

5th Symposium on Foundations of Responsible Computing

FORC 2024, June 12–14, 2024, Harvard University,
Cambridge, MA, USA

Edited by

Guy N. Rothblum



Editors

Guy N. Rothblum 

Apple, Cupertino, CA, USA
guy.rothblum@gmail.com

ACM Classification 2012

Theory of computation; Mathematics of computing; Security and privacy; Computing methodologies;
Applied computing; Social and professional topics

ISBN 978-3-95977-319-5

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/978-3-95977-319-5>.

Publication date

June, 2024

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>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0):
<https://creativecommons.org/licenses/by/4.0/legalcode>.



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work
under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/LIPIcs.FORC.2024.0

ISBN 978-3-95977-319-5

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.

Editorial Board

- Luca Aceto (Reykjavik University, IS and Gran Sasso Science Institute, IT)
- Christel Baier (TU Dresden, DE)
- Roberto Di Cosmo (Inria and Université de Paris, FR)
- Faith Ellen (University of Toronto, CA)
- Javier Esparza (TU München, DE)
- Daniel Král' (Masaryk University, Brno, CZ)
- Meena Mahajan (*Chair*, Institute of Mathematical Sciences, Chennai, IN)
- Anca Muscholl (University of Bordeaux, FR)
- Chih-Hao Luke Ong (University of Oxford, GB)
- Phillip Rogaway (University of California, Davis, US)
- Eva Rotenberg (Technical University of Denmark, Lyngby, DK)
- Raimund Seidel (Universität des Saarlandes, Saarbrücken, DE and Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, DE)
- Pierre Senellart (ENS, Université PSL, Paris, FR)

ISSN 1868-8969

<https://www.dagstuhl.de/lipics>

■ Contents

Preface	
<i>Guy N. Rothblum</i>	0:vii
Organizers	
.....	0:ix

Regular Papers

Effects of Privacy-Inducing Noise on Welfare and Influence of Referendum Systems	
<i>Suat Evren and Praneeth Vepakomma</i>	1:1–1:20
Incentivized Collaboration in Active Learning	
<i>Lee Cohen and Han Shao</i>	2:1–2:20
Can Copyright Be Reduced to Privacy?	
<i>Niva Elkin-Koren, Uri Hacoheh, Roi Livni, and Shay Moran</i>	3:1–3:18
Balanced Filtering via Disclosure-Controlled Proxies	
<i>Siqi Deng, Emily Diana, Michael Kearns, and Aaron Roth</i>	4:1–4:23
Distribution-Specific Auditing for Subgroup Fairness	
<i>Daniel Hsu, Jizhou Huang, and Brendan Juba</i>	5:1–5:20
Modeling Diversity Dynamics in Time-Evolving Collaboration Networks	
<i>Christopher Archer and Gireeja Ranade</i>	6:1–6:21
Drawing Competitive Districts in Redistricting	
<i>Gabriel Chuang, Oussama Hanguir, and Clifford Stein</i>	7:1–7:22
Score Design for Multi-Criteria Incentivization	
<i>Anmol Kabra, Mina Karzand, Tosca Lechner, Nati Srebro, and Serena Wang</i>	8:1–8:22
Privacy Can Arise Endogenously in an Economic System with Learning Agents	
<i>Nivasini Ananthkrishnan, Tiffany Ding, Mariel Werner, Sai Praneeth Karimireddy,</i> <i>and Michael I. Jordan</i>	9:1–9:22

Extended Abstract

Online Algorithms with Limited Data Retention	
<i>Nicole Immorlica, Brendan Lucier, Markus Mobius, and James Siderius</i>	10:1–10:8



■ Preface

The Symposium on Foundations of Responsible Computing (FORC), now in its fifth year, is a forum for mathematically rigorous research in computation and society writ large. The Symposium aims to catalyze the formation of a community supportive of the application of theoretical computer science, statistics, economics, and other relevant analytical fields to problems of pressing and anticipated societal concern.

Thirty-nine papers were selected to appear at FORC 2024, held at Harvard University in Cambridge, MA on June 12–14, 2024. These papers were selected by the program committee, with the help of additional expert reviewers, out of fifty-two submissions. FORC 2024 offered two submission tracks: archival-option (giving authors of selected papers the option to appear in this proceedings volume) and non-archival (providing a showcase for FORC-relevant work that will appear or has recently appeared in another venue). Ten archival-option and twenty-nine non-archival submissions were selected for the program.

The program committee awarded the FORC 2024 Best Paper Award to the paper “Balanced Filtering via Disclosure-Controlled Proxies” by Siqi Deng, Emily Diana, Michael Kearns and Aaron Roth. The FORC 2024 Best Student Paper Award was given to two papers: “Drawing Competitive Districts in Redistricting” by Gabriel Chuang, Oussama Hanguir and Clifford Stein, and “Distribution-Specific Auditing For Subgroup Fairness” by Daniel Hsu, Jizhou Huang and Brendan Juba.

Thank you to entire program committee and to the external reviewers for their hard work during the review process. It has been an honor and a pleasure to work together with you to shape the program of this young conference. Finally, I would like to thank our generous sponsors: the Simons Collaboration on the Theory of Algorithmic Fairness for their conference support.

Guy Rothblum
Tel Aviv, Israel
April 23, 2024



■ Organizers

Program Committee

Rediet Abebe
Moshe Babaioff
Ran Canetti
Aloni Cohen
Lee Cohen
Yuval Dagan
Ronen Gradwohl
Jason Hartline
Christopher Jung
Dan Linna
Pasin Manurangsi
Shay Moran
Moni Naor
Sofya Raskhodnikova
Guy N. Rothblum (chair)
Uri Stemmer
Kunal Talwar
Eliad Tsfadia
Jonathan Ullman
Gal Yona
Steven Wu
Tijana Zrnic

Steering Committee

Avrim Blum
Cynthia Dwork (co-chair)
Sampath Kannan
Jon Kleinberg
Shafi Goldwasser
Kobbi Nissim
Toni Pitassi
Omer Reingold (co-chair)
Guy N. Rothblum
Salvatore Ruggieri
Salil Vadhan
Adrian Weller



Effects of Privacy-Inducing Noise on Welfare and Influence of Referendum Systems

Suat Evren¹  

Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

Praneeth Vepakomma 

MIT Institute for Data, Systems and Society (IDSS), Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

Abstract

Social choice functions help aggregate individual preferences while differentially private mechanisms provide formal privacy guarantees to release answers of queries operating on sensitive data. However, preserving differential privacy requires introducing noise to the system, and therefore may lead to undesired byproducts. Does an increase in the level of privacy for releasing the outputs of social choice functions increase or decrease the level of *influence* and *welfare*, and at what rate? In this paper, we mainly address this question in more precise terms in a referendum setting with two candidates when the celebrated randomized response mechanism is used. We show that the level of privacy is inversely proportional to society's welfare and influence.

2012 ACM Subject Classification Security and privacy → Economics of security and privacy; Applied computing → Economics; Theory of computation → Dynamic programming

Keywords and phrases Welfare, influence, social choice functions, differential privacy, randomized response

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.1

Related Version *Full Version*: <https://arxiv.org/abs/2201.10115>

Funding Research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Suat Evren: Partially supported by MIT UROP Direct Funding and MIT Schwarzman SERC Scholarship.

Praneeth Vepakomma: Researcher was partially supported by MBZUAI Startup Grant, ADIA Lab Fellowship, MIT Schwarzman SERC Scholarship and Meta PhD Research Fellowship.

Acknowledgements We would like to thank Alex Pentland, Ramesh Raskar, and Ashwin Sah for helpful comments, support and feedback.

1 Introduction

Differential privacy [6] provides a compelling privacy guarantee to ensure that the outcome of a query over any dataset is substantially not influenced based on the presence or absence of an individual's record. This form of privacy has recently been studied in the context of social choice theory [26, 15, 12]. A predominant strategy to achieve differential privacy in general

¹ Corresponding author.



even outside the context of social choice theory is to introduce noise or randomization to the system. One of the issues that has been widely studied in this context is the loss of accuracy upon introducing noise and the trade-off between accuracy and increasing levels of privacy preservation. This has been commonly referred to as the *privacy-accuracy* or *privacy-utility trade-off*. Recent work has involved the formalization of other trade-offs such as the trade-off between privacy and fairness [4]. In this work, we analyze two other trade-offs. We show that introducing noise to privatize systems that aggregate the preferences of individuals may affect several other fundamental phenomena such as *influence* and *welfare*.

Does an increase in the level of privacy for releasing the outputs of social choice functions, increase or decrease the level of influence and welfare, and at what rate? In this paper, we mainly address this question in more precise terms and affirmatively answer that this relation is inversely-proportional and shares specific corresponding rates for the popular ρ -correlated randomized response mechanism of privatization when used in a referendum setting with two candidates.

The noisy mechanism that we propose and analyze in regards to influence and welfare in this paper is based on a simple coin-flipping perturbation of the input as follows. Let ρ be an exogenous constant in $[0, 1]$ and let each original vote made in the ballot take a value of either 1 or -1 . The randomized response records each original vote in the ballot as it is with a probability ρ while with probability $1 - \rho$, it ignores the original vote and instead records it as either a 1 or -1 with a uniformly random pick. The resulting probability space is known as ρ -correlated distribution or *noisy distribution* in the field of analysis of Boolean functions, and it is referred to as the *randomized response* mechanism in the field of differential privacy.² We show that this mechanism preserves ordinal relations between the influences of voters for *any* social choice function. Therefore, if Alice had more influence before than Bob, she will still continue to have more influence.

In the field of analysis of Boolean functions, the notion of the *influence* of a voter is used to measure the power of an individual on the final result of a social choice function. We extend this definition of influence to our probabilistic setting where noise is introduced for privacy, and term this new notion of influence as *probabilistic influence*. Similarly, we define *welfare* to address the second issue of capturing how *ideal* a voting rule is. First, we define it for deterministic functions and then we extend this definition to any probabilistic mechanism. We then show the effect of our privacy inducing randomized response on the welfare of the system. In particular, we show that it preserves the ordinal relations between the welfare of voting systems. That is, if a social choice function f had a greater welfare than g in the deterministic setting after the randomized response M_ρ is applied based on the exogenous parameter ρ , the welfare of $M_\rho f$ will continue to be greater than that of $M_\rho g$.

In this setting, we share precise statements connecting the noising probabilities ρ used in the mechanism M_ρ , their effect on level of privacy ϵ which in turn results in a specific level of influence and welfare expressed in terms of ρ . We precisely show that as the level of privacy increases, the welfare and influence happen to decrease at correspondingly specific rates. Arguably, having a higher welfare in a voting system is desirable and therefore we shine light on this new trade-off between privacy and welfare. In terms of influence, it is questionable whether a decrease in influence with an increase in privacy is desirable or not. We believe it depends on the context, and therefore in this case, we do not refer to it as a trade-off but instead call it a scaling law. However, as we show in Section 5, welfare of the society is equal to total influence of the society under the monotonicity assumption.

² For a survey of the field of analysis of Boolean functions, see [22]. For a survey of the field of differential privacy, see [7].

1.1 Contributions

We contribute towards bridging differential privacy and social choice theory by deriving the following results on the effect of randomized response over influence, welfare, and accuracy.

1. **Privacy-Influence scaling law:** A notion of *influence* is widely used in the analysis of Boolean functions to study social choice functions. We extend the notion of influence to the noisy setting, and call it *probabilistic influence*. We then show a result relating the trade-off between ρ -correlated distribution based differential privacy and probabilistic influence. We show that such privatization changes the influence of every single voter by a factor of $\frac{1+\rho^2}{2}$. Thus, the randomized response preserves the ordinal relations between influences of agents while scaling them by a factor depending on ρ while still ensuring their privacy is preserved.
2. **Privacy-Welfare trade-off:** We define *welfare* $W(f)$ of a social choice function f and extend the definition to probabilistic mechanisms. Then, we show that $W(M_\rho f) = \rho \cdot W(f)$, i.e. the randomized response scales the welfare by a factor of ρ , whereby preserving the ordinal relations between the welfare of social choice functions.
3. **Accuracy analysis:** We restrict the analysis of *accuracy*³ of our mechanism to social choice functions, i.e. the functions with range $\{-1, 1\}$. We give the accuracy for Dictatorship, Majority, AND, and OR functions. For dictatorship, AND, and OR functions, we provide a theoretical analysis of accuracy. For the Majority function, we give an asymptotic accuracy when n goes to ∞ based on the existing results in the literature. We also give an exact analysis of accuracy for the Majority function for small n by using a computational method that involves dynamic programming.

1.2 Organization

The rest of the paper is organized as follows. In Section 2, we provide further motivation and background. In Section 3, we formally describe the differentially private randomized response mechanism. In Section 4, we introduce the notion of probabilistic influence, and give one of our main results that influence scales down by the same constant for every individual. In Section 5, we introduce the concept of welfare for general probabilistic mechanisms, and analyze it for randomized response. We shed light into the connection between influence and welfare, and give our second main result that randomized response scales down welfare by the same factor for any given social choice function. In Section 6, we provide an analysis of the accuracy for the randomized response mechanism. In Section 7, we discuss the possible future work and the limitations of this paper, and we conclude. Some preliminaries from social choice theory are provided in Appendix B. All of the proofs are relegated to Appendix A.

2 Motivation

To intuitively expand on the potential relation between privacy and influence, consider an instance where it might be the case that introduction of noise for the sake of obtaining privacy results in undesired shifts of the power held by different individuals in deciding society's final outcome. For example, say that a voter Alice would have had more impact on the outcome than Bob in a case where there is no privatization. It could as well be the case

³ It is common to refer to *accuracy* with the name *utility* in the differential privacy literature. However, since this term is overloaded also in the economics and social choice theory literature with different meanings, we will opt to call it accuracy throughout our analysis to prevent possible confusion.

that the power balance shifts to Bob having more impact than Alice after a privacy-inducing noise is introduced. We conclusively show that this cannot be the case as the influence scales down for every voter with the increasing level of privacy by the same constant in the case of the popular randomized response privacy mechanism.

Second, regarding the potential relation between privacy and welfare, consider an instance where it may be the case that upon introduction of noise, the chosen social choice function that was originally used to aggregate the individual preferences into a final outcome ends up not being ideal anymore. Hence, it may instead be desirable to switch to another social choice function. For example, suppose that a system uses the majority function to decide which one of the two candidates is elected in the deterministic case. However, the majority function could be severely affected in some instances upon introduction of noise, and another function could end up being a *better* choice. We show that as the privacy increases in the randomized response mechanism, the welfare of each social choice function scales down proportionally under our definition of welfare, which is similar to the notions used in mathematical social choice theory. This implies that if a function is a welfare maximizer before introducing noise, it still is a welfare maximizer after the introduction of the noisy mechanism. These two results are especially useful, as they imply that the designers of the initial deterministic social choice mechanism do not have to be concerned about whether their design is robust to the introduction of noise in terms of influence and welfare.

We now discuss the work that has been done regarding influence and welfare in the context of social choice theory. Influences have long been studied in discrete Fourier analysis and theoretical computer science. The notion of influence was first introduced in [23] and it was first systematically studied in [3]. Some other novel works related to influences in the context of social choice theory include, but are not limited to, the KKL Theorem [14] and the Majority is Stablest Theorem [20]. We extend the notion of influence to the noisy setting and call it *probabilistic influence*, and prove a direct linear relation between deterministic influence and probabilistic influence.

The question of the ideal voting rule has long been a matter of discussion in social choice theory. When there are only two candidates, the answer is relatively simple as the majority function seems to be the most ideal voting rule. It is known that majority is the only social choice function that is anonymous and monotone among all two-candidate voting rules [19]. For more than two candidates, different objectives may result in different voting rules, or even in impossibility results [1, 2, 11, 9, 10]. Various aspects of utilitarian voting is studied in [13]. Finding the best function in computationally efficient ways has been studied in the recent field of computational social choice theory. There is a line of work [16, 17] that aims to maximize welfare given each voter's utility for candidates in a "distortion framework" in which there is a lack of information about voter's utilities. In that framework, a typical approach is to attempt to maximize the worst-case objective.

To the best of our knowledge, a definition of welfare that is closest to ours is the one given by O'Donnell ([22], page 51). Although the author does not explicitly define welfare of a social choice function, there is an affine relation between the expected value of their objective function and the way we define welfare. However, our main conceptual contribution is that our definitions are extended to hold for probabilistic mechanisms and we analyze the effects of privacy on influence and welfare. O'Donnell proves that among all two-candidate voting rules, majority is the unique maximizer of welfare, whose proof is essentially based on [27]. Our main objective is not to find the function that maximizes the welfare; that is rather a simple question. In fact, we show that majority is the unique welfare maximizer as well in an almost identical way to [22]. The primary motivation of the paper is to show that if a voting rule is better in the deterministic setting, it is still better after the privacy-inducing noise is introduced.

3 Model: Randomized Response and Privacy Guarantee

There are three main reasons as to why we chose the randomized response as the privacy-preserving mechanism to focus our attention. First, it is simple, in addition to being one of the earliest, and yet one of the most popularly used privacy-preserving mechanisms to date, be it in the classic form or as a variant of it. As an example, RAPPOR [8] is a recent popular real-world use-case of randomized response, otherwise classically used a few decades ago [28, 18]. Second, the mechanism is based on perturbations of the input which allows it to be applied to *any* social choice function. This enables us to talk about the ordinal relations between the welfare of potential social choice functions before and after the mechanism is applied. Third, ρ -correlated distributions are well studied in mathematical social choice theory [22].

Our randomized mechanism is an input-perturbing mechanism. That is, the mechanism introduces noise to the votes in the ballot so that one can use any social function afterward, yet the same privacy guarantee will continue to hold due to the post-processing property of differential privacy [5]. Randomized response introduces noise by utilizing a simple coin-flip scheme that is based on the following distribution that is widely used in the analysis of Boolean functions.

► **Definition 1.** Let $\rho \in [0, 1]$ and $x \in \{-1, 1\}^n$ be fixed. y is called ρ -correlated with x if for every $i \in [n]$, $y_i = x_i$ with probability ρ and uniformly distributed with probability $1 - \rho$, and it is denoted by $y \sim N_\rho x$.

Note the symmetry in the definition of ρ -correlation. We formalize this symmetry in the following fact, which we will often use in the proofs of our results.

► **Observation 2.** $x \sim \{-1, 1\}^n, y \sim N_\rho x$ if and only if $y \sim \{-1, 1\}^n, x \sim N_\rho y$. If $x \sim \{-1, 1\}^n, y \sim N_\rho x$, we say (x, y) is a ρ -correlated uniformly random pair.

In the literature, ρ -correlated distribution is sometimes referred to as *noisy distribution*. A famous analogy for this definition is as follows. Suppose the votes are recorded by a *noisy* machine. That is, the machine records each ballot correctly with probability ρ , and blurs the ballot with probability $1 - \rho$ and instead records it at uniform random. As a result, the vote gets misrecorded with probability $(1 - \rho)/2$. In fact, our mechanism corresponds to this noisy machine. Hence, we will call it by the generic name *randomized response*, or ρ -correlated *randomized response* when we need to specify ρ and denote a mechanism that applies it by M_ρ as defined below.⁴ It is worth noting that ρ -correlated *randomized response* is in essence just like *randomized response* [28], a classic scheme that inspired several privacy mechanisms.

► **Definition 3.** Let $f : \{-1, 1\}^n \rightarrow \mathbb{R}$ be any function. For every $x \in \{-1, 1\}^n$, the randomized response $M_\rho f(x)$ outputs $f(y)$ where $y \sim N_\rho x$.

Now that we formally defined the randomized response mechanism, we can give the formal definition of differential privacy in our context.

► **Definition 4** (ϵ -Differential Privacy [6]). A randomized voting mechanism $\mathcal{A} : \{-1, 1\}^n \rightarrow \{-1, 1\}$ is ϵ -differentially private if for all pair of neighboring voting profiles $\mathbf{x}, \mathbf{x}' \in \{-1, 1\}^n$ that differ in exactly one bit and for all $\mathbf{s} \in \{-1, 1\}$,

$$\Pr[\mathcal{A}(\mathbf{x}) = \mathbf{s}] \leq e^\epsilon \Pr[\mathcal{A}(\mathbf{x}') = \mathbf{s}]$$

⁴ Note the subtle distinction between M_ρ and N_ρ . The former is a randomized query function, i.e. a random variable; whereas the latter denotes a probability distribution.

The above definition of differential privacy is specific to our context. For the general definition of differential privacy and a broad survey of the field, see [7]. The randomized response mechanism preserves ε -differential privacy. The following result holds for any Boolean function f .

► **Proposition 5.** *For any $\rho \in [0, 1]$, randomized response $M_\rho f$ preserves $\log(\frac{1+\rho}{1-\rho})$ -differential privacy regardless of the function $f : \{-1, 1\}^n \rightarrow \mathbb{R}$. (or, $(\varepsilon, 0)$ -differential privacy when $\rho \leq 1 - \frac{2}{\exp(\varepsilon)+1}$).*

Proof. Proof is relegated to Appendix A.1. ◀

► **Remark 6.** The equality case is satisfied if f is a dictatorship, which implies that the bound $\log(\frac{1+\rho}{1-\rho})$ is tight. That is, when f is a dictatorship, $M_\rho f$ is not ε -differentially private for any $\varepsilon < \log(\frac{1+\rho}{1-\rho})$. In fact, it can be shown that a social choice function f satisfies the equality case if and only if there is a triple (r, b, i) where $r \in \mathbb{R}, b \in \{-1, 1\}, i \in [n]$ such that $\emptyset \neq \{z \in \{-1, 1\}^n | f(z) = r\} \subseteq \{z \in \{-1, 1\}^n | z_i = b\}$.

The reason our mechanism preserves differential privacy for any Boolean function f is that the mechanism is input-perturbing. In this sense, we could instead present the mechanism as $M_\rho : \{-1, 1\}^n \rightarrow \{-1, 1\}^n$ and write $f \circ M_\rho$ instead of $M_\rho f$. Then we could prove the analogous version of Proposition 5, and by using the post-processing property of differential privacy, we would again obtain Proposition 5. In fact, one can see that in the proof, we also re-prove the post-processing property, seemingly for no reason. However, the reason we choose to give the mechanism altogether after post-processing with f is to make the all equality cases in the above remark apparent. Once post-processing is applied black-box, whether the privacy result is robust is not clear anymore. For example, consider any constant function f , e.g. $f(x) = 1$ for any $x \in \{-1, 1\}^n$. In this case, $M_\rho f$ is not only $\log(\frac{1+\rho}{1-\rho})$ -differentially private but 0-differentially private. On the other hand, as Remark 6 implies, the privacy guarantee in Proposition 5 is tight, which we would not be able to show without an explicit proof.

4 Probabilistic Influence

Influence of a voter is a notion that is used to measure the power of an individual on a deterministic social choice function. Influences of Boolean functions have long been studied in computer science and the field of analysis of Boolean functions starting with [3]. The *influence* of a voter in a voting system is defined to be the probability of the change in outcome when the voter changes their vote *ceteris paribus*. For example, in the case of a dictatorship, the dictator has influence 1 while every other voter has influence 0. In the majority function with $n = 2k + 1$ voters, each voter's influence is the same and equal to $\binom{2k}{k}/2^{2k}$.

We use $x_{i \rightarrow 1} = (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n)$ to denote the case where the i -th voter chooses to vote for 1, and every other voter follows x . Similarly, we denote the alternate case where the i -th voter chooses to vote for -1 and every other voter follows x by $x_{i \rightarrow -1} = (x_1, \dots, x_{i-1}, -1, x_{i+1}, \dots, x_n)$. Using this notation, influence in the deterministic setting is defined as follows.

► **Definition 7.** *For $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$, the influence of elector i is defined as*

$$I_i[f] = \mathbb{P}_{x \sim \{-1, 1\}^n} [f(x_{i \rightarrow 1}) \neq f(x_{i \rightarrow -1})].$$

The total influence of the function f is defined to be

$$I[f] = \sum_{i=1}^n I_i[f].$$

A similar notion can be introduced in the probabilistic setting where the randomized response $M_\rho f(x)$ is applied. To do so, we consider the case where everybody casts their votes, following which $M_\rho f(x)$ is applied and the voter i changes their vote. That is, we leave all the noisy versions of the votes cast by everyone as is except for the elector i 's vote. For this particular vote, we re-run the randomized response on coordinate i . The probability of result being different is called the *probabilistic influence* of coordinate i . We now introduce the formal definition of the proposed probabilistic influence, which applies not only to social choice functions with range $\{-1, 1\}$ but to all Boolean functions with range in \mathbb{R} as follows. In the notation of the following definition, $y_i \sim N_\rho(1)$ refers to the case where voter i chooses to vote for 1 while $z_i \sim N_\rho(-1)$ refers to the case where voter i chooses to vote for -1 .

► **Definition 8.** Let $f : \{-1, 1\}^n \rightarrow \mathbb{R}$ and the probabilistic influence of coordinate i in a mechanism $M_\rho f(x)$ is defined as

$$I_i[M_\rho f] = \mathbb{E}_{x \sim \{-1, 1\}^n, \forall j \neq i, z_j = y_j = x_j, y_i \sim N_\rho(1), z_i \sim N_\rho(-1)} \left[\left(\frac{f(y) - f(z)}{2} \right)^2 \right].$$

The total influence of the mechanism $M_\rho f$ is defined to be

$$I[M_\rho f] = \sum_{i=1}^n I_i[M_\rho f].$$

We showed in Proposition 5 that our probabilistic voting mechanism preserves ϵ -differential privacy. Inducing such privacy requires probabilistic mechanisms as opposed to using deterministic functions. For example, in the majority voting with $2k + 1$ voters, if the votes are split k to $k + 1$, then changing only one bit in the input may change the outcome of the voting mechanism. Thus, it is not differentially private. Similarly, no deterministic Boolean function can preserve differential privacy unless it is a constant function.

On the other hand, introducing noise may cause several issues in the voting system, one of which is the accuracy of the mechanism, which we will discuss in more detail in Section 6. Another possible issue is that when noise is introduced, we might be altering the voting system in favor of a particular voter. For example, voter A might have more influence relative to voter B in the system now even if that was not the case before. For symmetric social choice functions, it is natural to expect that the randomized response mechanism would have the same effect for any voter since the noise is also symmetric. However, it is not as trivial for arbitrary social choice functions. Yet, we show that each voter's probabilistic influence is proportional to her influence in the deterministic setting, which is one of our main results.

► **Theorem 9.** Let $\rho \in [0, 1]$ be any real number and $f : \{-1, 1\}^n \rightarrow \mathbb{R}$ be any function. For every $i \in [n]$, $I_i[M_\rho f] = \frac{1+\rho^2}{2} I_i[f]$.

Proof. Proof is relegated to Appendix A.2. ◀

This result shows that the randomized response preserves the ordinal relations between influences of the voters regardless of the original social choice function being used. In other words, if voter A had greater influence than another voter B , she will still have a greater influence on the system after the noise is introduced.

5 Welfare

In this section, we introduce a formal definition of *welfare* of social choice functions. Then we extend this definition to probabilistic mechanisms, and we show that the randomized response preserves the ordinal relations between the welfare of social choice functions.

5.1 Welfare of Deterministic Voting Systems

[24] argues in his *Social Contract* that an ideal voting rule should maximize the number of votes that agree with the outcome. For a more comprehensive discussion on this, see [25]. [22] proves that the majority function is the unique ideal function based on Rousseau's perception of the ideal voting rule without formally introducing welfare. Perhaps, when he proved this result, he had some form of welfare in his mind, especially because he uses the letter w to denote the number of votes that agrees with the outcome. In this section, we will formally define welfare, which will be slightly different than what the w notation of O'Donnell describes. In particular, we define *welfare* of a social choice function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ as the average difference between the number of votes that agree with the outcome and the number of votes that do not agree with the outcome under the impartial culture assumption.

► **Definition 10.** Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ and $x \in \{-1, 1\}^n$, and let $w_x(f) = |\{i; x_i = f(x)\}| - |\{i; x_i \neq f(x)\}|$. *Welfare of the social choice function f is defined to be*

$$W(f) = \mathbb{E}_x[w_x(f)].$$

We can still prove that the majority function is the unique maximizer of welfare when n is odd by using a similar method as in the proof of Theorem 2.33 in [22].

► **Proposition 11.** *When n is odd, the unique maximizer of $W(f)$ is the majority function.*

Proof. Proof is relegated to A.3. ◀

Without further assessment, it is not possible to say whether we prefer total influence to be larger or smaller for the welfare of society in a voting system. As we show in the following result, if the social choice function is monotone – that is if a voter changes her vote in favor of a candidate, then this candidate should be weakly better off – then these two notions collide.

► **Proposition 12.** *Let f be any monotone social choice function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$. Then, $W(f) = I[f]$.*

Proof. Proof is relegated to Appendix A.4. ◀

This result has implications beyond being a simple identity, making the case that if we want to achieve a greater social welfare while adhering to monotone social choice functions, we must choose a function with a greater total influence.

5.2 Welfare of Noisy Mechanisms

To capture the same notion for the probabilistic functions as well, we similarly define welfare of a randomized mechanism applied on a social choice function as follows. Note that the following definition is not only for the randomized response M_ρ , but any mechanism defined on social choice functions.

► **Definition 13.** Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$, $x \in \{-1, 1\}^n$, and M be any mechanism. Let $w_x(Mf) = |\{i; x_i = Mf(x)\}| - |\{i; x_i \neq Mf(x)\}|$. *Welfare of the mechanism M with the social choice function f is defined to be*

$$W(Mf) = \mathbb{E}_{x, M}[w_x(Mf)]$$

where the expectation is both over x and the mechanism M .

We showed in Theorem 9 that although introducing ρ -correlated noise in a voting system has negative effects on influences, it does not provide an unfair advantage to any agent. Another possible undesired byproduct of a randomized mechanism could be that the effect of randomization on the welfare of a particular voting system is more severe compared to the other voting systems. For example, we showed in Proposition 11 that the majority function is the unique welfare maximizer. It could be the case that after we introduce noise, it is more likely in the majority function that the outcome will change. Within this context, the following result implies that every voting system is equally affected by the input-perturbing randomized response mechanism. Therefore the randomized response preserves the ordinal relations between the welfare of two-candidate voting systems.

► **Theorem 14.** *Let f be any social choice function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$. Then, $W(M_\rho f) = \rho \cdot W(f)$.*

Proof. Proof is relegated to Appendix A.5 ◀

This result, together with Proposition 11, implies that the majority function is the unique welfare maximizer also after the noise is introduced in applying the randomized response mechanism.

