that a single classifier would have varying classification accuracies across subpopulations with different costs of manipulation. [16] defined a concept called "social burden" of a classifier to be the sum of the minimum effort any qualified individual has to put in to be classified as 1. Thus, the subpopulations with higher costs of manipulation would have worse social burden and might be at a disadvantage. In such situations, intuitively, one would think that subsidizing the costs for the disadvantaged population might help. [9] showed that cost subsidy for disadvantaged individuals can sometimes lead to worse outcomes for the disadvantaged group.

In the present paper, we observe that the addition of noise, counter-intuitively, can help Jury's accuracy as well as serve the fairness concerns. There are many examples in game theory where loss of information helps an individual in strategic setting, for example, [6]. [11, 10] also studies the role of hiding information to serve fairness. [7] has a brief discussion at the end of the paper on making manipulated data more informative through addition of noise to the features (this was put online a couple of months after the first version of our paper was made online).

Another work related to Theorem 3 of the present paper is [13], which studies the scope of truthful mechanisms when the agents incur certain costs for misreporting their true type. In particular, the paper gives conditions, on the misreporting costs, that allow the revelation principle to hold, that is, any mechanism can be implemented by a truthful mechanism, where all the agents reveal their true types. The main difference between [13] and our paper is that the former allows the use of monetary transfers to the agents to develop truthful mechanisms and such transfers don't impact the objective value of the mechanism.

1.3 Organization

We formalize the model used for strategic classification in Section 2. In Section 3, we show how randomness helps in achieving better accuracy and efficiency. We also characterize the classifiers that achieve optimal efficiency for *simple* cost functions. In Section 4, we investigate the stability of randomized classifiers. In Section 5, we investigate the role of noisy features in strategic classification.

2 Preliminaries

In this paper, we concern ourselves with classification based on a one-dimensional feature space $\mathcal{X}$. In many of the examples, our feature space $\mathcal{X} \subseteq \mathbb{R}$ is discrete, hence, we use sum ($\sum$) in many of the definitions, but, these definitions are well-defined when $\mathcal{X}$ is taken to be continuous (for e.g., $\mathbb{R}$) by replacing sum ($\sum$) with integrals ($\int$) and probability distributions with probability density functions. We use the notation $\mathcal{N}(z, \sigma)$ to denote the gaussian distribution with mean z and standard deviation σ. We say a function $f : \mathcal{X} \to \{0, 1\}$ is a threshold function (classifier) with threshold τ if

$$f(x) = \begin{cases} 1 & \text{if } x \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

We also use $1_{x \geq \tau}$ to denote a threshold function (classifier) with threshold τ. Sometimes, we will use $1_{x > \tau}$ that classifies x as 1 if and only if $x > \tau$.

2.1 The Model

Let $\mathcal{X}$ be the set of features. Let $\pi : \mathcal{X} \to [0, 1]$ be the probability distribution over the feature set realized by the individuals. Let $h : \mathcal{X} \to [0, 1]$ be the true probability of an individual being qualified (1) given the feature. We also refer to it as the true qualification function. Let $c(x, x')$ be the cost incurred by an individual to manipulate their feature from x to x' (We also use words, change and move, to refer to this manipulation). The classification is modeled as a sequential game where a Jury publishes a classifier (possibly probabilistic) $f : \mathcal{X} \to [0, 1]$ and contestants (individuals) can change their features (after seeing f) as long as they are ready to incur the cost of change. The previous papers in the area considered the task of strategic classification when the published classifier is deterministic binary classifier. Here, we formalize the Stackelberg prediction game for probabilistic classifiers.

Given f, we define the best response of a contestant with feature x[1], as follows

$$\Delta_f(x) = \text{argmax}_{y \in (\{x\} \cup \{x' | (f(x') - f(x)) > c(x,x')\})}(f(y)) \tag{1}$$

We will denote it by Δ when f is clear from the context. $\Delta(x)$ might not be well defined if there are multiple values of y that attains the maximum. In those cases, $\Delta(x)$ is chosen to be the smallest y amongst them. In words, you jump to another feature only if the cost of jumping is less than the advantage in being classified as 1.

We define the Jury's utility for publishing f ($U(f)$) as the classification accuracy with respect to $h(x)$. Thus, Jury's utility for publishing f is

$$\begin{aligned} U(f) &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x))] \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot (2h(x) - 1) + 1 - h(x)] \end{aligned}$$

We define $C(f) = \sum_{x \in \mathcal{X}} \pi(x)[h(x) \cdot c(x, \Delta_f(x))]$ to be the cost of strategy for a published classifier f.

We define the efficiency of the classifier f $(E(f))^2$ as follows:

$$\begin{aligned} E(f) &= U(f) - C(f) \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x))] - \sum_{x \in \mathcal{X}} \pi(x)[h(x) \cdot c(x, \Delta(x))] \\ &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))] \end{aligned}$$

The focus of this paper is to demonstrate what role randomness and noise can play in strategic classification and not to give algorithms for learning the optimal or most efficient strategic classifier. We can present the ideas even by making the following assumptions on the cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

1. $c(x, x') \geq 0,\ \forall x, x' \in \mathcal{X}$.
2. $c(x', x) = 0,\ \forall x, x' \mid h(x') \geq h(x)$, that is, jumping to a lesser qualified feature is free.
3. $c(x, x'') \leq c(x, x') + c(x', x''),\ \forall x, x', x'' \in \mathcal{X}$, that is, the costs are sub-additive.
4. $c(x, x') \leq c(x, x''),\ \forall x, x', x'' \mid h(x'') \geq h(x')$, that is, jumping to a lesser qualified feature is easier.
5. $c(x', x'') \leq c(x, x''),\ \forall x, x', x'' \mid h(x') \geq h(x)$, that is, jumping from a lesser qualified feature is harder.

The last two points are implied by the first three, we wrote them as separate points for completeness. We call the cost function *simple* if it satisfies all the above assumptions.

By the virtue of the definition of simple cost functions, without loss of generality, we assume that h is monotonically increasing with the feature x, that is, $\forall x, x' \in \mathcal{X},\quad x' \geq x \implies h(x') \geq h(x)$.

Next, we mention a special kind of cost function that satisfies the assumptions: $c(x, x') = \max(a(x') - a(x), 0)$ where the function $a : \mathcal{X} \to \mathbb{R}$ is monotonically increasing in x, that is, $x' \geq x \implies a(x') \geq a(x)$.

[1] Such a best response model has been studied in the literature, for example, [17].

[2] We defined efficiency as $U(f) - C(f)$ for the simplicity of the presentation. Defining efficiency as $U(f) - \beta \cdot C(f)$ (for some $\beta > 0$) doesn't effect the theorems except for Theorem 3, which is no longer true for $\beta < 1$.

Given a cost function c, let

$$\mathrm{Lip}_1(c) = \{f \mid f : \mathcal{X} \to [0,1], f(x') - f(x) \leq c(x,x') \; \forall x, x' \in \mathcal{X}\}$$

Given the cost function c, we say f satisfies the Lipschitz constraint if $f \in \mathrm{Lip}_1(c)$. Note that any classifier $f \in \mathrm{Lip}_1(c)$ is monotonically increasing with x, that is, $x' \geq x \implies f(x') \geq f(x)$. This is because $\forall x' \geq x, f(x) - f(x') \leq c(x', x) = 0$. And $\forall x \in \mathcal{X}, \Delta_f(x) = x$, that is, no one changes their feature if f is the published classifier.

In Section 5, we generalize this model to the setting where the features are a noisy representation of an individual's private signal. An individual can make efforts to change their private signal but can't control the noise. The Jury only see the features and classifies an individual based on that. In Section 5, the fairness notion, we will concern ourselves with, is the classification accuracy of the published classifier across subpopulations.

3 Committed Randomness Helps both Utility and Efficiency

In this section, we compare the optimal utility and efficiency achieved by a deterministic binary classifier to a probabilistic classifier. Consider the following two scenarios:

Scenario 1: The Jury commits to using a binary classifier $f : \mathcal{X} \to \{0,1\}$. The best response function $\Delta_f : \mathcal{X} \to \mathcal{X}$, Jury's utility from publishing f ($U(f)$) and efficiency of the classifier f ($E(f)$) are defined as in Section 2.

Scenario 2: The Jury publishes a probabilistic classifier $f : \mathcal{X} \to [0,1]$ and commits to it. The best response function $\Delta_f : \mathcal{X} \to \mathcal{X}$, Jury's utility from publishing f ($U(f)$) and efficiency of the classifier f ($E(f)$) are as defined in Section 2. Note that this is equivalent to when Jury publishes a list of deterministic classifiers and chooses a classifier uniformly at random from them. Contestants update their feature without knowing which classifier gets picked up at the end.

The following example illustrates how randomization helps in getting strictly better utility and efficiency:

Let $\mathcal{X} = \{1,2\}$ and each feature contains half of the population. Let

$$h(x) = \begin{cases} 1 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Let the cost of changing the feature from 1 to 2 be 0.5. The the randomized classifier f defined as follows:

$$f(x) = \begin{cases} 1 & \text{if } x = 2 \\ 0.5 & \text{if } x = 1 \end{cases}$$

achieves an accuracy of 0.75. The contestants at $x = 2$ are happy as they are already being classified as 1 with probability 1. For the contestants at $x = 1$, $f(2) - f(1) = 0.5 = c(1,2)$ and hence, they don't have an incentive to manipulate their feature. As all the contestant retain their true features, the efficiency of f is also equal to 0.75. As the feature space is bounded, there are only three options for a deterministic classifier: keep the threshold at 1 and classify everyone as 1; keep the threshold at 2 and you end up classifying everyone as 1, as the contestants at 1 change their feature to 2; classify everyone as 0. All these classifiers have 0.5 accuracy and at most 0.5 efficiency.

In the mathematical example given above, the randomized classifier was set up such that none of the contestants had any incentive to change their feature. In the next subsection, we show that the most efficient classifier always looks like "this" for "simple" cost functions.

That is, if the cost function c satisfies the assumptions made in Section 2, then for every true qualification function h, there exists a function $f_h \in \mathrm{Lip}_1(c)$ that achieves the optimal efficiency.

3.1 Most Efficient Classifier for Simple Cost Functions

Recall, $E(f) = \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))]$. Let $E^* = \max_{f:\mathcal{X}\to[0,1]} \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta(x)) \cdot h(x) + (1 - f(\Delta(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta(x))]$.

▶ **Theorem 3.** *For every monotone true qualification function $h : \mathcal{X} \to [0, 1]$, probability distribution $\pi : \mathcal{X} \to [0, 1]$ over the features, simple cost function c, there exists $g \in Lip_1(c)$ such that $E(g) = E^*$.*

Proof. Let f be an efficiency maximizing classifier. We argue that $g : \mathcal{X} \to [0, 1]$ defined as

$$g(x) = \max_y\{f(y) - c(x, y)\}$$

is in $\mathrm{Lip}_1(c)$ and satisfies $E(g) \geq E(f)$. Let $\delta_f(x) = \mathrm{argmax}_y\{f(y) - c(x, y)\}$. When f is clear from the context, we will drop the subscript on δ. Using definition of δ, $g(x) \in [0, 1]$ as $\forall x, y \in \mathcal{X},\ f(y)-c(x, y) \leq f(y) \leq 1$ $(c(x, y) \geq 0)$ and $\max_y\{f(y)-c(x, y\} \geq f(x)-c(x, x) \geq 0$. For all $x, x' \in \mathcal{X}$,

$$\begin{aligned} g(x') - g(x) &= f(\delta(x')) - c(x', \delta(x')) - f(\delta(x)) + c(x, \delta(x)) \\ &= f(\delta(x')) - c(x, \delta(x')) - f(\delta(x)) + c(x, \delta(x)) + (c(x, \delta(x')) - c(x', \delta(x'))) \\ &\leq c(x, \delta(x')) - c(x', \delta(x')) \leq c(x, x') \qquad \text{(sub-additivity)} \end{aligned}$$

The first inequality follows the definition of δ, that is, $\forall y \in \mathcal{X}, f(\delta(x)) - c(x, \delta(x)) \geq f(y) - c(x, y)$. Therefore, $f(\delta(x')) - c(x, \delta(x')) - f(\delta(x)) + c(x, \delta(x)) \leq 0$. The second inequality follows from the fact that the cost function c is simple and satisfies the sub-additivity condition. This proves that $g \in \mathrm{Lip}_1(c)$. This implies, as observed previously, $\forall x \in \mathcal{X}, \Delta_g(x) = x$. Next, we show that $E(g) \geq E(f)$ and hence $E(g) = E^*$. Efficiency of the classifier g is

$$\begin{aligned} E(g) &= \sum_{x\in\mathcal{X}} \pi(x)[g(\Delta_g(x)) \cdot h(x) + (1 - g(\Delta_g(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta_g(x))] \\ &= \sum_{x\in\mathcal{X}} \pi(x)[2 \cdot g(x) \cdot h(x) - g(x) - h(x) + 1] \end{aligned}$$

Efficiency of the classifier f is

$$\begin{aligned} E(f) &= \sum_{x\in\mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot h(x) + (1 - f(\Delta_f(x)) \cdot (1 - h(x)) - h(x) \cdot c(x, \Delta_f(x))] \\ &= \sum_{x\in\mathcal{X}} \pi(x)[2f(\Delta(x)) \cdot h(x) - f(\Delta(x)) - h(x) + 1 - h(x) \cdot c(x, \Delta(x))] \end{aligned}$$

$E(g) - E(f) = \sum_{x\in\mathcal{X}} \pi(x)[(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x))]$

▷ Claim 4. $\forall x,\ \ [(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x))] \geq 0.$

Please refer to Appendix A for the proof of the claim. It's straightforward to see that $E(g) - E(f) \geq 0$ using the above claim. Therefore, we showed a classifier $g \in \mathrm{Lip}_1(c)$ such that $E(g) = E^*$. ◀

In words, *when we are concerned with the efficiency of the published classifier, the optimal is achieved by a probabilistic classifier that has zero cost of strategy and gives individuals no incentive to change their feature.*

4 Are Randomized Classifiers in Equilibrium from Jury's Perspective?

As discussed in the Section 1, there are many obstacles to implementing a randomized classifier in the strategic setting. In this section, we illustrate the instability caused by the use of randomized classifiers (which becomes increasingly important while considering multiple classifiers). In Section 3, we saw that a randomized classifier can achieve better accuracy and efficiency than any binary classifier. While maximizing efficiency, we further showed that the optimally efficient classifier is such that every contestant reveals their true feature. Once the Jury knows the contestants' true features, she can be greedy and classify the individuals using a threshold function with $\tau = \min\{x \mid h(x) \geq \frac{1}{2}\}$ as the threshold to achieve the best accuracy. Therefore, unless the Jury commits to using randomness, she has an incentive of not sticking to the promised randomized classifier. The question is: what's the best accuracy/efficiency achieved by a classifier that is in equilibrium even from Jury's perspective? We formalize this equilibrium concept as follows (the true qualification function h and the cost function c are fixed):

1. Jury publishes a randomized classifier $f : \mathcal{X} \to [0,1]$.
2. Contestants, knowing f, changes their feature from x to $\Delta_f(x)$.
3. f is in equilibrium from Jury's perspective if given that the contestants changed their features according to the best response function Δ_f, f achieves the best classification accuracy, that is, for all classifiers $g \in \mathcal{X} \to [0,1]$,

$$\sum_{x \in \mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot h(x) + (1 - f(\Delta_f(x)) \cdot (1 - h(x))] \qquad (2)$$
$$- \sum_{x \in \mathcal{X}} \pi(x)[g(\Delta_f(x)) \cdot h(x) + (1 - g(\Delta_f(x)) \cdot (1 - h(x))] \geq 0$$

Using next theorem, we show that for any randomized classifier that is in equilibrium from Jury's perspective, there exists a binary classifier that achieves at least the same accuracy.