6 Accuracy Analysis

There is one significant drawback of the randomized response privatization mechanism in consideration. It is hard to analyze the accuracy of releasing the output of social choice functions upon privatizing it with the randomized response. Although our main objective in this work is not about the analysis of accuracy, we will dedicate a section to the analysis of accuracy for the sake of completeness. As a first pass, we easily find a *generic* lower-bound on accuracy of the randomized response, but it ends up to be so low that it makes it redundant. Therefore, we restrict our analysis to *specific* social choice functions. We theoretically provide results on accuracy for dictatorship, AND, and OR functions.⁵ In addition, we give a tight lower bound as well as an upper bound for the accuracy of majority function. We also give an algorithm to calculate exact accuracy of majority function by using dynamic programming via memoization. The dynamic programming approach avoids the need to make calculations over every entry in the power-set and instead is much more efficient, while still resulting in an exact solution for computing the accuracy. Our definition of accuracy is in-fact the average of accuracy under the impartial culture assumption. That is,

$$\text{Acc}(M_\rho f) = \mathbb{P}_{x \sim \{-1, 1\}^n} [M_\rho f(x) = f(x)].$$

Now, we define the *noise operator*, also referred to as the noisy Markov operator, which is a linear operator on the set of Boolean functions. This operator will be useful for accuracy calculations.

► **Definition 15.** *For any $\rho \in [0, 1]$, the noise operator T_ρ is the linear operator on the set of functions $f : \{-1, 1\} \rightarrow \mathbb{R}$ defined by*

$$T_\rho f(x) = \mathbb{E}_{y \sim N_\rho x} [f(y)].$$

⁵ For formal definitions of these widely known social choice functions, see Appendix B.

Before we start our analysis, let us also give the definition of *noise stability*.

► **Definition 16.** For any $\rho \in [0, 1]$ and $f : \{-1, 1\}^n \rightarrow \mathbb{R}$, ρ -correlated noise stability of f is given by

$$Stab_\rho(f) = \mathbb{E}_{\substack{x \sim \{-1, 1\}^n \\ y \sim N_\rho(x)}} [f(x) \cdot f(y)]$$

There is a linear relation between the noise stability of a function and accuracy of the randomized response on this function. Note that $M_\rho f(x) \cdot f(x) = 1$ if $M_\rho f(x) = f(x)$, $M_\rho f(x) \cdot f(x) = -1$ otherwise. Thus,

$$2 \cdot Acc(M_\rho f) - 1 = 2 \cdot \mathbb{P}_{\substack{x \sim \{-1, 1\}^n \\ y \sim N_\rho(x)}} [f(y) = f(x)] - 1 = \mathbb{E}_{\substack{x \sim \{-1, 1\}^n \\ y \sim N_\rho(x)}} [f(y) \cdot f(x)] = Stab_\rho(f). \quad (1)$$

Also, note that

$$Stab_\rho(f) = \mathbb{E}_{\substack{x \sim \{-1, 1\}^n \\ y \sim N_\rho(x)}} [f(x) \cdot f(y)] = \mathbb{E}_{x \sim \{-1, 1\}^n} [f(x) T_\rho f(x)]. \quad (2)$$

The reason we feel the need to write accuracy in terms of stability is that in the field of Analysis of Boolean functions most results are given in terms of stability for convenience. Yet, we use stability explicitly only when we analyze the accuracy of the majority function.

6.1 Majority

In this section, we will give the asymptotic accuracy for Maj_n function where n is an odd number that goes to infinity.

► **Lemma 17** (Proposition 10, [21]). For any $\rho \in [0, 1]$, $Stab_\rho[Maj_n]$ is a decreasing function of n where n is an odd number, with

$$\frac{2}{\pi} \arcsin(\rho) \leq Stab_\rho[Maj_n] \leq \frac{2}{\pi} \arcsin(\rho) + O\left(\frac{1}{\sqrt{1 - \rho^2} \sqrt{n}}\right).$$

By using the fact that accuracy is equal to $\frac{1}{2} + \frac{1}{2} Stab_\rho(f)$ due to Equation (1), we get that

$$\frac{1}{2} + \frac{1}{\pi} \arcsin(\rho) \leq Acc[M_\rho(Maj_n)] \leq \frac{1}{2} + \frac{1}{\pi} \arcsin(\rho) + O\left(\frac{1}{\sqrt{1 - \rho^2} \sqrt{n}}\right). \quad (3)$$

Despite this fact being quite useful, there is no convenient way to calculate the exact value of accuracy of the randomized response on Majority function. Hence, we compute it using dynamic programming via memoization in the following section.

6.1.1 Algorithm to compute the exact accuracy for small n

We now provide a dynamic programming algorithm with memoization to compute the accuracy of the randomized response. In particular, we give the algorithm to calculate the accuracy of the *threshold functions*, that are of the form

$$f_\theta(x) = \begin{cases} 1 & \text{if } \sum_{i \in [n]} x_i > \theta \\ -1 & \text{if } \sum_{i \in [n]} x_i \leq \theta \end{cases}$$

Note that $Maj_n = f_0(\cdot)$ where it takes care of ties by considering them as if -1 is the winner. In general, we work with the odd number of voters when we talk about the majority function. But as a simple trick, we will compute it for any n based on the generic definition of the threshold function we gave above since it makes the algorithm less involved.

We now state the noise operator $T_\rho f_{\theta_0}(x)$ as introduced in Definition 15 when applied to threshold functions as a way to quantify the expected accuracy as

$$T_\rho f_{\theta_0}(x) = \mathbb{E}_{y \sim N_\rho x} [\mathbb{1}(y_1 + \dots + y_n > \theta_0)].$$

Let x_{-n} denote x without the last bit. In particular, if $x = (x_1, x_2, \dots, x_{n-1}, x_n)$, then $x_{-n} = (x_1, x_2, \dots, x_{n-1})$. Note that $x_{-n} \in \{-1, 1\}^{n-1}$ while $x \in \{-1, 1\}^n$. Then, the stability can be defined using two calls of recursion as follows

$$T_\rho f_{\theta_0}(x) = \frac{1+\rho}{2} T_\rho f_{\theta_0-x_n}(x_{-n}) + \frac{1-\rho}{2} T_\rho f_{\theta_0+x_n}(x_{-n})$$

That is because

$$\begin{aligned} & \mathbb{E}_{y \sim N_\rho x} [\mathbb{1}(y_1 + \dots + y_n > \theta_0)] \\ &= \mathbb{E}_{y_n \sim N_\rho x_n} [\mathbb{E}_{y_{-n} \sim N_\rho x_{-n}} [\mathbb{1}(y_1 + \dots + y_{n-1} > \theta_0 - y_n) \mid y_n]] \\ &= \frac{1+\rho}{2} \mathbb{E}_{y_{-n} \sim N_\rho(x_{-n})} [\mathbb{1}(y_1 + \dots + y_{n-1} > \theta_0 - x_n)] \\ &\quad + \frac{1-\rho}{2} \mathbb{E}_{y_{-n} \sim N_\rho(x_{-n})} [\mathbb{1}(y_1 + \dots + y_{n-1} > \theta_0 + x_n)] \\ &= \frac{1+\rho}{2} T_\rho f_{\theta_0-x_n}(x_{-n}) + \frac{1-\rho}{2} T_\rho f_{\theta_0+x_n}(x_{-n}) \end{aligned}$$

To summarize, this dynamic programming with memoization algorithm is as shown in Algorithm 1 below. In terms of notation we denote a specific dictionary (in terms of popular programming terminology of dictionary data types) as Dictionary: $\{(\rho, n, s, \theta) = T_\rho f_{\theta_0}(x) \text{ for some } x \text{ s.t. } \text{sum}(x) = s\}$.

Our approach is to use this proposed recursive relation with an appropriate initial condition to exactly compute the noise operator $T_\rho f(x)$. Then, by using Equation (2), we calculate the Stability of the function. Finally, by using the linear relation between stability and accuracy from Equation (1), we compute the exact accuracy. This dynamic programming approach avoids having to make 2^n computations, given that $x \sim \{-1, 1\}^n$. Note that, $T_\rho f_{\theta_0}(x) = T_\rho f_{\theta_0}(z)$ if $\text{sum}(x) = \text{sum}(z)$. Therefore we iterate over i from 1 to n to represent vectors with i number of 1's. Then as the rest of entries are -1 , and since the length of the array is n , this approach can model the exact sum of all possible vectors. Since the calculation of the stability is one-to-one with respect to sums, we store the intermediate results in a dictionary indexed by this sum. As there are $\binom{n}{i}$ vectors that can be represented this way, we just compute once per each i and multiply it by $\binom{n}{i}$. This enables us to model all possible vectors efficiently but allows us to not have to compute the intermediate results every time via our recursive approach.

In Figure 1, we plot the accuracy curves of the randomized response mechanism with varying values of ρ applied to the majority function as the number of voters increases. Note that as n goes to ∞ , the accuracy asymptotically approaches to $\frac{1}{2} + \frac{1}{\pi} \arcsin(\rho)$ as implied by Equation (3).

■ **Algorithm 1** Proposed dynamic programming algorithm with memoization.

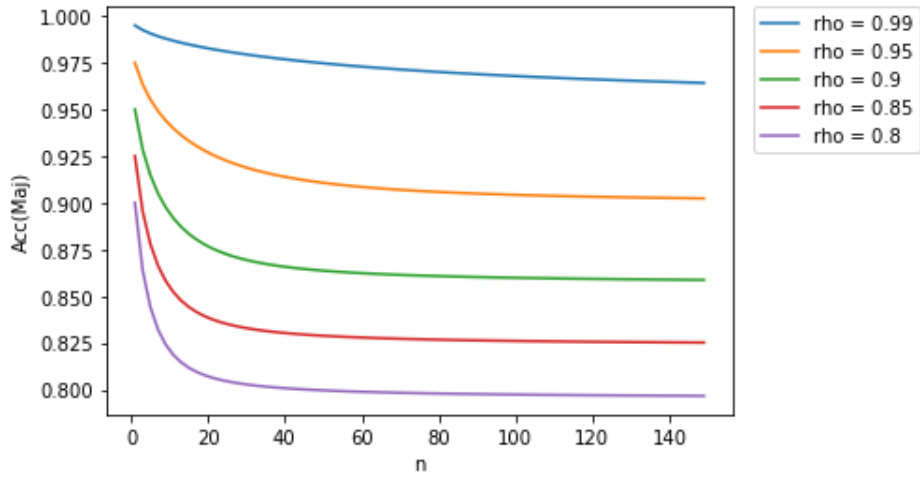
```

Result: Accuracy
Initialization;
Define Dictionary:  $\{(\rho, n, s, \theta) = T_\rho f_{\theta_0}(x) \text{ for some } x \text{ s.t. } \text{sum}(x) = s\}$ 
Def  $T_\rho f_{\theta_0}(x)$  :
 $s = \text{sum}(x)$ ;
if  $(\rho, n, s, \theta_0)$  is in dictionary then
|   return dictionary  $[(\rho, n, s, \theta_0)]$ ;
else
|   Using 2 recursive calls in summands, compute:
|
|       
$$\alpha = \frac{1 + \rho}{2} T_\rho f_{\theta_0 - x_n}(x_{-n}) + \frac{1 - \rho}{2} T_\rho f_{\theta_0 + x_n}(x_{-n})$$

|
|   Save  $(\rho, n, s, \theta_0) = \alpha$  to dictionary
end
Def  $\text{Acc}_\rho(f_\theta)$  :
total = 0
for  $i \leftarrow 1$  to  $n + 1$  do
|
|       
$$\text{total} += \binom{n}{i} \cdot f_{\theta_0}(x) \cdot T_\rho f_\theta(x) \text{ for some } x \text{ s.t. } x \text{ has } i \text{ different } +1 \text{ bits}$$

|
end
return  $\frac{1}{2} + \frac{\text{total}/2^n}{2}$ 

```



■ **Figure 1** The accuracy curves of the randomized response mechanism with varying values of ρ applied to the majority function as the number of voters increases.

6.2 Dictatorship

Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ be the dictatorship of voter- i , that is $f(x) = 1$ if and only if $x_i = 1$.

Then, for any given $x \in \{-1, 1\}^n$,

$$\mathbb{P}[M_\rho f(x) = f(x)] = \mathbb{P}_{y \sim N_\rho(x)}[f(y) = f(x)] = \mathbb{P}_{y_i \sim N_\rho(x_i)}[y_i = x_i] = \frac{1 + \rho}{2}.$$

Hence, the average accuracy is also equal to $\frac{1+\rho}{2}$.

6.3 AND_n and OR_n

We will first make the calculations for AND_n and the results will be analogous due to symmetry. We will make use of Fact 2 in the analysis.

First, we start with a generic calculation that holds for any social choice function f . In the calculations in this section, our probability space is $x \sim \{-1, 1\}^n$, $M_\rho f(x) \sim f(y)$ where $y \sim N_\rho x$.

Note that by Fact 2,

$$\mathbb{P}_{x, M_\rho}[M_\rho f(x) = 1] = \mathbb{P}_x[f(x) = 1].$$

$$\mathbb{P}[M_\rho f(x) = f(x)] = \mathbb{P}[M_\rho f(x) = 1 \wedge f(x) = 1] + \mathbb{P}[M_\rho f(x) = -1 \wedge f(x) = -1]$$

and

$$\begin{aligned} \mathbb{P}[M_\rho f(x) = -1 \wedge f(x) = -1] &= 1 - \mathbb{P}[M_\rho f(x) = 1 \vee f(x) = 1] \\ &= 1 - \mathbb{P}[M_\rho f(x) = 1] - \mathbb{P}[f(x) = 1] + \mathbb{P}[M_\rho f(x) = 1 \wedge f(x) = 1] \\ &= 1 - 2 \cdot \mathbb{P}[f(x) = 1] + \mathbb{P}[M_\rho f(x) = 1 \wedge f(x) = 1]. \end{aligned}$$

Thus for any social choice function f ,

$$\mathbb{P}[M_\rho f(x) = f(x)] = 1 - 2 \cdot \mathbb{P}[f(x) = 1] + 2 \cdot \mathbb{P}[M_\rho f(x) = 1 \wedge f(x) = 1]$$

For $f = AND_n$,

$$\mathbb{P}[f(x) = 1] = \prod_{i \in [n]} \mathbb{P}[x_i = 1] = 2^{-n},$$

and

$$\mathbb{P}[M_\rho f(x) = 1 \wedge f(x) = 1] = \mathbb{P}[f(x) = 1] \cdot \mathbb{P}[M_\rho f(x) = 1 | f(x) = 1] = 2^{-n} \cdot \left(\frac{1 + \rho}{2}\right)^{-n}.$$

Hence, the accuracy of M_ρ for AND_n function is equal to $1 - 2^{-n+1}(1 - (\frac{1+\rho}{2})^n)$, whose limit goes to 1 as n goes to ∞ . Due to symmetry, accuracy analysis is the same for OR_n function.

7 Conclusion

The main objective in this work is to study the privacy-welfare trade-off and the relation between privacy and probabilistic influence. The proposed definition of welfare happens to hold for any mechanism while on the other hand, the defined probabilistic influence is only specific to the randomized response mechanism. In fact, a more general definition of influence could be coined and a similar property could potentially be observed. We leave out this potential generalization of influence to future work. In terms of welfare, the analysis

done in this paper can be replicated in a similar style to other popular privatization schemes such as the Laplace and exponential mechanisms. The privacy-accuracy trade-off of the current mechanism for the majority function may also be further improved. Note that Dictatorship, AND, and OR functions satisfy the equality condition in Proposition 5 as discussed in Remark 6. Thus, the accuracy-privacy analyses for these functions are tight. On the other hand, for a given ρ , the asymptotic accuracy of majority is tight whereas the privacy result is a possibly loose upper bound.

Also, our definitions of influence and welfare assume that the votes are unbiased, that is, they consider everybody to be equally likely to vote for -1 or $+1$. In fact, these definitions can be further generalized to cover the same concept, but for the case of biased voting. For example, one can extend the definitions to be p -biased for a given $p \in [-1, 1]$, that is the expected value of each vote is p instead of 0. p -biased distribution is also well-studied in the field of Analysis of Boolean functions.

Finally, our voting model in this paper is a classical referendum model with two candidates. However, in most real-world applications, we generally have multiple candidates and we have to aggregate the rankings. If there is a Condorcet winner in a voting system, then the results regarding two-candidate elections can be directly applied in the multiple-candidate setting. Yet, in many cases, there is no Condorcet winner. Restricting the number of candidates to two has the primary advantage that both the definitions and analyses of welfare and influence naturally follow. We believe that extending the definitions and the tools developed in this paper to multiple-candidate settings would be interesting.

In a broader perspective, we study the effect of using privacy inducing randomized responses in the voting process. We construct a relation between the level of privacy and the resulting level of influence of voters involved in the voting system and the welfare of the chosen social choice function. An insightful takeaway that we can deduce from the derived relationships in this paper is that the ordering of voters' influences and the ordering of welfare amongst the considered social choice functions remain unchanged upon introducing noise via the celebrated randomized response mechanism. Existing works have extensively studied the relationship between privacy and the resulting accuracy in preserving the output of the query that was privatized. At a high level we are the first to shed light on the relationship between privacy and other important phenomena of influence and welfare. We hope that this bridge we have proposed between the two important fields of differential privacy and social choice theory will be further studied and extended as part of future works.

References

- 1 Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.
- 2 Kenneth J Arrow. *Social choice and individual values*. Yale university press, 1951.
- 3 Michael Ben-Or and Nathan Linial. Collective coin flipping, robust voting schemes and minima of banzhaf values. *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 408–416, 1985.
- 4 Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.
- 5 Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.
- 6 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

- 7 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.
- 8 Úlfar Erlingsson, Vasył Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.
- 9 Mark B. Garman and Morton I. Kamien. The paradox of voting: Probability calculations. *Behavioral Science*, 13(4):306–316, 1968.
- 10 Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.
- 11 Georges-Théodule Guilbaud. Les théories de l'intérêt général et le problème logique de l'agrégation. *Revue économique*, 63(4):659–720, 2012.
- 12 Michael Hay, Liudmila Elagina, and Gerome Miklau. Differentially private rank aggregation. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, pages 669–677, 2017.
- 13 Claude Hillinger. The case for utilitarian voting. *Homo Oeconomicus*, 22(3):295–321, 2005.
- 14 J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *[Proceedings 1988] 29th Annual Symposium on Foundations of Computer Science*, pages 68–80, 1988.
- 15 David Timothy Lee. Efficient, private, and eps-strategyproof elicitation of tournament voting rules. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- 16 Debmalya Mandal, Ariel D Procaccia, Nisarg Shah, and David Woodruff. Efficient and thrifty voting by any means necessary. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- 17 Debmalya Mandal, Nisarg Shah, and David P. Woodruff. Optimal communication-distortion tradeoff in voting. In *Proceedings of the 21st ACM Conference on Economics and Computation, EC '20*, pages 795–813, New York, NY, USA, 2020. Association for Computing Machinery.
- 18 Naurang S Mangat. An improved randomized response strategy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):93–95, 1994.
- 19 Kenneth O. May. A set of independent necessary and sufficient conditions for simple majority decisions. *Econometrica*, 20(4):680–684, 1952.
- 20 Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: Invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010.
- 21 Ryan O'Donnell. Hardness amplification within np. *Journal of Computer and System Sciences*, 69(1):68–94, 2004. Special Issue on Computational Complexity 2002.
- 22 Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.
- 23 Lionel Penrose. The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, pages 109(1):53–57, 1946.
- 24 Jean-Jacques Rousseau. Du contrat social. *Marc-Michel Rey*, 1762.
- 25 Melissa Schwartzberg. Voting the general will: Rousseau on decision rules. *Political Theory*, 36(3):403–423, 2008. URL: <http://www.jstor.org/stable/20452639>.
- 26 Shang Shang, Tiance Wang, Paul Cuff, and Sanjeev Kulkarni. The application of differential privacy for rank aggregation: Privacy and accuracy, 2014. [arXiv:1409.6831](https://arxiv.org/abs/1409.6831).
- 27 Robert Tittsworth. Correlation properties of cyclic sequences. *PhD thesis, CalTech*, 1962.
- 28 Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

A Proofs

A.1 Proof of Proposition 5

Proof. Let r be any element in the range of $M_\rho f$. Let $Z = \{z \in \{-1, 1\}^n \mid f(z) = r\}$. Let x and x' differ only at x_i for some $i \in [n]$.

$$\frac{\mathbb{P}[M_\rho f(x) = r]}{\mathbb{P}[M_\rho f(x') = r]} = \frac{\sum_{z \in Z} \mathbb{P}_{y \sim N_\rho x}[y = z]}{\sum_{z \in Z} \mathbb{P}_{y \sim N_\rho x'}[y = z]} = \frac{\sum_{z \in Z} \prod_{j \in [n]} \mathbb{P}_{y_j \sim N_\rho x_j}[y_j = z_j]}{\sum_{z \in Z} \prod_{j \in [n]} \mathbb{P}_{y_j \sim N_\rho x'_j}[y_j = z_j]}.$$

The first equality is upon considering all cases of output of the randomized response resulting in a $z \in Z$. Then by definition that would result in the function f evaluated on this output z to be r . The second equality is due to the independence assumption across the voters choices. Now, for any $z \in Z$,

$$\mathbb{P}_{y_j \sim N_\rho x_j}[y_j = z_j] = \begin{cases} \frac{1+\rho}{2} & \text{if } x_j = z_j \\ \frac{1-\rho}{2} & \text{if } x_j \neq z_j \end{cases} \quad \text{and} \quad \mathbb{P}_{y_j \sim N_\rho x'_j}[y_j = z_j] = \begin{cases} \frac{1+\rho}{2} & \text{if } x'_j = z_j \\ \frac{1-\rho}{2} & \text{if } x'_j \neq z_j \end{cases}$$

This is because $\frac{1-\rho}{2}$ is the probability of a misrecorded vote and $1 - \frac{1-\rho}{2} = \frac{1+\rho}{2}$ is the probability otherwise. More explicitly, with probability $1 - \rho$, it chooses to blur the ballot and the blurring is then done by picking uniformly out of the two options of $\{-1, 1\}$ with probability 0.5 each, out of which one pick would result in no change to the vote and the other would result in a misrecorded vote. Also, for any $j \neq i$,

$$\mathbb{P}_{y_j \sim N_\rho x_j}[y_j = z_j] = \mathbb{P}_{y_j \sim N_\rho x'_j}[y_j = z_j].$$

Thus,

$$\frac{1 - \rho}{1 + \rho} \leq \frac{\sum_{z \in Z} \prod_{j \in [n]} \mathbb{P}_{y_j \sim N_\rho x_j}[y_j = z_j]}{\sum_{z \in Z} \prod_{j \in [n]} \mathbb{P}_{y_j \sim N_\rho x'_j}[y_j = z_j]} \leq \frac{1 + \rho}{1 - \rho},$$

which completes the proof. ◀

A.2 Proof of Theorem 9

Proof. Using conditional probability, we get that

$$\begin{aligned} I_i[M_\rho f] &= \mathbb{E}_{x \sim \{-1, 1\}^n, \forall j \neq i, z_j = y_j = x_j, y_i \sim N_\rho(1), z_i \sim N_\rho(-1)} \left[\left(\frac{f(y) - f(z)}{2} \right)^2 \right] \\ &= \mathbb{P}_{y_i \sim N_\rho(1), z_i \sim N_\rho(-1)}[y_i = 1, z_i = -1] \cdot \mathbb{E}_{x \sim \{-1, 1\}^n} \left[\left(\frac{f(x_{i \rightarrow 1}) - f(x_{i \rightarrow -1})}{2} \right)^2 \right] \\ &\quad + \mathbb{P}_{y_i \sim N_\rho(1), z_i \sim N_\rho(-1)}[y_i = -1, z_i = 1] \cdot \mathbb{E}_{x \sim \{-1, 1\}^n} \left[\left(\frac{f(x_{i \rightarrow 1}) - f(x_{i \rightarrow -1})}{2} \right)^2 \right] \end{aligned}$$

Noting that

$$\begin{aligned} \mathbb{P}_{y_i \sim N_\rho(1), z_i \sim N_\rho(-1)}[y_i = 1, z_i = -1] &= \left(\frac{1 + \rho}{2} \right)^2, \\ \mathbb{P}_{y_i \sim N_\rho(1), z_i \sim N_\rho(-1)}[y_i = -1, z_i = 1] &= \left(\frac{1 - \rho}{2} \right)^2, \end{aligned}$$

and that

$$\mathbb{E}_{x \sim \{-1,1\}^n} \left[\left(\frac{f(x_{i \rightarrow 1}) - f(x_{i \rightarrow -1})}{2} \right)^2 \right] = I_i[f],$$

we get that

$$I_i[M_\rho f] = \frac{1 + \rho^2}{2} I_i[f]. \quad \blacktriangleleft$$

A.3 Proof of Proposition 11

Proof. First, let us fix x . Note that

$$w_x(f) = f(x) \cdot \sum_{i \in [n]} x_i.$$

Since $f(x) \in \{-1, 1\}$, $f(x) \cdot \sum_{i \in [n]} x_i$ is maximized when $f(x) = \text{sign}(\sum_{i \in [n]} x_i)$. Hence, $W(f)$ is maximized if $\forall x \in \{-1, 1\}^n$, $f(x) = \text{sign}(\sum_{i \in [n]} x_i)$, which is exactly the definition of the majority function. \blacktriangleleft

► **Remark 18.** Note that we used the condition that n is odd to ensure that sign function is well-defined. If n was even, then the maximizers of $W(f)$ are again the majority functions where it does not matter who is elected if it is tied.

A.4 Proof of Proposition 12

In the proof of this result, we use discrete Fourier analysis. It is a well-known result from the field of analysis of Boolean functions, that every function $f : \{-1, 1\}^n \rightarrow \mathbb{R}$ can be uniquely expressed as a multilinear polynomial,

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \chi_S(x)$$

where for any $S \subseteq [n]$

$$\chi_S(x) = \prod_{i \in S} x_i.$$

This expression is called the Fourier expansion of f , and the real number $\widehat{f}(S)$ is called the Fourier coefficient of f on S . Collectively, the coefficients are called the Fourier spectrum of f . The following is an essential result from discrete Fourier Analysis.

► **Lemma 19** (Plancherel's Theorem). *For any functions $f, g : \{-1, 1\}^n \rightarrow \mathbb{R}$,*

$$\mathbb{E}_{x \sim \{-1,1\}^n} [f(x)g(x)] = \sum_{S \subseteq [n]} \widehat{f}(S) \widehat{g}(S).$$

It is possible to neatly calculate many features of f including the influences in terms of Fourier coefficients.

► **Lemma 20** (Proposition 2.21, [22]). *Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ be a monotone function and let the Fourier spectrum of f be $f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \chi_S(x)$. Then, for any $i \in [n]$,*

$$I_i[f] = \widehat{f}(\{i\}).$$

It is also possible to calculate the welfare in terms of the Fourier coefficients by taking one step further from the proof of Proposition 11.

► **Lemma 21.** *Let f be any social choice function $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$. Then, $W(f) = \sum_{i \in [n]} \hat{f}(\{i\})$.*

Proof. By the definition of welfare,

$$W(f) = \mathbb{E}_x[w_x(f)] = \mathbb{E}_x[f(x) \cdot \sum_{i \in [n]} x_i] = \sum_{i \in [n]} \hat{f}(\{i\})$$

where the last equation follows from Lemma 19. ◀

We are ready to finish the proof.

Proof of Proposition 12. The proof follows immediately from Lemma 20 and Lemma 21. ◀

A.5 Proof of Theorem 14

Proof. We prove this identity by using a double-counting method and linearity of expectation. Fix f . For any $i \in [n]$, let $1_{i,x,\rho}$ be the indicator random variable defined as follows:

$$1_{i,x,\rho} = \begin{cases} 1 & \text{if } M_\rho f(x) = x_i \\ -1 & \text{if } M_\rho f(x) \neq x_i \end{cases}$$

where the randomization is due to the randomized response. Note then when x is given and $\rho = 1$, there is no randomization because $M_\rho f(x) = f(x)$ with probability 1. Therefore, $1_{i,x,1}$ is a deterministic function. For the sake of simplicity, we will abuse the notation and write $1_{i,x}$ instead of $1_{i,x,1}$ in the deterministic case. Then,

$$w_x(M_\rho f) = \sum_{i \in [n]} 1_{i,x,\rho} \quad \text{and} \quad w_x(f) = \sum_{i \in [n]} 1_{i,x}$$

Thus,

$$W(M_\rho f) = \mathbb{E}_{M_\rho, x}[w_x(M_\rho f)] = \mathbb{E}_{x, M_\rho} \left[\sum_{i \in [n]} 1_{i,x,\rho} \right] = \sum_{i \in [n]} \mathbb{E}_{x, M_\rho} [1_{i,x,\rho}]$$

and so

$$W(f) = \sum_{i \in [n]} \mathbb{E}_x [1_{i,x}].$$

Now, we will show that for any $i \in [n]$,

$$\mathbb{E}_{x, M_\rho} [1_{i,x,\rho}] = \rho \cdot \mathbb{E}_x [1_{i,x}].$$

First, note that

$$\mathbb{E}_{x, M_\rho} [1_{i,x,\rho}] = \mathbb{P}_{\substack{x \sim \{-1,1\}^n \\ y \sim N_\rho x}} [f(y) = x_i] - \mathbb{P}_{\substack{x \sim \{-1,1\}^n \\ y \sim N_\rho x}} [f(y) \neq x_i].$$

By using

$$\mathbb{P}_{\substack{x \sim \{-1,1\}^n \\ y \sim N_\rho x}} [f(y) = x_i] + \mathbb{P}_{\substack{x \sim \{-1,1\}^n \\ y \sim N_\rho x}} [f(y) \neq x_i] = 1,$$

we get that

$$\mathbb{E}_{x, M_\rho}[1_{i, x, \rho}] = 2 \cdot \mathbb{P}_{\substack{x \sim \{-1, 1\}^n \\ y \sim N_\rho x}}[f(y) = x_i] - 1.$$

By Fact 2, we can replace $x \sim \{-1, 1\}^n, y \sim N_\rho x$ with $y \sim \{-1, 1\}^n, x \sim N_\rho y$. Thus, by using conditional probability,

$$\begin{aligned} \mathbb{E}_{x, M_\rho}[1_{i, x, \rho}] &= 2 \cdot \mathbb{P}_{\substack{y \sim \{-1, 1\}^n \\ x \sim N_\rho y}}[f(y) = x_i] - 1 \\ &= 2(\mathbb{P}_{x \sim N_\rho y}[x_i = y_i] \cdot \mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) = y_i] \\ &\quad + \mathbb{P}_{x \sim N_\rho y}[x_i = -y_i] \cdot \mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) = -y_i]) - 1 \\ &= (1 + \rho) \cdot \mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) = y_i] + (1 - \rho) \cdot \mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) \neq y_i] - 1 \\ &= \rho \cdot (\mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) = y_i] - \mathbb{P}_{y \sim \{-1, 1\}^n}[f(y) \neq y_i]) \\ &= \rho \cdot \mathbb{E}_x[1_{i, x}] \end{aligned}$$

which completes the proof. ◀

B Social Choice Functions

In this paper, we exclusively focus on social choice functions with two alternatives. There are many ways to interpret these functions. It can be considered as a two-candidate election or as a referendum in the context of political science. It can also be interpreted as a classifier in the context of Machine Learning. In this paper, we will generally give the interpretations in the context of two-candidate elections.

In general, we work with the Boolean functions defined as $f : \{-1, 1\}^n \rightarrow \mathbb{R}$, and we denote the bit i of the input x by x_i for any $i \in [n]$. However, we define welfare only for *social choice functions*, that is the Boolean functions whose ranges are $\{-1, 1\}$. We analyze accuracy only for the following specific social choice functions.

- **Majority:** Suppose that n is an odd number. The majority function of n agents/voters is denoted by Maj_n and defined as

$$f(x) = \text{sign}\left(\sum_{i \in [n]} x_i\right)$$

for any $x \in \{-1, 1\}^n$ where $\text{sign} : \mathbb{R} \rightarrow \{-1, 0, 1\}$ is the function such that

$$\text{sign}(a) = \frac{a}{|a|}$$

for any $a \in \mathbb{R}, a \neq 0$ and $\text{sign}(0) = 0$.

- **Dictatorship:** For a given number n and $i \in [n]$, the dictatorship of voter- i is defined as

$$f(x) = x_i$$

for any $x \in \{-1, 1\}^n$.

- **AND_n:** The AND_n function outputs 1 if there is unanimity on 1, outputs -1 otherwise. Namely,

$$f(x) = \begin{cases} 1 & \text{if } \forall i \in [n], x_i = 1 \\ -1 & \text{otherwise} \end{cases}$$

1:20 Privacy-Welfare-Influence in Referendums

- **OR_n**: The OR_n function outputs 1 if at least one voter votes for 1, and outputs -1 otherwise. In other words, it outputs -1 if there is unanimity on -1 , outputs 1 otherwise. Namely,

$$f(x) = \begin{cases} -1 & \text{if } \forall i \in [n], x_i = -1 \\ 1 & \text{otherwise} \end{cases}$$

Note that, in this paper, we assume *the impartial culture assumption*, that is the voters are not affected by each other and they vote independently uniform at random between two candidates.

Incentivized Collaboration in Active Learning

Lee Cohen¹ ✉

Stanford University, CA, USA

Han Shao ✉

Toyota Technological Institute of Chicago, IL, USA

Abstract

In collaborative active learning, where multiple agents try to learn labels from a common hypothesis, we introduce an innovative framework for incentivized collaboration. Here, rational agents aim to obtain labels for their data sets while keeping label complexity at a minimum. We focus on designing (strict) *individually rational* (IR) collaboration protocols, ensuring that agents cannot reduce their expected label complexity by acting individually. We first show that given any optimal active learning algorithm, the collaboration protocol that runs the algorithm as is over the entire data is already IR. However, computing the optimal algorithm is NP-hard. We therefore provide collaboration protocols that achieve (strict) IR and are comparable with the best known tractable approximation algorithm in terms of label complexity.

2012 ACM Subject Classification Social and professional topics → Computing / technology policy

Keywords and phrases pool-based active learning, individual rationality, incentives, Bayesian, collaboration

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.2

Related Version *Full Version:* <https://arxiv.org/abs/2311.00260>

Funding This work was supported in part by the National Science Foundation under grants 2212968 and 2216899, by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness, by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. The views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

Acknowledgements We would like to thank Avrim Blum for several useful discussions.

1 Introduction

Active learning has emerged as a powerful paradigm in which labels of selected data points are sequentially queried from a large pool of unlabeled data, referred to as the unlabeled pool. The primary objective is to minimize labeling effort to find a classifier that exhibits low error on fresh data points from the same data source, known as generalization error. Typically, if the pool is large enough, a classifier that performs well on the pool can also achieve low generalization error through uniform convergence.

Active learning has also been studied in the distributed setting, where the unlabeled pool is scattered across multiple machines (called agents), (e.g., [45, 2]). While active learning has demonstrated promising results, traditional approaches often operate in isolation, neglecting the potential benefits of collaboration among agents should they agree to collaborate. In this paper, we propose a novel framework for incentivized collaboration active learning, where agents can collaboratively explore their data pools to discover a common target function.

¹ This work was done when the author was at Toyota Technological Institute of Chicago.



The motivation for collaboration in active learning stems from real-life scenarios where collaboration and collective intelligence yield improved outcomes, e.g., when agents collect data from the same distribution, and can easily end up labeling the same or very similar points. This redundancy leads to unnecessary and inefficient utilization of resources, as the labeling is often done by experts. Additionally, more data can be translated to improved accuracy, prompting agents to pool their resources and employ a more powerful model.

The incentive-driven nature of our framework aligns with the reality of collaboration in the real world. When agents are incentivized to collaborate only when their expected labeling complexity decreases, it reflects the real-life scenario where individuals are motivated to engage in cooperative endeavors if they perceive a clear benefit, such as reduced effort, faster, and better outcomes. In this work, we focus on a specific notion of incentives, where agents already have access to a baseline algorithm and they are motivated to join the collaboration if their label complexity is smaller than running the baseline algorithm on their own.

Consider, for example, the case of a new drug (e.g., Paxlovid for Covid-19[40]), that has different efficacy on patients with different features. While individual hospitals can test the drug on their patients in an active learning fashion by executing their preferred baseline algorithm, collaborating efficiently with other hospitals, each with their own patients, often leads to a better prognosis.

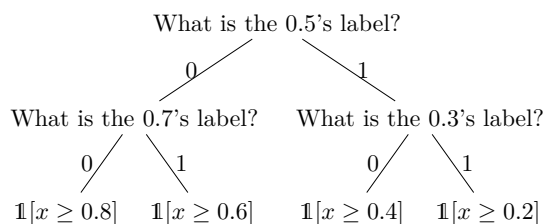
However, if the incentives of the hospitals are not maintained, i.e., the effort of some hospitals is increased, the collaboration may be compromised. By emulating this collaboration within the active learning framework, we unlock the potential of collective intelligence to enhance the learning process. Besides, imagine that several data labeling companies have to recover the labels of unlabeled images assigned to them. Each data labeling company would like to collaborate with other companies to recover the labels of all images while minimizing the query complexity and not increasing their burden.

Our basic model is as follows: there are k agents, each with their own set of unlabeled data points, and a single hypothesis class with a prior on the hypotheses, which all agents are aware of. We assume realizability, meaning that there exists an underlying ground truth labeling function called the target function, labeling all the data points, and that the hypothesis class encompasses such a target function. We refer readers to the discussion for more information about this assumption in the context of active learning.

The agents reach a consensus on an arbitrary baseline algorithm for pool-based active learning (e.g., the best tractable approximation algorithm). To select whether or not to join the collaboration, the agents need to evaluate their utility from joining the collaboration. Since the goal of each individual agent (regardless of the collaboration) is to minimize their expected query complexity, the most natural cost function is the expected query complexity. To ensure each individual benefits from their collaboration, we establish a collaboration protocol that guarantees that each agent cannot reduce their expected label complexity by running the baseline algorithm individually. This concept is referred to as individual rationality (IR). Our objective is to design an IR collaboration protocol that minimizes the overall labeling queries.

There are cases in which collaboration is not necessarily beneficial. For example, each agent has non-zero points on a different axis and the hypothesis class contains every possible halfspace. If the prior distribution is uniform over all labelings, then no agent can reduce their label complexity by joining the collaboration.

Clearly, if each agent has the same set of points, the label complexity of each agent can decrease to $1/k$ of its original label complexity if the collaboration protocol equally splits the labeling burden. Even if agents do not share the same set of points, they can still benefit from collaboration, as we show in the next example.



■ **Figure 1** The query tree of binary search for thresholds.

► **Example 1.** Consider the scenario with 1-dimensional thresholds $\mathcal{H} = \{\mathbb{1}[x \geq \alpha] \mid \alpha = 0.2, 0.4, 0.6, 0.8\}$ and a uniform prior distribution over \mathcal{H} . Suppose agent 1 has points $\{0.25, 0.5, 0.75\}$ and agent 2 has points $\{0.3, 0.45, 0.55, 0.7\}$. When running binary search collaboratively, each agent only performs one labeling query, as illustrated as a search tree in Fig 1. On the other hand, if they were to run binary search independently, each agent would need to query 2 labels. Thus, collaboration can effectively reduce the label complexity for each agent by 1.

Bayesian Assumption. The reason why we have a Bayesian assumption regarding the hypothesis class is that without it, querying all the labels to discover the target hypothesis can be inevitable, even for a simple class of linear separators in \mathcal{R}^2 (see, e.g., Claim 1 in [21]). It is worth noting that as in [21], we *do not require the prior distribution to align with nature*. Instead, the prior distribution serves as a measure for average case analysis. Having a prior belief in our model has the following clear assumption. If the algorithm reaches a point where the remaining consistent hypotheses largely agree on the unlabeled data, it is reasonable to stop and output one of these remaining hypotheses [28]. In a non-Bayesian setting, it does not make sense to operate this way.

Game Theory Interpretation. The agreed-upon baseline algorithm induces a sort of (not private) values for agents- each agent has its (negative) individual labeling complexity as value. The collaboration protocol can be then interpreted as a mechanism: Initially, the collaboration protocol (principal) is introduced to the agents, and each agent can understand it and have confidence in the principal's commitment to implementing it faithfully. Subsequently, the agents either rely on their trust in the algorithm's IR property or have the ability to verify it autonomously. Lastly, the agents behave rationally by joining the collaboration only if it is IR.

We remark that there is an interesting parallelism between our IR collaboration algorithms and truthful mechanisms. It is well known that Vickrey–Clarke–Groves (VCG) mechanism is a truthful mechanism that maximizes social welfare, but since it is hard to compute and to approximate [15], the optimal outcome is replaced by a sub-optimal outcome of an approximation algorithm, and the resulting mechanism is not necessarily truthful. The goal is therefore relaxed to design an efficient approximation algorithm that returns a truthful mechanism.

Contributions and Organization. We formalize the model in Section 2. In Section 3, we demonstrate that any optimal algorithm is individually rational when the baseline is itself. This implies that optimizing for optimality ensures individual rationality for all baseline algorithms.

However, computing (or even approximating) the optimal algorithm is known to be NP-hard. To address this, we then show that the best available tractable approximation algorithm, the greedy algorithm [34, 21], is not individually rational when the baseline is itself. We demonstrate this by presenting an example where joining the collaboration increases the labeling complexity of an agent from $O(1)$ to $\Omega(n)$. In response, we introduce a general approach that can transform any arbitrary baseline algorithm into an IR collaborative algorithm. This conversion ensures that the total label complexity remains competitive with running the baseline algorithm on the entire data set. Furthermore, in Section 4 we present a scheme that converts any IR collaborative algorithm into a strict IR one, guaranteeing the label complexity is strictly lower by joining the collaboration under mild assumptions. When the baseline algorithm is both efficient and approximately optimal, our (strict) IR algorithms efficiently achieve label complexity that is approximately optimal.

1.1 Related Work

The most related work is the recent work of [51], which studies individual rationality in collaborative active learning in a Gaussian Process. While their notion of IR is similar to ours, we focus on query complexity in binary classification. [27] studied incentive compatibility in active learning, where there is a single agent that responds to a learner’s query strategically. Our work is situated at the junction of Learning in the presence of strategic behavior and active learning.

Learning in the presence of strategic behavior

encompasses a vast body of research, including [8, 52, 31]. We are particularly driven by prior research in this area, and how to create learning algorithms that incentivize agents to participate while maximizing the overall welfare. For example, *incentivized exploration* in Multi Arm Bandits [35, 37, 38, 39, 20, 19, 5, 4, 6, 32, 33, 42, 7, 48, 47] or MDPs [46], where the principal recommends actions to the agents (in order to explore different alternatives), but the agents ultimately decide whether to follow the given recommendation. This raises the issue of incentives in addition to the exploration-exploitation trade-off. In particular, [3] study this problem in the context of fairness with a group-based regret notion. They show that regret-optimal bandit algorithms can be unfair and design a nearly optimal fair algorithm. Incentivizing agents to share their data has been studied by [50] in federated bandits.

Federated learning

has gained popularity as a method to foster collaboration among large populations of learning agents among else for incentivizing participation and fairness purposes [11, 36, 24, 23, 26, 25, 49]. Our work also addresses fairness, in the sense that if a collaborative algorithm is individually rational, it is fair for all the participating agents. Another related line of research is *kidney exchange* [41, 1, 9, 10, 14, 22], where the goal is to find a maximum match in a directed graph (representing transplant compatibilities between patient–donor pairs). In this problem, incentives arise in the form of individual rationality when different hospitals have different subsets of patient–donor pairs, and will not join the collaboration if the number of pairs matched by the collaboration is lower than the number of pairs matched they could pair on their own.

Active learning

There are two basic models in active learning—stream-based [28] (where the learner has to determine immediately whether to query the label of the current instance or discard it), and pool-based, which is the basis for our model. Pool-based active learning investigates scenarios in which a learner is confronted with an array of unlabeled data points and the goal is to recover a target function by querying the labels of these points (see [30] for a survey). Active learning has been studied in the context of other societal desiderata such as fairness [44, 16], and safety [17].

To our knowledge, no research has amalgamated these fields to explore strategic constraints in the context of active learning. This is where our work makes a valuable contribution.

2 Preliminaries and Model

Throughout the work, we consider the binary classification problem. Let \mathcal{X} denote the input space, $\mathcal{Y} = \{0, 1\}$ denote the label space, and $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ denote the hypothesis class.² We focus on the *realizable* setting in this work, namely, there exists a target hypothesis $h^* \in \mathcal{H}$ correctly labels every point. In the pool-based active learning setting [30], given a collected unlabeled data set $X = \{x_1, \dots, x_m\}$, the learning goal is to recover the labels of X . Now just suppose the pool of unlabeled data x_1, \dots, x_m is available. The possible labelings of these points form a subset of $\{0, 1\}^m$, called the effective hypothesis class, which is

$$\widehat{H} = \{h(X) \mid h \in \mathcal{H}\},$$

where $h(X) = (h(x_1), \dots, h(x_m))$ is the labeling of X by h . Note that $|\widehat{H}| \leq 2^m$ and $|\widehat{H}| = \mathcal{O}(m^d)$ if the VC dimension of \mathcal{H} is d .

In this work, we focus on the Bayesian setting [21], where the target hypothesis is chosen in advance from some prior distribution π over \widehat{H} . Namely, without any additional information, for any labeling $h \in \widehat{H}$, the probability that h is the correct labeling of X is $\pi(h)$. Since we can eliminate any hypothesis h with $\pi(h) = 0$ before starting to query for labels, we assume w.l.o.g. that $\pi(h) > 0$ for all effective hypotheses in \widehat{H} .

We remark that assuming that the unlabeled data X is collected from some distribution D_x , which is essentially a distribution D projected onto its input space, and that this distribution D_x can be accurately classified by a hypothesis in H with VC dimension d , standard generalization guarantees apply when the prior π over H is uniform (see [21] for more details).

Standard active learning model. In the standard pool-based active learning setting, a single agent owns the pool of unlabeled data X . The agent, who knows both \widehat{H} and π , can query the labels of points in X , and her goal is to recover the labeling of X (or to find the target hypothesis) by querying as few points as possible.

A *standard query algorithm* receives as input the prior distribution π and unlabeled data set, X . In each iteration $t = 1, 2, \dots$, given the history up to time t ,

$$\mathcal{F}_t = ((x_1, y_1), \dots, (x_{t-1}, y_{t-1})) \in (X \times \{0, 1\})^{t-1},$$

² Results in this work can be directly extended to any active learning problem that can be formalized using a hypothesis class, e.g., multiclass classification.

it selects a point x_t to query and observes its label, y_t . The algorithm stops when all the labels of X are recovered. Alternatively, the algorithm stops when for every two hypotheses $h_1, h_2 \in \widehat{H}$ consistent with \mathcal{F}_t (meaning that $h_1(x_\tau) = h_2(x_\tau) = y_\tau$ for all $\tau = 1, \dots, t-1$), $h_1(X) = h_2(X)$.

Collaborative active learning model. In the collaborative setting, we assume there is more than one agent. Formally, there are k agents and each agent i has an individual unlabeled data set X_i such that they together compose the pool, i.e., $\cup_{i \in [k]} X_i = X$, and each can query points from their own set X_i (but cannot query points which are not in their set). The goal of each agent is to recover the true labeling of their own set while performing as few queries as possible. The collaboration protocol, also called principal, who knows $\{X_1, \dots, X_k\}$, \widehat{H} and π , decides which point should be queried at each iteration, and her goal is to recover all the labels of X using as few queries as possible. We remark that since data points belong to agents, queries of any point $x \in X$ can only be performed by agents whose data set contains x .

The query algorithm in the collaborative setting is similar to that in the standard setting, except that the algorithm needs to coordinate among the agents and decide which agent will query each point as some data points might belong to more than one agent. In this setting, agents can decide to join the collaboration or learn individually at the beginning of the learning. But if they join the collaboration, they commit to follow the instructions of the query algorithm. Therefore, given a prior distribution π over \widehat{H} and a set of agents who would join the collaboration, w.l.o.g. denoted as $\{X_1, \dots, X_\kappa\}$ for some $\kappa \in [k]$, at time $t = 1, 2, \dots$, a *collaborative query algorithm* asks agent $i_t \in [\kappa]$ such that $x_t \in X_{i_t}$ to query point x_t , and observes its label, y_t ; the algorithm stops when the labels of points in $\cup_{i \in [\kappa]} X_i$ are completely recovered.

It is straightforward to check that standard query algorithms are a special case of collaborative query algorithms when there is a single agent, i.e., $k = 1$. Additionally, a standard algorithm can also be run over multiple agents by considering the union of their data $\cup_{i \in [\kappa]} X_i$ as a single agent. Hence, we omit “standard” or “collaborative” in a query algorithm when it is clear from the context how many agents are involved.

For any collaborative algorithm \mathcal{A} , given an input π and any collection of unlabeled data sets $X_1, \dots, X_\kappa \subseteq X$ of size $\kappa \geq 1$, we denote by $Q(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\}, h)$ the label complexity (number of label queries) of $\mathcal{A}(\pi, \{X_1, \dots, X_\kappa\})$ when the target hypothesis is h . For randomized algorithms, the label complexity is taken expectation over the randomness of the algorithm. We define the label complexity as follows.

► **Definition 2 (Label complexity).** *Given any fixed unlabeled pool and effective hypothesis class (X, \widehat{H}) , for any algorithm \mathcal{A} , prior distribution π over \widehat{H} and any collection of unlabeled data sets $X_1, \dots, X_\kappa \subseteq X$ of size $\kappa \geq 1$, the label complexity of \mathcal{A} with $(\pi, \{X_1, \dots, X_\kappa\})$ as input, denoted by $Q(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\})$, is the expected number of label queries when h is drawn from the prior π , i.e.,*

$$Q(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\}) = \mathbb{E}_{h \sim \pi} [Q(\mathcal{A}, h, \{X_1, \dots, X_\kappa\})] .$$

For each agent $i \in [\kappa]$ in the collaboration, we let $Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\})$ denote the expected number of queries performed by agent i .

For any $(\pi, \{X_1, \dots, X_\kappa\})$, let $Q^*(\pi, \{X_1, \dots, X_\kappa\}) = \min_{\mathcal{A}} Q(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\})$ denote the optimal query complexity. An algorithm \mathcal{A} is said to be *optimal* if $Q(\mathcal{A}, \pi, \{X_1, \dots, X_\kappa\}) = Q^*(\pi, \{X_1, \dots, X_\kappa\})$ for any prior distribution π and X_1, \dots, X_κ .

Rational agents. We assume that agents have access to a baseline algorithm and are able to run it on their own local data. Agents can decide to join the collaboration or run the baseline individually at the beginning of the learning. If they join the collaboration, they commit to follow the instructions of the query algorithm. Each agent is incentivized to join the collaboration if she could perform fewer label queries (assuming that all others join the collaboration) by pulling out and running the baseline \mathcal{A} individually. Formally,

► **Definition 3** (Individual rationality). *In a collaborative learning problem with prior distribution π and k agents $\{X_1, \dots, X_k\}$, given a baseline algorithm \mathcal{A} , a collaborative algorithm \mathcal{A}' is individually rational (IR) if*

$$Q_i(\mathcal{A}', \pi, \{X_1, \dots, X_k\}) \leq Q(\mathcal{A}, \pi, \{X_i\}), \forall i \in [k]. \quad (1)$$

We say \mathcal{A}' is *strictly individually rational* (henceforth, SIR) if

$$Q_i(\mathcal{A}', \pi, \{X_1, \dots, X_k\}) < Q(\mathcal{A}, \pi, \{X_i\}), \forall i \in [k].$$

We remark that in addition to their own sets, each agent knows all the unlabeled data sets, $\{X_1, \dots, X_k\}$, and the prior distribution π , otherwise, they will not be able to compute $Q_i(\mathcal{A}', \pi, \{X_1, \dots, X_k\})$. The principal also has access to $\{X_1, \dots, X_k\}$ and π , and can therefore make sure these constraints are satisfied.

It is worth noting that our model can also accommodate different individual baseline algorithms. We briefly discuss this in Section 5.

An alternative interpretation of the problem in a game theoretic framework is as follows: each agent has a strategy space of two strategies, joining the collaboration and not. The utility of an agent that performs Q queries is $-Q$. If the algorithm \mathcal{A} is IR, then the case of all agents joining the collaboration is a Nash equilibrium (since switching to not joining will not increase their utility). If \mathcal{A} is SIR, then all agents joining the collaboration is a strict Nash equilibrium.

3 Construction of IR Collaborative Algorithms

When agents are limited to a poor baseline algorithm, e.g., randomly selecting points to query, the principal can simply incentivize agents to collaborate by using a superior algorithm that requires fewer labeling efforts. We therefore start by considering optimal baseline algorithms in Section 3.1. If we are able to find an IR collaborative algorithm for an optimal baseline algorithm, OPT, then it must be IR w.r.t. all baseline algorithms. We demonstrate that, surprisingly, the optimal algorithm OPT is IR given that the baseline algorithm is OPT itself. Since computing an optimal algorithm is known to be NP-hard, we continue by considering the best-known approximation algorithm, the greedy algorithm. In Section 3.2, we show that given the greedy algorithm as baseline, the collaboration protocol that runs the greedy algorithm is not IR. Then in Section 3.3, we provide a general scheme that transforms any baseline algorithm into an IR algorithm while maintaining a comparable label complexity.

3.1 Optimality Implies Universal Individual Rationality

Incorporating individual rationality as an additional constraint to optimality usually requires additional effort in certain settings, e.g., in online learning by [13]. However, in our specific setting, optimality does not contradict the individual rationality property. That is, an optimal algorithm will not increase any agent's label complexity to benefit other agents. In fact, optimizing for optimality implies achieving individual rationality for all baseline algorithms.

► **Theorem 4.** *For any optimal collaborative algorithm OPT , we have*

$$Q_i(OPT, \pi, \{X_1, \dots, X_k\}) \leq Q(OPT, \pi, \{X_i\}) = Q^*(\pi, \{X_i\}), \forall i \in [k].$$

Therefore, OPT is IR w.r.t. any baseline algorithm.

We prove the theorem by contradiction. If OPT is not IR for the baseline being OPT , then there exists an agent i such that $Q(OPT, \pi, \{X_i\}) < Q_i(OPT, \pi, \{X_1, \dots, X_k\})$. In this case, we can construct a new algorithm by first running OPT over $\{X_i\}$ (to recover the labels of X_i) and then running $OPT(\pi, \{X_1, \dots, X_k\})$ and replacing agent i 's queries with the recovered labels of X_i . This new algorithm incurs a strictly smaller label complexity than OPT , which is a contradiction to the optimality of OPT . The formal proof is deferred to Appendix A. Unfortunately, computing an optimal query algorithm is not just NP-hard, but also hard to approximate within a factor of $\Omega(\log(|\hat{H}|))$ [29, 18]. One of the most popular heuristics to find an approximated solution is greedy.

3.2 The Greedy Algorithm is Not Individually Rational

For standard Bayesian active learning, [34, 21] presented a simple greedy algorithm called generalized binary search (GBS), which chooses a point leading to the most balanced partition of the set of hypotheses consistent with the history. More specifically, at time step t , given the history $\mathcal{F}_t = ((x_1, i_1, y_1), \dots, (x_{t-1}, i_{t-1}, y_{t-1}))$, let $VS(\mathcal{F}_t) = \{h \in \hat{H} | h(x_\tau) = y_\tau, \forall \tau \in [t-1]\}$ denote the set of hypotheses consistent with the history \mathcal{F}_t (often called the version space associated with \mathcal{F}_t). Given \mathcal{F}_t and $(\pi, \{X_1, \dots, X_k\})$ as input, GBS will query

$$x_t = \arg \max_{x \in \cup_{i \in [k]} X_i} \min(\pi(\{h \in VS(\mathcal{F}_t) | h(x) = 1\}), \pi(\{h \in VS(\mathcal{F}_t) | h(x) = 0\}))$$

at time t . When referring to GBS as a collaborative algorithm, we complement it with an arbitrary tie-breaking rule for selecting i_t , as GBS itself does not specify how to choose which agent to query. GBS is guaranteed to achieve competitive label complexity with the optimal label complexity.

► **Lemma 5** (Optimality of GBS, Theorem 3 of [21]). *For any prior distribution π over \hat{H} and k agents $\{X_1, \dots, X_k\}$, the label complexity of GBS satisfies that*

$$Q(GBS, \pi, \{X_1, \dots, X_k\}) \leq 4Q^*(\pi, \{X_1, \dots, X_k\}) \ln\left(\frac{1}{\min_{h \in \hat{H}} \pi(h)}\right).$$

The greedy algorithm GBS not only achieves approximately optimal label complexity, but it is also computationally efficient, with a running time of $\mathcal{O}(m^2 |\hat{H}|)$. As GBS is the best-known efficient approximation algorithm, it is natural to think that agents would adopt GBS as a baseline.

As we have shown that the optimal algorithm is IR w.r.t. itself, the next natural question is: *Is GBS (as collaboration protocol) individually rational w.r.t. GBS itself?*

We answer this question negatively, even in the case of two agents. Even worse, we present an example in which an agent's label complexity is $\Omega(n)$ when participating in the collaboration, but only $\mathcal{O}(1)$ when not participating.

► **Theorem 6.** *For the algorithm of GBS, there exists an instance of (X_1, X_2, π) with $|X_1| = n$, in which agent 1 incurs a label complexity of $Q_1(GBS, \pi, \{X_1, X_2\}) = \Omega(n)$ when participating the collaboration and can achieve $Q(GBS, \pi, \{X_1\}) = \mathcal{O}(1)$ when not participating.*

Intuitively, at each time step, GBS only searches for an x_t which leads to the most balanced partition of the version space, which does not necessarily lead to the optimal point to query. Given additional label information from the other agent, GBS possibly choose a worse point to query. In addition, the label complexity of GBS is upper bounded by the optimal label complexity multiplied by a logarithmic factor. It is possible that the agent achieves a smaller multiplicative factor by running GBS individually and a larger factor in the collaboration. To prove the theorem, we construct an instance in which there exists a hypothesis with a prior probability of $1/4$, such that if GBS runs on $\{X_1, X_2\}$ and this hypothesis is the target, GBS will query almost all the points in X_1 (in a particular order) before returning this hypothesis. We show this part by induction. Additionally, we compute the query tree by running GBS solely on X_1 and use it to show that in this case, GBS has an expected query complexity of $O(1)$. The full construction of the instance and the proof of Theorem 6 is deferred to Appendix B.

3.3 A Scheme of Converting Algorithms to IR Algorithms

Given that the greedy algorithm has been proven to be not individually rational w.r.t. itself, we raise the following question: *Is it possible to develop a general scheme that can generate an IR algorithm given any baseline algorithm?* In this section, we propose such a scheme that addresses this question. Moreover, given a baseline algorithm \mathcal{A} , the resulting IR algorithm can achieve a label complexity comparable to implementing the baseline algorithm over all agents, i.e., $\mathcal{A}(\pi, \{X_1, \dots, X_k\})$. It is important to note that we aim for the label complexity to be comparable to $Q(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$ rather than $\sum_{i \in [k]} Q(\mathcal{A}, \pi, X_i)$, as the latter holds true by individual rationality. Given an efficient approximately optimal algorithm as baseline (e.g., GBS), our scheme can provide an algorithm that simultaneously exhibits individual rationality, efficiency, and approximately optimal label complexity.

For any baseline algorithm \mathcal{A} , we define a new algorithm $B2IR(\mathcal{A})$, which runs \mathcal{A} as a subroutine. Basically, we first calculate the label complexity of agent i both when she is in collaboration with all the other agents, i.e., $Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$, and when she is not in collaboration, i.e., $Q(\mathcal{A}, \pi, \{X_i\})$, for all $i \in [k]$. By doing so, we can distinguish which agents can benefit from collaboration when running \mathcal{A} and which cannot. We denote the set of agents who cannot benefit from collaboration with all others when running \mathcal{A} as $S = \{i | Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}) > Q(\mathcal{A}, \pi, \{X_i\})\}$. For those who do not benefit from the collaboration, we just run \mathcal{A} on their own data. For those who benefit from collaborating with the others together, we run \mathcal{A} over all agents $[k]$ — Only whenever $\mathcal{A}(\pi, \{X_1, \dots, X_k\})$ asks to query the label of a point belonging to some $i \in S$, since we already recovered the labels of X_i , we just feed $\mathcal{A}(\pi, \{X_1, \dots, X_k\})$ with this label without actually asking agent i to query. The detailed algorithm is described in Algorithm 1.