▶ **Theorem 5.** *Given a monotone true qualification function h, probability distribution π over the features, and a simple cost function c, let $f^* : \mathcal{X} \to \{0,1\}$ be the classifier that optimizes Jury's utility over the deterministic classifiers under Stackelberg equilibrium. Let $f : \mathcal{X} \to [0,1]$ be a randomized classifier such that $U(f) > U(f^*)$, then f is not in an equilibrium from Jury's perspective (the notion defined above).*

Please refer to Appendix B for the proof.

Disclaimer: f' as defined above might also not be in equilibrium from Jury's perspective. The above theorem illustrates the following point: *Jury doesn't benefit from randomized classifiers without creating instability in the system.*

Can we somehow exploit this power of randomness while overcoming the obstacles to randomized classification? The answer is yes – make the features noisy.

5 Noisy Features Give the System Free Randomness

We formalize the setting with noisy features as follows: every individual has a private signal $y \in \mathcal{X}$. The true qualification function $h : \mathcal{X} \to [0,1]$ depends on y, that is, $h(y)$ is the probability of an individual being qualified (1) given that its private signal is y. Given a private signal y, a feature is drawn randomly from the distribution $p_y : \mathcal{X} \to [0,1]$, that is, $p_y(x)$ is the probability that an individual's feature is x when their private signal is y. If $\mathcal{X} = \mathbb{R}$, the right intuition for p_y is it being $\mathcal{N}(y, \sigma)$ where $\mathcal{N}(y, \sigma)$ is the gaussian distribution with mean y and standard deviation σ. Let $\pi : \mathcal{X} \to [0,1]$ be the probability distribution over the private signals y realized by the individuals.

Let $c(y, y')$ be the cost incurred by the contestant to change their private signal from y to y'. The contestants can put effort to change their private signals but the feature would still be drawn randomly using the updated private signal.

The classification is again modeled as a sequential game where a Jury publishes a deterministic classifier $f : \mathcal{X} \to \{0, 1\}$. We restricts ourselves to deterministic classifiers due to the observations made in Section 4. Contestants change their private signals as long as they are ready to incur the cost of change. Given a private signal y, let $q_f(y)$ denote the probability of a contestant, with private signal y, being classified as 1 when f is the classifier. Therefore, $q_f(y) = \sum_{x \in \mathcal{X}} p_y(x) \cdot f(x)$.

Given f, the best response of a contestant with private signal y is given as,

$$\Delta_f(y) = \text{argmax}_{z \in \{y\} \cup \{y' | q_f(y') - q_f(y) > c(y,y')\}} (q_f(z)) \tag{3}$$

We will denote it by Δ when f is clear from the context. $\Delta(y)$ might not be well defined if there are multiple values of z that attains the maximum. In those cases, $\Delta(y)$ is chosen to be the smallest z amongst them. In words, you jump to another private signal only if the cost of jumping is less than the advantage in being classified as 1. Even though f is deterministic, due to noisy features, the effective classifier given the private signal y (q_f) is probabilistic. Therefore, we will see below that the noise allows us similar advantages as that of a probabilistic classifier.

The accuracy and efficiency of the classifier f are defined as follows:

$$U(f) = \sum_{y \in \mathcal{X}} \pi(y)[q_f(\Delta(y)) \cdot h(y) + (1 - q_f(\Delta(y)) \cdot (1 - h(y))]$$

$$E(f) = \sum_{y \in \mathcal{X}} \pi(y)[q_f(\Delta(y)) \cdot h(y) + (1 - q_f(\Delta(y)) \cdot (1 - h(y))] - \sum_{y \in \mathcal{X}} \pi(y)[h(y) \cdot c(y, \Delta(y))]$$

We assume that h is monotonically increasing with y and the cost function c is simple. Next, we will demonstrate how noisy features can lead fairer outcomes and even increase Jury's accuracy.

5.1 Noisy Features achieve Fairer Equilibriums

Consider two subpopulations A and B. For simplicity, these subpopulations are a partition of the individuals in the universe. Let s_A denote the probability an individual from the universe is in subpopulation A. Similarly, s_B $(s_A = 1 - s_B)$. Let $h_A : A \to [0, 1]$ be the true qualification function for the subpopulation A. Similarly, h_B. Let $c_A : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the cost function for the subpopulation A, that is, $c_A(y, y')$ is the cost of changing the private signal from y to y' for an individual in A. Similarly, c_B is defined. Let $\pi_A : A \to [0, 1]$ and π_B be the probability distribution over the private signals realized by the subpopulations A and B respectively.

Given a published deterministic classifier $f : \mathcal{X} \to \{0, 1\}$, the best response of the contestant in subpopulation A with private signal y $(\Delta_f^A(y))$ is defined using c_A as the cost function. Similarly, for subpopulation B, let $\Delta_f^B(y)$ denote the best response of the contestant in subpopulation B with private signal y and when the published classifier is f. We use $U_A(f)$ and $U_B(f)$ to denote the accuracy of the classifier f on the respective subpopulations.

We consider the setting where $h_A = h_B = h$ and $\pi_A = \pi_B = \Pi$, but the cost functions c_A and c_B are different. In this section, we use the symbol Π to denote the probability distribution over the private signals to avoid confusion with the Archimedes' constant π.

In our first example, we show that even though the subpopulations are identical with respect to their qualifications, different costs can lead to unfair classification when classification is based on private signals. Through our second example, we show that the use of noisy features, for strategic classification, can lead to increase in the overall accuracy of classification as well as give fair classification. We evaluate the fairness of a classifier f quantitively using the difference between the accuracies, that is, $|U_A(f) - U_B(f)|$.

Let's start with the example. $\mathcal{X} = \mathbb{R}$. Let the true qualification function for both the subpopulations be as follows: $h(y) = \begin{cases} 1 & \text{if } y > d \\ \frac{y}{2d} + \frac{1}{2} & \text{if } y \in [-d, d] \\ 0 & \text{if } y < -d \end{cases}$, where d is a fixed large enough positive real number. Let the probability density function on the private signals realized by the subpopulations be as follows: $\Pi(y) = \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t}$, that is, the gaussian distribution with mean 0 and standard deviation t. Again, t is fixed positive real number. We assume $d >> t$.

Let σ_A and σ_B be positive real numbers. The cost function for a subpopulation $S \in \{A, B\}$ is defined as follows (with $(y' - y)^+ = \max\{y' - y, 0\}$):

$$c_S(y, y') = \frac{(y' - y)^+}{\sqrt{2\pi}\sigma_S} \tag{4}$$

We start with the setting where the features are the private signals and not a noisy representation of them.

Remark: If the Jury is allowed to publish different classifiers for the two subpopulations, then she can achieve "the best possible accuracy" on both the subpopulations. It's easy to see that the classifier $f_S : \mathcal{X} \to \{0, 1\}$, defined as follows, achieves as much accuracy as a classifier under no strategic manipulation of the features can achieve on the subpopulation $S \in \{A, B\}$: $f_S(y) = \begin{cases} 1 & \text{if } y \geq \sqrt{2\pi}\sigma_S \\ 0 & \text{otherwise} \end{cases}$.

All the contestants in a subpopulation S, with $0 < y < \sqrt{2\pi}\sigma_S$ report their private signals to be $\sqrt{2\pi}\sigma_S$ as cost of this change is < 1 whereas the advantage gained in the probability of being classified as 1 is 1. For all the contestants with private signal $y \leq 0$, the cost of change is too high (≥ 1) and thus, they report their true private signals. Therefore, the classifier f_S ends up classifying everyone with private signal $y > 0$ as 1 which is the accuracy maximizing classification under the "no strategic manipulation" setting.

How strategic classification leads to unfairness: When $\sigma_A \neq \sigma_B$, the optimal classifiers for the subpopulations A and B are different and hence, when we choose a single classifier for both the subpopulations, we are bound to loose on the accuracy of at least one of the subpopulations. Through an example (Theorem 6), we suggest that: *while maximizing the overall accuracy over the universe, the minority group might be at a disadvantage irrespective of whether their costs to change the private signals are higher or lower than the majority subpopulation.* Without loss of generality, we assume that A is the minority subpopulation, that is, $s_A \leq s_B$. In many real life scenarios, the Jury would publish a single classifier for both the subpopulations either because A is a protected group and the Jury is not allowed to discriminate based on the subgroup membership or because the Jury has not yet identified these subpopulations and the differences in their cost functions.

▶ **Theorem 6.** *Let A and B be two subpopulations such that the true qualification functions, h_A, h_B, the probability density functions, π_A, π_B and the cost functions c_A, c_B are as instantiated above.*

Assuming $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$, let f^* be the deterministic classifier that maximizes Jury's utility ($U(f)$), if $s_A < s_B$ and $\sigma_A \neq \sigma_B$ (the cost functions are different), then $U_A(f^*) < U_B(f^*)$, that is, the minority is at a disadvantage, even though their qualifications were identical ($h_A = h_B$, $\pi_A = \pi_B$).

Please refer to Appendix C for the proof.

Next we show that, when the features are appropriately noisy, the optimal classifier from Jury's perspective is fair to the subpopulations. The intuition is as follows: if the noise is large enough such that none of contestants in either of the subpopulations want to manipulate their private signals, then the cost differences become irrelevant and hence, the optimal classifier achieves equal accuracy on both the subpopulations. You would think that this addition of noise would compromise Jury's utility. Subsequently, we show that adding noise might also improve the overall accuracy of the Jury's optimal classifier, therefore, addition of noise can make everyone happier. The latter is a continuation to the results at the start of Section 5 about the usefulness of noise to the Jury under strategic classification.

Noisy features lead to fairer outcomes: Now, we analyze the setting with noisy features and prove the following theorem. The true qualification function h, cost functions (c_A and c_B) and the probability density function Π are as defined for the first example. Let $\sigma = \max\{\sigma_A, \sigma_B\}$. Given a private signal y, the features x are distributed according to the gaussian with mean y and standard deviation σ. The probability density function for the feature x given the private signal y is $p_y(x) = \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$.

▶ **Theorem 7.** *Let A and B be two subpopulations such that the true qualification functions, h_A, h_B, the probability density functions, π_A, π_B and the cost functions c_A, c_B are as instantiated above. When the features are drawn with a gaussian noise of mean 0 and standard deviation σ, such that, $\sigma \geq \sigma_A, \sigma_B$, if f^* is the deterministic classifier that maximizes Jury's utility ($U(f)$), then f^* is fair, that is, $U_A(f^*) = U_B(f^*)$.*

Please refer to Appendix D for the proof.

Theorem 7 would hold for when we are concerned with multiple subpopulations as long as $\sigma \geq \sigma_S$ for every relevant subpopulation S. In words, using noisy features we *can* ensure that the best response of a Jury, maximizing her own utility, is fair to all the subpopulations that are identical in terms of qualifications but different in terms of the costs to manipulate the private signals, as long as the costs of manipulation for a subpopulation are not too small.

Noisy features can also improve Jury's utility: Next, we show that further in some cases, *the addition of noise to the features is not only beneficial for ensuring fairness but might also achieve better overall accuracy under strategic classification compared to when a noiseless signal is used.*

Retaining the instantiations of h_A, h_B, π_A, π_B, c_A, c_B and σ as above, consider the following two scenarios: 1. Jury bases her classifier on the private signal y. 2. The features are drawn with a gaussian noise of mean 0 and standard deviation σ and Jury bases her classifier on the features (x).

Let f_0^* and f_σ^* be the optimal classifiers under strategic classification in the two scenarios respectively. Let $U(f_0^*)$ be the overall classification accuracy (Jury's utility) under Scenario 1 and $U(f_\sigma^*)$ be the overall classification accuracy (Jury's utility) under Scenario 2. We assume that the subpopulations are equally populated, that is, $s_A = s_B$ for simplicity of calculations in the next theorem.

▶ **Theorem 8.** *There exists qualification functions, h_A, h_B, the probability density functions over the private signals, π_A, π_B, the cost functions c_A, c_B and $\sigma > 0$ such that, $U(f^*_\sigma) > U(f^*_0)$, that is, the Jury gets better classification accuracy when the features are drawn with a gaussian noise of mean 0 and standard deviation σ. Here, the subpopulations have identical qualifications ($h_A = h_B$, $\pi_A = \pi_B$) but different cost functions.*

Please refer to Appendix E for the proof. This theorem corroborates the idea that not only the subpopulations, but even the Jury might prefer noisy features. In the above example, for simplicity, we assumed $s_A = s_B$. Therefore, the optimal classifier was fair even in the noiseless setting. But a slight tweak in s_A so that $s_A < s_B$ wouldn't change Jury's utility, in Scenario 1, by much and thus, would give an example where the noiseless setting has both unfairness and lesser overall classification accuracy.

In this paper, we study the interaction of noise with strategic classification through some simple examples, and leave the task of generalizing these results for future research.