► **Theorem 7.** *For any baseline algorithm \mathcal{A} , the algorithm $B2IR(\mathcal{A})$ satisfies the following properties:*

- **IR property:** $B2IR(\mathcal{A})$ is individually rational w.r.t. the baseline algorithm \mathcal{A} .
- **Efficiency:** $B2IR(\mathcal{A})$ runs in $\mathcal{O}(kT_{\mathcal{A},Q} + mT_{\mathcal{A},0})$ time, where $T_{\mathcal{A},0}$ is the time of computing (i_t, x_t) at each time t for \mathcal{A} and $T_{\mathcal{A},Q}$ is the maximum time of computing $Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$ for an agent i , unlabeled data $\{X_1, \dots, X_k\}$, and algorithm \mathcal{A} .
- **Label complexity:** $Q(B2IR(\mathcal{A}), \pi, \{X_1, \dots, X_k\}) \leq Q(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$.

Algorithm 1 B2IR.

- 1: **input:** A query algorithm \mathcal{A} , set $\{X_1, \dots, X_k\}$ and prior π over \widehat{H}
 - 2: For each $i \in [k]$, calculate $Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$ and $Q(\mathcal{A}, \pi, \{X_i\})$.
 - 3: Let $S \leftarrow \{i | Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}) > Q(\mathcal{A}, \pi, \{X_i\})\}$ and $X_S \leftarrow \cup_{i \in S} X_i$ // the agents who do not benefit from collaboration
 - 4: **for** each $i \in S$ **do** $Y_i \leftarrow$ Run \mathcal{A} over $\{X_i\}$ // recover the labels for agent i
 - 5: **for** $t = 1, \dots$ **do**
 - 6: $(i_t, x_t) \leftarrow$ the querying agent and the query point from $\mathcal{A}(\pi, \{X_1, \dots, X_k\})$
 - 7: **if** $i_t \in S$ **then** Feed the label of x_t from Y_{i_t} // we already recovered the labels of X_S
 - 8: **else** Ask agent i_t to query the label of x_t
 - 9: **end for**
-

The proof follows the algorithm description immediately. Note that when the baseline is GBS, we have $T_{\text{GBS},0} = \mathcal{O}(m)$. We can compute $Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\})$ by simulating over all effective hypotheses $h \in \widehat{H}$. For each h , we will query at most m rounds. Therefore, we have $T_{\text{GBS},Q} = \mathcal{O}(m^2 |\widehat{H}|)$ and we can run B2IR(GBS) in $\mathcal{O}(km^2 |\widehat{H}|)$ time. Using GBS as the baseline, we derive the following corollary.

► **Corollary 8.** *Given GBS as the baseline, B2IR(GBS) is IR; runs in $\mathcal{O}(km^2 |\widehat{H}|)$ time; and satisfies that $Q(\text{B2IR}(\text{GBS}), \pi, \{X_1, \dots, X_k\}) \leq 4Q^*(\pi, \{X_1, \dots, X_k\}) \ln(\frac{1}{\min_{h \in \widehat{H}} \pi(h)})$.*

4 Converting Algorithms to SIR Algorithms

In Section 3, we provided a generic scheme for constructing an IR algorithm given any baseline algorithm. In this section, we focus on constructing SIR algorithms given IR algorithms. Since strict individual rationality requires that agents strictly benefit from collaboration, this is impossible without further assumptions. For example, consider a set of agents who only have one single independent point in their own sets and a prior distribution that is uniform over all labelings. In this case, each agent, regardless of whether she collaborates or not, has a label complexity of 1 and cannot *strictly* benefit from collaboration as the other agents cannot obtain information about her data.

Now, let us consider a notion weaker than SIR, called *i*-partially SIR, in which only agent i strictly benefits from the collaboration, and any other agent $j \neq i$ does not get worse by joining the collaboration. More formally,

► **Definition 9** (Partially SIR algorithms). *For any baseline algorithm \mathcal{A} , for all $i \in [k]$, an algorithm \mathcal{O}_i is *i*-partially SIR, if \mathcal{O}_i satisfies that*

$$Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}) < Q(\mathcal{A}, \pi, \{X_i\}),$$

and

$$Q_j(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}) \leq Q(\mathcal{A}, \pi, \{X_j\}), \forall j \in [k] \setminus \{i\}.$$

If we are given an *i*-partially SIR algorithm \mathcal{O}_i for each i , then we can construct a SIR algorithm by running a mixture of an IR algorithm \mathcal{A}' (e.g., B2IR(\mathcal{A}) in Algorithm 1) and $\{\mathcal{O}_i | i \in [k]\}$ with the label complexity a little (arbitrarily small) higher than that of \mathcal{A}' .

► **Lemma 10.** *For any baseline algorithm \mathcal{A} , given an IR algorithm \mathcal{A}' and partially SIR algorithms $\{\mathcal{O}_i | i \in [k]\}$, for any $\varepsilon > 0$, let $\mathcal{A}''_\varepsilon$ be the algorithm of running \mathcal{A}' with probability $(1 - \frac{\varepsilon}{n})$ and running \mathcal{O}_i with probability $\frac{\varepsilon}{kn}$. Then $\mathcal{A}''_\varepsilon$ satisfies the following properties.*

- *SIR property:* \mathcal{A}'_ε is SIR with respect to the baseline algorithm \mathcal{A} .
- *Label complexity:* $Q(\mathcal{A}'_\varepsilon, \pi, \{X_1, \dots, X_k\}) \leq Q(\mathcal{A}, \pi, \{X_1, \dots, X_k\}) + \varepsilon$.

The proof is straightforward from the definition, and we include it in Appendix C for completeness. Since a SIR algorithm is also i -partially SIR for all $i \in [k]$, constructing a SIR algorithm is equivalent to constructing a set of partially SIR algorithms $\{\mathcal{O}_i | i \in [k]\}$. Therefore, the problem of constructing a SIR algorithm is reduced to constructing partially SIR algorithms $\{\mathcal{O}_i | i \in [k]\}$.

For the remainder of this section, we will present the SIR results for an optimal baseline algorithm in Section 4.1, where we propose a sufficient and necessary assumption for the existence of SIR algorithms and then provide a SIR algorithm. This algorithm is SIR w.r.t. any baseline algorithm but again, computationally inefficient. In Section 4.2, we provide a general scheme that transforms any baseline algorithm into a SIR algorithm.

4.1 A Universal SIR Algorithm for Any Baseline Algorithm

Constructing a universal SIR algorithm w.r.t. any baseline is equivalent to constructing a SIR algorithm for an optimal baseline. For the existence of SIR algorithms given any optimal baseline algorithm, we propose the following assumption, which is sufficient and necessary. We include the proof for the necessity of this assumption in Appendix D. The sufficiency of this assumption will be verified immediately after we construct a SIR algorithm.

► **Assumption 1.** We assume that for any $i \in [k]$, the optimal label complexity of agent i given the information regarding the labels of all other agents is strictly smaller than that without this additional information, i.e., $Q^*(\pi, \{X_i\}) - \mathbb{E}_{h \sim \pi} [Q^*(\pi_{h,-i}, \{X_i\})] > 0$, where $\pi_{h,-i}$ is the posterior distribution of π after observing $\{(x, h(x)) | x \in \cup_{j \neq i} X_j\}$.

According to Lemma 10, we can construct a SIR algorithm by constructing a set of partially SIR algorithms $\{\mathcal{O}_i | i \in [k]\}$.

Let \mathcal{O}_i be the algorithm of running an optimal algorithm OPT over $(\pi, \{X_j | j \neq i\})$ first, then given the query-label history of $\{(x, h(x)) | x \in \cup_{j \neq i} X_j\}$ for some $h \in \widehat{H}$, run OPT over $(\pi_{h,-i}, \{X_i\})$. Then it immediately follows that \mathcal{O}_i is i -partially SIR from Assumption 1. Let OPT'_ε denote the algorithm of running OPT with probability $(1 - \frac{\varepsilon}{n})$ and running \mathcal{O}_i with probability $\frac{\varepsilon}{kn}$ for all $i \in [k]$. By Lemma 10, we have

► **Corollary 11.** Under Assumption 1, for any $\varepsilon > 0$, OPT'_ε is SIR w.r.t. OPT and satisfies

$$Q(\text{OPT}'_\varepsilon, \pi, \{X_1, \dots, X_k\}) \leq Q^*(\pi, \{X_1, \dots, X_k\}) + \varepsilon.$$

In addition, OPT'_ε is SIR w.r.t. any baseline algorithm \mathcal{A} as

$$Q_i(\text{OPT}'_\varepsilon, \pi, \{X_1, \dots, X_k\}) < Q^*(\pi, \{X_i\}) \leq Q(\mathcal{A}, \pi, \{X_i\}).$$

4.2 A Scheme of Converting Algorithms to SIR Algorithms

As mentioned before, computing an optimal algorithm is NP-hard. Assumption 1 assumes that collaboration can strictly benefit agents when the collaboration protocol can compute the optimal algorithm given $\pi_{h,-i}$. Hence, the assumption does not take the computational issue into consideration and thus might not be enough for the existence of an efficient SIR algorithm w.r.t. an efficient approximation algorithm like GBS.

Instead, we propose prior-independent assumption that is sufficient for the existence of efficient SIR algorithms when we are given an efficient baseline and an efficient IR algorithm w.r.t. the baseline. Basically, we assume that, there exists an effective hypothesis $h \in \widehat{H}$,

given the information that all other agents are labeled by h , the number of labelings of X_i consistent with the label information is strictly smaller than the total number of labelings of X_i by \widehat{H} . Formally, for any $i \in [k]$, let $X_{-i} = \cup_{j \neq i} X_j$ denote the union of all agents' data except agent i . Let $H(X_i) = \{h'(X_i) | h' \in \widehat{H}\}$ denote the effective hypothesis class of X_i , i.e., all labelings of X_i . For any $h \in \widehat{H}$, let $H(X_i|h) = \{h'(X_i) | h'(X_{-i}) = h(X_{-i}), h' \in \widehat{H}\}$ denote the subset which are consistent with all other agents being labeled by h .

► **Assumption 2.** For all $i \in [k]$, there exists an $h \in \widehat{H}$ s.t. the number of labelings of X_i consistent with $(X_{-i}, h(X_{-i}))$ is strictly smaller than the number of labelings by \widehat{H} , i.e., $|H(X_i|h)| < |H(X_i)|$.

Intuitively, Assumption 2 means that for every agent i , there exists an hypothesis h such that when $h^* = h$, the cardinality of the set of hypotheses consistent with $(X_{-i}, h(X_{-i}))$ is strictly smaller than $|\widehat{H}|$. We will show that this assumption is sufficient for the *existence* of algorithms satisfying SIR property. Without it, it is unclear if there exist SIR algorithms. The assumption can be easily verified by iterating each $h \in \widehat{H}$ (this is polynomial in $|\widehat{H}|$ and m).

Notice that each deterministic query algorithm \mathcal{A} can be represented as a binary tree, $\mathcal{T}_{\mathcal{A}}$ whose internal nodes at level t are queries (“what is the x_t 's label?”), and whose leaves are labelings as illustrated in Figure 1. Under Assumption 2, we can prune the query tree of $\mathcal{A}(\pi, \{X_i\})$ by removing all subtrees whose leaves are all in $H(X_i) \setminus H(X_i|h)$. We do not need to construct this pruned tree when we implement the algorithm. At time t , we just need to generate an x_t from $\mathcal{A}(\pi, \{X_i\})$, then check if this node should be pruned by checking if all the hypotheses $H(X_i|h)$ agree on the label of x_t . If this is true, it means that we have already recovered the label of x_t and thus we just need to feed the label to the algorithm without actually querying x_t again. Then we can construct a i -partially SIR algorithm \mathcal{O}_i by running $B2IR(\mathcal{A})$ over $(\pi, \{X_j | j \neq i\})$ to recover the labeling of X_{-i} first, then running pruned version of $\mathcal{A}(\pi, \{X_i\})$. Note that the implementation also works when \mathcal{A} is randomized.

Consider Example 1, where the hypothesis class $\mathcal{H} = \{x \geq \alpha | \alpha = 0.2, 0.4, 0.5, 0.6, 0.8\}$ with a uniform prior, agent 1 with points $X_1 = \{0.25, 0.5, 0.75\}$ and agent 2 with points $X_2 = \{0.3, 0.45, 0.55, 0.7\}$. When agent 1 runs a binary search, the query tree has 0.5 as a root, then if 0.25 if $h^*(0.5) = 1$, and 0.75 otherwise. Now, algorithm \mathcal{O}_1 runs a binary search on X_2 and obtains all the labels of the points in X_2 . The hypothesis $h = 1(x \geq 0.5)$ holds $|H(X_1|h)| = 1 < |\widehat{H}(X_1)| = 4$. When h is the labeling function, 0.3 and 0.45 are labeled as negative, and 0.55 and 0.7 are labeled as positive. Then, \mathcal{O}_1 will only need agent 1 to query 0.5 as the labels of 0.25 and 0.75 can be inferred and they are pruned in the query tree.

► **Lemma 12.** *Under Assumption 2, the algorithm \mathcal{O}_i constructed above is i -partially SIR and runs in time $\mathcal{O}(\mathcal{T}_{B2IR(\mathcal{A})} + m(|\widehat{H}| + T_{\mathcal{A},0}))$ time, where $\mathcal{T}_{B2IR(\mathcal{A})}$ is the running time of $B2IR(\mathcal{A})$ and $T_{\mathcal{A},0}$ is the time of computing (i_t, x_t) at each time t for \mathcal{A} .*

The proof of Lemma 12 is deferred to Appendix E.

We can then construct an algorithm $\mathcal{A}'_{\varepsilon}$ by running $B2IR(\mathcal{A})$ with probability with probability $(1 - \frac{\varepsilon}{n})$ and running \mathcal{O}_i constructed in the above way with probability $\frac{\varepsilon}{kn}$ for all $i \in [k]$. By combining Lemmas 10 and 12, we derive the following theorem. Then, combining it with Corollary 8, we derive a SIR algorithm for GBS as baseline GBS.

► **Theorem 13.** *Under Assumption 2, for any baseline algorithm \mathcal{A} , for any $\varepsilon > 0$, $\mathcal{A}'_{\varepsilon}$ is SIR and satisfies*

$$Q(\mathcal{A}'_{\varepsilon}, \pi, \{X_1, \dots, X_k\}) \leq Q(B2IR(\mathcal{A}), \pi, \{X_1, \dots, X_k\}) + \varepsilon.$$

In addition, Algorithm $\mathcal{A}'_{\varepsilon}$ runs in $\mathcal{O}(\mathcal{T}_{B2IR(\mathcal{A})} + m(|\widehat{H}| + T_{\mathcal{A},0}))$ time.

► **Corollary 14.** *Given GBS as the baseline, Algorithm GBS'_ε is SIR; runs in $\mathcal{O}(km^2|\widehat{H}|)$ time; and satisfies that*

$$Q(GBS'_\varepsilon, \pi, \{X_1, \dots, X_k\}) \leq 4Q^*(\pi, \{X_1, \dots, X_k\}) \ln \left(\frac{1}{\min_{h \in \mathcal{H}} \pi(h)} \right) + \varepsilon.$$

5 Discussion

In this paper, we have initiated the study of collaboration in active learning in the presence of incentivized agents. We first show that an optimal collaborative algorithm is IR w.r.t. any baseline algorithm while approximate algorithms are not. Then we provide meta-algorithms capable of producing IR/SIR algorithms given any baseline algorithm as input.

Our model and algorithms can also allow different agents to have different baseline algorithms – Whenever the principal plans to run an algorithm on a union of datasets, she can simply check which baseline algorithm has the lowest expected query complexity on this union, and run it. When she needs to run an algorithm on a dataset of an individual, she can simply run their baseline. This way the (S)IR is preserved.

There are a few problems we leave open. First, relaxing the assumption that each agent i has full knowledge of X_{-i} (e.g., due to privacy concerns). Second, relaxing the assumption that agents provide reliable labels. Third, deriving results in non-realizable settings. Realizability is a standard assumption in learning theory at large and particularly within active learning as highlighted in the classical machine learning theory textbook by [43] and the active learning theory survey by [30]. Moreover, realizability has also been adopted in the collaborative learning setting (e.g., [12]). The reason is that without realizability, additional complications might arise in collaboration such as why collaboration would yield benefits. In general, we believe that additional assumptions would be required to relax this assumption. Forth, towards a more game theory orientation, it would be interesting to design collaborative algorithms in a setting where agents can form coalitions. Finally, finding a necessary and sufficient assumption(s) for the existence of efficient SIR algorithms will be an interesting direction (and we have found a sufficient one in this work).

References

- 1 Itai Ashlagi and Alvin E. Roth. Individual rationality and participation in large scale, multi-hospital kidney exchange. In *Proceedings 12th ACM Conference on Electronic Commerce (EC-2011)*, 2011.
- 2 Nicolas Aussel, Sophie Chabridon, and Yohan Petetin. Combining federated and active learning for communication-efficient distributed failure prediction in aeronautics. *arXiv preprint arXiv:2001.07504*, 2020.
- 3 Jackie Baek and Vivek F. Farias. Fair exploration via axiomatic bargaining. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22034–22045, 2021.
- 4 Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown, and Moshe Tennenholtz. Fiduciary bandits. In *International Conference on Machine Learning*, 2019.
- 5 Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Economic recommendation systems: One page abstract. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC*, page 757, 2016.
- 6 Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. Social learning and the innkeeper’s challenge. In *Proceedings of the 2019 ACM Conference on Economics and Computation*,

- EC '19, page 153–170, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3328526.3329569.
- 7 Kiarash Banihashem, MohammadTaghi Hajiaghayi, Suho Shin, and Aleksandrs Slivkins. Bandit social learning: Exploration under myopic behavior, 2023. [arXiv:2302.07425](https://arxiv.org/abs/2302.07425).
 - 8 Omer Ben-Porat and Rotem Torkan. Learning with exposure constraints in recommendation systems. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2023.
 - 9 Avrim Blum, Ioannis Caragiannis, Nika Haghtalab, Ariel D. Procaccia, Eviatar B. Procaccia, and Rohit Vaish. Opting into optimal matchings. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, page 2351–2363, USA, 2017. Society for Industrial and Applied Mathematics.
 - 10 Avrim Blum and Paul Gölz. Incentive-compatible kidney exchange in a slightly semi-random model. In *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*. ACM, 2021.
 - 11 Avrim Blum, Nika Haghtalab, Richard Lanus Phillips, and Han Shao. One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, 2021*.
 - 12 Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/186a157b2992e7daed3677ce8e9fe40f-Paper.pdf.
 - 13 Avrim Blum and Thodoris Lykouris. Advancing Subgroup Fairness via Sleeping Experts. In Thomas Vidick, editor, *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 55:1–55:24, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.ITCS.2020.55.
 - 14 Avrim Blum and Yishay Mansour. Kidney exchange and endless paths: On the optimal use of an altruistic donor. *CoRR*, abs/2010.01645, 2020. [arXiv:2010.01645](https://arxiv.org/abs/2010.01645).
 - 15 David Buchfuhrer, Shaddin Dughmi, Hu Fu, Robert Kleinberg, Elchanan Mossel, Christos H. Papadimitriou, Michael Schapira, Yaron Singer, and Christopher Umans. Inapproximability for vcg-based combinatorial auctions. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*. SIAM, 2010.
 - 16 Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, and Kevin Jamieson. Fair active learning in low-data regimes. *CoRR*, abs/2312.08559, 2023. doi:10.48550/ARXIV.2312.08559.
 - 17 Romain Camilleri, Andrew Wagenmaker, Jamie H. Morgenstern, Lalit Jain, and Kevin G. Jamieson. Active learning with safety constraints. In *NeurIPS*, 2022.
 - 18 Venkatesan T Chakaravathy, Vinayaka Pandit, Sambuddha Roy, Pranjal Awasthi, and Mukesh Mohania. Decision trees for entity identification: Approximation algorithms and hardness results. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 53–62, 2007.
 - 19 Yeon-Koo Che and Johannes Hörner. Optimal design for social learning, 2013.
 - 20 Lee Cohen and Yishay Mansour. Optimal algorithm for bayesian incentive-compatible exploration. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 135–151, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3328526.3329581.
 - 21 Sanjoy Dasgupta. Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 17, 2004.
 - 22 John P. Dickerson, Ariel D. Procaccia, and Tuomas Sandholm. Failure-aware kidney exchange. *Manag. Sci.*, 2019.

- 23 Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, page 1639–1656, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533221.
- 24 Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.
- 25 Kate Donahue and Jon Kleinberg. Fairness in model-sharing games. In *Proceedings of the ACM Web Conference 2023, WWW '23*, page 3775–3783, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3543507.3583483.
- 26 Kate Donahue and Jon M. Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021.
- 27 Federico Echenique and Siddharth Prasad. Incentive compatible active learning. *arXiv preprint arXiv:1911.05171*, 2019.
- 28 Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 1997.
- 29 Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. *Advances in Neural Information Processing Systems*, 23, 2010.
- 30 Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7, 2014.
- 31 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, 2016.
- 32 Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration with heterogeneous agents. In *The World Wide Web Conference, WWW '19*, page 751–761, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3308558.3313649.
- 33 Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation, EC '20*, page 647–648, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3391403.3399487.
- 34 S Rao Kosaraju, Teresa M Przytycka, and Ryan Borgstrom. On an optimal split tree problem. In *Algorithms and Data Structures: 6th International Workshop, WADS'99 Vancouver, Canada, August 11-14, 1999 Proceedings*, pages 157–168. Springer, 2002.
- 35 Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *J. of Political Economy*, 122:988–1012, October 2014. Preliminary version appeared in *ACM Conf. on Economics and Computation*, 2014.
- 36 Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. *Collaborative Fairness in Federated Learning*, pages 189–204. Springer International Publishing, Cham, 2020. doi:10.1007/978-3-030-63076-8_14.
- 37 Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15*, pages 565–582, 2015.
- 38 Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16*, pages 661–661, New York, NY, USA, 2016. ACM. doi:10.1145/2940716.2940755.
- 39 Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing Bandits: Learning Under Competition. In *Innovations in Theoretical Computer Science Conference (ITCS 2018)*, 2018.

- 40 Ronza Najjar-Debbiny, Naomi Gronich, Gabriel Weber, Johad Khoury, Maisam Amar, Nili Stein, Lee Hilary Goldstein, and Walid Saliba. Effectiveness of Paxlovid in Reducing Severe Coronavirus Disease 2019 and Mortality in High-Risk Patients. *Clinical Infectious Diseases*, 2022.
- 41 Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Kidney Exchange*. *The Quarterly Journal of Economics*, 2004.
- 42 Mark Sellke and Aleksandrs Slivkins. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity. In *Proceedings of the 22nd ACM Conference on Economics and Computation, EC '21*, page 795–796, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3465456.3467549.
- 43 Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- 44 Jie Shen, Nan Cui, and Jing Wang. Metric-fair active learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, 2022.
- 45 Pengcheng Shen, Chunguang Li, and Zhaoyang Zhang. Distributed active learning. *IEEE Access*, 4:2572–2579, 2016.
- 46 Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning, 2023. arXiv:2103.00360.
- 47 A. Slivkins. *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019. URL: <https://books.google.com/books?id=6ViCzQEACAAJ>.
- 48 Aleksandrs Slivkins. Incentivizing exploration via information asymmetry. *ACM Crossroads*, 24(1):38–41, 2017. doi:10.1145/3123744.
- 49 Xinran Wang, Qi Le, Ahmad Faraz Khan, Jie Ding, and Ali Anwar. A framework for incentivized collaborative learning. *arXiv preprint arXiv:2305.17052*, 2023.
- 50 Zhepei Wei, Chuanhao Li, Haifeng Xu, and Hongning Wang. Incentivized communication for federated bandits. *arXiv preprint arXiv:2309.11702*, 2023.
- 51 Xinyi Xu, Zhaoxuan Wu, Arun Verma, Chuan Sheng Foo, and Bryan Kian Hsiang Low. FAIR: fair collaborative active learning with individual rationality for scientific discovery. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, 2023.
- 52 Hanrui Zhang, Yu Cheng, and Vincent Conitzer. Efficient algorithms for planning with participation constraints. In *Proceedings of the 23rd ACM Conference on Economics and Computation, EC '22*, page 1121–1140, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3490486.3538280.

A Proof of Theorem 4

Proof. For any randomized optimal algorithm OPT, every realization of the internal randomness of OPT must have the same query complexity and be optimal. Otherwise, there exists a realization with query complexity smaller than OPT, which conflicts with that OPT is optimal. Therefore, it suffices to prove the theorem for deterministic optimal collaborative algorithms.

We will prove the theorem for deterministic algorithms by contradiction. Suppose that there exists a deterministic optimal collaborative algorithm OPT that is not individually rational w.r.t. running itself as baseline. Hence, there exists an agent $i \in [k]$ such that $Q_i(\text{OPT}, \pi, \{X_1, \dots, X_k\}) > Q(\text{OPT}, \pi, \{X_i\})$. In this case, we can construct another algorithm \mathcal{A}' with smaller label complexity, which will contradict the optimality of OPT.

The basic idea of \mathcal{A}' is to run OPT over $(\pi, \{X_i\})$ first and to recover the labels of X_i . Then, \mathcal{A}' simulates $\text{OPT}(\pi, \{X_1, \dots, X_k\})$ and asks $\text{OPT}(\pi, \{X_1, \dots, X_k\})$ what point to query. But whenever $\text{OPT}(\pi, \{X_1, \dots, X_k\})$ asks to query the label of some point in X_i ,

since we already know the labeling of X_i , we can just feed $\text{OPT}(\pi, \{X_1, \dots, X_k\})$ with these labels without actually asking agent i to query them.

Thus, the label complexity of \mathcal{A}' is

$$\begin{aligned} Q(\mathcal{A}', \pi, \{X_1, \dots, X_k\}) &= Q(\text{OPT}, \pi, \{X_i\}) + \sum_{j:j \neq i} Q_j(\text{OPT}, \pi, \{X_1, \dots, X_k\}) \\ &< Q_i(\text{OPT}, \pi, \{X_1, \dots, X_k\}) + \sum_{j:j \neq i} Q_j(\text{OPT}, \pi, \{X_1, \dots, X_k\}) \\ &= Q(\text{OPT}, \pi, \{X_1, \dots, X_k\}) = Q^*(\pi, \{X_1, \dots, X_k\}), \end{aligned}$$

where the first inequality holds due to that OPT is not IR and the last equality holds since OPT is optimal. Since $Q^*(\pi, \{X_1, \dots, X_k\}) \leq Q(\mathcal{A}', \pi, \{X_1, \dots, X_k\})$ by definition, there is a contradiction. \blacktriangleleft

B Proof of Theorem 6

Proof. The construction is inspired by [21]. Consider $k = 2$ and let the unlabeled data set of agent 1 be

$$X_1 = \{(0, 1, 0), (0, 2, 0), (0, 0, 1), (0, 0, 2), \dots, (0, 0, n)\}$$

for some $n \in \mathbb{N}_+$.

Let the unlabeled data set of agent 2 be

$$X_2 = \{(1, 0, 0)\}.$$

Let the unlabeled pool $X = X_1 \cup X_2$. Let $h_{i,j,l}$ denote the hypothesis which labels $(i, 0, 0), (0, j, 0), (0, 0, l)$ as 1 and the rest as 0.

Let the hypothesis class be $\mathcal{H} = \{h_{i,j,l} | i \in \{0, 1\}, j \in [2], l \in [n]\}$.

Let the prior distribution $\pi_0 = \pi$ be defined as follows:

$$\begin{cases} \pi(h_{0,0,0}) = \frac{1}{4} \\ \pi(h_{0,j,l}) = \frac{1}{4 \cdot 3^l} & \text{for } j = 1, 2, l = 1, \dots, n-1 \\ \pi(h_{0,j,n}) = \frac{1}{8 \cdot 3^{n-1}} & \text{for } j = 1, 2 \\ \pi(h_{1,1,l}) = \frac{1}{3^l} & \text{for } l = 1, \dots, n-1 \\ \pi(h_{1,1,n}) = \frac{1}{2 \cdot 3^{n-1}}. \end{cases}$$

Now we show that the label complexity of agent 1 in the collaboration is $Q_1(\text{GBS}, \pi, \{X_1, X_2\}) = \Omega(n)$. While the label complexity of running GBS itself is $Q(\text{GBS}, \pi, \{X_1\}) = \mathcal{O}(1)$.

Label complexity of agent 1 in the collaboration

Let VS denote the version space. And for any point x , let $\text{VS}_x^+ = \{h \in \text{VS} | h(x) = 1\}$ denote the subset of the version space which labels x by 1. Similarly, let $\text{VS}_x^- = \{h \in \text{VS} | h(x) = 0\}$.

Now let us consider the length of the path in the query tree when the target hypothesis is $h_{0,0,0}$.

A-priori (before starting to query), for point $(1, 0, 0)$, we have

$$\pi(\text{VS}_{(1,0,0)}^+) = \sum_{l=1}^n \pi(h_{1,1,l}) = \frac{1}{3} + \frac{1}{3^2} + \dots + \frac{1}{3^{n-1}} + \frac{1}{2 \cdot 3^{n-1}} = \frac{1}{2}.$$

2:18 Incentivized Collaboration in Active Learning

For point $(0, 1, 0)$, we have

$$\pi(\text{VS}_{(0,1,0)}^+) = \sum_{l=1}^n (\pi(h_{0,1,l}) + \pi(h_{1,1,l})) = \sum_{l=1}^n \pi(h_{0,1,l}) + \pi(\text{VS}_{(1,0,0)}^+) > \frac{1}{2}.$$

For point $(0, 2, 0)$, we have

$$\pi(\text{VS}_{(0,2,0)}^+) = \sum_{l=1}^n \pi(h_{0,2,l}) = \frac{1}{4 \cdot 3} + \frac{1}{4 \cdot 3^2} + \dots + \frac{1}{4 \cdot 3^{n-1}} + \frac{1}{8 \cdot 3^{n-1}} = \frac{1}{8}.$$

For other points $(0, 0, l)$ for $l \in [n-1]$, we have $\pi(\text{VS}_{(0,0,l)}^+) = \frac{1}{4 \cdot 3^l} + \frac{1}{3^l} = \frac{5}{4 \cdot 3^l} < \frac{1}{2}$ and for $(0, 0, n)$, we have $\pi(\text{VS}_{(0,0,n)}^+) < \pi(\text{VS}_{(0,0,n-1)}^+) < \frac{1}{2}$.

Therefore, the algorithm $\text{GBS}(\pi, \{X_1, X_2\})$ will query $(1, 0, 0)$ at time 1. Suppose the label of $(1, 0, 0)$ is 0 since we consider the path corresponding to $h_{0,0,0}$ as the target hypothesis.

Now we show that $\text{GBS}(\pi, \{X_1, X_2\})$ will query points $(0, 0, 1), (0, 0, 2), \dots, (0, 0, n)$ sequentially by induction.

At time 1, the version space is $\text{VS} = \{h_{0,0,0}\} \cup \{h_{i,j,l} \in \widehat{H} \mid i = 0\}$. We list $\pi(h_{0,j,l})$ for $j \in \{1, 2\}$ and $l \in [n]$ in Table 1 for illustration.

■ **Table 1** Table of $\pi(h_{0,j,l})$ for $j \in \{1, 2\}$ and $l \in [n]$.

	$(0, 1, 0)$	$(0, 2, 0)$
$(0, 0, 1)$	$\frac{1}{4 \cdot 3}$	$\frac{1}{4 \cdot 3}$
$(0, 0, 2)$	$\frac{1}{4 \cdot 3^2}$	$\frac{1}{4 \cdot 3^2}$
\dots	\dots	\dots
$(0, 0, n)$	$\frac{1}{8 \cdot 3^{n-1}}$	$\frac{1}{8 \cdot 3^{n-1}}$

Then we can compute that

$$\pi(S_{(0,1,0)}^+) = \pi(S_{(0,2,0)}^+) = \frac{1}{4 \cdot 3} + \frac{1}{4 \cdot 3^2} + \dots + \frac{1}{4 \cdot 3^{n-1}} + \frac{1}{8 \cdot 3^{n-1}} = \frac{1}{8},$$

$$\pi(S_{(0,0,1)}^+) = \frac{1}{6} > \pi(S_{(0,0,l)}^+),$$

$$\pi(S_{(0,0,l)}^+) \leq \pi(S_{(0,0,2)}^+) = \frac{1}{18},$$

for all $l \geq 2$.

Thus, the algorithm $\text{GBS}(\pi, \{X_1, X_2\})$ will choose $(0, 0, 1)$ at time 2.

Suppose that at time $t = 2, 3, \dots, l$, $\text{GBS}(\pi, \{X_1, X_2\})$ has picked $(0, 0, 1), \dots, (0, 0, l-1)$ and all are labeled 0.

Now we show that $\text{GBS}(\pi, \{X_1, X_2\})$ will pick $(0, 0, l)$ at time $t = l+1$. The version space at the beginning of time $l+1$ is $\text{VS} = \{h_{0,0,0}\} \cup \{h_{i,j,p} \in \widehat{H} \mid p \geq l\}$. We can compute that

$$\pi(S_{(0,0,l)}^+) = \frac{1}{2 \cdot 3^l} > \pi(S_{(0,0,p)}^+)$$

for all $p > l$, and that

$$\pi(S_{(0,1,0)}^+) = \pi(S_{(0,2,0)}^+) = \frac{1}{4 \cdot 3^l} + \frac{1}{4 \cdot 3^{l+1}} + \dots + \frac{1}{4 \cdot 3^{n-1}} + \frac{1}{8 \cdot 3^{n-1}} = \frac{1}{8 \cdot 3^{l-1}}.$$

Hence, $\text{GBS}(\pi, \{X_1, X_2\})$ will pick $(0, 0, l)$.

Therefore, we proved that when the target hypothesis is $h_{0,0,0}$, $\text{GBS}(\pi, \{X_1, X_2\})$ will query

$(1, 0, 0), (0, 0, 1), (0, 0, 2), \dots, (0, 0, n)$ sequentially.

Thus, we have that $Q_1(\text{GBS}, \pi, \{X_1, X_2\}, h_{0,0,0}) = n + 1$, and $Q_1(\text{GBS}, \pi, \{X_1, X_2\}) \geq \frac{n+1}{4}$ as $\pi(h_{0,0,0}) = \frac{1}{4}$.

Label complexity of agent 1 when she runs the (GBS) baseline individually

Now we show that $Q(\text{GBS}, \pi, \{X_1\}) = \mathcal{O}(1)$. Since X_1 does not contain $(1, 0, 0)$, both $h_{0,j,l}$ and $h_{1,j,l}$ label X_1 identically. Every effective hypothesis over X_1 can be written as $h_{*,j,l}$ with $\pi(h_{*,j,l}) = \pi(h_{0,j,l}) + \pi(h_{1,j,l})$, which is listed in Table 2.

■ **Table 2** Table of $\pi(h_{*,j,l})$ for $j \in \{1, 2\}$ and $l \in [n]$.

	$(0, 1, 0)$	$(0, 2, 0)$
$(0, 0, 1)$	$\frac{1}{4 \cdot 3} + \frac{1}{3}$	$\frac{1}{4 \cdot 3}$
$(0, 0, 2)$	$\frac{1}{4 \cdot 3^2} + \frac{1}{3^2}$	$\frac{1}{4 \cdot 3^2}$
\dots	\dots	\dots
$(0, 0, n)$	$\frac{1}{8 \cdot 3^{n-1}} + \frac{1}{2 \cdot 3^{n-1}}$	$\frac{1}{8 \cdot 3^{n-1}}$

Notice that if we know that the label of $(0, 0, l)$ is positive for some l , then the version space has at most 2 effective hypotheses, $h_{*,1,l}$ and $h_{*,2,l}$. In this case, the algorithm needs at most 2 more queries.

At time $t = 1$, we have

$$\pi(S_{(0,0,1)}^+) = \frac{1}{4 \cdot 3} + \frac{1}{3} + \frac{1}{4 \cdot 3} = \frac{1}{2},$$

$$\pi(S_{(0,0,l)}^+) < \pi(S_{(0,0,1)}^+), \forall l \geq 2,$$

$$\pi(S_{(0,2,0)}^+) = \frac{1}{4 \cdot 3} + \frac{1}{4 \cdot 3^2} + \dots + \frac{1}{4 \cdot 3^{n-1}} + \frac{1}{8 \cdot 3^{n-1}} = \frac{1}{8},$$

$$\pi(S_{(0,1,0)}^+) = \pi(S_{(0,2,0)}^+) \cdot 5 = \frac{5}{8}.$$

Therefore, $\text{GBS}(\pi, \{X_1\})$ will query $(0, 0, 1)$ at $t = 1$.

We complete the proof by exhaustion. If $(0, 0, 1)$ is labeled as 1, then the algorithm needs at most two more queries as aforementioned.

If $(0, 0, 1)$ is labeled as 0, then $h_{*,1,1}$ and $h_{*,2,1}$ will be removed from the version space and $\text{GBS}(\pi, \{X_1\})$ will query $(0, 1, 0)$ at $t = 2$ then.

If the label is 1, the version space is reduced to $\{h_{*,1,l} | l = 2, \dots, n\}$ and $\text{GBS}(\pi, \{X_1\})$ will query $(0, 0, 2), (0, 0, 3), \dots$ sequentially until receiving a positive label.

If the label of $(0, 1, 0)$ is 0, $\text{GBS}(\pi, \{X_1\})$ will query $(0, 2, 0)$ at time $t = 3$. If the label of $(0, 2, 0)$ is 1, then it is similar to the case of $(0, 1, 0)$ being labeled 1 and the algorithm will query $(0, 0, 2), (0, 0, 3), \dots$ sequentially.

If the label of $(0, 2, 0)$ is 0, we know the target hypothesis is $h_{0,0,0}$ and we are done.

Hence we have $Q(\text{GBS}, \pi, \{X_1\}) \leq \sum_{l=1}^n (\pi(h_{*,1,l}) + \pi(h_{*,2,l})) \cdot (3 + l) + \pi(h_{0,0,0}) \cdot 3 = \sum_{l=1}^{n-1} \frac{1}{2 \cdot 3^{l-1}} \cdot (3 + l) + \frac{1}{4} \cdot 3 = \mathcal{O}(1)$. ◀

C Proof of Lemma 10

Proof. SIR property: Since \mathcal{A}' and $\{\mathcal{O}_i | i \in [k]\}$ are IR and agent i can strictly benefit from \mathcal{O}_i , we have $Q_i(\mathcal{A}'_\varepsilon, \pi, \{X_1, \dots, X_k\}) < Q(\mathcal{A}, \pi, \{X_i\})$ for all $i \in [k]$.

Label complexity: The label complexity of \mathcal{A}'_ε is

$$\begin{aligned} Q(\mathcal{A}'_\varepsilon, \pi, \{X_1, \dots, X_k\}) &= (1 - \frac{\varepsilon}{n})Q(\mathcal{A}', \pi, \{X_1, \dots, X_k\}) + \frac{\varepsilon}{kn} \sum_{i=1}^k Q(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}) \\ &\leq (1 - \frac{\varepsilon}{n})Q(\mathcal{A}', \pi, \{X_1, \dots, X_k\}) + \varepsilon. \end{aligned}$$

Then we are done. \blacktriangleleft

D Proof of necessity of Assumption 1

Proof of necessity. Suppose that there exists an SIR algorithm \mathcal{A}'' when the baseline algorithm is optimal. We therefore have $Q_i(\mathcal{A}'', \pi, \{X_1, \dots, X_k\}) < Q^*(\pi, \{X_i\})$ by definition. We claim that \mathcal{A}'' must satisfy $Q_i(\mathcal{A}'', \pi, \{X_1, \dots, X_k\}) \geq \mathbb{E}_{h \sim \pi} [Q^*(\pi_{h, -i}, \{X_i\})]$. This is because we can construct another algorithm \mathcal{B} by running \mathcal{A}'' over all other agents except agent i , i.e., running \mathcal{A}'' over (π, X_{-i}) with $X_{-i} = \{X_j | j \neq i\}$ first to recover the labels of all other agents X_{-i} . Then, \mathcal{B} simulates \mathcal{A}'' over $(\pi, \{X_1, \dots, X_k\})$ without actually querying any point in X_{-i} (similarly to Algorithm 1). In this case, the label complexities of agent i are identical for algorithms \mathcal{B} and \mathcal{A}'' , i.e., $Q_i(\mathcal{B}, \pi, \{X_1, \dots, X_k\}) = Q_i(\mathcal{A}'', \pi, \{X_1, \dots, X_k\})$. Since $Q^*(\pi_{h, -i}, \{X_i\})$ is the optimal label complexity of agent i given the label information of X_{-i} , we have $Q_i(\mathcal{B}, \pi, \{X_1, \dots, X_k\}) \geq Q^*(\pi_{h, -i}, \{X_i\})$. Therefore, we have $Q^*(\pi, \{X_i\}) > Q_i(\mathcal{A}'', \pi, \{X_1, \dots, X_k\}) \geq Q^*(\pi_{h, -i}, \{X_i\})$. \blacktriangleleft

E Proof of Lemma 12

Proof. First, note that \mathcal{O}_i is IR as B2IR(\mathcal{A}) is IR, and pruning the query tree does not increase label complexity. For any $i \in [k]$, suppose that there exists an hypothesis $h \in \widehat{H}$ s.t. $|H(X_i|h)| < |H(X_i)|$. Then in the query tree of $\mathcal{A}(\pi, \{X_i\})$, either all leaves are inconsistent with $h(X_{-i})$ or there exists one internal node v who has exactly one subtree with all leaves inconsistent with $h(X_{-i})$. This node v as well as the corresponding subtree are pruned in \mathcal{O}_i and thus the leaves in the other subtree rooted at v have their depth reduced by at least 1. Now, there exists an hypothesis $h'' \in \widehat{H}$ such that $h''(X_i) \in H(X_i|h)$ and $h''(X_i)$ is in the other subtree. Since $h''(X_{-i}) = h(X_{-i})$, when the underlying hypothesis is h'' , the pruned tree given $h''(X_{-i})$ is the same as that given $h(X_{-i})$. Hence, we have $Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}, h'') \leq Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}, h'') - 1$.

Then we have

$$\begin{aligned} Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}) &= \mathbb{E}_{h \sim \pi} [Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}, h)] \\ &= \pi(h'')Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}, h'') + (1 - \pi(h''))\mathbb{E}_{h \sim \pi | h \neq h''} [Q_i(\mathcal{O}_i, \pi, \{X_1, \dots, X_k\}, h)] \\ &\leq \pi(h'')(Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}, h'') - 1) + (1 - \pi(h''))\mathbb{E}_{h \sim \pi | h \neq h''} [Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}, h)] \\ &< Q_i(\mathcal{A}, \pi, \{X_1, \dots, X_k\}), \end{aligned}$$

where the last inequality holds due to $\pi(h'') > 0$ since w.l.o.g., we assumed $\pi(h) > 0$ for all $h \in \widehat{H}$ in Section 2. \blacktriangleleft

Can Copyright Be Reduced to Privacy?

Niva Elkin-Koren ✉

Faculty of Law, Tel Aviv University, Israel

Uri Hacoheh ✉

Faculty of Law, Tel Aviv University, Israel

Roi Livni ✉

School of Electrical Engineering, Tel Aviv University, Israel

Shay Moran ✉

Departments of Mathematics and Computer Science, Technion, Haifa, Israel

Abstract

There is a growing concern that generative AI models will generate outputs closely resembling the copyrighted materials for which they are trained. This worry has intensified as the quality and complexity of generative models have immensely improved, and the availability of extensive datasets containing copyrighted material has expanded. Researchers are actively exploring strategies to mitigate the risk of generating infringing samples, with a recent line of work suggesting to employ techniques such as differential privacy and other forms of algorithmic stability to provide guarantees on the lack of infringing copying. In this work, we examine whether such algorithmic stability techniques are suitable to ensure the responsible use of generative models without inadvertently violating copyright laws. We argue that while these techniques aim to verify the presence of identifiable information in datasets, thus being privacy-oriented, copyright law aims to promote the use of original works for the benefit of society as a whole, provided that no unlicensed use of protected expression occurred. These fundamental differences between privacy and copyright must not be overlooked. In particular, we demonstrate that while algorithmic stability may be perceived as a practical tool to detect copying, such copying does not necessarily constitute copyright infringement. Therefore, if adopted as a standard for detecting an establishing copyright infringement, algorithmic stability may undermine the intended objectives of copyright law.

2012 ACM Subject Classification Social and professional topics → Copyrights

Keywords and phrases Copyright, Privacy, Generative Learning

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.3

Related Version *Full Version*: <https://arxiv.org/abs/2305.14822>

Funding This research was funded in part by and ISF Grant (2188\20), an ERC grant (FOG, 101116258) and ERC grant (GENERALIZATION, 10139692). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. In addition, the research leading to these results was supported by TILabs Tel-Aviv University Innovation Labs.

Acknowledgements We thank Bruria Friedman for research assistance.

1 Introduction

Recent advancements in machine learning have sparked a wave of new possibilities and applications that could potentially transform various aspects of our daily lives and revolutionize numerous professions through automation. However, training such algorithms heavily relies on extensive content which may include copyrighted materials. Under U.S. copyright law, copyright protection subsists in original content of authorship fixed in any tangible medium of expression [55], excluding any “idea, procedure, process, system, method



© Niva Elkin-Koren, Uri Hacoheh, Roi Livni, and Shay Moran;
licensed under Creative Commons License CC-BY 4.0

5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 3; pp. 3:1–3:18



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

3:2 Can Copyright Be Reduced to Privacy?

of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.” [55, § 102(b)]. The unauthorized copying of copyrighted works may amount to copyright infringement [55, § 106] unless permitted by exceptions and limitations provided by law ([55, §107-122], and [50]). Consequently, identifying and, determining when and how content can be used within this framework without infringing upon individuals’ legal rights has become a pressing challenge. Foundation Models and generative AI (GenAI), trained on gigantic datasets, exacerbate this challenge. One area where this issue arises prominently is in the operation of generative models, which take human-produced content – much of it copyrighted as input and are expected to generate “-similar” content. For instance, consider a machine trained on images and then generates new images that resemble the ones it was trained on. In this context, the fundamental question arises:

When does the content generated by a machine (output content) infringe copyright in the training set (input content)?

This question is not purely theoretical, as various aspects of this problem have become subjects of legal disputes in recent years. In 2022, a class action was filed against Microsoft, GitHub, and OpenAI, claiming that their code-generating systems, Codex and Copilot, infringed copyright in the licensed code that the system was allegedly trained on [13]. Similarly, in another class action, against Stable Diffusion, Midjourney, and DeviantArt, plaintiffs argue that by training their system on web-scraped images, the defendant infringes millions of artists’ rights [3]. Allegedly, the images produced by these systems, in response to prompts provided by the systems’ users, are derived solely from the training images, which belong to plaintiffs, and, as such, are considered unauthorized derivative works of the plaintiffs’ images [55, § 106 (2)].

A preliminary question is whether it is lawful to make use of copyrighted content in the course of training [36, 23, 34]. There are compelling arguments to suggest that such intermediary copying might be considered fair use [36]. For example, Google’s Book Search Project – entailing the mass digitization of copyrighted books from university library collections to create a searchable database of millions of books – was held by US courts to be fair use [22]. Then, there is a claim that generative models reproduce protected copyright expressions from the input content on which the model was trained. However, to claim that the output of a generative model infringes her copyright, a plaintiff must prove not only that the model had access to her copyrighted work, but also that the alleged copy is substantially similar to her original work [53, 8]

Identifying what constitutes “substantial similarity,” and unlawful copying remains a pressing challenge. Recent studies have proposed measurable metrics to quantify copyright infringement [59, 5, 51, 9]. One approach, [59, 5] asserts that a machine generating output content substantially similar to an input content does not infringe that input content copyright if the machine would have reasonably generated the same output content even without accessing the input content. This argument can be illustrated as follows: Suppose that Alice outputs content A and Bob claims it plagiarizes content B. Alice might argue that she never saw content B, and would reason that this means she did not infringe Bob’s copyright. However, since Alice must have observed some content, a second line of defense could be that “**had** she never saw B” she would still be likely to produce A. The above argument was exemplified by [5] who interprets differential-privacy in the above manner. Subsequently, [59] presented a certain generalization, in the form of a *near-free access* (NAF) notion that can potentially allow a more versatile notion of copyright protection. Both applications draw on algorithmic stability notions used in privacy research.

However, certain crucial traits of copyright law make it challenging to reduce the problem to a question of privacy. An essential element of copyright law in the United States is utilitarian rationale, seeking to promote the creation and deployment of creative works [11, 41]. It is crucial, then, that any interpretation of copyright, or for that matter any quantifiable measure for copyright, will be aligned with these objectives. In particular, while the law delineates exclusive rights to the creators of original expressions, it must ensure sufficient creative space for current and future creators [49]. For this reason, several criteria exist in copyright law, specifically allowing breathing room for subsequent authors to draw upon copyrighted content. These criteria distinguish copyright law from privacy as defined by algorithmic stability notions. First, copyright is limited in time, and once protection has expired, the copyright content enters the public domain and is free for all to use without authorization [37]. This issue, though, can be modeled by distinguishing between private and public data (or protected and non-protected data). Second, and more importantly, copyright law excludes specific subject (e.g. ideas, methods of operation, facts), since they are regarded as raw materials needed for cultural expression. According to the US Supreme Court, “originality” is the “sine qua non” of copyright. [20] Thus, only the original elements within copyrighted works are legally protected by copyright law. Unoriginal elements (e.g., ideas, facts) are never protected. Privacy, in contrast, protects content and not expression, which in turn can be misaligned with the original objectives of copyright law.

This point cannot be overestimated. Copyright law not only allows subsequent authors to draw upon the unoriginal, and thus unprotected, elements of copyrighted works (unlike in privacy) but also encourages subsequent authors to do so [37, 18]. Because copyright protection only applies to some elements within copyrighted works (i.e. expression) while deliberately excluding others (i.e., ideas) courts need to delineate the scope of legal protection when deciding copyright disputes. As a result, the scope of copyright protection varies not only among different works but also among different elements within a single work [54].

Third, in a stark distinction from privacy, copyright law also encourages using the original (and thus protected) elements of copyrighted works in certain circumstances. These include de minimis quotations, transformative uses serving different purposes compared to the purpose of the original work (such as parodies), and other types of “fair uses” such as learning and research [43]. The fair use doctrine serves as a check on copyright, to ensuring it does not stifle the very creativity copyright law seeks to foster. Fair use is also considered one of the safety valves that allows copyright protection to coexist with freedom of expression [42].

For all these reasons, privacy notations are both over-inclusive and under-inclusive from a copyright perspective. They are over-inclusive because they withhold much more from subsequent authors than copyright law necessitates, consequently undermining the objectives of copyright law. At the same time, by focusing on content rather than original expression, privacy notations are also under-inclusive because they allow (in some cases) unlawful access to original copyrighted expression. This could happen, for example, if Alice’s model did not access input content B, but did access input content C that incorporated original expression deriving (lawfully or not) from input content B.

In this study we initiate a discussion about the challenges involved in providing a rigorous definition capturing the concept of copyright. We commence with a technical discussion, comparing different proposed notions of copyright (in particular, differential privacy and NAF) and examining their close connection to algorithmic stability. Subsequently, we argue that any approach following this line of reasoning encounters significant obstacles in modeling copyright as understood within the legal context. In more detail, we argue that algorithmic stability strategies fail to account for some principles of copyright law

that intend to preserve copyright law’s delicate balance. We identify several major gaps between algorithmic stability strategies and copyright doctrine. Accordingly, we argue, that if algorithmic stability techniques are adopted as a standard for copyright infringement, they may undermine the intended goals of copyright law. We further propose a different approach to using quantified measures in copyright disputes that could better reconcile copyright trade-offs.

1.1 Related Work

A growing number of researchers in recent years have explored how to address legal problems by applying computer science theories and methods. This literature seeks to narrow the gap between the vague and abstract concepts used by law and mathematical models, and to offer more rigor, coherent, and scalable definitions for issues such as privacy [14], fairness and discrimination. [15, 30] In the context of generative models, [9] and [25] have explored whether generative diffusion models memorize protected works that appeared in the models’ training set. Their approach indicates the mere possibility of unauthorized copying by GenAI models. However, as discussed, memorizing of the input content does not necessarily equate to copyright infringement. To evaluate infringement we must consider other measurable metrics and quantified measures for copyright key limiting concepts.

There is also active and thought-provoking discussion on how ML technologies are reshaping our understanding of copyright within the realm of law. [2] explores the question of whether AI system outputs should be subject to copyright protection. [23, 36] examine the implications of copyright law’s notions of authorship and learning for literary machines. Our Focus, though, is on the legitimacy of using copyrighted materials by models that generate similar output content.

The works of [5] and [59], which rely on privacy/privacy-like notions, are the main focus of our work. An alternative approach taken by [51] proposes a framework to test the substantial similarity of a model’s output content by comparing Kolmogorov-Levin complexity with and without access to the copyrighted input content. However, one has to distinguish between protected expressions and non-protected ideas; this crucial challenge is overlooked by their approach. Another work by [19] suggests using generative learning techniques to assess creativity. Such approaches may prove valuable, as we indicate in Section 4, but only if they are designed to align with copyright principles. Lastly, [27] seek to develop strategies to be applied to generative models to ensure they satisfy the same fair use standard as in human discretion. The application of this solution may not be possible, though, in cases where little to no open source or fair use data is readily available.

2 Algorithmic stability as a surrogate for copyright

In this section, we focus on introducing and discussing two notions of algorithmic stability: near-access-freeness (NAF) and differential privacy (DP); these two notions were specifically investigated in the realm of training methods aimed at safeguarding copyrighted data.

NAF and DP adhere to a shared form of stability: they ensure that the resulting model, denoted as q , satisfies a safety condition with respect to each copyrighted data instance, denoted as c . This safety condition guarantees the existence of a “safe model”, denoted by q_c , which does not infringe the copyright of data c , and importantly, q exhibits sufficient similarity to q_c . Consequently, both NAF and DP guarantee that p itself does not violate the copyright of the respective data instance c .

Formally, we consider a standard setup of an unknown distribution \mathcal{D} , and a generative algorithm A . The algorithm A , gets as an input a training set of i.i.d samples $S = \{z_1, \dots, z_m\} \in Z^m \sim D^m$, and outputs a model $p_S^A = A(S)$, which is a distribution supported on Z . For simplicity, we will assume here that Z is a discrete finite set, but of arbitrary size. [59] consider a more general variant in which the output posterior is dependent on a “prompt” x , and A outputs a mapping $p^{(A_S)}(\cdot|x)$ that may be regarded as a mapping from prompts to posteriors. For our purposes there is no loss in generality in assuming that p is “promptless”, and our results easily extend to the promptful case, by thinking of each prompt as inducing a different algorithm when we hard-code the prompt into the algorithm.

Differential Privacy

A is said to be (α, β) -differentially private [16] if for every pair of input datasets S, S' that differ on a single datapoint, we have that for every event E :

$$\mathbb{P}(A(S) \in E) \leq e^\alpha \mathbb{P}(A(S') \in E) + \beta \text{ and } \mathbb{P}(A(S') \in E) \leq e^\alpha \mathbb{P}(A(S) \in E) + \beta \quad (1)$$

The concept of privacy, viewed as a measure of copyright, can be explained as follows: Let's consider an event, denoted as E , which indicates that the generative model produced by A violates the copyright of a protected content item c . The underlying assumption is that if the model has not been trained on c , the occurrence of event E is highly improbable. Thus, we can compare the likelihood of the event E when c is present in the sample S with the likelihood of E when c is not included in a neighboring sample S' (which is otherwise identical to S). If A satisfies the condition stated in equation Equation (1), then the likelihood of event E remains extremely low, even if c happened to be present once in its training set.

Near Access Freeness

There are several shortcomings of the notion of differential privacy that have been identified. Some of these are reiterated in Section 3. [59] proposed the notion of Near-Access Freeness (NAF) that relaxes differential privacy in several aspects. Formally, NAF (or more accurately NAF w.r.t safe function safe and Δ_{max} is defined as follows: First, we assume a mapping safe that assigns to each protected content c a model q_c which is considered safe in the sense that it does not breach the copyright of c . The function safe, for example, can assign c to a model that was trained on a sample that does not contain c . Several safe functions have been suggested in [59].

A model p is considered α -NAF if the following inequality holds simultaneously for every protected content c and every z :

$$p(z) \leq e^\alpha q_c(z). \quad (2)$$

The intuition behind NAF is very similar to the one behind DP, however there are key differences that can, in principle, help it circumvent the stringency of DP.

1. The first difference between NAF and DP is that the NAF framework allows more flexibility by picking the 'safe' function. Whereas DP is restricted to a safe model corresponding to training the learning algorithm on a neighboring sample excluding the content c .
2. A second difference is the fact that NAF is one sided (see Equation (2)), in contrast with DP which is symmetric (see Equation (1)). Note that one-sidedness is indeed more aligned with the requirement of copyright which is non-symmetric.

3. NAF makes the distinction between content-safety and model-safety [59]. In more detail, the NAF notion requires that the output model is stable. This is in contrast with privacy that requires stability of the posterior distribution over the output models. In this sense the notion of NAF is more akin to *prediction differential privacy* [14] than to differential privacy.
4. Finally, NAF poses constraints on the model outputted by the learning algorithm (each constraint corresponds to a prespecified *safe model*). This is in contrast with privacy which does not restrict the output model, but requires stability of the posterior distributions over output models. This distinction may seem minor but it can lead to peculiarities. For example, an algorithm that is completely oblivious to its training set and that always outputs original content can still violate the requirements of NAF. To see this, imagine that our learning rule outputs a model q that always generates the same content z which is completely original and not similar to any protected content c . However, depending on the safe models q_c it can be the case that the model q is not similar to any of them.

These differences, potentially, allow NAF to circumvent some of the hurdles for using DP as a notion for copyright. For example, the one-sidedness seems sufficient for copyright and may allow models that are discarded via DP. Also, the distinction between model-safety and content-safety can, for example, allow models that may memorize completely the training set as long as a content they output does not provide a proof for such memorization. Next, the fact that NAF is defined by a set of constraints, and not a property of the learning algorithm, allows one to treat breaches of Equation (2) as soft “flagging” and not necessarily as hard constraints. This advantage is further discussed in Section 4. Finally, perhaps most distinguishable, is the possibility to use general safety functions that can capture copyright breaches more flexibly. We next discuss the implications of these refinements, and the question of model safety vs. content safety in NAF and in DP.

Model safety vs. Content safety

Our first result is a parallel to Theorem 3.1 in [59] in the context of DP stability. Theorem 3.1 in [59] shows how to efficiently transform a given learning rule A to a learning rule B which is NAF-stable, provided that A tends to output similar generative models when given inputs that are identically distributed. We state and prove a similar result by replacing NAF stability with DP stability, which demonstrates that the notion of DP can be relaxed, analogously to NAF, to require only content safety under proper assumptions:

Recall that the total variation distance between any two distributions is defined as: $\|q_1 - q_2\| = \frac{1}{2} \sum |q_1(x) - q_2(x)| = \sup_E (q_1(E) - q_2(E))$,

► **Proposition 1.** *Let A be an algorithm mapping samples S to models q_S^A such that $\mathbb{E}_{S_1, S_2} [\|q_{S_1}^A - q_{S_2}^A\|] \leq \alpha$, where $S_1, S_2 \sim D^m$ are two independent samples. Then, there exist an (ϵ, δ) DP algorithm B that receives a sample $S_B \sim D^{m_{priv}}$ such that if $m_{priv} = \tilde{O}\left(\frac{m}{\eta\epsilon} \log 1/\delta\right)$ and $S_A \sim D^m$ then: $\mathbb{E}_{S_A, S_B} [\|\mathbb{E}[q_{S_B}^B] - q_{S_A}^A\|] \leq \frac{2\alpha}{1+\alpha} + O(\eta)$. Where the expectation within, is taken over the randomness of B .*

The premise in the above theorem is identical to that in Theorem 3.1 in [59] and captures the property that A provides similar outputs on identically distributed inputs. The obtained algorithm B is DP-stable and at the same time it has a similar functionality like A in sense that its output model q^B generates content z which in expectation is distributed like contents generated by q^A .