6 Discussion

The problem of classification (and the strategic classification problem it entails) is of tremendous importance both practically (affecting pretty much every industry) and theoretically (with implications ranging from algorithms to policy and law). Therefore, clarifying the role randomness plays in this specific family of games is an important goal. Just as in games, randomness may lead to better solution in strategic classification. Moreover, in many important settings (such as college admissions in some jurisdictions), the classifier is required to be deterministic by law – which is not a handicap for algorithmic classification, but is a handicap for strategic one. In addition, we proved that, in many natural cases, any randomized classifier (based on one-dimension) that achieves strictly better accuracy than the optimal deterministic one is not stable from the classifier's standpoint, thus illustrating the difficulty of implementing a randomized classifier in a more complicated scenario with multiple classifiers (such as college admissions). This motivates the use of noisy features as a commitment device, which can improve both accuracy and fairness, and is also practically possible (for example by restricting the types of information available to the classifier).

References

1 Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. *arXiv preprint*, 2016. `arXiv:1610.08210`.

2 Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555. ACM, 2011.

3 Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26. ACM, 2018.

4 Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.

5 Kfir Eliaz and Ran Spiegler. The model selection curse. *American Economic Review: Insights*, 2018.

6 Richard Engelbrecht-Wiggans. On the value of private information in an auction: ignorance may be bliss. *BEBR faculty working paper; no. 1242*, 1986.

7 Alex Frankel and Navin Kartik. Improving information from manipulable data. *arXiv preprint*, 2019. `arXiv:1908.10330`.

8 Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122. ACM, 2016.

9 Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268. ACM, 2019.

10 Nicole Immorlica, Katrina Ligett, and Juba Ziani. Access to population-level signaling as a source of inequality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 249–258. ACM, 2019.

11 Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.

12 Andrew Kephart and Vincent Conitzer. Complexity of mechanism design with signaling costs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 357–365, 2015.

13 Andrew Kephart and Vincent Conitzer. The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 85–102, 2016.

14 Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844. ACM, 2019.

15 John Miller, Smitha Milli, and Moritz Hardt. Strategic adaptation to classifiers: A causal perspective. *arXiv preprint*, 2019. `arXiv:1910.10362`.

16 Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239. ACM, 2019.

17 Christopher A Wilkens, Ruggiero Cavallo, Rad Niazadeh, and Samuel Taggart. Mechanism design for value maximizers. *arXiv preprint*, 2016. `arXiv:1607.04362`.

A Proof of Claim 4

Recalling, $g(x) = f(\delta(x)) - c(x, \delta(x))$. Using definition of δ, we know that

$$g(x) = f(\delta(x)) - c(x, \delta(x)) \geq f(\Delta(x)) - c(x, \Delta(x)) \tag{5}$$

And, using definition of Δ, we can show that

$$f(\Delta(x)) \geq g(x) \tag{6}$$

This is because, either $f(\delta(x)) - c(x, \delta(x)) = f(x)$ and as $f(\Delta(x)) \geq f(x)$, we get the inequality. Or, $f(\delta(x)) - c(x, \delta(x)) > f(x)$, which implies that x has an incentive to change its feature to $\delta(x)$. Therefore, by the definition of Δ, $f(\Delta(x)) \geq f(\delta(x)) \geq f(\delta(x)) - c(x, \delta(x))$. The expression in the claim can be rewritten as

$$\begin{aligned}&(g(x) - f(\Delta(x))) \cdot (2h(x) - 1) + h(x) \cdot c(x, \Delta(x)) \\ &\qquad = (g(x) - f(\Delta(x))) \cdot (h(x) - 1) + h(x) \cdot (g(x) - f(\Delta(x)) + c(x, \Delta(x)))\end{aligned}$$

As $g(x) - f(\Delta(x)) \leq 0$ from Equation 6 and $g(x) - f(\Delta(x)) + c(x, \Delta(x)) \geq 0$ from Equation 5, the inequality follows from the fact that $0 \leq h(x) \leq 1$. This proves the claim.

B Proof of Theorem 5

Equation 2 implies that for all classifiers $g \in \mathcal{X} \to [0,1]$,

$$\begin{aligned} &\sum_{x \in \mathcal{X}} \pi(x)[(f(\Delta_f(x)) - g(\Delta_f(x))) \cdot (2h(x)-1)] \geq 0 \\ \implies &\sum_{y \in \mathcal{X}} (f(y) - g(y)) \cdot \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) \geq 0 \end{aligned}$$

Therefore, if f is in equilibrium from the Jury's perspective, for all $y \in \mathcal{X}$ such that $f(y) \in (0,1)$, $\sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) = 0$ otherwise Jury can choose $g(y) = 1$ (or 0) depending on whether $\sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) > 0$ (or < 0) to increase her accuracy. Therefore, accuracy of the classifier f is given by

$$\begin{aligned} U(f) &= \sum_{x \in \mathcal{X}} \pi(x)[f(\Delta_f(x)) \cdot (2h(x)-1) + (1-h(x))] \\ &= \sum_{y \in \mathcal{X}} f(y) \cdot \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) + \sum_{x \in \mathcal{X}} \pi(x)(1-h(x)) \\ &= \sum_{y:f(y)=1} \sum_{x:\Delta_f(x)=y} \pi(x)(2h(x)-1) + \sum_{x \in \mathcal{X}} \pi(x)(1-h(x)) \end{aligned}$$

Consider a binary classifier $f' : \mathcal{X} \to \{0,1\}$ defined as follows: $f(x) \in [0,1) \implies f'(x) = 0$ and $f(x) = 1 \implies f'(x) = 1$. We can show that $U(f') \geq U(f)$. The contestants who change their features when f' is the published classifier is a subset of $\{x \in \mathcal{X} \mid f(\Delta_f(x)) \in (0,1]\}$ and as $\sum_{x:f(\Delta_f(x)) \in (0,1)} \pi(x)(2h(x)-1) = 0$, the accuracy of f' can only increase. This is because: $\forall x \in \mathcal{X}$ if $f(\Delta_f(x)) = 0$, then $f'(\Delta_{f'}(x)) = 0$ as otherwise if x changed its feature under f', it had an incentive to change under f too.

If $x' > x$, $f(\Delta_f(x')), f(\Delta_f(x)) \in (0,1)$ and x changes its feature under f', then x' has the incentive to change too as $c(x',x) = 0$, and hence, the subset of $\{x \in \mathcal{X} \mid f(\Delta_f(x)) \in (0,1)\}$ that change their features under f' can only do a positive addition to the utility (h is monotonically increasing with x and $\sum_{x:f(\Delta_f(x)) \in (0,1)} \pi(x)(2h(x)-1) = 0$). And, the contestants (x) who changed their features under f such that $f(\Delta_f(x)) = 1$ would also change their features under f' such that $f'(\Delta_{f'}(x)) = 1$ (as $f'(x) \leq f(x)$) and are already included in the calculation of $U(f)$.

C Proof of Theorem 6

Jury publishes a deterministic classifier and as there's no noise involved, without loss of generality, we can assume that f is a threshold classifier on the space $\mathcal{X}$ (as c_A and c_B are simple cost functions). This assumption is justified in Section 3. Given the classifier $f : \mathcal{X} \to \{0,1\}$ with threshold τ, the best response of a contestant in the subpopulation $S \in \{A, B\}$ is given as follows:

$$\Delta_f^S(y) = \begin{cases} y & \text{if } y \geq \tau \\ \tau & \text{if } \tau - \sqrt{2\pi}\sigma_S < y < \tau \\ y & \text{if } y \leq \tau - \sqrt{2\pi}\sigma_S \end{cases}$$

The accuracy of the classifier f for the subpopulation S is given as follows:

$$U_S(f) = \int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h(y)-1) + (1-h(y))]dy$$

Let $c = \int_{-\infty}^{\infty} \Pi(y)[(1 - h(y))]dy$ which is independent of the subpopulation and the classifier. Therefore, $U_S(f) = \left(\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h(y) - 1)]dy\right) + c$.

For the convenience of calculations, we will replace $h(y)$ with the following function,

$$h'(y) = \frac{y}{2d} + \frac{1}{2}$$

As d is large and Π is a gaussian centered at 0, this change barely affects the utility values. To be precise, the difference in the utility calculations for any classifier f while using h' instead of h is bounded by

$$\begin{aligned}\left|\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot 2(h(y) - h'(y))]dy\right| \leq& 2\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y))|h(y) - h'(y)|]dy \\ \leq& 2\int_{-\infty}^{\infty} \Pi(y) \cdot |h(y) - h'(y)|dy \\ =& 4\int_{d}^{\infty} \Pi(y) \cdot \left(\frac{y}{2d} - \frac{1}{2}\right) dy \\ \leq& 2\int_{d}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi t}} \cdot \frac{y}{d}\, dy \quad = 2\frac{te^{-\frac{d^2}{2t^2}}}{\sqrt{2\pi d}}\end{aligned}$$

As we take d ($d >> t$) to be large enough, we would be able to ignore this difference. From now onwards, we use h' as the "true qualification function".

Therefore, the accuracy of the classifier f over the subpopulation $S \in \{A, B\}$ can be approximated by

$$\begin{aligned}U_S(f) &= \left(\int_{-\infty}^{\infty} \Pi(y)[f(\Delta^S(y)) \cdot (2h'(y) - 1)]dy\right) + c = \left(\int_{-\infty}^{\infty} \Pi(y) \cdot f(\Delta^S(y)) \cdot \frac{y}{d}\, dy\right) + c \\ &= \left(\int_{\tau-\sqrt{2\pi}\sigma_S}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi t}} \cdot \frac{y}{d}\, dy\right) + c = \frac{t}{\sqrt{2\pi d}} e^{-(\tau-\sqrt{2\pi}\sigma_S)^2/2t^2} + c\end{aligned}$$

The second last equality follows from the definition of Δ_f^S and the fact that f classifies everyone, with the updated private signal greater than or equal to τ, as 1 and 0 otherwise.

The overall accuracy of the classifier f is given by

$$\begin{aligned}U(f) &= s_A \cdot U_A(f) + s_B \cdot U_B(f) \\ &= s_A \cdot \frac{t}{\sqrt{2\pi d}} e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot \frac{t}{\sqrt{2\pi d}} e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2} + c \qquad (7)\end{aligned}$$

It's clear from the expression that the accuracy for the subpopulation A is maximized at $\tau_A = \sqrt{2\pi}\sigma_A$ and that of B is maximized at $\tau_B = \sqrt{2\pi}\sigma_B$. Consider the case when $s_A < s_B$. As $\tau_A \neq \tau_B$, and $U_B(f)$ has a larger weight in the expression, intuitively, while optimizing the overall accuracy, τ would try to achieve better accuracy for the subpopulation B, irrespective of whether $\sigma_A > \sigma_B$ or $\sigma_A < \sigma_B$, leading to unfairness across the subpopulations (A being at a disadvantage).

It's complicated to calculate the optimal τ, below we give a proof of the fact that the optimal τ would be such that $U_A(f) < U_B(f)$. To find the optimal value of τ, we differentiate $U(f)$ with respect τ as follows:

$$\begin{aligned}\frac{dU(f)}{d\tau} &= s_A \cdot \frac{dU_A(f)}{d\tau} + s_B \cdot \frac{dU_B(f)}{d\tau} \\ &= -\frac{1}{\sqrt{2\pi}td}\left(s_A \cdot (\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot (\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau-\sqrt{2\pi}\sigma_B)^2/2t^2}\right)\end{aligned}$$

Therefore, $\frac{dU(f)}{d\tau} = 0$

$$\implies s_A \cdot (\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau - \sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot (\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau - \sqrt{2\pi}\sigma_B)^2/2t^2} = 0$$
$$\implies \left| \frac{(\tau - \sqrt{2\pi}\sigma_A) \cdot e^{-(\tau - \sqrt{2\pi}\sigma_A)^2/2t^2}}{(\tau - \sqrt{2\pi}\sigma_B) \cdot e^{-(\tau - \sqrt{2\pi}\sigma_B)^2/2t^2}} \right| > 1 \qquad (s_B > s_A)$$

As $ze^{-\frac{z^2}{2t^2}}$ is maximized at $z = t$, as long as $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$ (implying $|\tau - \sqrt{2\pi}\sigma_S| \leq t$ for $S \in \{A, B\}$), the overall accuracy is maximized at a threshold τ such that $|\tau - \sqrt{2\pi}\sigma_A| > |\tau - \sqrt{2\pi}\sigma_B|$ and hence, $U_A(f^*) < U_B(f^*)$, where f^* is the optimal classifier from Jury's perspective. The assumption, $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$, can be interpreted as the subpopulations being different but not extremely different, which is reasonable assumption in many real life scenarios.

D Proof of Theorem 7

Again, we will replace the function h with h' (as in proof of Theorem 6) while loosing an insignificant amount in all the calculations ($d >> t, \sigma$). Let $\Pi' : \mathcal{X} \to [0, 1]$ be the probability density function over the features realized by each of the subpopulations. Let $H(x)$ ($H : \mathcal{X} \to [0, 1]$) represent the probability of an individual being qualified (1) given that the Jury sees feature x. These functions are same for both the subpopulations. As the Jury only sees the feature and not the private signal, her accuracy is information-theoretically limited by these functions as we will describe below. Firstly, $\Pi' : \mathcal{X} \to [0, 1]$ is given as follows:

$$\begin{aligned}
\Pi'(x) = \int_{-\infty}^{\infty} \Pi(y) \cdot p_y(x) dy &= \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t} \cdot \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dy \\
&= \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t} \cdot \frac{e^{-(y - \frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})}}{\sqrt{2\pi}\sigma} dy \\
&= \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(y - \frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2t^2}{\sigma^2+t^2})} dy \\
&= \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma} \sqrt{2\pi \frac{\sigma^2 t^2}{\sigma^2 + t^2}} = \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi(\sigma^2 + t^2)}}
\end{aligned}$$

Therefore, the probability density function over the features realized by the subpopulations, with $\mathcal{N}(0, \sigma)$ gaussian noise, is itself a gaussian with mean 0 and $\sqrt{(\sigma^2 + t^2)}$ standard deviation.

The qualification function given the features, H, is given as follows:

$$H(x) = \frac{1}{\Pi'(x)} \int_{-\infty}^{\infty} \Pi(y) \cdot p_y(x) \cdot h(y) dy$$

We replace h with h', thus replacing H with H' as defined below:

$$
\begin{aligned}
H'(x) = \frac{1}{\Pi'(x)} \int_{-\infty}^{\infty} \Pi(y) \cdot p_y(x) \cdot h'(y) dy &= \frac{1}{\Pi'(x)} \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2t^2}}}{\sqrt{2\pi}t} \cdot \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \cdot (\frac{y}{2d} + \frac{1}{2}) dy \\
&= \frac{1}{2} + \frac{1}{\Pi'(x)} \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t} \cdot \frac{e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2 t^2}{\sigma^2+t^2})}}{\sqrt{2\pi}\sigma} \cdot \frac{y}{2d} \;\; dy \\
&= \frac{1}{2} + \frac{1}{2d \cdot \Pi'(x)} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(y-\frac{xt^2}{\sigma^2+t^2})^2/(2\frac{\sigma^2 t^2}{\sigma^2+t^2})} \cdot y \;\; dy \\
&= \frac{1}{2} + \frac{1}{2d \cdot \Pi'(x)} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi}t \cdot \sqrt{2\pi}\sigma} \cdot \sqrt{2\pi \frac{\sigma^2 t^2}{\sigma^2+t^2}} \cdot \frac{xt^2}{\sigma^2+t^2} \\
&= \frac{1}{2} + \frac{t^2}{\sigma^2+t^2} \frac{x}{2d}
\end{aligned}
$$

Therefore, when there's no strategic manipulation, Jury would classify any individual with feature $x > 0$ as 1 and 0 otherwise. This is because, $H'(x) > \frac{1}{2}$ if and only if $x > 0$ and the Jury would classify a feature as 1 if and only if, in expectation, the individuals with that feature are more likely to be qualified. This is true irrespective of whether an individual is from the subpopulation A or B because these subpopulations are identical in terms of qualifications, that is, $h_A = h_B = h$ and $\pi_A = \pi_B = \Pi$.

We show that for the cost functions defined above, if Jury publishes $f = 1_{x>0}$, as the classifier, then none of the contestants in both the subpopulations A and B have an incentive to change their private signal (under $\mathcal{N}(0, \sigma)$ gaussian noise). Hence, the Jury gets the best possible accuracy from these features and the classification is fair. For a subpopulation $S \in \{A, B\}$, let $q_f^S(y)$ denote the probability of a contestant, with private signal y, being classified as 1 when f is the classifier. Therefore,

$$
q_f^S(y) = \int_{-\infty}^{\infty} f(x) \cdot p_y(x) dx = \int_0^{\infty} \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx
$$

For a subpopulation $S \in \{A, B\}$, let's calculate the advantage that a contestant, with private signal y, gets by changing its signal to y' ($y' > y$, otherwise $q_f^S(y') \leq q_f^S(y)$):

$$
\begin{aligned}
q_f^S(y') - q_f^S(y) &= \int_0^{\infty} \frac{e^{-\frac{(x-y')^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx - \int_0^{\infty} \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx \quad = \int_{-y'}^{\infty} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx - \int_{-y}^{\infty} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx \\
&= \int_{-y'}^{-y} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx \quad \leq \int_{-y'}^{-y} \frac{1}{\sqrt{2\pi}\sigma} dx \quad = \frac{y'-y}{\sqrt{2\pi}\sigma}
\end{aligned}
$$

As $\sigma = \max\{\sigma_A, \sigma_B\}$ and recalling the definitions of the cost functions c_A and c_B (Equation 4), we get that

$$
q_f^A(y') - q_f^A(y) \leq c_A(y, y') \qquad \text{and} \qquad q_f^B(y') - q_f^B(y) \leq c_B(y, y')
$$

Therefore, none of the contestants in any of the subpopulations have an incentive to change their private signals. The accuracy of the classifier f on the subpopulation A is given as

$$\begin{aligned}
U_A(f) &= \left(\int_{-\infty}^{\infty} \Pi(y)[q_f^A(\Delta_f^A(y)) \cdot (2h(y)-1)]dy\right) + c \\
&= \left(\int_{-\infty}^{\infty} \Pi(y) \int_0^{\infty} \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx \cdot (2h(y)-1)dy\right) + c \\
&= \left(\int_0^{\infty} \left(\int_{-\infty}^{\infty} \Pi(y) \frac{e^{-\frac{(x-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \cdot (2h(y)-1)dy\right) dx\right) + c \\
&= \left(\int_0^{\infty} \Pi'(x) \cdot (2H(x)-1)dx\right) + c
\end{aligned}$$

Replacing H with H' without loosing much in the approximation, we get that

$$U_A(f) = \left(\int_0^{\infty} \frac{e^{-\frac{x^2}{2(\sigma^2+t^2)}}}{\sqrt{2\pi(\sigma^2+t^2)}} \cdot \frac{t^2}{\sigma^2+t^2} \frac{x}{d} dx\right) + c = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$$

Similarly for $U_B(f)$ and hence, $U(f) = U_B(f) = U_A(f) = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$.

E Proof of Theorem 8

We retain the instantiations of h_A, h_B, π_A, π_B, c_A, c_B and σ as above. As seen above, in Scenario 2, $1_{x>0}$ is the classifier that optimizes Jury's utility and hence, $U(f_\sigma^*) = \frac{t^2}{\sqrt{2\pi(\sigma^2+t^2)} \cdot d} + c$. Actually, it's approximately equal to this but the error is extremely small ($e^{-\Omega(d)}$, $d >> t, \sigma$).

In Scenario 1, the utility of any threshold classifier (f) with τ as the threshold is given by Equation 7 (without loss of generality, we can optimize over threshold classifiers). Therefore,

$$U(f) = s_A \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau - \sqrt{2\pi}\sigma_A)^2/2t^2} + s_B \cdot \frac{t}{\sqrt{2\pi}d} e^{-(\tau - \sqrt{2\pi}\sigma_B)^2/2t^2} + c$$

When $s_A = s_B = \frac{1}{2}$ and we assume that $|\sigma_A - \sigma_B| \leq \frac{t}{\sqrt{2\pi}}$, it's easy enough to see that the above expression is maximized at $\tau = \frac{\sqrt{2\pi}\sigma_A + \sqrt{2\pi}\sigma_B}{2}$. Therefore, the optimal classification accuracy in Scenario 1, is

$$U(f_0^*) = \frac{t}{\sqrt{2\pi}d} e^{-(\frac{\sqrt{2\pi}\sigma_A - \sqrt{2\pi}\sigma_B}{2})^2/2t^2} + c$$

For $\sigma_B = \sigma$, $\sigma_A = 0.1\sigma$, $t = 0.9\sqrt{2\pi}\sigma$, $U(f_\sigma^*) > U(f_0^*)$.

Bounded-Leakage Differential Privacy

Katrina Ligett

Computer Science Department, Hebrew University of Jerusalem, Israel
katrina@cs.huji.ac.il

Charlotte Peale

Stanford University, Stanford, CA, USA
cpeale@stanford.edu

Omer Reingold

Computer Science Department, Stanford University, Stanford, CA, USA
reingold@stanford.edu

Abstract

We introduce and study a relaxation of differential privacy [3] that accounts for mechanisms that leak some additional, bounded information about the database. We apply this notion to reason about two distinct settings where the notion of differential privacy is of limited use. First, we consider cases, such as in the 2020 US Census [1], in which some information about the database is released exactly or with small noise. Second, we consider the accumulation of privacy harms for an individual across studies that may not even include the data of this individual. The tools that we develop for bounded-leakage differential privacy allow us reason about privacy loss in these settings, and to show that individuals preserve some meaningful protections.

2012 ACM Subject Classification Theory of computation → Theory of database privacy and security

Keywords and phrases differential privacy, applications, privacy, leakage, auxiliary information

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.10

Funding *Katrina Ligett*: This work was supported in part by Israel Science Foundation (ISF) grant #1044/16, United States Air Force and DARPA under contracts FA8750-16-C-0022 and FA8750-19-2-0222, and the Federmann Cyber Security Center in conjunction with the Israel national cyber directorate. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.
Omer Reingold: Supported in part by NSF Award IIS-1908774 and by VMWare fellowship.

Acknowledgements Part of this work was done while the first and third authors were visiting the Simons Institute for the Theory of Computing.

1 Introduction and Related Work

Differential privacy [3], a notion of the stability of computations, has emerged as the gold standard of mathematically rigorous privacy protection for computations on statistical databases, in part because of its appealing interpretations from the perspective of individuals in the database. For example, an economic interpretation of differential privacy holds that individuals' future utilities will be harmed by at most a small constant factor by providing their true data, rather than lying, to the mechanism [5]. Also appealing is an interpretation that shows that an individual's truthful provision of data to a differentially private mechanism will not substantially change a Bayesian observer's posterior beliefs versus the beliefs that would result if that individual provided false data [10].

One clear advantage of differential privacy over other approaches to defining privacy is that it is not necessary to reason about auxiliary information in order to give differential privacy guarantees. Composition attacks [7] leveraging access to such "outside information"

1st Symposium on Foundations of Responsible Computing (FORC 2020).
Editor: Aaron Roth; Article No. 10; pp. 10:1–10:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

are a common Achilles' Heel of ad hoc privacy notions. Differential privacy, as it is a property of the *mechanism*, rather than the mechanism's output, is a guarantee that holds regardless of the presence of auxiliary information. However, as observed by Dwork and Naor [4] and Kifer and Machanavajjhala [11, 12], a differentially private release of data against a background of prior data releases can still result in substantial privacy harms. In particular, Dwork and Naor show essentially that innocuous information can always be combined with statistical information to yield a privacy breach. These negative results regarding auxiliary information are used by Dwork and Naor as justification for differential privacy's focus on *relative* guarantees – harms relative to the harm you would have experienced had you not participated in the study or lied about your data.

In this work, we return to the question of auxiliary information, which we term "leakage." We develop a formal framework in which to study differential privacy in the presence of leakage and argue about non-trivial properties of such bounded-leakage differential privacy. Our research is motivated by two applications:

Application: 2020 Census

One context in which auxiliary information has recently gained attention is the 2020 US Census. The majority of data releases, including synthetic data, for the 2020 US Census, are slated to be released subject to differential privacy [1, 8]. However, the Census, by agreement with the Department of Justice, will provide "exact counts,"[1] known as *invariants*, for certain statistics [8, 9, 2]. As Garfinkel et al. [9] allude to, "there is no well-developed theory for how differential privacy operates in the presence of such invariants."

One particularly nefarious problem with auxiliary information emerges if the function that produced the auxiliary information might share randomness with the differentially private algorithm. If so, all guarantees of differential privacy might be lost. Our definitions directly apply to this setting and help clarify the impact of such auxiliary information (and in particular, the role of such shared randomness).

Application: Big-World Privacy

One of the benefits of differential privacy is that it gives a way to quantify the privacy losses of individuals whose data were included in a database input into a mechanism, and even allows quantification of how privacy losses "add up" across multiple mechanisms. In its most basic form, differential privacy tells us that for every ϵ-differentially private study an individual participates in, she incurs an ϵ privacy loss. However, it is also relevant to ask whether a study that an individual *does not* participate in could also degrade that individual's privacy. Consider, for example, a study that included everyone in a certain city who had a particular disease. Then, this would mean that the information that a particular individual from that city did not participate in the study also tells us that this individual doesn't have the disease, and could potentially degrade the individual's privacy in unexpected ways.

If we cannot necessarily quantify an individual's overall privacy loss using only the studies they have participated in, what should we do? One possible solution is to treat all individuals in the world as belonging to a single huge database D. A differentially private mechanism M that computes over some subset of the population would be considered instead to be a composition of two mechanisms: first, a selection function for determining which individuals to include in the study, and then the subsequent differentially private mechanism. This concatenated mechanism, however, need not be differentially private if the selection function

[1] "Exact counts" are of course not really an exact population count, but reflect many sources of error, including non-participation and potentially also Census techniques intended to infer missing data.

is not differentially private. In fact, the selection function could adversarially choose a subset of the database in order to cause sensitive data about an individual to be encoded into the likely outputs of the subsequent differentially private mechanism.

Even if the selection of studies that include the data of a particular individual is independent for other individuals, then the inclusion or exclusion of the data of individual i in study j may depend on i's data. It seems that we cannot avoid the conclusion that the only bound that differential privacy can give us about i's privacy loss is the composition of the ϵ privacy losses for every ϵ-differentially private study that has ever occurred, *even for studies in which i did not participate*. Our framework allows us to separate the privacy loss into that which is incurred from exclusion from particular studies and the loss that is incurred from participation in ϵ-differentially private studies. Once we acknowledge the leakage of an upper-bound on the number of studies i participate in, the privacy loss is incurred as a function of this upper bound rather than as a function of all studies.

Our Contributions

In this paper, we present (Section 2) a definition for *bounded-leakage differential privacy*, a relaxed variant of differential privacy that quantifies the privacy that is maintained by a mechanism despite bounded, additional leakage of information by some "leakage function." We investigate (Section 3) the connections between standard differential privacy and this new variant, and give conditions for when the bounded-leakage privacy of a mechanism can imply something about its differential privacy, and vice versa.

Differentially private mechanisms are known to satisfy some appealing, simple properties such as privacy conservation under post-processing and quantifiable privacy loss for groups. We show that bounded-leakage privacy satisfies the same post-processing results as standard differential privacy (Theorem 11), and prove an analogous result for the group privacy of bounded-leakage differentially private pairs of mechanisms and leakage functions (Theorem 22). Additionally, we show that explicitly "leaking" the value of the leakage function does not affect the bounded-leakage differential privacy of a mechanism/function pair (Corollary 14).

There are numerous results about the composition of differentially private mechanisms, both in a non-adaptive setting where the mechanisms and databases queried are chosen before execution, and additionally in an adaptive setting where intermediate outputs may affect the choice of future queries. We define adaptive composition for bounded-leakage privacy and, using a new reduction technique, we show that any general adaptive or non-adaptive composition bound that holds for differentially private mechanisms must also hold for the class of bounded-leakage private mechanism/function pairs, with the same privacy parameters (Section 4.4).

It is conceivable that leaking additional information about a mechanism could also affect the utility of its outputs. The exponential mechanism for differentially private mechanisms presents a way to construct a differentially private mechanism with a high utility guarantee on its output. We define an analogous mechanism for the bounded-leakage privacy setting, such that given a leakage function and a utility function, we can construct a bounded-leakage private mechanism with high utility guarantees for each possible output of the leakage function (Section E).