Safety functions

We now turn to a discussion on the potential behind the use of different safety functions. The crucial point (which we discuss in great detail in Section 3 below) is that a satisfactory “copyright definition” *must* allow algorithms to be highly influenced, even by their input content which is *protected*. This reveals a stark contrast with algorithmic stability: it is easy to see that DP does not allow such influence. Indeed, the whole philosophy behind privacy is that a model is “safe” if it did not observe the private example (in particular was not influenced by it).

This raises the question of whether the greater flexibility of the NAF model can provide better aligned notions of safety. In fact, if it is allowed to be influenced by protected data, one might even want to consider safe models that have *intentionally* observed a certain content and derived out of it the derivatives that are not protected.

The next result, though, shows that there is a *no free lunch* phenomenon. For every protected content c , we can either only consider safe models that observed c and are influenced by it, or only safe models that *never* observed it and were *not* influenced by it. In other words, if a protected content c influenced its safe model q_c then it must influence all safe models $q_{c'}$ for all protected contents c' . We further elaborate on the implication of this result in Section 4.

Below, q_1 and q_2 should be thought of as safe models, and p as the model outputted by the NAF learning algorithm. (So, in particular p should satisfy Equation (2) w.r.t q_1 and q_2 .) This result complements Theorem 3.1 in [59] which shows that NAF can be satisfied in the sharded-safety setting when the two safe models are close in total-variation. The proof is left to Appendix A.1.

► **Proposition 2.** *Let q_1 and q_2 be two distributions such that $\|q_1 - q_2\| \geq \alpha$, then for any distribution p we have that for some z : $p(z) \geq \frac{1}{2(1-\alpha)} \min\{q_1(z), q_2(z)\}$.*

3 The gap between algorithmic stability and copyright

So far, we have provided a technical comparison between existing notions in the CS literature aimed at provable copyright protection. While the technical notion of privacy may seem closely related, as observed through NAF, there are differences. Accordingly, there is room for more refined definitions that could capture these essential differences. While algorithmic stability approaches hold promise in helping courts assess copyright infringement cases (an issue we further discuss in Section 4), they cannot serve as a definitive test for copyright infringement. To see that, we next discuss the issue of copyright from a legal perspective. From this perspective, formal algorithmic stability approaches are both over inclusive and under-inclusive. Consequently, we will organize this section based on these challenges.

3.1 Over-inclusiveness

Here we focus on a concern that algorithmic stability approaches may filter out lawful output content that does not infringe copyright in the input content. Because non-infringing output content is lawful, employing algorithmic stability approaches as filters to generative models may needlessly limit their production capabilities, and, thereby, undermine the ultimate objectives of copyright law. Copyright law intends to foster the creation of original works of authorship by securing incentives to authors and, at the same time, ensuring the freedom of current and future authors to use and build upon existing works. The law derives from

3:8 Can Copyright Be Reduced to Privacy?

the U.S Constitutional authority: “To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” [11]

However, promoting progress is often at odds with granting unlimited control over copyrighted materials. This is why copyright law sets fundamental limits on the rights granted to authors. Promoting progress is inconsistent with an unrestricted right to prevent every unauthorized use because creators and creative processes are embedded in cultural contexts. Creative processes often requires ongoing interactions with preexisting materials, whether through learning and research, engagement with prior art to generating new interpretations, or using a shared cultural language and applying existing styles to make works of authorship more comprehensible. Consequently, using copyrighted materials becomes a crucial input in any creative discourse [10, 17].

For this reason, unlike the mandate of the algorithmic stability approaches, copyright law does not require output contents not to draw on input contents to be lawful. On the contrary, there are many cases where copyright law explicitly allows output contents to draw heavily on input contents without raising infringement concerns. In such cases, allowing input contents to impact output contents is not only something copyright law permits, but it is also something copyright law encourages. Doing so, as Jessica Litman put it, “is not parasitism; it is the essence of authorship.” [37]

Copyright law allows output contents to substantially draw an input contents in three main cases, which we next explore: (1) When an input content is in the public domain, (2) When an input content is copyrighted but incorporates aspects excluded from copyright protection, and (3) When the use of the protected aspects of the input content is lawful.

When input content is in the public domain

Input content may be unprotected because its copyright term has lapsed. Copyrights are limited in duration (though relatively long duration, which in most countries will last the life of the author plus seventy years). Once the copyright term expires, input content enters the public domain and can freely be used and impact output content without risking copyright infringement [37]. Public domain materials may also contain anything that is not copyrightable, such as natural resources. For instance, if two photographers are taking pictures of the same person, some similarity between those pictures is likely due to how this person looks, which is in the public domain. Other elements such as an original composition, or the choices made regarding lighting conditions and the exposure settings used in capturing the photograph, might be considered copyrighted expression. If the generative model only uses the former in the output content, it may not constitute an infringement.

When an input content incorporates unprotected aspects

Input content with a valid copyright term enjoys “full” legal protection, but it too is limited in scope. As provided by the copyright statute, “[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.” [55]. By this principle, output content may substantially draw on input content without infringing copyright in the latter, as long as such taking is limited to the input’s content unprotected elements.

- **Procedures, processes, systems and methods of operation** Copyright protection does not extend to “useful” or “functional” aspects of copyrighted works such as procedures, systems, and methods of operation. These aspects of an input content are freely accessible

for an output content to draw upon. For example, in the seminal case of *Baker vs. Selden*, the Supreme Court allowed Baker to create a book covering an improved book-keeping system while drawing heavily on the charts, examples, and descriptions used in Selden's book without infringing Selden's copyright [7]. As the court explained, these aspects that Baker took from Selden's work are functional methods of operations and as such are not within the domain of copyright law. Similarly, in *Lotus v. Borland*, the United States Court of Appeals for the First Circuit allowed Borland to copy Lotus's menu command hierarchy for its spreadsheet program, *Lotus 1-2-3*. The court ruled that Lotus menu command hierarchy was not copyrightable because they form methods of operation [39] - Consequently, if a generative model simply extracts procedures, processes, systems and methods from the training set it may not infringe copyright.

- **Ideas** Copyright protection is limited to concrete “expressions” and does not cover abstract “ideas.” Thus, in *Nicholas v. Universal*, the United States Court of Appeals for the Second Circuit allowed Universal to incorporate many aspects of Anne Nichols' play *Abie's Irish Rose*, in their film *The Cohens and Kellys* [58]. The court explained that the narratives and characters that Universal used (“a quarrel between a Jewish and an Irish father, the marriage of their children, the birth of grandchildren and a reconciliation”), were “too generalized an abstraction from what she wrote. . . [and, as such]. . . only a part of her [unprotected] ‘ideas.’” [58] When a generative model simply extract ideas from copyrighted materials, rather than replicating expressive content from their training data, it does not trigger copyright infringement.
- **Facts** Copyright protection also does not extend to facts. For example, in *Nash v. CBS.*, the court ruled that CBS. could draw heavily from Jay Robert Nash's books without infringing his copyright [44]. As the court explained, the hypotheses that Nash rose speculating the capture of the gangster John Dillinger and the evidence he gathered (such as the physical differences between Dillinger and the corpse, the planted fingerprints, and photographs of Dillinger and other gangsters in the 1930s) were all unprotected facts that Nash could not legally appropriate. Consequently, generative models which simply memorize facts do not infringe copyright law.

When the use of the protected aspects of the input content was lawful

Even when the protected elements of an input content (“expressions” rather than the “ideas”) are impacting an output content, such impact may be legally permissible. There are two main categories of lawful uses: *de minimis* copying and fair use.

- **De minimis copying** Copyright law allows *de minimis* copying of protected expression. I.e. copying of an insignificant amount that has no substantial impact on the rights of the copyright owner or their economic value. Similarly, “[w]ords and short phrases, such as names, titles, and slogans, are uncopyrightable.”[45]. However, *de minimis* copying of protected expression may be unlawful if it captures the heart of the work [28]. E.g. phrases like “E.T. Phone Home.” [56]
- **Fair Use** Copyright law also allows copying of protected expression if it qualifies as fair use. The U.S fair use doctrine, as codified in § 107 of the U.S Copyright Act of 1976, is yet another legal standard to carve out an exception for an otherwise infringing use after weighing a set of four statutory factors. The four statutory factors are: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work [55].

3:10 Can Copyright Be Reduced to Privacy?

Importantly, the fair use claimant need not satisfy each factor for the use to qualify as fair use [12]. Nor are the four factors meant to set out some kind of mathematical equation whereby, if at least three factors favor or disfavor fair use, that determines the result [43]. Rather, the factors serve as guidelines for holistic, case-by-case decision. In that vein, in its preamble paragraph, § 107 provides a list of several examples of the types of uses that can qualify as fair use. The examples, which include “criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, [and] research,”[55] are often thought to be favored uses for qualifying for fair use. Importantly, however, the list of favored uses is not dispositive. Rather, fair use’s open-ended framework imposes no limits on the types of uses that courts may determine “fair” [12].

When the factors strongly favor a finding of fair use, even output contents that are heavily impacted by copyrighted input contents may be excused from copyright infringement. For example, in *Campbell v. Acuff-Rose*, although the rap music group 2 Live Crew copied significant portions of lyrics and sound from Roy Orbison’s familiar rock ballad “Oh, Pretty Woman” [12]. The Supreme Court denied liability in this case, based on the premise that the 2 Live Crew’s derivative work was considered a “parody” of Orbison’s original work, and, therefore, constituted fair use. Similarly, in *The Authors Guild v. Google*, the court defended Google’s mass digitization of millions of copyrighted books to create a searchable online database as fair use, because it considered Google’s venture to be socially desirable [22] as explained by [47], concluding that the copying of expressive works for non-expressive purposes should not be counted as a copyright infringement.

3.2 Under-Inclusiveness

Algorithmic stability approaches are under exclusive because they might fail to filter out unlawful output content that infringes copyright in the input content. As explained, algorithmic stability approaches find infringement only when the output content heavily draws on input content. The law of copyright infringement, however, is not so narrow. Copyright law only requires that the output content heavily draw on the protected expression originating from an input content to find infringement. Such expression need not come from the input content itself; it may come from other sources including copies, derivatives or snippets of the original input content [33].

To illustrate this point, consider the fact pattern in the U.S Supreme Court case *Warhol vs. Goldsmith* [60]. In that case, the portrait photographer Lynn Goldsmith accused Andy Warhol of infringing copyrights in a photograph she took of the American singer Prince. Goldsmith authorized Warhol to use her photograph as an “artistic reference” for creating a single derivative illustration (see Figure 1, bottom right most picture). Still, she did not approve nor imagine that Warhol had, in fact, made 16 different derivatives from the original photograph. Warhol’s collection of Prince portraits, also known as the Prince series, is depicted in Figure 1, right side.

For our purposes, assume the Prince Series’ portraits served as input for a generative machine. Suppose the machine’s output content draws heavily on Goldsmith’s protected expression that is baked into the Prince Series’ portraits. In that case, the machine’s output content may infringe Goldsmith’s copyright in original photograph (Figure 1, left side), even if the machine did not have access to Goldsmith’s original photograph. Moreover, this risk will not be eliminated even if the Supreme Court decided that the Prince Series’ portraits themselves are non-infringing because they constitute fair use.

Simply put, copying from a derivative work – whether authorized by the copyright owner or not – may infringe copyright in the original work on which the derivative work is based. This situation is prevalent in copyright practice, especially in music. In modern music

copyright cases, plaintiffs usually show access to the original copyrighted work (musical composition) by showing access to a derivative work of that original work (sound recording). Plaintiffs are not required to demonstrate that the defendants also had access to the original sheet music nor that they could actually read musical notes.

Lastly, output content can also infringe copyright in input content by accessing parts or snippets of the input content even without accessing the input content in its entirety. This concern was raised recently in *The Authors Guild v. Google*, a case dealing with the legality of the Google Book Search Library Partner project [22]. As part of this project, Google scanned and entered many copyrighted books into their searchable database but only provided “snippet views” of the scanned pages in search results to their users. The plaintiff in the case argued that Google facilitated copyright infringement by allowing users to aggregate different snippets and reconstruct infringing copies of their original works. The court ended up dismissing this claim, but only because Google took affirmative steps to prevent such reconstruction by limiting the number of available snippets and by blacklisting certain pages.

To sum up, there are numerous instances where copyright law permits (even encourages) an output content to draw on an input content. The more substantial unprotected aspects of input content, and the more likely it is that using the input content’s protectable aspects is considered lawful, the more expansively can the output content draw upon the input content without fearing copyright infringement. At the same time, there are cases where copyright law outlaws an output even if it did not draw upon an input content, provided that it did draw on protected expression originating from that content. The more original the input content, and the more copies, derivatives, or snippets of that original content exist in the model datasets, the more likely the output content is to infringe copyrights in that input content. Therefore, any strategy for detecting or mitigating copyright infringement must account for these crucial copyright distinctions.

4 Discussion

Algorithmic stability approaches, when used to establish proof of copyright infringement are either too strict or too lenient from a legal perspective. Due to this misfit, applying algorithmic stability approaches as filters for generative models will likely to distort the delicate balance that copyright law aims to achieve between economic incentives and access to creative works.

The purpose of this article is to illuminate this misfit. This is not to say that algorithmic approaches in general and algorithmic stability approaches, in particular, have no value to the legal profession. Quite the opposite. Computer science methodologies significantly benefit the judicial table: the capability to process large volumes of information and assist policymakers in making more informed decisions. Many areas in law involve applying murky “standards” as opposed to rigid “rules.” [31]. As discussed, copyright law extensively uses legal standards, such as idea/expression distinction, or fair use principles, to restrict the scope of protection accorded to copyrighted works. Consequently, copyright infringement cannot be boiled down to a binary computational test.

The true value of computer science methodologies to the legal profession is not necessarily to convert murky standards into rigid rules (e.g., by constructing a definitive binary test for copyright infringement), but, instead, to make legal standards less murky. A rich body of scholarship explores the ills of vaguely-defined legal standards, especially in the context of intellectual property [46, 4, 48, 21, 40] Algorithmic stability approaches, if applied with caution, may introduce new quantifiable methods for applying legal standards more clearly

3:12 Can Copyright Be Reduced to Privacy?

and predictably. Such methods could help measure vague legal concepts such as “fairness” “privacy,” and, in the copyright context – “originality”, and at the same time facilitate the ongoing development of legal and social norms [24]. However, to ensure these methods are beneficial, it is vital to acknowledge the limitations of applying algorithmic stability approaches to copyright.

Stability is not safe

The NAF framework, which allows a rich class of safety functions, has the potential to circumvent some of the challenges presented, but may still be limited and we now wish to discuss this in further details. RL is supported by an ERC Grant (FOG

To utilize the NAF framework, the first basic question one needs to address is *Given a protected content c how should we choose the safe model $\text{safe}(c)$?* It seems natural to include models that are not heavily influenced by c since otherwise this might allow copyright breaching. However, such choice of $\text{safe}(c)$ leads to the discussed limitations encountered by algorithmic-stability approaches such as DP. It is true that some aspects, such as content safety vs. model safety, can be better aligned through the definition of NAF but also, as Proposition 1 shows, through variants of DP. Overall, there is room, then, to further investigate the different possible models for copyright, within such an approach, but we should take into account the limitations presented in Section 3.

Perhaps a more exciting application of NAF, then, is to consider notions of safety that allow some influence by c . e.g. to enable generating parodies, fair-use, de minimis copying, etc. We consider then safety functions that now *do* have access to c , and exploit this access to enable only allowed influence. Here we face a different challenge. Suppose that $q_c, q_{c'}$ are such a safe models for contents c and c' respectively. If $q_{c'}$ and q_c are far away, then Proposition 1 shows that there is no hope to output a NAF model. But even if q_c and $q_{c'}$ are not far away, but suppose that $q_{c'}$ ignores content c , then for any content z that is influenced by c we may assume that:

$$q_c(z) \gg q_{c'}(z).$$

But, if p is a NAF model, we must also have due to Equation (2) with respect to c' and z :

$$q_c(z) \gg p(z).$$

In other words, the NAF model censors permissible content z even though it is safe. This happens because z is an improbable event in model $q_{c'}$. Not because z breaches copyright of c' but because it is influenced by c , and content that is influenced by c is discarded by safe models that had no access to c . It follows, then, that all safe models must treat protected content in a similar manner, and $q_{c'}$ must also be influenced by c if we expect the NAF model to make any use of it. Hence, it is unclear if a more refined notion of safe may help circumvent the hurdles of applying the privacy approach for establishing a copyright infringement. This suggests, though, to perhaps consider a relaxed variant of NAF in which a content is discarded by a safe model only when certain links between the protected content and the generated content are established.

It seems, then, that an algorithmic approach that assists jurists in understanding such links between existing works of authorship, study their hidden interconnection, and quantify their originality holds a great promise. In other words, rather than constructing binary legal rules (e.g., aiming to devise a definitive test for copyright infringement), algorithmic stability approaches could facilitate new quantifiable methods for applying legal standards, such as



■ **Figure 1** The Prince series.

measuring originality [24]. From this perspective, originality is evaluated by the semantic distance between the elements of a measured expressive work and similar elements found in the corpus of the training content. The more salient the expressive elements within the larger corpus of pre-existing content, the less likely these elements are to be considered original by copyright law, and the more likely copyright law is to legitimize drawing upon them by the output content.

Research in this area is still in its infancy but holds outstanding potential for the copyright system [52, 26]. Algorithmic approaches that focus on the element level rather than the content level, and are applied not as binary tests for apprising infringement but as tools for measuring copyright originality may greatly empower the legal profession. As the extensive body of legal scholarship has long acknowledged, the originality standard in copyright law, along with many of its related doctrines for delineating scope (such as the “idea-expression dichotomy”), is inherently vague and uncertain [37, 35, 29]. Such vagueness leads to inconsistent judicial precedent, deters permissible uses of copyrighted material, and undermines the goals of copyright law [21, 42, 38, 48, 57].

References

- 1 Omer Angel and Yinon Spinka. Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*, 2019.
- 2 Clark D Asay. Independent creation in a world of ai. *FIU L. Rev.*, 14:201, 2020.
- 3 AVs.S. Andersen et al v. Stability AI Ltd. et al, Docket No. 3:23-cv-00201 (N.D. Cal. Jan 13, 2023), 2023.
- 4 Yochai Benkler. Free as the air to common use: First amendment constraints on enclosure of the public domain. *NyuL Rev.*, 74:354, 1999.
- 5 Olivier Bousquet, Roi Livni, and Shay Moran. Synthetic data generators—sequential and private. *Advances in Neural Information Processing Systems*, 33:7114–7124, 2020.

3:14 Can Copyright Be Reduced to Privacy?

- 6 Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 369–380, 2016.
- 7 Bvs.S. Baker v. Selden, 101 U.S 99, 1879.
- 8 Bvs.S. Brown Bag Software v. Symantec Corp., 960 F.2d 1465, 1472 (9th Cir. 1992)), 1992.
- 9 Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Schwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.
- 10 Julie E Cohen. *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press, 2012.
- 11 U.S CONST. U.S CONST. art. I, 8, cl. 8, ().
- 12 Cvs.A. Campbell v. Acuff-Rose Music, Inc., 510 U.S 569, 578, 1994.
- 13 Dvs.G. DOE 1 et al v. GitHub, Inc. et al class action, 2022.
- 14 Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Conference On Learning Theory*, pages 1693–1702. PMLR, 2018.
- 15 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, 2012.
- 16 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.
- 17 Niva Elkin-Koren. Cyberlaw and social change: A democratic approach to copyright law in cyberspace. *Cardozo Arts & Ent. LJ*, 14:215, 1996.
- 18 Niva Elkin-Koren. Copyright in a digital ecosystem: a user-rights approach. *Forthcoming in RUTH OKEDIJI, COPYRIGHT IN AN AGE OF LIMITATIONS AND EXCEPTIONS (2015)*, 2015.
- 19 Giorgio Franceschelli and Mirco Musolesi. Deepcreativity: measuring creativity with deep learning techniques. *Intelligenza Artificiale*, 16(2):151–163, 2022.
- 20 Fvs.R. Feist Publ’ns, Inc. v. Rural Tel. Serv. Co., 499 US 340, 345, 1991.
- 21 James Gibson. Risk aversion and rights accretion in intellectual property law. *Yale LJ*, 116:882, 2006.
- 22 Avs. Google. Authors Guild v. Google, Inc., 804 F.3d 202, 207–08, 225 (2d Cir. 2015)), 2015.
- 23 James Grimmelman. Copyright for literate robots. *Iowa L. Rev.*, 101:657, 2015.
- 24 Uri Y Hacohen and Niva Elkin-Koren. Copyright regenerated: Harnessing genai to measure originality and copyright scope. *Harvard Journal of Law & Technology*, 2024.
- 25 Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.07758*, 2022.
- 26 Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.
- 27 Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.
- 28 Hvs.R. Harper & Row v. Nation Enterprises, 471 U.S 539, 1985.
- 29 Richard H Jones. The myth of the idea/expression dichotomy in copyright law. *Pace L. Rev.*, 10:551, 1990.
- 30 Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- 31 Louis Kaplow. Rules versus standards: An economic analysis. *Duke Law Journal*, 42(3):557–629, 1992.

- 32 Aleksandra Korolova, Krishnamurthy Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009.
- 33 Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin”bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.
- 34 Legislation and Legal Counsel (Civil Law). *OPINION: USES OF COPYRIGHTED MATERIALS FOR MACHINE LEARNING*. State of Israel Ministry of Justice, 2022.
- 35 Mark A Lemley. Our bizarre system for proving copyright infringement. *J. Copyright Soc’y USA*, 57:719, 2009.
- 36 Mark A Lemley and Bryan Casey. Fair learning. *Tex. L. Rev.*, 99:743, 2020.
- 37 Jessica Litman. The public domain. *Emory Lj*, 39:965, 1990.
- 38 Jessica Litman. Billowing white goo. *Colum. JL & Arts*, 31:587, 2007.
- 39 L.vs.B. Lotus Dev. Corp. v. Borland Int’l, Inc., 49 F.3d 807, 815 (1st Cir. 1995); Lotus Dev. Corp. v. Borland Int’l, Inc., 516 U.S 233 (1996) , 1996.
- 40 Peter S Menell and Michael J Meurer. Notice failure and notice externalities. *Journal of Legal Analysis*, 5(1):1–59, 2013.
- 41 Mvs.S. Mazer v. Stein, 347 U.S 201, 219] , 1954.
- 42 Neil Weinstock Netanel. *Copyright’s paradox*. Oxford University Press, 2008.
- 43 Neil Weinstock Netanel. Making sense of fair use. *Lewis & Clark L. Rev.*, 15:715, 2011.
- 44 N.vs.C. Nash v. CBS, Inc., 899 F.2d 1537 (7th, cir., 1990), 1990.
- 45 U.S Copyright Office. Works Not Protected by Copyright, 2021.
- 46 Gideon Parchomovsky and Alex Stein. Originality. *Va. L. Rev.*, 95:1505, 2009.
- 47 Matthew Sag. The new legal landscape for text mining and machine learning. *J. Copyright Soc’y USA*, 66:291, 2018.
- 48 Pamela Samuelson. The copyright grab. *Wired Magazine*, 4, 1996.
- 49 Pamela Samuelson. Reconceptualizing copyright’s merger doctrine. *J. Copyright Soc’y USA*, 63:417, 2016.
- 50 Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023.
- 51 Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing human ingenuity: A quantitative framework for coyright law’s substantial similarity. *arXiv preprint arXiv:2206.01230*, 2022.
- 52 Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, 2022.
- 53 Svs.M. Sid & Marty Krofft TV Prod., Inc. v. McDonald’s Corp., 562 F.2d 1157, 1164 (9th Cir. 1977), 1977.
- 54 Svs.W. SAS Institute Inc. v. World Programming Ltd., 64 F. Supp. 3d 755, 762 , 2014.
- 55 U.S.C. 17 U.S.C. § 102(b), 2006.
- 56 Uvs.K. Universal City Studios v. Kamar Industries, Inc., 217 USPQ. (BNA) 1165 (S.D Tex 1982), 1982.
- 57 Siva Vaidhyanathan. Copyrights and copywrongs. In *Copyrights and Copywrongs*. New York University Press, 2001.
- 58 N vs. U. Nichols v. Universal Pictures Corporation, 45 F.2d 119, (2st Cir., 1930) , 1930.
- 59 Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.
- 60 Wvs.G. Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith (Docket 21–869)]. .

A Proofs**A.1 Proof of Proposition 2**

Suppose that

$$\|q_1 - q_2\| \geq \alpha.$$

In particular there exists an event E such that:

$$q_2(E) \leq q_1(E) - \alpha \leq 1 - \alpha.$$

Let p be some distribution. We assume that $p(E) \geq 1/2$ (otherwise, replace E with its complement and q_1 and q_2 replace roles). Thus, we have that:

$$p(E) \geq \frac{1}{2} \geq \frac{1}{2(1-\alpha)} q_2(E).$$

In particular, for some $z \in E$, the result follows.

A.2 Proof of Proposition 1

The proof relies on a coupling Lemma, taken from [1]. Recall that, given a collection of distribution measures Q , a coupling can be thought of as a collection of random variables $X = (X_q)_{q \in Q}$, whose marginal distributions are given by q . I.e. $\mathbb{P}(X_q = x) = q(x)$:

► **Lemma 3** (A special case of Thm 2 in [1]). *Let Q be the collection of all posteriors over a finite domain \mathcal{X}^1 . There exists a coupling such that for every $q, q' \in Q$:*

$$\mathbb{P}(X_q \neq X_{q'}) \leq \frac{2\|q - q'\|}{1 + \|q - q'\|}.$$

The second Lemma we rely on is a private heavy hitter mechanism, described as follows:

► **Lemma 4** ([32, 6]). *Let Z be a finite data domain. For some*

$$k \geq \Omega\left(\frac{\log 1/\eta\beta\delta}{\eta\epsilon}\right),$$

there exists an (ϵ, δ) -DP algorithm hist , such that with probability $(1 - \beta)$ on an inputs $S = \{z_1, \dots, z_k\}$ outputs a mapping $a \in [0, 1]^Z$, such that, for every $z \in Z$,

$$|a(z) - \text{freq}_S(z)| \leq \eta.$$

In particular, if $\text{freq}_S(z) > 0$, then $a(z) > 0$.

Where we denote by $\text{freq}_S(z) = \frac{|i:z_i=z|}{|S|}$.

We next move on to prove the claim. Let X be the coupling from Lemma 3. Our private algorithm works as follows:

1. First, we take $\beta = \eta$, and set

$$k = \Omega\left(\frac{\log 1/\eta^2\delta}{\eta\epsilon}\right).$$

To be as in Lemma 4.

¹ which are all absolutely continuous w.r.t the uniform distribution

2. Divide S , the input sample, to k , disjoint datasets S_1, \dots, S_k of size m . Each data set, via A , defines a model $q_{S_i}^A$.
3. Next, we define the random sample

$$S_X = \{X_{q_{S_1}^A}, X_{q_{S_2}^A}, \dots, X_{q_{S_k}^A}\} \in Z^K.$$

4. Apply the mechanism in Lemma 4 and output $a \in [0, 1]^Z$ such that, w.p. $1 - \eta$, for all $z \in Z$:

$$|a(z) - \text{freq}_{S_X}(z)| \leq \eta.$$

5. Let p be any arbitrary distribution such that for every $z \in Z$:

$$|a(z) - p(z)| \leq \eta \tag{3}$$

(if no such distribution exists p is any distribution). and output

$$q_S^B = p.$$

Notice that each sample z_j affects only a single sub-sample S_i and in turn only a single random variable $X_{q_{S_i}^A}$. The histogram function a is then (ϵ, δ) -DP w.r.t to its input S . The output p , by processing is also private. We obtain, then, that the above algorithm is (ϵ, δ) -private.

We next set out to prove that $p = q_S^B$ is close in TV distance to $q_{S_A}^A$ in expectation. For ease of notation let us denote $X_i = X_{q_{S_i}^A}$. Notice that, with probability $(1 - \eta)$, for every z :

$$|a(z) - \text{freq}_{S_X}(z)| \leq \eta,$$

in particular, there is a p that satisfies the requirement in Item 5 (i.e. freq_{S_X} defines such a distribution) and Equation (3) is satisfied. We then have that for every z :

$$\left| p(z) - \frac{1}{k} \sum \mathbf{1}[X_i = z] \right| \leq |p(z) - a(z)| + \left| a(z) - \frac{1}{k} \sum \mathbf{1}[X_i = z] \right| \leq 2\eta. \tag{4}$$

We now move on to bound the total variation between the model $\mathbb{E}[q_S^B]$ and q_{S_A} , where expectation is taken over the randomness of B .

To show this, we will use the reverse inequality of the coupling Lemma, in particular if (\hat{X}_B, \hat{X}_A) is a coupling of q_S^B and $q_{S_A}^A$ (where S and S_A are now fixed), then:

$$\|\mathbb{E}[q_S^B] - q_{S_A}^A\| \leq \mathbb{P}(\hat{X}_B \neq \hat{X}_A). \tag{5}$$

Our coupling will work as follows, first we output $p = q_S^B$ and sample $\hat{X}_B \sim p$, and we let $\hat{X}_A = X_{q_{S_A}^A}$. This defines a coupling (\hat{X}_B, \hat{X}_A) . Applying Equation (4), with $z = \hat{X}_A$, exploiting the fact that Equation (4) holds with probability at least $1 - \eta$:

$$\begin{aligned} \mathbb{P}(\hat{X}_B \neq \hat{X}_A) &\leq \frac{1}{k} \sum_{i=1}^k \mathbb{P}(X_i \neq X_{q_{S_A}^A}) + \eta \\ &\leq 2\eta + \eta. \end{aligned}$$

And we have that:

$$\mathbb{P}(\hat{X}_B \neq \hat{X}_A) \leq \frac{1}{k} \sum_{i=1}^k \mathbb{P}(X_i \neq X_{q_{S_A}^A}) + 3\eta \leq \frac{1}{k} \sum_{i=1}^k \frac{2\|q_{S_i}^A - q_{S_A}\|}{1 + \|q_{S_i}^A - q_{S_A}\|} + 3\eta.$$

3:18 Can Copyright Be Reduced to Privacy?