Finally, (Section 6) we use the bounded-leakage differential privacy framework to study the Census and Big-World applications, and show that it is possible to formally bound privacy harms in these settings.

Due to space constraints, almost all proofs appear in the appendices.

Additional Related Work

Kasiviswanathan and Smith [10] also provide a formal treatment of auxiliary information, in the context of their Bayesian interpretation of differential privacy. [11] explore a notion of privacy in the context of auxiliary information that is similar to ours, but more limited in the (deterministic) prior releases they consider. The Pufferfish framework [12] is an extremely general privacy framework that allows for reasoning about the conclusions drawn by specific types of attackers. [12] explicitly explores composition of private mechanisms with non-private mechanisms in their Sections 9 and 10, and gives general statements based on the conditional probability distribution of one release given the other. This appears to have analogy in the exploration of independence that we do in Section 3. Given the generality of the Pufferfish framework, it is likely capable of describing our notion of bounded-leakage differential privacy. Our notion is focused on the behavior of differential privacy in the presence of auxiliary information rather than on stronger notions that are resilient to adversaries with auxiliary information. By focusing on a more specific notion, we are able to build up a richer set of properties and consequences of the notion.

2 Model

The standard definition of differential privacy promises that the distribution of results of a randomized mechanism run on any database does not change too much if we change any particular individual's data in that database. Formally, we will represent a database that holds data about n individuals as a tuple from the set $\mathcal{X}^n$ where $\mathcal{X}$ is a data universe of possible characteristics. We call two databases $x, x' \in \mathcal{X}^n$ *neighboring* or *adjacent* if they differ in the data of one individual, and will denote this as $x \sim x'$. Using this notation, the notion of standard differential is formally defined as follows:

▶ **Definition 1** ((ϵ,δ)-Differential Privacy (DP) [3]). *Let $\mathcal{X}$ be some data universe, O an output space, and R a space of random inputs. Given a mechanism $M : \mathcal{X}^n \times R \to O$, we say that M is* (ϵ,δ)-differentially private *if for all neighboring databases $x \sim x' \in \mathcal{X}^n$ and all subsets $S \subseteq O$, we have that*

$$Pr_{r\in R}[M(x,r) \in S] \leq e^{\epsilon} Pr_{r\in R}[M(x',r) \in S] + \delta.$$

Building off of this definition, we introduce a new definition for a variant of differential privacy that we call *bounded-leakage differential privacy.* Intuitively, bounded-leakage differential privacy asserts that, given a database and a mechanism to be run on it, if an additional piece of information about the database were leaked, then the output of the mechanism when run on the database would not leak much more information than what was already leaked.

▶ **Definition 2** ((ϵ,δ)-Bounded-Leakage Differential Privacy (blDP)). *Let $\mathcal{X}$ be some data universe, O_M an output space, O_P a countable output space, and R a space of random inputs. Given a mechanism $M : \mathcal{X}^n \times R \to O_M$, and a* leakage function *$P : \mathcal{X}^n \times R \to O_P$, we say that M is* (ϵ,δ)-bounded-leakage differentially private *with respect to P if for all neighboring databases $x \sim x' \in \mathcal{X}^n$, all $S \subseteq O_M$, and $o \in O_P$, we have that either*

$$Pr_{r\in R}[P(x',r) = o] \cdot Pr_{r\in R}[P(x,r) = o] = 0$$

or

$$Pr_{r\in R}[M(x,r) \in S | P(x,r) = o] \leq e^{\epsilon} Pr_{r\in R}[M(x',r) \in S | P(x',r) = o] + \delta.$$

For the rest of this paper, given any mechanism-function pair (M, P), we will by default denote the output space of M as O_M and the output space of P as O_P.

3 The Relationship Between blDP and DP

3.1 When Does blDP Imply DP?

Bounded-leakage differential privacy is clearly a weaker notion than standard differential privacy due to the fact that the robustness constraint across two neighboring databases is only required to hold conditioning on the value of the leakage function. The following example illustrates one scenario in which we can have perfect bounded-leakage differential privacy, but no meaningful differential privacy.

▶ **Example 3** (perfect blDP does not imply DP for the mechanism)**.** Consider any arbitrary mechanism M. The pair (M, M) satisfies perfect bounded-leakage differential privacy because for any databases $x \sim x'$, $S \subseteq O_M$, and $o \in O_P$ such that $\Pr_r[M(x,r) = o] \cdot \Pr_r[M(x',r) = o] \neq 0$, we will always have

$$\Pr_r[M(x,r) \in S | M(x,r) = o] = \Pr_r[M(x',r) \in S | M(x',r) = o] = \begin{cases} 1 & \text{if } o \in S \\ 0 & \text{otherwise} \end{cases}$$

and so $\Pr_r[M(x,r) \in S | M(x,r) = o] = \Pr_r[M(x',r) \in S | M(x',r) = o]$. Therefore, (M, M) satisfies perfect blDP, but M need not have any sort of differential privacy guarantee.

We should note that this example depends on the fact that blDP is defined such that the privacy mechanism and associated leakage function share the same random input.

However, we *can* say something about the differential privacy of the mechanism in a blDP pair if we additionally know that the leakage function is differentially private.

▶ **Theorem 4** (blDP with a DP leakage function)**.** *Suppose $M : \mathcal{X}^n \times R \to O_M$ is a mechanism satisfying (ϵ_1, δ_1)-blDP with respect to a leakage function $P : \mathcal{X}^n \times R \to O_P$. Additionally, suppose that P satisfies $(\epsilon_2, 0)$-DP. Then M satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP.*

3.2 When Does DP Imply blDP?

In the previous subsection, we noted that intuitively, blDP seems like a weaker notion than DP. We saw that blDP for a privacy mechanism and leakage function pair does not guarantee anything about the DP of the privacy mechanism. Following this intuition further, we might also expect that having a mechanism/function pair where the mechanism is DP should guarantee that the pair satisfies blDP. However, this is actually not necessarily the case. Consider the following example where a perfectly DP mechanism loses all blDP when paired with a particular leakage function:

▶ **Example 5** (perfect DP does not guarantee blDP)**.** Consider the mechanism $M : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^n$. Then, for $x, r \in \{0,1\}^n$, define M to be $M(x,r) = x \oplus r$. Under this definition, M's output is uniformly distributed over $\{0,1\}^n$, making it perfectly DP.

Now, define the leakage function $P : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^n$ to be $P(x,r) = r$. Conditioning on P being a particular value means that the randomness used in M will be fixed. For any $x \neq x'$, we will always have $x \oplus r \neq x' \oplus r$ and therefore $M(x,r) \neq M(x',r)$.

Thus, (M, P) does not satisfy blDP for any practical values of ϵ and δ, despite the fact that M is perfectly DP.

Why does our intuition fail here? We can identify three properties of the above example that lead to this unexpected result:

(1) P released information about its random input.
(2) M and P shared the same random input.
(3) The output of M depended on the output of P.

In fact, one can show that if the example above were changed such that *any* of these three properties did not hold, then we would get the opposite result and the differential privacy of M *would* imply bounded-leakage differential privacy for the pair.

Informally, we see that conclusions about the blDP of a pair will depend on how correlated the mechanism and the leakage function are.

▶ **Definition 6.** *We call a mechanism-function pair* (M, P) perfectly independent *if for any database* x *,* $S \subseteq O_M$*, and output* $o \in O_P$ *such that* $\Pr_r[P(x,r) = o] \neq 0$*, we have*

$$\Pr_r[M(x,r) \in S | P(x,r) = o] = \Pr_r[M(x,r) \in S].$$

In other words, the output of M is completely uncorrelated with the value of P when given the same database and random input.

We consider two common examples of perfectly independent leakage functions. One might be where M and/or P are completely deterministic, such as a function that releases an exact summary statistic. A second interesting case is where P uses "fresh randomness," disjoint from M's computation. M and P may share the same random input string, but if the random bits they depend on are completely separate, then their outputs will be perfectly independent.

If a mechanism-function pair is perfectly independent, then a DP mechanism *does* imply blDP for the pair. The following lemma follows immediately from combining the definition of perfect independence with the definitions of DP and blDP, and its proof is omitted:

▶ **Lemma 7.** *If* (M, P) *is a perfectly independent mechanism-function pair and* M *satisfies* (ϵ, δ)*-DP, then* M *must also satisfy* (ϵ, δ)*-blDP with respect to* P*.*

Perfectly independent pairs constitute the class of mechanism-function pairs with completely uncorrelated outputs. On the other end of the spectrum, we have pairs such as the one presented in Example 5 where the output of M is a deterministic function of the output of P. However, there are mechanism-function pairs that lie between these two extremes for which we would also like to derive blDP bounds. Such pairs can be thought of as mechanisms that are only partially dependent on the output of their associated leakage functions (or vice versa).

Where might we see partially independent mechanism-function combinations in practical applications of blDP? One possible scenario is a study that computes its output based on a random sample of individuals chosen from the provided database. For such a study, the leakage function may need to depend on the randomness used to pick the sample. This might be due to space reasons – for example, the study might discard other data after picking the random sample – or for accuracy reasons. In either case, the result of such a leakage function will be somewhat correlated with the randomness used to select the sample in the original study, but may also employ its own randomness as well, giving us a leaked value that partially depends on the randomness used in the original study.

We can quantify this generalized notion of partial independence as follows:

▶ **Definition 8** ((ϵ, δ)-independence)**.** *Consider a mechanism-function pair* (M, P)*. We say that* M *and* P *are* (ϵ, δ)-independent *if for every database* x*, subset* $S \subseteq O_M$*, and output* $o \in O_P$ *such that* $\Pr_r[P(x,r) = o] \neq 0$*, we have that*

$$\Pr_r[M(x,r) \in S | P(x,r) = o] \leq e^{\epsilon} \Pr_r[M(x,r) \in S] + \delta$$

and

$$\Pr_r[M(x,r) \in S] \leq e^{\epsilon} \Pr_r[M(x,r) \in S | P(x,r) = o] + \delta.$$

Note that by this definition, (0, 0)-independent pairs are perfectly independent. Using this definition of dependence, we can get a more general version of Theorem 7 which wells us what sort of blDP bounds we can expect from a partially independent pair with a DP mechanism. This is presented in the following theorem, and we note that by substituting in $\epsilon' = \delta' = 0$, we get the result of Lemma 7 as a corollary.

▶ **Theorem 9.** *Suppose we have a mechanism-function pair (M, P) where the outputs of M and P are (ϵ', δ')-independent of one another. If M satisfies (ϵ, δ)-DP, then (M, P) must satisfy $(\epsilon + 2\epsilon', (e^{\epsilon'+\epsilon} + 1)\delta' + e^{\epsilon'}\delta)$-blDP.*

4 Properties

We can show that bounded-leakage differential privacy satisfies many of the properties satisfied by standard differential privacy [3], plus additional desirable properties.

4.1 Post-Processing

We begin by observing that bounded-leakage differential privacy is closed under convex combination.

▶ **Lemma 10.** *A convex combination of (ϵ, δ)-blDP mechanisms with respect to some leakage function $P : \mathcal{X}^n \times R \to O_P$ is also an (ϵ, δ)-blDP mechanism with respect to the same P.*

With this observation in hand, it quickly follows that blDP is preserved under post-processing of the privacy mechanism. The proof is very similar to the proof for post-processing in the DP setting, and is omitted.

▶ **Theorem 11** (Post-processing). *If $M : \mathcal{X}^n \times R \to O_M$ is an (ϵ, δ)-blDP mechanism with respect to $P : \mathcal{X}^n \times R \to O_P$ and $f : O_M \to O'$ is any arbitrary mapping from O_M to an output space O', then $f \circ M$ is also an (ϵ, δ)-blDP mechanism with respect to P.*

4.2 Group Privacy

Group privacy is an important property of differential privacy that quantifies how privacy degrades between databases that differ in the data of more than one individual. In standard differential privacy, group privacy properties hold due to the fact that given any two databases, we can construct a "path" of adjacent databases and apply differential privacy properties to each pair in the path.

Group privacy is a bit more complicated in the case of bounded-leakage differential privacy due to the fact that we cannot always construct a path between two databases such that the leakage function maintains the same value between all pairs of databases along the path. Example 18 shows a situation where bounded-leakage privacy between pairs of databases cannot imply anything about the group privacy of the same mechanism-function pair. However, we can still state properties about group blDP when for every output o of P, we can find a path between the two databases in question such that for every database x on the path, we have $\Pr[P(x, r) = o] > 0$; we see this in Definitions 20 and 21, and Theorem 22.

4.3 What if the Value of P is Leaked?

The definition of blDP we have presented conditions probability on values of P, but the value of P is never explicitly revealed or “leaked.” However, intuitively, we would like our definition to satisfy the property that even if the value of of P is explicitly revealed (that is, the output of P is included in the mechanism output), the pair retains its bounded-leakage privacy. The following theorem shows that this property holds.

▶ **Theorem 12** (Value of P leaked). *Let $M : \mathcal{X}^n \times R \to O_M$ be a mechanism that satisfies (ϵ, δ)-blDP with respect to the leakage function $P : \mathcal{X}^n \times R \to O_P$. Consider another mechanism, $M' : \mathcal{X}^n \times R \to O_M \times O_P$ such that M' returns the output of M concatenated with the output of P; that is, $M'(x, r) := M(x, r)||P(x, r)$. Then M' also satisfies (ϵ, δ)-blDP with respect to P.*

We can combine this theorem with the independence results of Lemma 7 or Theorem 9 to get Corollaries 13 and 14, respectively

▶ **Corollary 13.** *Suppose M is an (ϵ, δ)-DP mechanism and P is a leakage function such that the outputs of M and P are perfectly independent. Then the concatenation of M and P, $M'(x, r) = M(x, r)||P(x, r)$, satisfies (ϵ, δ)-blDP with respect to P.*

▶ **Corollary 14.** *Suppose M is an (ϵ, δ)-DP mechanism and P is a leakage function such that the outputs of M and P are (ϵ', δ')-independent. Then the concatenation of M and P, $M'(x, r) = M(x, r)||P(x, r)$, satisfies $(\epsilon + 2\epsilon', (e^{\epsilon'+\epsilon} + 1)\delta' + e^{\epsilon'}\delta)$-blDP with respect to P.*

4.4 Composition

We derive some properties of the composition of multiple blDP mechanism-function pairs. There are two types of composition that we consider: non-adaptive composition, in which the sequence of mechanism-function pairs is fixed in advance; and adaptive composition, in which the choice of future mechanism-function pairs might depend on the results returned by previous mechanisms.

We use a unified reduction technique to obtain results for both settings.

▶ **Definition 15** (DP reduction mechanism). *Given a mechanism $M : \mathcal{X}^n \times R \to O_M$, a leakage function $P : \mathcal{X}^n \times R \to O_P$, some output $o \in O_P$, and two databases $x_0, x_1 \in \mathcal{X}^n$, we define the* DP reduction mechanism *for $(M, P), o, x_0, x_1$ to be the mechanism $M_{P,o}^{x_0,x_1} : \mathcal{X}^n \times (R_{o,x_0} \times R_{o,x_1}) \to O_M \cup \{\text{“null”}\}$, where R_{o,x_b} is defined as the subset of the random input space R such that $R_{o,x_b} = \{r \in R : P(x_b, r) = o\}$. Then, given any $x \in \mathcal{X}^n$ and $(r_0, r_1) \in R_{o,x_0} \times R_{o,x_1}$, $M_{P,o}^{x_0,x_1}(x, (r_0, r_1))$ is defined as*

$$M_{P,o}^{x_0,x_1}(x, (r_0, r_1)) = \begin{cases} \textit{“null”} & \textit{if } \Pr_r[P(x_0, r) = o] \Pr_r[P(x_1, r) = o] = 0 \\ M(x_0, r_0) & \textit{if } x = x_0 \\ M(x_1, r_1) & \textit{otherwise} \end{cases}$$

In order to use this mechanism in our reduction proofs, we need it to satisfy two important properties. The first (Proposition 23) is that the distribution of $M_{P,o}^{x_0;x_1}$ on inputs x_0 and x_1 should match the distribution of M conditioned on P outputting o in the blDP setting for those inputs. The second (Proposition 24) states that if (M, P) satisfies bounded-leakage privacy and x_0 and x_1 are neighboring databases, $M_{P,o}^{x_0;x_1}$ must be differentially private.

In Section C, we show how to use this reduction to translate results on non-adaptive composition of differentially private mechanisms to results for blDP. Section D shows the analogous reduction for adaptive composition, yielding the following theorem:

▶ **Theorem 16.** *For all $\epsilon, \delta, \delta' \geq 0$, the class of (ϵ, δ)-blDP mechanisms satisfies $(\epsilon', k\delta + \delta')$-blDP under k-fold adaptive composition for $\epsilon' = \epsilon\sqrt{2k \ln(1/\delta')} + k\epsilon(e^{\epsilon-1})$.*

5 Tools for Achieving blDP

Our Lemma 7 and Theorem 9 give tools for understanding when existing differentially private algorithms can be used to achieve guarantees of blDP. In addition, in Section E, we establish a blDP variant of the exponential mechanism [13]. The exponential mechanism is a foundational differentially private algorithm that performs exponentially weighted sampling from a space of outcomes with weights chosen according to a utility function over databases and outputs. The standard exponential mechanism cannot be directly applied to the blDP scenario because we have no guarantees about how a particular utility function will depend on the additional leaked function P. As an example, if P is the standard deviation of entries in the database and we are seeking to output a result that is close to the average of the database, the value of P will affect the distribution of how "useful" particular outputs are based on their distance from the true mean. To address this, we introduce a notion of a coupled utility function, and show that an analogous mechanism enjoys blDP.

6 Applications

Now that we have presented a definition of bounded-leakage privacy and explored some of its properties, we consider some applications of this definition.

6.1 2020 Census: Releasing Additional Information About a Dataset

In some situations where a differentially private study is run, there may be additional releases of information about the underlying private dataset. This could happen unintentionally via a leak or some sort of adversarial attack, or it could be intentional if those running the study choose to release select pieces of information without noise (or with lower noise levels), such as releasing the number of outliers, or a summary statistic such as the standard deviation or average of the data surveyed.

In this situation, we would like to understand what sort of privacy is maintained after such a leak, and whether the release of such information combined with the results of a differentially private study could degrade the participants' privacy in unexpected ways.

If the leaked information is a deterministic function that depends only on the database or a randomized function that uses independent randomness from that used in the differentially private mechanism, then the differentially private mechanism and this additional function will be perfectly independent. Applying Theorem 7 tells us that releasing this additional information guarantees bounded-leakage privacy with the same bounds as the DP mechanism in the original study, and so other than revealing that the database used in the study was such that it produced the additional statistic in question, the privacy of the results of the original study does not degrade further. This gives a formal language for reasoning about, for example, the privacy properties of the 2020 US Census, where some statistics will be revealed without any noise, and other computations will be subject to differential privacy [9].

6.2 Big World Privacy: Controlling Privacy Degradation due to Absence from Studies

Recall the Big World Privacy problem from the Introduction. Bounded-leakage differential privacy can aid in reasoning about how privacy degrades across many studies, some of which an individual may not have participated in.

Suppose that a sequence of k studies has been run on various subsets of the entire population D. In the Introduction, we discussed modeling each study as the composition of two mechanisms, one being a standard differential-privacy mechanism $M^{(j)}$, and the other a participation function, which we will denote $f_{par}^{(j)}$. Using this definition, the result of the jth study is computed by first getting the output of $f_{par}^{(j)}(D, r_{par}^{(j)}) = D^{(j)}$, which will be a subset of D containing only the data of those chosen to participate. After getting $D^{(j)}$, we compute $M^{(j)}(D^{(j)}, r^{(j)})$ to get the final result of study j. We will assume that each $M^{(j)}$ satisfies ϵ-DP.

When the participation function is arbitrary, then no level of privacy can be maintained. For example, for a particular study of average height, we can either choose a group of NBA players or a group of toddlers. If this decision is based on a sensitive property of individual i, this property will be completely revealed. We will therefore make the simplifying assumption that the participation of an individual i is independent of all other individuals. Our results will hold in more general settings as well, but as the above example demonstrates, some assumption of this sort needs to be made.

With this assumption we are guaranteed that if a mechanism is (ϵ, δ)-DP, then each pair of databases with different i values but the same participation in other individuals is a pair of neighboring databases, and therefore satisfies (ϵ, δ)-DP. The study will be a convex combination of the results of the privacy mechanism on such pairs, and so must also satisfy (ϵ, δ)-DP.

Now, suppose that we would like to reason about the privacy loss of a particular individual i if some of her participation data is leaked. Consider a leakage function $P_i(D, r_{par}(i), r_p)$ that takes in the giant database D, the randomness used to decide i's participation in each study, $r_{par}(i)$, and some additional randomness r_p. We define this function such that it outputs some $t \in \mathbb{Z}$, with $0 \leq t \leq k$, such that t is an upper bound on the total number of studies that our individual i participated in. There are numerous real-world situations where such a bound might be released, such as the observation that "i only participated in studies conducted in the U.S." or that some number of the studies were conducted before i was born.

Let $\overline{M}(D, r_{par}, \overline{r})$ denote the concatenation of all k studies. Conditioning on this leakage function, we can show that the following result holds:

▶ **Theorem 17.** *For any $t \leq k$, subset $S \subseteq O_{\overline{M}}$, and database D_i that differs from D only in i's data, we have that*

$$\Pr_{r_{par}, \overline{r}}[\overline{M}(D, r_{par}, \overline{r}) \in S | P_i(D, r_{par}(i), r_p) = t]$$

$$\leq e^{2t\epsilon} \Pr_{r_{par}, \overline{r}}[\overline{M}(D_i, r_{par}, \overline{r}) \in S | P_i(D_i, r_{par}(i), r_p) = t] + 2t\delta.$$

This bound arises from the standard additive bound (Theorem 32) for the composition of $2t$ (ϵ, δ)-DP mechanisms, but any existing composition bound for the same setting could be substituted in to get an analogous result. It should be noted that this result is a bit different from our standard definition of blDP due to the fact that the resulting privacy bound $(2t\epsilon)$ depends on the value of the leakage function (t). We include the proof of this result in Appendix F. As previously noted, standard differential privacy only tells us that i incurs an ϵ privacy loss for every single study in $\overline{M}$. However, considering the same question in the bounded-leakage differential privacy setting allows us to conclude that the privacy loss of i is bounded by the number of studies she may have participated in.

7 Future Directions

We hope that the notion of bounded-leakage differential privacy will aid in the rigorous analysis of privacy guarantees in settings where differential privacy does not hold. This initial exploration suggests many additional avenues to pursue. It would be interesting to develop additional mechanisms that enjoy blDP, and to apply the notion in new domains. Additionally, one might consider variations on the definition of blDP, for example a variant that allows weakened privacy if the probability of the function P attaining a particular set of values is very small.

References

1 John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

2 Aref N Dajani, Amy D Lauger, Phyllis E Singer, Daniel Kifer, Jerome P Reiter, Ashwin Machanavajjhala, Simson L Garfinkel, Scot A Dahl, Matthew Graham, Vishesh Karwa, et al. The modernization of statistical disclosure limitation at the us census bureau. In *Washington, DC: US Census Bureau. Available at:* `https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf`, 2017.

3 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

4 Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

5 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

6 Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

7 Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 265–273, 2008.

8 Simson L Garfinkel. Modernizing disclosure avoidance: Report on the 2020 disclosure avoidance system as implemented for the 2018 end-to-end test, 2018. URL: `https://www.census.gov/about/cac/sac/meetings/2017-09-meeting.html`.

9 Simson L Garfinkel, John M Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 133–137, 2018.

10 Shiva P Kasiviswanathan and Adam Smith. On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), 2014.

11 Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

12 Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

13 Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

A Missing Proofs from Section 3

Proof of Theorem 4. Consider any neighboring databases $x \sim x'$ and a subset $S \subseteq O_M$. Let $O' = \{o \in O_P : \Pr_r[P(x,r) = o] \Pr_r[P(x',r) = o] \neq 0\}$.

By the definition of O' and the fact that P satisfies $(\epsilon_2, 0)$-DP, we have that

$$\Pr_r[M(x,r) \in S, P(x,r) \in O_P \setminus O'] = 0$$

and therefore

$$\begin{aligned}\Pr_r[M(x,r) \in S] &= \sum_{o \in O'} \Pr_r[M(x,r) \in S | P(x,r) = o] \Pr_r[P(x,r) = o] \\ &\leq e^{\epsilon_1} \left(\sum_{o \in O'} \Pr_r[M(x',r) \in S | P(x',r) = o] \Pr_r[P(x,r) = o] \right) + \delta_1 \\ &\leq e^{\epsilon_1 + \epsilon_2} \Pr_r[M(x',r) \in S] + \delta_1.\end{aligned}$$

So M satisfies $(\epsilon_1 + \epsilon_2, \delta_1)$-DP. ◀

▶ Remark. One can extend this theorem to account for a non-zero δ_2 value, at a cost of an additional $(1 + |O'|)\delta_2$ in the δ.

Proof of Theorem 9. Consider any subset $S \subseteq O_M$, neighboring databases $x \sim x'$, and output $o \in O_P$ such that $\Pr_r[P(x,r) = o] \Pr_r[P(x',r) = o] \neq 0$. Combining the definitions of (ϵ', δ')-independence and (ϵ, δ)-DP gives us

$$\begin{aligned}\Pr_r[M(x,r) \in S | P(x,r) = o] &\leq e^{\epsilon'} \Pr_r[M(x,r) \in S] + \delta' \\ &\leq e^{\epsilon + \epsilon'} \Pr_r[M(x',r) \in S] + e^{\epsilon'}\delta + \delta' \\ &\leq e^{\epsilon + 2\epsilon'} \Pr_r[M(x',r) \in S | P(x',r) = o] + (e^{\epsilon' + \epsilon} + 1)\delta' + e^{\epsilon'}\delta\end{aligned}$$

and therefore (M, P) satisfies $(\epsilon + 2\epsilon', (e^{\epsilon' + \epsilon} + 1)\delta' + e^{\epsilon'}\delta)$-blDP. ◀

B Some Missing Proofs from Section 4

Proof of Lemma 10. Suppose we have some mechanism $M : \mathcal{X}^n \times R \to O_M$ that is a convex combination of the mechanisms $M_1, ..., M_k : \mathcal{X}^n \times R \to O_M$ such that for each $1 \leq i \leq k$, we have $\Pr_r[M = M_i] = a_i$ for some $a_1, ..., a_k \geq 0$ such that $\sum_{i=1}^k a_i = 1$. Suppose that each M_i satisfies (ϵ, δ)-blDP with respect to a function $P : \mathcal{X}^n \times R \to O_P$.

Now, consider any neighboring databases $x \sim x'$, $S \subseteq O$, and $o \in O_P$ such that $\Pr_r[P(x,r) = o] \Pr_r[P(x',r) = o] \neq 0$. Then we have that

$$\begin{aligned}\Pr_r[M(x,r) \in S | P(x,r) = o] &= \sum_{i=1}^k \Pr[M = M_i] \Pr_r[M_i(x,r) \in S | P(x,r) = o] \\ &\leq \sum_{i=1}^k a_i (e^{\epsilon} \Pr_r[M_i(x',r) \in S | P(x',r) = o] + \delta) \\ &= e^{\epsilon} \left(\sum_{i=1}^k a_i \Pr_r[M_i(x',r) \in S | P(x',r) = o] \right) + \delta \left(\sum_{i=1}^k a_i \right) \\ &= e^{\epsilon} \Pr_r[M(x',r) \in S | P(x',r) = o] + \delta,\end{aligned}$$

and so M satisfies (ϵ, δ)-blDP with respect to P. ◀

▶ **Example 18** (A blDP mechanism that fails group privacy)**.** The parity function paired with any arbitrary privacy mechanism acts as a simple example of how a mechanism/function pair can fail to satisfy any sort of group privacy guarantee. If we think of the database as being represented as a vector in $\{0,1\}^n$, and choose the leakage function to be the parity of that vector, then any neighboring databases will have different parity. So, any arbitrary mechanism trivially satisfies perfect blDP with respect to the parity function. However, we can make no guarantees about non-neighboring databases that may share the same parity, and so can make no guarantees about group blDP for the same mechanism/function pair.