And,

$$\begin{aligned}
 \mathbb{E}_{S_A, S} \| \mathbb{E}[q_S^B] - q_{S_A} \| &\leq \mathbb{E}_{S_A, S} \frac{1}{k} \sum_{i=1}^k \left[\frac{2 \| q_{S_i}^A - q_{S_A} \|}{1 + \| q_{S_i}^A - q_{S_A} \|} \right] + 3\eta \\
 &\leq \mathbb{E}_{S_1, S_2 \sim S} \left[\frac{2 \| q_{S_1}^A - q_{S_2} \|}{1 + \| q_{S_1}^A - q_{S_2} \|} \right] + 3\eta \\
 &\leq \left[\frac{2 \mathbb{E}[\| q_{S_1}^A - q_{S_2} \|]}{1 + \mathbb{E}[\| q_{S_1}^A - q_{S_2} \|]} \right] + 3\eta && \text{concavity of } \frac{2x}{1+x} \\
 &\leq \left[\frac{2\alpha}{1+\alpha} \right] + 3\eta && \text{monotonicity } \frac{2x}{1+x}
 \end{aligned}$$

Balanced Filtering via Disclosure-Controlled Proxies

Siqi Deng¹ ✉

Amazon AWS AI, Palo Alto, CA, USA

Emily Diana ✉

Toyota Technological Institute at Chicago, IL, USA

Michael Kearns ✉

University of Pennsylvania, Philadelphia, PA, USA

Amazon AWS AI, Palo Alto, CA, USA

Aaron Roth ✉

University of Pennsylvania, Philadelphia, PA, USA

Amazon AWS AI, Palo Alto, CA, USA

Abstract

We study the problem of collecting a cohort or set that is *balanced* with respect to sensitive groups when group membership is unavailable or prohibited from use at deployment time. Specifically, our deployment-time collection mechanism does not reveal significantly more about the group membership of any individual sample than can be ascertained from base rates alone. To do this, we study a learner that can use a small set of labeled data to train a proxy function that can later be used for this filtering or selection task. We then associate the range of the proxy function with sampling probabilities; given a new example, we classify it using our proxy function and then select it with probability corresponding to its proxy classification. Importantly, we require that the proxy classification does not reveal significantly more information about the sensitive group membership of any individual example compared to population base rates alone (i.e., the level of disclosure should be controlled) and show that we can find such a proxy in a sample- and oracle-efficient manner. Finally, we experimentally evaluate our algorithm and analyze its generalization properties.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Security and privacy → Human and societal aspects of security and privacy

Keywords and phrases Algorithms, Sampling, Ethical/Societal Implications

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.4

Related Version *Full Version*: <https://arxiv.org/abs/2306.15083>

1 Introduction

There are a variety of situations in which we would like to select a cohort or set that is *balanced* or *representative* (having an approximately equal number of samples from different groups) with respect to race, sex, or other sensitive attributes – but, we cannot explicitly select based on these attributes. This could be because the attributes are sensitive so were never collected, they could be redacted from the information we see, they could be too resource intensive to collect, or selecting based on these attributes could be illegal.

Consider the context of college admissions. Out of many qualified applicants, a college may prioritize racial diversity when deciding upon the final cohort to admit. However, in the United States Supreme Court decision for *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*, it was determined that “Harvard’s and UNC’s

¹ Corresponding author



[race-conscious] admissions programs violate the Equal Protection Clause of the Fourteenth Amendment” [36]. How might a college select a racially diverse cohort with affirmative action prohibited?

Our approach is based on training a proxy classifier – in the form of a decision-tree – with the following properties: (1) The set of points classified at each leaf should not be strongly correlated with the protected attribute (the *disclosure-control* part) and (2) The set of distributions on the protected attributes induced at each leaf should be such that the uniform distribution on protected attributes is in their convex hull (the *balancing* part). The second condition allows us to assign sampling probabilities to the leaves such that if we accept each example with probability corresponding to its proxy classification, in expectation the selected cohort will be balanced with respect to the protected attribute.²

1.1 Related Work

The proxy problem is a subject of ongoing debate in the philosophy of science and causal inference literatures (e.g. [18, 3, 32, 24, 41, 29, 31, 9]), and our work engages with this literature methodologically – we do not believe that it is our role to take a philosophical or legal stance but rather to broaden the set of available tools. Using proxy variables for sensitive attributes in settings where diversity or equity is a concern has been standard practice, yet in many cases, existing features are chosen for the proxies (such as surname, first name, or geographic location [13, 40, 44]). Rather than using an existing feature as a proxy, we propose deliberately *constructing* a proxy. Several works take this perspective – in [10], for example, the authors produce a proxy that can be used during training to build a fair model downstream. But often, proxies for protected attributes are explicitly intended to be good predictors for those attributes; it is not clear that using an accurate “race predictor” is an acceptable solution to making decisions in which race should not be used (and is often explicitly prohibited). Our primary point of departure is that we train a model to make classifications that are *minimally correlated* with the protected attribute.

While our intended use cases are primarily curation or cohort selection, one may also use our method for collecting balanced data sets for machine learning applications. However, we recommend caution in these scenarios, as our approach does not give guarantees about the level of distortion of the final filtered data set. In order to provide comparisons to existing empirical techniques, however, we do measure our approach against a common data pre-processing technique, SMOTE (Synthetic Minority Oversampling Technique) [8]. Other re-sampling methods for data balancing include ADASYN [17], MIXUP [43], SMOTE adaptations [26, 4, 5, 11, 15, 23]) and cluster-based approaches that under-sample disproportionately represented classes [16, 42, 33, 21, 34]. In the causal literature, propensity score re-weighting [22] is also a popular approach to account for group size differences. Each of these techniques, however, requires access to the sensitive attribute. Our approach’s primary point of departure is that we do not use the sensitive attribute – or a direct prediction or imputation of it – at the final collection time when we are deploying our method.

² A natural first approach is to add noise to the predictor for the protected attribute, against which we compare. However, we are motivated by the need to have strategies that never involve training a classifier for the protected attribute, especially if it could be used outside of the intended system.

1.2 Limitations and Discussion

Our contributions are twofold. First, for situations where balance is desired but disclosure is not a concern, we introduce a sampling scheme optimized to collect a balanced cohort. Second, for when disclosure is a concern, we present a method to produce a proxy function for the balanced selection task that is *guaranteed* not to be too disclosive. Below, we discuss important considerations having to do with appropriate usage of our methodology, limitations, and areas for expansion.

- **The sensitive attribute is still used to inform the proxy, and our approach relies on accessing a small sample of data with this attribute:** The proxy training algorithm we propose is not blind to the sensitive attributes, which it must access during *training*. Rather, the proxy does not use these attributes at the time of *deployment*.³ We emphasize that there is no contradiction between (1) being able to obtain (once) a small data set labeled with sensitive attributes and then using it to train a classification algorithm (in this case, our proxy model) and (2) having the inability to collect or use sensitive attributes when gathering the bulk of one’s data. This is especially true when the final selection criterion is not closely correlated with the sensitive attribute, which is one of our primary objectives. In the algorithmic fairness literature in particular, there is a substantial and growing body of work on learning classifiers that satisfy fairness constraints by sensitive attributes but that do not use these attributes at test time (e.g. [1, 20, 27, 28]). *These methods still require access to the attributes at training time.* The distinction between using sensitive attributes at train versus test time is essential. In certain financial applications in the United States, using race or gender at test time (i.e., when making lending decisions) is illegal. But it is not illegal to use these attributes at training time to audit models for statistical bias and to remove it if found. The distinction in our case is similar: We use these attributes to find a statistical selection criterion but do not use sensitive attributes of individuals to make selection decisions about them.
- **The filtered cohort or data set will likely exhibit within-group distortion:** This is an important consideration that should be taken into account when using our method. Our theoretical guarantees provide bounds on the level of balance and disclosure when measured with respect to the sensitive attributes, but they do not guarantee that the distribution over selected individuals matches that of the true population. In fact, this is perhaps a necessary effect of our process and in many cases may be natural. For example, in the context of college admissions or interview selection, a university or firm is intentionally selecting a pool that is *not* representative of the base population. The use cases for which this quality may create the greatest challenge is in curating data sets for training machine learning models. While our method can improve *representation* in data sets, it will not necessarily lead to improvements in downstream fairness of models trained on the balanced data. We provide a detailed analysis of this in Appendix B.⁴
- **Affirmative Action and Legal Challenges:** We do not propose our method as a way to circumvent the intent of legislation, nor do we make claims regarding the legal or moral appropriateness of its usage in any particular affirmative action setting. Rather, we view it as a tool that can be used, when permitted legally, in settings where diversity or

³ It would be impossible to give an algorithm making no use of the protected attribute during deployment or training and yet promising any sort of balance – it would have to behave identically on any distributions with the same marginals over non-protected attributes, even if they differed on the protected attribute.

⁴ One note, however, is that our method does allow for balancing multiple attributes at a time – therefore, one could ask for a cohort that has the same number of positive and negative examples in each group.

balance is desired but when the sensitive attribute can or should be used only minimally. For example, in the college admissions example, it is also undesirable to use explicit race-based predictors in lieu of observing the sensitive attribute. However, students can include race considerations in their admissions essays, which admissions officers see. Given the stakes of college admissions and the strategic behavior of both sides (applicants and schools), it is very likely that an ad-hoc system will still develop to indirectly make use of racial information for the sake of diversity. In a situation such as this, our method provides a controlled way to achieve such desiderata without unintentionally revealing too much information or making use of highly correlated proxies.

2 Model and Preliminaries

Let $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ be an arbitrary data domain and \mathcal{P} be the probability distribution over Ω . \mathcal{P}_x will refer to the marginal distribution over \mathcal{X} , \mathcal{P}_z will refer to the marginal distribution over \mathcal{Z} , and $\mathcal{P}_{z|x}$ will refer to the conditional distribution over $\mathcal{Z}|\mathcal{X}$. Each data point is a triplet $\omega = (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $x \in \mathcal{X}$ is the non-sensitive feature vector and $y \in \mathcal{Y} = \{0, 1\}$ is the label. The label y is not required in training or applying our filtering method, but it will be used in the analysis of downstream fairness effects of the filtering process provided in Appendix B. In the paper body, therefore, we omit it for clarity.

Unless otherwise specified, we take group membership as disjoint such that $z \in \mathcal{Z} = [K]$ is an integer indicating sensitive group membership, but our framework can easily be extended to the case where group membership need not be disjoint. We consider the uniform distribution, U , to be our target distribution over sensitive attributes, where $U = (\frac{1}{K}, \dots, \frac{1}{K})$. We also provide a brief extension to the intersecting case below.

We imagine we can sample unlimited data from \mathcal{P}_x , but the corresponding value z can only be obtained from self-report, authorized agencies, or human annotation. We use r_k to denote the base rate $\Pr_{z \sim \mathcal{P}_z}[z = k]$ in the underlying population distribution. We also assume that we can obtain a limited sample of data D of n samples $\{(x_i, z_i)\}_{i=1}^n \subset \Omega$ for which we can observe the true sensitive attribute z . We would like to use this sample to collect a much larger set $S \subset \Omega$ such that even if we cannot observe the sensitive attributes, S is balanced with respect to z .

Formally, we define a balanced set as follows, where $\Pr_{z \sim S}[z = k] = \frac{1}{|S|} \sum_{i=1}^{|S|} \mathbf{1}_{z_i=k}$ is the empirical distribution of z drawn uniformly from S .

► **Definition 1 (Balance).** *A set S is balanced with respect to K disjoint groups $z \in [K]$ if $\Pr_{z \sim S}[z = k] = \frac{1}{K} \forall k$.*

Due to finite sampling, Definition 1 will rarely be met, even if the underlying *distribution* is uniform over sensitive attributes. Therefore, we also discuss approximate balance:⁵

► **Definition 2 (β -Approximate Balance).** *A set S is β -approximately balanced with respect to K disjoint groups $z \in [K]$ if $\|(\Pr_{z \sim S}[z = 1], \dots, \Pr_{z \sim S}[z = K]) - (1/K, \dots, 1/K)\|_2 \leq \beta$.*

Note that β -approximate balance involves the *distribution* of sensitive groups in S : as this distribution deviates farther from the uniform, the imbalance, β , increases.

In the intersecting groups case, we let the sensitive attribute domain \mathcal{Z} be a binary vector of length K , such that $z \in \mathcal{Z} = \{0, 1\}^K$. We will assume \mathcal{Z} is composed of G group classes $\{Z_i\}_{i=1}^G$ (e.g., sex, race, etc), and each group class Z_i has K_i groups. Thus, the vector z will

⁵ We use the L_2 norm due to operational reasons, as it allows us to make useful geometric arguments.

indicate all possible group memberships, where the length of the vector of group memberships is $K = \sum_{i=1}^G K_i$. We also replace U with $U_{\text{int}} = (\frac{1}{K_1}, \dots, \frac{1}{K_1}, \dots, \frac{1}{K_G}, \dots, \frac{1}{K_G})$ to indicate the target distribution over intersecting sensitive attributes.

► **Definition 3** (Multi-Class Balance). *We will say that a set S is balanced with respect to K intersecting groups if, for any group class $\{Z_i\}_{i=1}^G$ composed of groups $\{Z_{i_j}\}_{j=1}^{K_i}$, $\Pr_{z \sim S}[z[Z_{i_j}] = 1] = \frac{1}{K_i} \forall j$.*

► **Definition 4** (β -Approximate Multi-Class Balance). *We say that a set S is β -approximately balanced with respect to K intersecting groups if*

$$\|(\Pr_{z \sim S}[z[Z_{1_1}] = 1], \Pr_{z \sim S}[z[Z_{1_2}] = 1], \dots, \Pr_{z \sim S}[z[Z_{G_{K_G-1}}] = 1], \Pr_{z \sim S}[z[Z_{G_{K_G}}] = 1]) - U_{\text{int}}\|_2 \leq \beta$$

► **Remark 5.** This definition aims to take into account the fact that for different group classes, there may be a different number of potential groups. For example, there may only be two sex groups but eight income groups. Asking that the representation of each of those categories be one-tenth of the final sample would not make sense. However, our definition does not prevent certain intersections being more represented than others. When the number of groups is small, this can always be dealt with by using the Cartesian product over group classes to enumerate intersectional groups.

In addition to desiring that our proxy allows us to select an approximately *balanced* cohort, we would also like the classification outcomes of the proxy not to be overly disclosive. We model this by asking that the posterior distribution on group membership is close to the prior distribution when conditioning on the outcome of the proxy classifier.

► **Definition 6** (α -Disclosive Proxy). *A proxy g is α -disclosive (or has disclosure level at most α) on set S if, for all sensitive groups k and proxy values i , $|\Pr_{z|x \sim S}[z = k|g(x) = i] - \Pr_{z \sim S}[z = k]| \leq \alpha$.*

For any proxy function g , we can analyze the distribution of sensitive groups amongst points mapped to each value k in the range of the proxy. Call this conditional distribution a_k , let l be the number of unique elements in the range of the proxy, and let A be the $l \times K$ matrix whose k^{th} row is a_k . Denote the convex hull, defined in Definition 8, of the rows of A by $C(A)$. Then, we can add a notion of *balance* into our proxy definition in the following way:

► **Definition 7** ((α, β) Proxy). *$g : \mathcal{X} \rightarrow \mathbb{N}$ is an (α, β) proxy if it is α -disclosive and $\inf_{U' \in C(A)} \|U' - U\|_2 \leq \beta$.*

Here, $\inf_{U' \in C(A)} \|U' - U\|_2$ indicates the Euclidean distance between U and the closest point in $C(A)$. We will slightly abuse notation and refer to this as the distance $\|C(A) - U\|_2$. The disclosure parameter α controls the amount of additional information the proxy gives about group membership, while the balance parameter β quantifies the minimum distance from uniform achievable with any acceptance probabilities for a given proxy. There do exist limitations on how small α can be if we need full balance. With K sensitive groups, the final frequency of each group must be $\frac{1}{K}$ if we desire $\beta = 0$: if there is a group with initial frequency f , there is no avoiding that $\alpha \geq |f - \frac{1}{K}|$. For some data sets, this unavoidably can be quite large. For example, consider a data set of $\frac{1}{3}$ men and $\frac{2}{3}$ women and assume the proxy g takes values 0 or 1. Of the samples mapped to $g = 0$, $\frac{1}{4}$ are men and $\frac{3}{4}$ are women. Of those mapped to $g = 1$, $\frac{1}{2}$ are men and $\frac{1}{2}$ are women. Then, $\alpha = \frac{1}{6}$, because $|\Pr[z = \text{men}|g = 0] - \Pr[z = \text{men}]| = |\frac{1}{3} - \frac{1}{2}| = \frac{1}{6}$. In this example, β would be 0, because the convex hull of the conditionals $\Pr[z|g(x)]$ contains $[\frac{1}{2}, \frac{1}{2}]$. Next, we provide definitions for a convex hull and stochastic vector.

► **Definition 8** (Convex Hull [7]). *The convex hull of a set of points S in K dimensions is the intersection of all convex sets containing S . For l points s_1, \dots, s_l , the convex hull C is given by the expression: $C \equiv \{\sum_{i=1}^l q_i s_i : q_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^l q_i = 1\}$*

► **Definition 9** (Stochastic Vector [6]). *$v = (v_i)_{i=1}^\ell$ is a stochastic vector if $\sum_{i=1}^\ell v_i = 1$ and $v_i \geq 0 \forall i$.*

Finally, we observe a necessary and sufficient condition for U to be in $C(A)$.

► **Lemma 10** (Inclusion in Convex Hull). *Let A be an $l \times K$ matrix and U be a $1 \times K$ vector with $\frac{1}{K}$ in each entry. There exists a stochastic vector q such that $qA = U$ if and only if $U \in C(A)$.*

Proof. Let a_i denote the i^{th} row of A . If U is in $C(A)$, then by Definition 8, there exists a non-negative vector q such that $\sum_{i=1}^l q_i a_i = U$ and $\sum_{i=1}^l q_i = 1$. Similarly, for any stochastic q , $qA \in C(A)$. Then if $qA = U$, $U \in C(A)$. ◀

Finally, we outline several key results that we will use in the derivations and proofs for our proxy training algorithm. We begin by considering a zero-sum game between two players, a Learner with strategies in S_1 and an Auditor with strategies in S_2 . The payoff function of the game is $W : S_1 \times S_2 \rightarrow \mathbb{R}_{\geq 0}$.

► **Definition 11** (Approximate Equilibrium [14]). *A pair of strategies $(s_1, s_2) \in S_1 \times S_2$ is said to be a ν -approximate minimax equilibrium of the game if the following conditions hold: $U(s_1, s_2) - \min_{s'_1 \in S_1} U(s'_1, s_2) \leq \nu$, $\max_{s'_2 \in S_2} U(s_1, s'_2) - U(s_1, s_2) \leq \nu$*

Freund and Schapire [14] show that if a sequence of actions for the players jointly has low regret, the uniform distribution over each player's actions forms an approximate equilibrium:

► **Theorem 1** (No-Regret Dynamics [14]). *Let S_1 and S_2 be convex, and suppose $W(\cdot, s_2) : S_1 \rightarrow \mathbb{R}_{\geq 0}$ is convex for all $s_2 \in S_2$ and $W(s_1, \cdot) : S_2 \rightarrow \mathbb{R}_{\geq 0}$ is concave for all $s_1 \in S_1$. Let $(s_1^1, s_1^2, \dots, s_1^T)$ and $(s_2^1, s_2^2, \dots, s_2^T)$ be sequences of actions for each player. If for $\nu_1, \nu_2 \geq 0$, the regret of the players jointly satisfies*

$$\sum_{t=1}^T W(s_1^t, s_2^t) - \min_{s_1 \in S_1} \sum_{t=1}^T W(s_1, s_2^t) \leq \nu_1 T \quad \max_{s_2 \in S_2} \sum_{t=1}^T W(s_1^t, s_2) - \sum_{t=1}^T W(s_1^t, s_2^t) \leq \nu_2 T$$

then the pair (\bar{s}_1, \bar{s}_2) is a $(\nu_1 + \nu_2)$ -approximate equilibrium, where $\bar{s}_1 = \frac{1}{T} \sum_{t=1}^T s_1^t \in S_1$ and $\bar{s}_2 = \frac{1}{T} \sum_{t=1}^T s_2^t \in S_2$ are the uniform distributions over the action sequences.

Additionally, we define a Cost Sensitive Classification (CSC) oracle over a classification model class \mathcal{H} , which we will use as an efficient subroutine in our algorithm.

► **Definition 12** (Weighted Cost-Sensitive Classification Oracle for \mathcal{H} [2]). *An instance of a Weighted Cost-Sensitive Classification problem, or a CSC problem, for the class \mathcal{H} , is given by a set of n tuples $\{w(x_i), x_i, c_i^0, c_i^1\}_{i=1}^n$ such that c_i^1 corresponds to the cost for predicting label 1 on sample x_i and c_i^0 corresponds to the cost for prediction label 0 on sample x_i . The weight of x_i is denoted by $w(x_i)$. Given such an instance as input, a $CSC(\mathcal{H})$ oracle finds a hypothesis $h \in \mathcal{H}$ that minimizes the total cost across all points: $h \in \operatorname{argmin}_{h' \in \mathcal{H}} \sum_{i=1}^n w(x_i) [h'(x_i)c_i^1 + (1 - h'(x_i))c_i^0]$.*

3 Computing Sampling Weights from a Proxy (QP Approach)

Now we introduce our first methodological contribution: a selection approach for producing a balanced set *given* a proxy. At a high level, our approach involves mapping each example to an acceptance probability. We construct such a mapping by labeling the range of the proxy $g : \mathcal{X} \rightarrow \mathbb{N}$ with acceptance probabilities and then selecting samples for our set by applying the proxy function to a sample and keeping it with probability corresponding to the element of the range of the proxy that the point maps to.

Recall the condition distribution matrix A , where each row represents the distribution of z values mapped to a given proxy value. Our goal is to find acceptance probabilities such that the induced distribution on retained points is uniform over the protected attributes. By Lemma 10, such probabilities exist if A contains the uniform distribution in its convex hull. Consider the system $qA = U$, where $U = (\frac{1}{K}, \dots, \frac{1}{K})$ and q must be a length ℓ stochastic vector. If there is a solution for q , we consider this a valid acceptance rate scheme and use it to derive the selection probabilities for our filtering problem. If there is not an exact solution (which will happen frequently) we take $\operatorname{argmin}_q \|qA - U\|_2$ as our best acceptance rate scheme. *Because this involves solving a quadratic program, we refer to the proxies and accompanying selection schemes produced by this approach as QP (Quadratic Program) proxies.*

■ Algorithm 1 Finding Acceptance Probabilities ρ .

Input: proxy g , $D = \{(x_i, z_i)\}_{i=1}^n$, number of sensitive groups K
for j in $\operatorname{Range}(g)$ **do**
 For k in $[K]$, let $a_k = \Pr_{z|x \sim D}[z = k | g(x) = j]$
 Let $\hat{r}_j = \Pr_{x \sim D}[g(x) = j]$
 Let A be the matrix with k^{th} row a_k and let $U = (\frac{1}{K}, \dots, \frac{1}{K})$
 $q = \operatorname{argmin}_q \|qA - U\|_2$ s.t. $q_i \geq 0$ and $\sum q_i = 1$
 For j in $\operatorname{Range}(g)$, set $\rho_j = \frac{q_j}{\hat{r}_j}$
 Let $C = \max_j \rho_j$ and normalize $\rho_j = \rho_j / C$
return ρ, A

■ Algorithm 2 Filtering with ρ .

Input: g, ρ, \mathcal{P}_x
 Draw $x \sim \mathcal{P}_x$ and compute $g(x)$
 With probability $\rho_{g(x)}$, accept x into sample

► **Lemma 13** (Filtering According to ρ). *Consider acceptance probabilities ρ and conditional distribution matrix A returned by Algorithm 1. Then, if $U \in C(A)$, filtering according to ρ as in Algorithm 2 induces a uniform distribution over protected attributes.*

Proof. We want to show that the distribution over sensitive attributes in the *filtered set* is uniform. We begin by expressing the distribution over sensitive attributes in the filtered set constructively, as the distribution obtained from sampling according to ρ . From there, we plug in our definitions of $a_{k,j}$ as the j^{th} element in the k^{th} row of the conditional distribution matrix of z values given proxy values and as well as our definition of \hat{r}_j as the marginal

probability that a proxy value is j . Finally, we use the result that $qA = U = (\frac{1}{K} \dots \frac{1}{K})$

$$\begin{aligned} \sum_{j \in \text{Range}(g)} \rho_j \Pr[z = k, g(x) = j] &= \sum_{j \in \text{Range}(g)} \rho_j \Pr[z = k | g(x) = j] \Pr[g(x) = j] \\ &= \sum_{j \in \text{Range}(g)} a_{k,j} \hat{\rho}_j \rho_j = \sum_{j \in \text{Range}(g)} a_{k,j} q_j = \frac{1}{K} \quad \blacktriangleleft \end{aligned}$$

4 Learning an (α, β) Proxy

We have discussed a proxy function $g : \mathcal{X} \rightarrow \mathbb{N}$ that maps samples to proxy groups and described the conditional distribution matrix A indicating the distribution of sensitive attributes *within* each proxy group. In Section 3, we showed how A can be used to derive acceptance probabilities for each group, such that under appropriate conditions, selecting according to these probabilities induces a uniform distribution over the protected attributes. Up until now, however, we have referenced A as fixed – we have used it to derive retention probabilities but have not described how it and the proxy can be generated. Recall that our proxy function $g \in \mathcal{G}$ takes the form of a decision tree, where each leaf is a *proxy group*. Therefore, each row in A , corresponding to the distribution over sensitive attributes in a given *proxy group*, also corresponds to the distribution over these attributes in a given *leaf*.

We grow our decision tree by sequentially making *splits* over the feature space – our tree will start as a stump and our matrix will have just one row, then we will split the tree into two leaves and the matrix will have two rows, and we will continue in this manner, splitting a leaf (and adding a row to the matrix) at each iteration. We will make these splits by employing a classification function from the pre-specified model class $\mathcal{H} \subseteq \{h : X \rightarrow \{0, 1\}\}$ assigned to each leaf. Because the two representations, as a matrix or a tree, afford different analytical advantages, we will continue to refer to both as we derive our algorithm. One advantage of the matrix representation is that it allows us to reason about the convex hull of a set of conditional distributions. Lemma 10 showed that there is a solution to $qA = U$ for a stochastic vector q if and only if U lies in the convex hull of A . Our goal will be to grow our tree (and the matrix A) so that the $\inf_{U' \in C(A)} \|U' - U\|_2$ shrinks at each iteration – until finally U is contained within (or sufficiently close to) $C(A)$.⁶

We begin with a geometric interpretation of $C(A)$ and describe how it changes as our tree and conditional distribution matrix expand. In particular, we grow a tree that has leaves V and keep track of the corresponding matrix A of sensitive attribute distributions conditional on their classification by the tree. We can always label the leaves of a tree with a binary sequence, so from now on we will identify each V with a binary sequence. Using this description, we derive sufficient conditions to decrease the Euclidean distance between $C(A)$ and U . We begin with several definitions that we will use to characterize $C(A)$.

► **Definition 14 (Vertex).** *Let R be a bijective mapping of vertices to a rows in A . Then $V \in \{0, 1\}^{\mathbb{N}}$ is a vertex of $C(A)$ if $R(V)$ corresponds to a row a_i such that $a_i \notin C(\{a_j\}_{j < i})$.*

Note that in our context this means that each *row* of A corresponds to a *vertex* of $C(A)$ as long as it cannot be represented as a convex combination of the other rows. Next, we introduce the function that is used at a node of the decision tree to partition samples into the left or right child. It will also be convenient in our algorithm to make use of randomized splitting functions, so we handle both cases.

⁶ Algorithms 1 and 2 and Lemma 13 extend easily to distributions other than the uniform.

► **Definition 15** (Splitting Function). We call $h_V \in \mathcal{H}$ a deterministic splitting function at vertex V . A randomized splitting function $\tilde{h}_V \in \Delta\mathcal{H}$ is a distribution supported on a finite set of deterministic splitting functions $\{h_V^i\}_{i=1}^n$ such that $\tilde{h}_V(x) = h_V^i(x)$ with probability $\frac{1}{n}$ for all i .

Each vertex V is paired with a splitting function \tilde{h}_V operating on samples mapped to V . To model the *expected* action of a randomized splitting function, we introduce the notion of *sample weights*, where the weight of a sample x at V is the probability that x reaches V in its random walk down the tree (as determined by the randomized splitting function). Here, $V \setminus 0$ indicates the parent of V if V ends in 0, and $V \setminus 1$ indicates the parent if V ends in 1. Note that because V is a binary sequence, we can apply the modulo operator with the binary representation of 2 to isolate the last digit.

► **Definition 16** (Sample Weights). The weight of a sample x at vertex V is defined as follows:
 $w_0(x) = 1$ and for $V \neq 0$, $w_V(x) = \begin{cases} w_{V \setminus 0}(x) \cdot \mathbb{E}[\tilde{h}_{V \setminus 0}(x)] & \text{if } V \bmod 2 = 0 \\ w_{V \setminus 1}(x) \cdot \mathbb{E}[1 - \tilde{h}_{V \setminus 1}(x)] & \text{if } V \bmod 2 = 1 \end{cases}$

We distinguish between V and the collection of weighted samples represented by V , l_V .

► **Definition 17** (Collection of Weighted Samples at V). Given randomized splitting functions $\{\tilde{h}_i\}_{i=0}^V$, the collection of weighted samples at V is denoted by $l_V = \{w_V(x), (x, z) : (x, z) \in S\}$.

► **Definition 18** (Vertex Split). A vertex split results from applying \tilde{h}_V to $x \in l_V$, where $l_{V0} = \{w_{V0}(x), (x, z) : (x, z) \in S\}$ and $l_{V1} = \{w_{V1}(x), (x, z) : (x, z) \in S\}$.

After V is split into $V0$ and $V1$, V is no longer a vertex, whereas $V0$ and $V1$ may be. So, the number of leaves in the tree, and therefore the number of rows in A , increased by at most 1.

4.1 Growing the Convex Hull and Learning a Splitting Function

Imagine that we have started to grow our proxy tree, but U is not in $C(A)$. We would like to expand $C(A)$ to contain U , and intuitively, we might like to expand $C(A)$ in the direction of U . One way to do so is to choose a vertex V to split into two vertices, $V1$ and $V0$. We assume that $V1$ is the split such that $R(V1) - R(V)$ is most in the direction of $U - U'$, where U' is the closest point in Euclidean distance to U in $C(A)$.