▶ Remark 19. The above is also an example of how certain leakage functions can cause blDP to be trivially satisfied when the probability that neighboring databases will leak the same output is zero. If the leakage function has even a small amount of noise, guaranteeing that neighboring databases always have non-zero probability to agree on the leakage (such as in an ϵ- DP with large ϵ), then blDP is guaranteed to be non-trivial. We also note that an alternative definition of blDP could restrict the behavior of the mechanism across *any* two databases inducing the same result under the leakage function, while scaling the corresponding constraint on probabilities according to the Hamming distance between the databases. This strictly stronger definition is worthy of further investigation.

▶ **Definition 20.** *Given a database universe $\mathcal{X}^n$, a* path *of length n between databases x and x' is a sequence of $n+1$ databases from $\mathcal{X}^n$, $x = x_0, x_1, ..., x_n = x'$, such that for all $i \in \{0, ..., n-1\}$, the databases x_i and x_{i+1} are adjacent.*

▶ **Definition 21.** *Given a function $P : \mathcal{X}^n \times R \to O_P$, a path $P = (x_0, ..., x_n)$ is* non-zero *for an output $o \in O_P$ if for all $i \in \{0, ..., n\}$, we have $\Pr_r[P(x_i, r) = o] > 0$.*

▶ **Theorem 22** (Group privacy)**.** *Consider a mechanism M that satisfies (ϵ, δ)-blDP with respect to a function P, and two databases x and x'. If for a particular output $o \in O_P$ there exists a non-zero path of length k between x and x', then for any $S \subseteq O_M$, we have*

$$\Pr_r[M(x,r) \in S | P(x,r) = o] \le e^{k\epsilon} \Pr_r[M(x',r) \in S | P(x',r) = o] + \delta \left(\frac{e^{k\epsilon} - 1}{e^{\epsilon} - 1} \right).$$

The proof of this theorem follows the same approach as the proof of group privacy in the standard DP setting, and is omitted here.

Proof of Theorem 12. Consider two neighboring databases $x \sin x'$, a subset $S \subseteq O_M \times O_P$, and an output $o \in O_P$ such that $\Pr_r[P(x,r) = o] \Pr_r[P(x',r) = o] \ne 0$. Then,

$$\Pr_r[M'(x,r) \in S | P(x,r) = o] = \Pr_r[M(x,r) \in S_o | P(x,r) = o]$$

where $S_o = \{y \in O_M : (y, o) \in S\}$, with the same holding true when x is replaced with x'.

Therefore, combining the blDP properties of (M, P) and the above equalities gives

$$\Pr_r[M'(x,r) \in S | P(x,r) = o] \le e^{\epsilon} \Pr_r[M'(x',r) \in S | P(x',r) = o] + \delta$$

and thus M' satisfies (ϵ, δ)-blDP with respect to P. ◀

▶ **Proposition 23.** *Consider any mechanism M, leakage function P, output o of P, and databases x_0 and x_1 such that $\Pr_r[P(x_0, r) = o] \cdot \Pr[P(x_1, r) = o] \ne 0$. Then for any subset $S \subseteq O_M$ and $b \in \{0,1\}$, we have that*

$$\Pr_{r \in (R_{o,x_0} \times R_{o,x_1})}[M_{P,o}^{x_0;x_1}(x_b, r) \in S] = \Pr_{r \in R}[M(x_b, r) \in S | P(x_b, r) = o].$$

▶ **Proposition 24.** *Suppose that a mechanism M satisfies (ϵ, δ)-blDP with respect to a leakage function P. Then, for any $o \in O_P$ and neighboring databases $x_0 \sim x_1$, the DP reduction mechanism $M_{P,o}^{x_0,x_1}$ satisfies (ϵ, δ)-DP.*

The proofs of Propositions 23 and 24 are omitted here for space reasons, but are easily verified.

▶ Remark 25. Another simpler construction can be used to prove a reduction in the reverse direction, reducing DP composition to the blDP setting.

C Details on Non-Adaptive Composition

The following theorem shows how to translate a non-adaptive composition theorem for differential privacy into one for blDP.

▶ **Theorem 26.** *For some $k \geq 1$, suppose that the following implication were to hold: if for any choice of k mechanisms $L_1, ..., L_k$ such that each L_i satisfies (ϵ_i, δ_i)-DP, then the composition of these mechanisms,*

$$L(x, (r_1, ..., r_k)) := L_1(x, r_1)||L_2(x, r_2)||...||L_k(x, r_k),$$

(where each r_i is chosen independently at random) would satisfy (ϵ', δ')-DP. Then, for any choice of k mechanism function pairs $(M_1, P_1), ..., (M_k, P_k)$ such that each M_i satisfies (ϵ_i, δ_i)-blDP with respect to P_i, if we define the composed functions

$$M(x, (r_1, ..., r_k)) := M_1(x, r_1)||...||M(x, r_k) \text{ and } P(x, (r_1, ..., r_k)) := P_1(x, r_1)||...||P_k(x, r_k),$$

then M must also satisfy (ϵ', δ')-blDP with respect to P.

Proof. Consider any neighboring databases $x_0 \sim x_1$, some subset $S \subseteq O_M$, and some $o = (o_1, ..., o_k) \in O_P$ such that $\Pr_r[P(x_0, r) = o] \Pr_r[P(x_1, r) = o] \neq 0$. We note that this requirement implies that $\Pr_r[P_i(x_0, r) = o_i] \Pr_r[P_i(x_1, r) = o_i] \neq 0$ for all i. Then, consider the k DP reduction functions, $(M_1)_{P_1,o_1}^{x_0,x_1}, ..., (M_k)_{P_k,o_k}^{x_0,x_1}$, defined in terms of each (M_i, P_i). By Proposition 24, each $(M_i)_{P_i,o_i}^{x_0,x_1}$ must satisfy (ϵ_i, δ_i)-DP.

Therefore, by our assumption, the composition

$$M'(x, (r_1, ..., r_k)) = (M_1)_{P_1,o_1}^{x_0,x_1}(x, r_1)||...||(M_k)_{P_k,o_k}^{x_0,x_1}(x, r_k)$$

must satisfy (ϵ', δ')-DP.

We can express any subset S as the sum of disjoint rectangles, so it suffices to assume that S is a rectangle. So, $S = S_1 \times S_2 \times ... \times S_k$ where each $S_i \subseteq O_{M_i}$. Then because each r_i is chosen independently, we know that for any $b \in \{0, 1\}$,

$$\Pr[M'(x_b, (r_1, ..., r_k)) \in S] = \prod_{i=1}^{k} \Pr_r[M_i(x_b, r) \in S_i | P_i(x_b, r) = o_i],$$

where because each P_i uses independent randomness as well,

$$\prod_{i=1}^{k} \Pr_r[M_i(x_b, r) \in S_i | P_i(x_b, r) = o_i] = \Pr_r[M(x_b, r) \in S | P(x_b, r) = o].$$

Combining the DP guarantee for M' and the above equality gives us

$$\Pr_r[M(x_0, r) \in S | P(x_0, r) = o] \leq e^{\epsilon'} \Pr_r[M(x_1, r) \in S | P(x_1, r) = o] + \delta'$$

Therefore the composed mechanism M must satisfy (ϵ', δ')-blDP with respect to P. ◀

This result tells us that any nonadaptive composition bounds for the DP setting can be extended to the blDP setting. In particular, we present a corollary below that is reached by applying this statement to a well-known composition theorem for DP.

We first recall the following theorem:

▶ **Theorem 27** ([5]). *Suppose $M_1, ..., M_k$ are mechanisms such that M_i satisfies (ϵ_i, δ_i)-DP. Then the composition of these mechanisms, $M(x,(r_1,...,r_k)) := M_1(x,r_1)||...||M_k(x,r_k)$, satisfies $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$-DP.*

The following corollary is a direct result of combining this theorem with Theorem 26:

▶ **Corollary 28.** *Suppose $(M_1, P_1), ..., (M_k, P_k)$ are mechanism-function pairs such that each M_i satisfies (ϵ_i, δ_i)-blDP with respect to P_i. Then the composition of these mechanisms,*

$$M(x,(r_1,...,r_k)) := M_1(x,r_1)||...||M_k(x,r_k)$$

satisfies $(\sum_{i=1}^k \epsilon_i, \sum_{i=1}^k \delta_i)$-blDP with respect to the composition of the P_is,

$$P(x,(r_1,...,r_k)) := P_1(x,r_1)||...||P_k(x,r_k).$$

Therefore, we can conclude that, like for differential privacy, the rate of bounded-leakage privacy loss as we increase the number of queries to the database, is at most linear.

D Adaptive Composition

It is also important to consider how bounded-leakage-privacy can be affected if the composed mechanisms are chosen adaptively based on the outputs of the previously chosen mechanisms. We analyze how this form of composition affects privacy via an experiment/adversary model. We define two experiments: Experiment 0 and Experiment 1, as follows:

Algorithm 1 Experiment b: blDP of Adaptive k-Fold Composition.

```
Input: a family of mechanism-function pairs $\mathcal{F} = \{(M_1, P_1), (M_2, P_2), ...\}$, and a
       probabilistic adversary $\mathcal{A}$.
Repeat $k$ times:
    $\mathcal{A}$ outputs some query $((x_0, x_1), (M_i, P_i), o_i)$ where $(x_0, x_1)$ is a pair of adjacent
      databases, $(M_i, P_i) \in \mathcal{F}$, and $o_i$ is some member of the output space of $P_i$.
    if $\Pr[P_i(x_0, r) = o_i] \cdot \Pr[P_i(x_1, r) = o_i] = 0$ then $\mathcal{A}$ receives "null".
    else $\mathcal{A}$ receives $M_i(x_b, r)$ for some random $r \in R$ such that $P_i(x_b, r) = o_i$.
```

Intuitively, the definition of bounded-leakage privacy states that if a mechanism-function pair has good bounded-leakage privacy, it should be hard to differentiate the outputs of the mechanism on two adjacent databases for some fixed function output. Extending this to the composition setting, it should still be difficult to distinguish which database was used even if we are able to get more information with multiple queries. Connecting this to the experiments, the adversary should have difficulty distinguishing between the outputs of Experiments 0 and 1 if our family of mechanisms and functions has good bounded-leakage privacy.

To formalize this, we define the "view" of the adversary in a particular experiment to be everything that the adversary sees or knows after the experiment, i.e. the contents of all the adversary's k queries and the responses to those queries. It should be noted that this will not include the random coin flips used to generate the responses to each query, nor whether $b = 0$

or 1. An adversary's view can be denoted by the values of all the queries and their responses, i.e., a "transcript" of the experiment, or just the values of the random coin flips that the adversary used and all of the responses. These are equivalent representations because an adversary's queries can always be reconstructed from the randomness that the adversary used and the responses that it received, and so we will use these two representations of the view interchangeably throughout.

Now that we have formalized this notion of an adversary's view, we can use our model to define bounded-leakage privacy under adaptive composition as follows:

▶ **Definition 29.** *We say that a family $\mathcal{F}$ of mechanism-function pairs satisfies* (ϵ, δ)-blDP under k-fold adaptive composition *if for every adversary $\mathcal{A}$, random variables V^b corresponding to the view of $\mathcal{A}$ in Experiment b, and subset of possible views V, we have that*

$$\Pr[V_0 \in V] \leq e^{\epsilon} \Pr[V_1 \in V] + \delta$$

This definition is designed to parallel the definition for adaptive composition of DP mechanisms given by Dwork and Roth [5]. We include the DP version here so that the two can be easily compared:

Algorithm 2 Experiment b: DP of Adaptive k-Fold Composition [5].

```
Input: A family $\mathcal{F}$ of mechanisms and a probabilistic adversary $\mathcal{A}$.
Repeat $k$ times:
    $\mathcal{A}$ outputs the query $((x_0, x_1), M_i)$ where $(x_0, x_1)$ is a pair of adjacent databases
      and $M_i \in \mathcal{F}$.
    $\mathcal{A}$ recieves $M_i(x_b, r)$ for some random $r \in R$.
```

▶ **Definition 30** ([5]). *We say that the family of mechanisms $\mathcal{F}$ satisfies (ϵ, δ)-DP under k-fold adaptive composition if for every adversary $\mathcal{A}$, we have $D_{\infty}^{\delta}(V_0 || V_1) \leq \epsilon$, where V_b denotes the view of A in the DP composition experiment, and $D_{\infty}^{\delta}(V_0||V_1)$ is the δ-approximate max divergence between V_0 and V_1, defined as*

$$D_{\infty}^{\delta}(V_0||V_1) = \max_{S \subseteq Supp(V_0):\Pr[V_0 \in V] \geq \delta} \left[\ln \frac{\Pr[V_0 \in S] - \delta}{\Pr[V_1 \in S]} \right].$$

We note that this max-divergence definition is equivalent to requiring that for all adversaries and subsets of views of the DP experiment, V, we have $\Pr[V_0 \in V] \leq e^{\epsilon} \Pr[V_1 \in V] + \delta$, which puts the definition in a more familiar form.

We will now connect this model to the standard DP setting. The following theorem states that any bounds for adaptive composition that can be shown to hold in the DP setting must also hold in the blDP setting:

▶ **Theorem 31.** *If a class of (ϵ, δ)-DP mechanism-function pairs satisfies (ϵ', δ')-DP under k-fold adaptive composition, then that class of (ϵ, δ)-blDP mechanisms satisfies (ϵ', δ')-blDP under k-fold adaptive composition.*

Similar to our nonadaptive composition result, the proof of this theorem uses the strategy of reducing the blDP setting to the DP setting by converting any arbitrary blDP adversary to an adversary in the DP setting with the same distribution of views. We can then argue that therefore the original blDP adversary must be constrained by the same bounds as the DP adversary.

Proof of Theorem 31. For the purposes of this proof, we will consider the views of the adversary in both the DP and blDP adaptive composition experiments to contain only the value of the adversary's random bits and the responses it receives for each query so that we can easily compare the views in the DP and blDP settings.

First, assume that the class of (ϵ, δ)-DP mechanisms satisfies (ϵ', δ')-DP under k-fold adaptive composition. Now, consider some adversary $\mathcal{A}^{blDP}$ for the blDP composition experiments for the class of (ϵ, δ)-blDP mechanism-function pairs.

Using $\mathcal{A}^{blDP}$, we construct an adversary for the DP composition experiment on (ϵ, δ)-DP mechanisms as follows: whenever $\mathcal{A}^{blDP}$ would output the query $((x_0, x_1), (M, P), o)$ given the current view of the experiment, $\mathcal{A}^{DP}$ outputs $((x_0, x_1), M_{P,o}^{x_0;x_1})$, where $M_{P,o}^{x_0;x_1}$ is the DP reduction mechanism for $(x_0, x_1), o$, and P.

By Proposition 24, $M_{P,o}^{x_0;x_1})$ satisfies (ϵ, δ)-DP and therefore $\mathcal{A}^{DP}$ is a valid adversary for the class of (ϵ, δ)-DP mechanisms.

Now, we want to show that given this definition of $\mathcal{A}^{DP}$, if V_b^{DP} is a random variable for the view of $\mathcal{A}^{DP}$ in the DP composition Experiment b and V_b^{blDP} is a random variable for the view of $\mathcal{A}^{blDP}$ in the blDP composition Experiment b, then we have $dist(V_b^{DP}) = dist(V_b^{blDP})$.

We split both views into their component random variables corresponding to the randomness of the adversaries and the responses in the experiment such that

$$V_b^{DP} = (R^{DP}, S_1^{DP}, ..., S_k^{DP}) \quad \text{and} \quad V_b^{blDP} = (R^{blDP}, S_1^{blDP}, ..., S_k^{blDP}),$$

where R^{DP} and R^{blDP} are random variables corresponding to the random bits of the adversaries, and each S_i is the response received for the ith query in the experiment.

We will use induction on the number of outputs to show that these distributions must be equal. First, because $\mathcal{A}^{DP}$ uses no additional randomness apart from the randomness of $\mathcal{A}^{blDP}$, we clearly have $dist(R^{DP}) = dist(R^{blDP})$. This forms our base case.

Now, suppose that for some i with $1 \leq i \leq k$, we have

$$dist((R^{DP}, S_1^{DP}, ..., S_{i-1}^{DP})) = dist((R^{blDP}, S_1^{blDP}, ..., S_{i-1}^{blDP})).$$

Then, for any partial view $(r, s_1, ..., s_i)$, we can rewrite $\Pr[(R^{DP}, ..., S_i^{DP}) = (r, ..., s_i)]$ as

$$\Pr[(R^{DP}, ..., S_{i-1}^{DP}) = (r, ..., s_{i-1})] \Pr[S_i^{DP} = s_i | (R^{DP}, ..., S_{i-1}^{DP}) = (r, ..., s_{i-1})],$$

where by fixing $(r, s_1, ..., s_{i-1})$, the ith query from $\mathcal{A}^{DP}$ is deterministically fixed to be some $((x_0, x_1), (M_i, P_i), o_i)$, and therefore the ith query of $\mathcal{A}^{DP}$ is fixed to be $((x_0, x_1), (M_i)_{P_i,o_i}^{x_0;x_1})$. By Proposition 23, if $\Pr_r[P_i(x_0, r) = o_i] \Pr_r[P_i(x_1, r) = o_i] = 0$, then $(M_i)_{P_i,o_i}^{x_0;x_1}(x_b, r)$ will output "null" with probability one. Otherwise, we have that

$$dist_r((M_i)_{P_i,o_i}^{x_0;x_1}(x_b, r)) = dist_{r:P(x_b,r)=o_i}(M_i(x_b, r))$$

and therefore in either case,

$$\begin{aligned} dist(S_i^{DP} | (R^{DP}, ..., S_{i-1}^{DP}) = (r, ..., s_{i-1})) &= dist_r((M_i)_{P_i,o_i}^{x_0;x_1}(x_b, r)) \\ &= dist(S_i^{blDP} | (R^{blDP}, ..., S_{i-1}^{blDP}) = (r, ..., s_{i-1})) \end{aligned}$$

By our inductive assumption, $dist((R^{blDP}, ..., S_{i-1}^{blDP})) = dist((R^{DP}, ..., S_{i-1}^{DP}))$. Putting these together, we must have $dist((R^{blDP}, ..., S_i^{blDP})) = dist((R^{DP}, ..., S_i^{DP}))$. This completes our inductive step, and therefore it follows by induction that

$$dist(V_b^{blDP}) = dist((R^{blDP}, ..., S_k^{blDP})) = dist((R^{DP}, ..., S_k^{DP})) = dist(V_b^{DP})$$

$$dist(V_b^{blDP}) = dist(V_b^{DP}).$$

Because by our initial assumption the class of (ϵ, δ)-DP mechanisms satisfies (ϵ', δ')-DP under k-fold adaptive composition, we must also have that for any subset of views V, we have

$$\Pr[V_0^{blDP} \in V] = \Pr[V_0^{DP} \in V] \leq e^{\epsilon'} \Pr[V_1^{DP} \in V] + \delta' = e^{\epsilon'} \Pr[V_1^{blDP} \in V] + \delta'$$

$$\Pr[V_0^{blDP} \in V] \leq e^{\epsilon'} \Pr[V_1^{blDP} \in V] + \delta'$$

Therefore the class of (ϵ, δ)-blDP mechanisms must also satisfy (ϵ', δ')-blDP under k-fold adaptive composition. ◀

Theorem 31 allows us to apply existing bounds for the adaptive composition of DP mechanisms to the blDP context. Recall the following composition bound for DP mechanisms:

▶ **Theorem 32** ([6]). *For all $\epsilon, \delta, \delta' \geq 0$, the class of (ϵ, δ)-DP mechanisms satisfies $(\epsilon', k\delta+\delta')$-DP under k-fold adaptive composition for:*

$$\epsilon' = \epsilon\sqrt{2k \ln(1/\delta')} + k\epsilon(e^{\epsilon-1}).$$

Combining the results of Theorem 31 and Theorem 32 gives us Theorem 16 as a corollary.

E The Exponential Mechanism

▶ **Definition 33.** *Given a set of outputs O_M and a set of outputs O_P, a* coupled utility function for O_M and O_P *is some function $u_{M,P} : \mathcal{X}^n \times O_M \times O_P \to \mathbb{R}$ that maps triples containing a database, an element of O_M, and an element of O_P to a real-valued score.*

We define this coupled utility function with the intent of defining O_M to be the output space of a particular mechanism, and O_P to be the output space of some associated function. This definition would allow us to define utility functions in the bounded-leakage setting that are inherently stronger than just considering a standard utility function conditioned on a particular output of the leaked function, because here we can have the utility function depend on the output of the associated function even if it is randomized.

We also want to define an analogous notion of utility sensitivity for coupled utilities.

▶ **Definition 34.** *Given a coupled utility function $u_{M,P} : \mathcal{X}^n \times O_M \times O_P \times \mathbb{R}$, we define a function corresponding to the sensitivity of $u_{M,P}$ conditioned on a particular element of O_P, $\Delta u_{M,P} : O_P \to \mathbb{R}$, such that for any $o \in O_P$,*

$$\Delta u_{M,P}(o) = \max_{y \in O_M} \max_{x \sim x'} |u_{M,P}(x, y, o) - u_{M,P}(x', y, o)|$$

This quantifies the sensitivity for our coupled utility given a particular value for the output of P.

Using this new concept of a coupled utility function, we can define a version of the exponential mechanism for blDP as follows:

▶ **Definition 35** (The Exponential Mechanism for blDP). *Given a set of outputs, O_M, a function $P : \mathcal{X}^n \times \mathcal{R} \to O_P$, and a coupled utility function $u_{M,P} : \mathcal{X}^n \times O_M \times O_P \to \mathbb{R}$, the* bounded-leakage exponential mechanism *$M_{E,P}(x, u_{M,P}, O_M, r)$ is defined such that if $P(x, r) = o$, then for any $y \in O_M$, the probability that the mechanism outputs y is proportional to*

$$\exp\left(\frac{\epsilon u_{M,P}(x, y, o)}{2\Delta u_{M,P}(o)}\right).$$

We can now show that in the same way that the standard exponential mechanism guarantees $(\epsilon, 0)$-DP, this version of the mechanism guarantees $(\epsilon, 0)$-blDP.

▶ **Theorem 36.** *The exponential mechanism for blDP satisfies $(\epsilon, 0)$-blDP with respect to its associated function P.*

Proof. Consider any two neighboring databases $x \sim x'$, some output o of P such that $\Pr_r[P(x,r) = o] \cdot \Pr_r[P(x',r) = o] \neq 0$, and some output $s \in O_M$. Applying the definition of the mechanism and utility sensitivity, we have that

$$\frac{\Pr_r[M_{E,P}(x, u_{M,P}, O_M, r) = s | P(x,r) = o]}{\Pr_r[M_{E,P}(x', u_{M,P}, O_M, r) = s | P(x',r) = o]}$$

is at most

$$\exp\left(\frac{\epsilon \Delta u_{M,P}(o}{2\Delta u_{M,P}(o)}\right) \frac{\sum_{y \in O_M} \exp\left(\frac{\epsilon(u_{M,P}(x,y,o) + \Delta u_{M,P}(o))}{2\Delta u_{M,P}(o)}\right)}{\sum_{y \in O_M} \exp\left(\frac{\epsilon u_{M,P}(x,y,o)}{2\Delta u_{M,P}(o)}\right)} = \exp(\epsilon)$$

Therefore, for any subset $S \subseteq O_M$, we have

$$\begin{aligned}\Pr_r[M_{E,P}(x, u_{M,P}, O_M, r) \in S | P(x,r) = o]& \\ &\leq \sum_{s \in S} e^{\epsilon} \Pr_r[M_{E,P}(x', u_{M,P}, O_M, r) = s | P(x',r) = o] \\ &= e^{\epsilon} \Pr_r[M_{E,P}(x', u_{M,P}, O_M, r) \in S | P(x',r) = o],\end{aligned}$$

and so $M_{E,P}$ satisfies $(\epsilon, 0)$-blDP with respect to P. ◀

As in the case of the standard exponential mechanism, we also want to show that our mechanism for the bounded-leakage case can give us some guarantee of "good" utility. In the case of the standard exponential mechanism, we recall the following theorem:

▶ **Theorem 37** ([13])**.** *Let M_E be the standard exponential mechanism for a set of outputs S and utility function u. For any database x, let $OPT_u(x) = \max_{y \in S} u(x,y)$, and $S_{OPT} = \{y \in O_M : u(x,y) = OPT_u(x)$. Then, for any $t \in \mathbb{R}$, we have that*

$$\Pr_r[u(M_E(x,u,S,r)) \leq OPT_u(x) - \frac{2\Delta u}{\epsilon}\left(\ln\left(\frac{|S|}{|S_{OPT}|}\right) + t\right)] \leq e^{-t}$$

By the properties of our exponential mechanism for bounded-leakage privacy, we can conclude the following analogous theorem in the blDP setting:

▶ **Theorem 38.** *Let $M_{E,P}$ be the standard bounded-leakage exponential mechanism for a set of outputs S, function P, and utility function $u_{S,P}$. For any database x and output o of P, let $OPT_{u_{S,P}}(x,o) = \max_{y \in S} u_{S,P}(x,y,o)$, and $S_{OPT} = \{y \in O_M : u_{S,P}(x,y,o) = OPT_u(x)$. Then, for any $t \in \mathbb{R}$, database x, and output o of P, we have that*

$$\Pr_{r:P(x,r)=o}[u_{S,P}(M_{E,P}(x, u_{S,P}, S, r)) \leq OPT_{u_{S,P}}(x,o) - \frac{2\Delta u_{S,P}(o)}{\epsilon}\left(\ln\left(\frac{|S|}{|S_{OPT}|}\right) + t\right)] \leq e^{-t}.$$

Proof. This result follows immediately from combining Theorem 37 and the observation that once the output of P is set, $M_{E,P}$ behaves like the standard exponential mechanism for output space S and utility $u(x,y) = u_{S,P}(x,y,o)$. ◀

F Additional Details on the BigWorld Application of blDP

In this section, we provide a proof for the result stated in Theorem 17. To begin, we have the following lemma:

▶ **Lemma 39.** *Let V_i be a leakage function that releases the exact number of studies that i participated in. Then, for any $S \subseteq O_M$, $t \leq k$, and $D \sim D_i$ such that D_i differs from D only in i's data, we have that*

$$\Pr_{\overline{r}, r_{par}}[\overline{M}(D, r_{par}, \overline{r}) \in S | V_i(D, r_{par}) = t] \leq e^{2t\epsilon} \Pr_{\overline{r}, r_{par}}[\overline{M}(D_i, r_{par}, \overline{r}) \in S | V_i(D_i, r_{par}) = t] + 2t\delta$$

Proof. We first note that because of our assumption that the participation of i is independent of the participation of all other individuals in D, we can consider the distribution of a particular study $M^{(j)}$'s output to be a convex combination of neighboring databases differing only in i's data.

If i does not participate in study j in either case, then neither mechanism can depend on i, so the distributions of possible results conditioned on v and v' will be equal. Otherwise, we apply the (ϵ, δ)-DP to conclude that

$$\Pr_{r, r_{par}}[M^{(j)}(f_{par}^{(j)}(D, r_{par}), r) \in S_j | V_i(D, r_{par}) = t]$$

$$\leq e^{\epsilon} \Pr_{r, r_{par}}[M^{(j)}(f_{par}^{(j)}(D_i, r_{par}), r) \in S_j | V_i(D_i, r_{par}) = t] + \delta$$

By the definition of our leakage function, the maximum number of j such that i participates in at least one of the two versions of each study is $2t$. So, $\overline{M}$ can be viewed as the composition of at most $2t$ (ϵ, δ)-blDP mechanisms, all with respect to V_i. Meanwhile all other mechanisms are perfectly blDP. Therefore, using the standard composition bound for blDP mechanisms (Corollary 28) we get the desired inequality. ◀

Proof of Theorem 17. With this result in hand, we can now consider our original leakage function P_i, which leaks an upper bound for the magnitude of the participation vector.

For any particular t, we will have that $\Pr_{\overline{r}, r_{par}}[\overline{M}(D, r_{par}, \overline{r}) \in S | P_i(D, r_{par}) = t]$ is a convex combination of the set of probabilities

$$\left\{ \Pr_{\overline{r}, r_{par}}[\overline{M}(D, r_{par}, \overline{r}) \in S | V_i(D, r_{par}) = j] \right\}_{0 \leq j \leq t}.$$

By Lemma 39, each of these satisfies $(2t\epsilon, 2t\delta)$-blDP, and so applying Lemma 10 gives the desired inequality. ◀

We should note that we applied the simplest composition bound for (ϵ, δ)-DP or blDP mechanisms in this case, but any bound could be substituted for an analogous result.