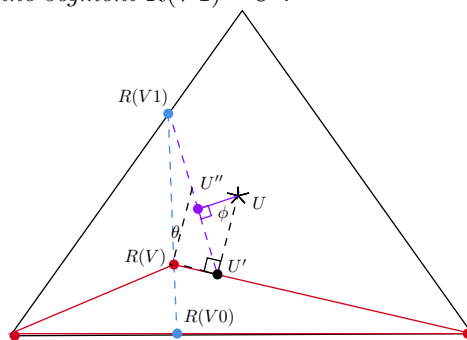
► **Definition 19** (Convex Hull Notation). Let θ be the angle between $R(V1) - R(V)$ and $U - U'$, and U'' be the closest point to U on the line segment $R(V1) - U'$:

$$U' = \arg \min_{U^* \in C(A)} \|U - U^*\|_2$$

$$\cos \theta = \frac{\langle R(V1) - R(V), U - U' \rangle}{\|R(V1) - R(V)\|_2 \|U - U'\|_2}$$

$$U'' = tU' + (1 - t)R(V1) \text{ where}$$

$$t = \operatorname{argmin}_{0 < t^* < 1} \|U - (t^*U' + (1 - t^*)R(V1))\|_2$$



We show that, given certain assumptions, we can lower bound how much this splitting process will decrease the distance from $C(A)$ to U . The first condition in Lemma 20 will be used to derive an objective function over which we can optimize to find a splitting function. The

second and third conditions limit the theory to the case where we can prove our progress lemma. The second condition says that the distance between $R(V)$ and $R(V1)$ has to be sufficiently large compared to the existing distance between the uniform distribution and its projection onto $C(A)$. The third condition is needed for the proof, allowing us to make arguments based on right triangles – it is satisfied when the second condition is met and the angle between $R(V1) - R(V)$ and $U - U'$ is not too large. As these conditions are potentially limiting theoretically, we verify that they are indeed frequently satisfied in the experiments.

► **Lemma 20** (Progress via Vertex Split). *When a vertex V is split, forming new vertices $V0$ and $V1$, the distance from the convex hull to U decreases by at least a factor of $1 - \gamma$ if*

$$\begin{aligned} \langle R(V1) - R(V), U - U' \rangle / \|U - U'\|_2 &\geq f(\gamma) \text{ and} \\ \|R(V1) - R(V)\|_2 &\geq (1 - \gamma)^{-1} \sqrt{2\gamma - \gamma^2} \|U - U'\|_2, \quad R(V1) - U' \perp U - U'' \text{ where} \end{aligned} \quad (1)$$

$$f(\gamma) := \sqrt{(2\gamma - \gamma^2) \left(2 - \|R(V) - U'\|_2^2 + 2\|R(V) - U'\|_2(1 - \gamma) \sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|_2^2 + 2} \right)}$$

► **Remark 21.** These are sufficient, but not necessary, conditions for a split to make sufficient progress. Empirically, we simply require that each split decreases the distance from the convex hull to U by at least a factor of $1 - \gamma$ for the algorithm to continue.

To summarize, these conditions ask that we split a vertex of the convex hull (equivalently a leaf of the proxy tree), so that the convex hull expands in the direction of the target vector. In other words, we want to split a leaf into the over-represented groups in one child and the under-represented groups in the other child, without violating the disclosure constraints. Lemma 1 also involves conditions that make sure that this split is sufficiently large to move the convex hull closer to the uniform rather than making minute progress. Having identified a sufficient condition for a split to make suitable progress toward containing the uniform distribution within the convex hull, we present a subroutine to find an α -proxy. We first express Equation (1) in a form amenable to use in a linear program:

► **Lemma 22** (Objective Function). *Let m_V be the number of samples in l_V and let h_V be the splitting function for vertex V . The condition $\frac{\langle R(V1) - R(V), U - U' \rangle}{\|U - U'\|_2} \geq f(\gamma)$ is equivalent to*

$$\begin{aligned} \sum_{i=1}^{m_V} w_V(x_i) h_V(x_i) (-Q + \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k)) &\leq 0 \\ \text{for } Q_{V,U',\gamma} := \|U' - U\| f(\gamma) + \frac{\sum_{i=1}^{m_V} w_V(x_i) \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k)}{\sum_{i=1}^{m_V} w_V(x_i)} \end{aligned}$$

Proof. We begin by expanding the scaled dot product between $R(V1) - R(V)$ and $U - U'$:

$$\frac{\langle R(V1) - R(V), U - U' \rangle}{\|U - U'\|} = \sum_{j=1}^{m_V} w_V(x_j) \sum_{k=1}^K \mathbb{1}_{z_j=k} \left(\frac{h_V(x_j)}{\sum_{j=1}^{m_V} w_V(x_j) h_V(x_j)} - \frac{1}{\sum_{j=1}^{m_V} w_V(x_j)} \right) \frac{U_k - U'_k}{\|U - U'\|}$$

Asking $\frac{\langle R(V1) - R(V), U - U' \rangle}{\|U - U'\|} \geq f(\gamma)$ is equivalent to asking $\frac{\langle R(V1) - R(V), U' - U \rangle}{\|U' - U\|} \leq f(\gamma)$ or:

$$\frac{\sum_{j=1}^{m_V} w_V(x_j) h_V(x_j) \sum_{k=1}^K \mathbb{1}_{z_j=k} (U'_k - U_k)}{\sum_{i=1}^{m_V} w_V(x_j) h_V(x_j)} \leq \|U' - U\| f(\gamma) + \frac{\sum_{j=1}^{m_V} w_V(x_j) \sum_{k=1}^K \mathbb{1}_{z_j=k} (U'_k - U_k)}{\sum_{j=1}^{m_V} w_V(x_j)} \quad (2)$$

Finally, the right-hand side is constant given V , U' , and γ . Therefore, we represent it by a constant $Q_{V,U',\gamma} := \|U' - U\|f(\gamma) + \frac{\sum_{j=1}^{m_V} w_V(x_j) \sum_{k=1}^K \mathbb{1}_{z_j=k}(U'_k - U_k)}{\sum_{j=1}^{m_V} w_V(x_j)}$. This allows us to rewrite Equation (2) as

$$\sum_{j=1}^{m_V} w_V(x_j) h_V(x_j) \left(-Q_{V,U',\gamma} + \sum_{k=1}^K \mathbb{1}_{z_j=k}(U'_k - U_k) \right) \leq 0 \quad \blacktriangleleft$$

We use Lemma 22 to form a cost-sensitive classification problem for vertex V , where the constraints make sure that any candidate proxy is no more than α -disclosive:

$$\begin{aligned} \min_{h_V \in \mathcal{H}} \sum_{i=1}^{m_V} w_V(x_i) h_V(x_i) \left(-Q_{V,U',\gamma} + \sum_{k=1}^K \mathbb{1}_{z_i=k}(U'_k - U_k) \right) \quad \text{s.t. } \forall k \quad (3) \\ \left| \frac{\sum_{i=1}^{m_V} w_V(x_i) h_V(x_i) \mathbb{1}_{z_i=k}}{\sum_{i=1}^{m_V} w_V(x_i) h_V(x_i)} - r_k \right| \leq \alpha \quad \text{and} \quad \left| \frac{\sum_{i=1}^{m_V} w_V(x_i) (1 - h_V(x_i)) \mathbb{1}_{z_i=k}}{\sum_{i=1}^{m_V} w_V(x_i) (1 - h_V(x_i))} - r_k \right| \leq \alpha \end{aligned}$$

Next, we will appeal to strong duality to derive the corresponding Lagrangian. We note that computing an approximately optimal solution to the linear program corresponds to finding approximate equilibrium strategies for both players in the game in which one player, the ‘‘Learner,’’ controls the primal variables and aims to minimize the Lagrangian value. The other player, the ‘‘Auditor,’’ controls the dual variables and seeks to maximize the Lagrangian value. If we construct our algorithm in such a way that it simulates repeated play of the Lagrangian game such that both players have sufficiently small regret, we can apply Theorem 1 to conclude that our empirical play converges to an approximate equilibrium of the game. Furthermore, our algorithm will be *oracle efficient*: it will make polynomially many calls to oracles that solve weighted cost-sensitive classification problems over \mathcal{H} .

To turn Program (3) into a form amenable to our two-player zero-sum game formulation, we expand \mathcal{H} to $\Delta\mathcal{H}$, allow our splitting function to be *randomized*, and take expectations over the objective and constraints with respect to deterministic splitting functions drawn according to \tilde{h}_V . Doing so yields the following CSC problem to be solved for vertex V :

$$\begin{aligned} \min_{\tilde{h}_V \in \Delta\mathcal{H}} \quad & \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) h_V(x_i) \left(-Q_{V,U',\gamma} + \sum_{k=1}^K \mathbb{1}_{z_i=k}(U'_k - U_k) \right) \\ \text{s.t.} \quad & \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) \tilde{h}_V(x_i) (\mathbb{1}_{z_i=k} - r_k - \alpha) \leq 0 \quad \forall k, \\ & \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) (1 - h_V(x_i)) (\mathbb{1}_{z_i=k} - r_k - \alpha) \leq 0 \quad \forall k, \quad (4) \\ & \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) h_V(x_i) (r_k - \mathbb{1}_{z_i=k} - \alpha) \leq 0 \quad \forall k, \\ & \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) (1 - h_V(x_i)) (r_k - \mathbb{1}_{z_i=k} - \alpha) \leq 0 \quad \forall k \end{aligned}$$

We solve this constrained optimization problem by simulation a zero-sum two-player game on the Lagrangian dual. Given dual variables $\lambda \in \mathbb{R}_{\geq 0}^{4K}$ such that $\|\lambda\|_2 \leq \lambda_{max}$ for some constant λ_{max} , the Lagrangian of Program (4) is:

$$\begin{aligned}
 L(\lambda, \tilde{h}_V) = \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^{m_V} w_V(x_i) & \left(-Q_{V,U',\gamma} h_V(x_i) + \sum_{k=1}^K h_V(x_i) \mathbb{1}_{z_i=k} (U'_k - U_k) + \right. \\
 & (\lambda_{k,1} h_V(x_i) + \lambda_{k,0} (1 - h_V(x_i))) (\mathbb{1}_{z_i=k} - r_k - \alpha) + \\
 & \left. (\lambda_{k,3} h_V(x_i) + \lambda_{k,2} (1 - h_V(x_i))) (r_k - \mathbb{1}_{z_i=k} - \alpha) \right)
 \end{aligned}$$

Given the Lagrangian, solving Program (4) is equivalent to solving the minimax problem $\min_{\tilde{h}_V \in \Delta \mathcal{H}} \max_{\lambda \in \mathbb{R}_{\geq 0}^{4K}} L(\lambda, \tilde{h}_V) = \max_{\lambda \in \mathbb{R}_{\geq 0}^{4K}} \min_{\tilde{h}_V \in \Delta \mathcal{H}} L(\lambda, \tilde{h}_V)$, where the minimax theorem holds because the range of the primal variable, i.e., $\Delta \mathcal{H}$ is convex and compact, the range of the dual variable, i.e., $\mathbb{R}_{\geq 0}^{4K}$ is convex, and the Lagrangian function L is linear in both primal and dual variables. Therefore, we focus on solving the minimax problem, which can be seen as a two-player zero-sum game between the primal player (the Learner) who is controlling \tilde{h}_V and the dual player (the Auditor) who is controlling λ . Using no-regret dynamics, we will have the Learner deploy its best response strategy in every round, which will be reduced to a call to $CSC(\mathcal{H})$ and let the Auditor with strategies in $\Lambda = \{\lambda : 0 \leq \lambda \leq \lambda_{max}\}$ play according to Online Projected Gradient Descent [45].

Our local algorithm for splitting a vertex is described in Algorithm 3, and its guarantee is given in Theorem 2. We note that the algorithm returns a distribution over \mathcal{H} . Given an action λ of the Auditor, we write $LC(\lambda)$ for the vector of costs for labeling each data point as 1. We view our costs as the inner product of the outputs of a deterministic splitting function h_V on the m_V points and corresponding cost vector. We define the cost for labeling an example 0 to be 0 for all x ($c^0(x) = 0$), and the cost for labeling an example 1 as:

$$\begin{aligned}
 c^1(x) = w_V(x) & (-Q_{V,U',\gamma} + \sum_{k=1}^K \mathbb{1}_{z_j=k} (U'_k - U_k) + (\lambda_{k,1} - \lambda_{k,0}) (\mathbb{1}_{z=k} - r_k - \alpha) \\
 & + (\lambda_{k,3} - \lambda_{k,2}) (r_k - \mathbb{1}_{z=k} - \alpha))
 \end{aligned}$$

■ **Algorithm 3** Learning a Splitting Function.

Input: $\{w_V(x_i), (x_i, z_i)\}_{i=1}^{m_V}$, model class \mathcal{H} , $CSC(\mathcal{H})$, α , ϵ , γ
 Set $\lambda_{max} = m(K-1)/K\epsilon + 2$ and $T = \lceil (2Km(1+\alpha)\lambda_{max}/\epsilon)^2 \rceil$
 Initialize $\lambda_k = 0 \forall k$
for $t = 1 \dots T$ **do**
 $h_t = \operatorname{argmin}_{h \in \mathcal{H}} \langle LC(\lambda), h \rangle$
 $\lambda_t = \lambda_{t-1} + t^{-1/2} (\nabla_{\lambda} L)^+$; If $\|\lambda_t\| > \lambda_{max}$, set $\lambda_t = \lambda_{max} \frac{\lambda_t}{\|\lambda_t\|}$
return $\tilde{h}_V :=$ uniform distribution over h_t^t

► **Theorem 2** (Learning an $(\alpha + \epsilon)$ -Disclosive Proxy). *Fix α , ϵ , suppose \mathcal{H} has finite VC dimension, and suppose $\exists \tilde{h}_V^* \in \Delta \mathcal{H}$ that is a feasible solution to Program (4). Then, Algorithm 3 returns a distribution \tilde{h}_V that is an ϵ -optimal solution to Program (4).*

Theorem 2 says that with appropriate conditions on the model class \mathcal{H} and access to $CSC(\mathcal{H})$, Algorithm 3 returns a model satisfying the conditions of Program 4 (i.e. produces an acceptable split) up to an additive factor of ϵ . A few requirements of this theorem may not hold in practice and thus motivate our experiments. The choice of a base model class \mathcal{H} impacts whether a feasible solution exists – typically more complex model classes will be more likely to contain a feasible solution, but this complexity will impact the generalization bounds. Also, the guarantee relies on Algorithm 3 having access to a cost sensitive classification oracle. In practice, we typically do not have such an oracle so must use a heuristic.

4.2 Decision Tree Meta-Algorithm

Finally, we use these results to greedily construct a proxy $g : \mathcal{X} \rightarrow \mathbb{N}$. We do this iteratively using a decision tree, where leaves correspond to proxy groups. We split the data into these leaves in such a way that when we consider the distribution of groups in each leaf, the uniform vector is contained in their convex hull. This allows us to select a balanced set in expectation. In addition, we require that the proxy be α -disclosive at every step. We grow the tree as follows, for some tolerance β : (1) If $\|U - C(A)\|_2 \leq \beta$, output the tree. (2) Otherwise, look for a leaf to split. If we find a suitable split, make it, and continue. If not, output the tree. To determine if a split is suitable, we use the results from Section 4.1: for fixed approximation factor ϵ , disclosivity budget $\alpha - \epsilon$, and progress parameter γ , a splitting function \tilde{h}_V must be an ϵ -approximate solution to Program 4 (and therefore no more than α -disclosive) at vertex V . If we can find such an \tilde{h}_V for at least $\frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}$ rounds, the decision tree will be an (α, β) proxy.

■ **Algorithm 4** Learning an (α, β) Proxy.

Input: $D = \{x_i, z_i\}_{i=1}^n$, $CSC(\mathcal{H})$, α , ϵ , γ , β

while $\inf_{U' \in C(A)} \|U' - U\|_2 > \beta$ **do**

 Apply Algorithm 3 to find feasible split (if no feasible split, terminate)

 Expand tree T and re-calculate A , $C(A)$

return T , A

► **Theorem 3** (Learning an (α, β) Proxy). *If the conditions of Lemma 20 are satisfied at every split, Algorithm 4 produces an (α, β) proxy in-sample within $\frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}$ rounds.*

Proof. Let A_{i^*} be the conditional distribution matrix returned by Algorithm 4 after i^* rounds. Our goal is to produce A_{i^*} such that $\|U - C(A_{i^*})\|_2 \leq \beta$. Let A_0 be the initial conditional distribution matrix, and observe that if we decrease the distance from the current conditional distribution matrix to U by a factor of $1 - \gamma$ each round, at round i , $\|U - C(A_i)\|_2 \leq (1 - \gamma)^i \|U - C(A_0)\|_2$. Further, recall that $\|U - C(A_0)\|_2 \leq \sqrt{2}$ because both U and $C(A_0)$ must lie in the unit simplex. Setting $\|U - C(A_{i^*})\|_2 \leq \beta$, we have $\beta \leq (1 - \gamma)^{i^*} \sqrt{2} \implies \frac{\beta}{\sqrt{2}} \leq (1 - \gamma)^{i^*} \implies i^* \geq \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}$. Then, after $i^* = \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}$ rounds, $\|U - C(A_{i^*})\|_2 \leq \beta$. Finally, because our linear program constrains splits to only those that guarantee α -disclosiveness, the final proxy must be α -disclosive in-sample. ◀

Theorem 3 allows us to upper bound the number of times that Algorithm 4 performs a split and, therefore, the number of unique proxy groups generated. The theorem's hypothesis states, informally, that it must be possible to find a splitting function at each round that makes both a sufficiently *large* split (i.e. the new vertex is sufficiently far from the old vertex compared to the current distance from the convex hull to the target uniform) and the split is sufficiently in the direction of the target. This theorem, in turn, allows us to state generalization bounds depending on both the number and size of each proxy group.

► **Theorem 4** (Generalization). *Let $\epsilon, \delta, \gamma > 0$ and G be the proxy class. Let there be K sensitive groups. If each proxy group has at least $\frac{1}{2\epsilon^2} \ln \frac{8K \cdot VC(G)(\ln \beta - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$ samples, with probability $1 - \delta$, an (α, β) proxy in-sample will be an $(\alpha + 2\epsilon, \beta + K\epsilon \sqrt{\frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}})$ proxy out-of-sample.*

Theorem 4 presents the number of samples needed in each proxy group to obtain a sufficiently small generalization gap in both the disclosure level and imbalance – it is based on the size of the *smallest proxy group in-sample*, which might get quite small in practice.

Furthermore, the generalization gap for β scales by an additive factor of the number of sensitive groups, K . Therefore, as our problem becomes more challenging, more samples are required to achieve a proxy that performs similarly out-of-sample compared to in-sample.

5 Experiments

Here, we test our two main methodological contributions. The first is to use Algorithms 1 and 2 to solve $\min_q \|qA - U\|_2$ subject to $q_i \geq 0 \forall i$ and $\sum_i q_i = 1$ and derive the corresponding acceptance probabilities ρ for the given proxy. The second is to additionally use Algorithms 3 and 4 to learn a decision-tree proxy *guaranteed* not to exceed a specified level of disclosure.

1. *QP Regression and Decision Tree Proxies*: We train a multinomial logistic regression model or decision tree to *directly predict* the sensitive attribute but select our acceptance probabilities by employing Algorithms 1 and 2.
2. (α, β) Proxy: We use Algorithm 4 to develop a proxy for a specified disclosure budget. We compare the performances of these proxy functions against those of two baselines:⁷
 1. *Naive Regression and Decision Tree proxies*: We train models to *directly predict* sensitive attributes then sample the same number of points from each predicted group, inducing a conditional distribution matrix of the distribution of sensitive attributes in each proxy group. We then calculate the degree of disclosure and imbalance of the sampled set.
 2. SMOTE [8]: We train a decision tree to *directly predict* groups and then, using these predictions as input for SMOTE, balance the data by synthesizing minority examples.

5.1 Data, Hyperparameters, and Compute Time

We evaluate the disclosure, α , and imbalance, β , obtained by each proxy filtering scheme on the Bank Marketing [25, 12], Adult [12], and Communities and Crime data sets [12, 35, 37, 38, 39, 30], for which we have 5, 4, and 12 sensitive attribute values, respectively. The Marketing data set consists of 45211 labeled samples with 48 non-sensitive attributes and a sensitive attribute of job type. The downstream classification goal is to predict whether a client will subscribe a term deposit based on a phone call marketing campaign of a Portuguese banking institution. The Adult data set consists of 48842 labeled samples with 14 non-sensitive attributes, and we select race as the sensitive attribute. The associated classification task is to determine whether individuals make over \$50K dollars per year. The Communities and Crime data set consists of 1594 samples with 132 non-sensitive features, race as the sensitive group, and the number of violent crimes per population as the prediction task.

For each experiment, we run trials with 20 different seeds, and for each seed, we input a grid of values with increments of 0.1 for the disclosure parameter, α , evenly spaced between 0 and 1. We then average over the seeds for each α and calculate empirical 95% confidence intervals (which are displayed as the shaded region around each line in the plots). Each data set is split into three parts of sizes 50%, 30%, and 20%. The first is used to train the proxy. The second is used first to test the filtering effects of the proxy out-of-sample and then to train a classification model on to study downstream performance. The third is the set upon which we apply these classifiers trained on filtered and unfiltered data to see how

⁷ For the Naive Proxies, QP Proxies, and SMOTE, we interpolate between a uniform and proxy-specific sampling strategy by post-processing: We predict z with the proxy and then, with probability $\eta \in [0, 1]$, uniformly re-assign the prediction. Finally, we apply Algorithm 2 to sample according to the post-processed proxy labels and plot the balance and disclosure of the corresponding data set *with respect to the post-processed proxy values*. We use a large point marker for the results without post-processing

the group-wise accuracy levels are affected. For brevity, we will refer to these three splits as the “Train” set, “Test” set, and “Post-Test” set, respectively. See Appendix B for an analysis of downstream fairness effects induced by our strategy.

On the Adult and Communities and Crime data sets, one run over the grid of α values typically took between 20 minutes and two hours for the (α, β) proxy. On the Marketing data set, running one full experiment over the grid of α values took about three hours. The parameter γ was set to 0.0001, the maximum height of the proxy tree was set to 15, and the learning process was stopped once the distance between the convex hull of the conditional distribution matrix and the uniform distribution fell below 0.05. As we used publicly available tabular data sets that has already been cleaned, there were no missing values.

Finally, the choice of oracle (the base model class for the (α, β) proxy) is heuristic – as we do not have a true cost-sensitive classification oracle for Algorithm 3, we choose two models that allow us to predict the cost of each example and then classify based on the cost’s sign. We experiment with a linear threshold function – the paired regression classifier (PRC) used in [19] and defined below – as well as the XGBoost Regressor model. We found that the PRC was simpler and seemed to perform at least as well as the XGBoost Regressor, so we relegate the analysis for the latter to Appendix B.

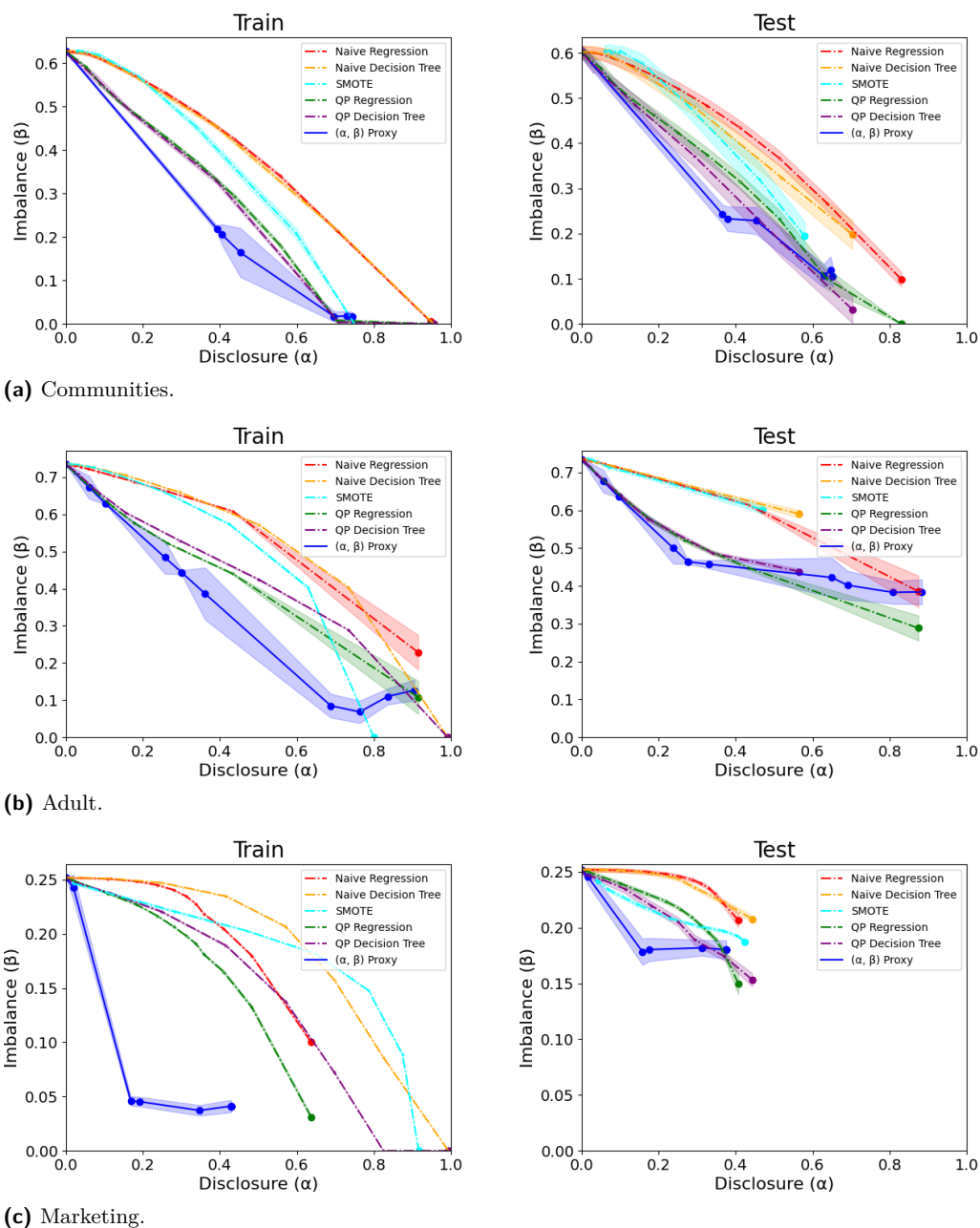
► **Definition 23** (Paired Regression Classifier [19]). *The paired regression classifier operates as follows: We form two weight vectors, z^0 and z^1 , where z_i^k corresponds to the penalty assigned to sample i in the event that it is labeled k . For the correct labeling of x_i , the penalty is 0. For the incorrect labeling, the penalty is the current sample weight of the point, w_V . We fit two linear regression models h^0 and h^1 to predict z^0 and z^1 , respectively, on all samples. Then, given a new point x , we calculate $h^0(x)$ and $h^1(x)$ and output $h(x) = \operatorname{argmin}_{k \in \{0,1\}} h^k(x)$.*

5.2 Results

In Figure 1, on the Communities data set, the (α, β) proxy Pareto-dominates the other approaches in sample, while the QP proxies Pareto-dominate SMOTE and the Naive proxies. All methods generalize well. On the Adult data set, the (α, β) proxy primarily dominates the remaining approaches in-sample. The generalization performance for all methods, but particularly the (α, β) proxy, is weaker on the Adult data set. This is likely because there are slightly more sensitive groups than in the Communities data set, and the acceptance probabilities were sparse. On the Marketing data set, the (α, β) and QP Decision Tree proxies exhibit favorable performance in-sample, driving the imbalance to just above zero at higher levels of disclosure. The plot on the test set shows a more modest improvement in balance for all methods. One source of variance in Figure 1 is the generalization performance by the (α, β) proxy. We believe this to be due to the size of the smallest proxy group being quite low (especially for the Marketing data set which has 12 sensitive groups). Recall that the generalization gap depends directly on this quantity. There is also nothing in our method to prevent a sparse sampling scheme. Empirically, we found that in cases where generalization results were weak, the acceptance probabilities were nonzero for only a handful of the final proxy groups. Addressing these weaknesses, if possible, could strengthen our approach.

5.3 Discussion and Future Work

Our primary conceptual point is that even though the final goal (balance) references the protected attributes, it is a condition on the aggregate composition of the final selected set. Therefore, achieving it does not necessarily require finding a predictor strongly correlated with the protected attribute. We emphasize that while the QP proxies (our secondary contribution) are appealingly simple and provide a range of disclosure levels *after* post-processing, they



■ **Figure 1** Trade-off of Disclosure and Balance of Proxies on Communities, Adult, and Marketing.

still involve explicitly training a classifier for the attribute. In contrast, the (α, β) proxy (our primary contribution) never involves training a classifier at any step of the process that is more disclosive than a pre-specified threshold. While this does not solve the challenging legal and technical problems associated with proxy use in high-stakes selection processes, it takes a step in this direction by permitting controlled trade-offs between balance and disclosure.

References

- 1 Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018. [arXiv:1803.02453](https://arxiv.org/abs/1803.02453).

- 2 Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. *CoRR*, abs/1905.12843, 2019. [arXiv:1905.12843](https://arxiv.org/abs/1905.12843).
- 3 Larry Alexander. What makes wrongful discrimination wrong? biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review*, 141(1):149–219, 1992. URL: <http://www.jstor.org/stable/3312397>.
- 4 Gustavo E. A. P. A. Batista, Ana Lúcia Cetertich Bazzan, and Maria Carolina Monard. Balancing training data for automated annotation of keywords: a case study. In *WOB*, 2003.
- 5 Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004. doi:10.1145/1007730.1007735.
- 6 Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, second edition, 1986.
- 7 Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- 8 Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.
- 9 Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 0(0):1–12, 2023. doi:10.1080/01621459.2023.2191817.
- 10 Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajardi. Multiaccurate proxies for downstream fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 1207–1239, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3531146.3533180.
- 11 Georgios Douzas, Fernando Bacao, and Felix Last. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, 465:1–20, October 2018. doi:10.1016/j.ins.2018.06.056.
- 12 Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL: <http://archive.ics.uci.edu/ml>.
- 13 Marc N. Elliott, Peter A. Morrison, Allen M. Fremont, Daniel F. McCaffrey, Philip M Pantoja, and Nicole Lurie. Using the census bureau’s surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83, 2009.
- 14 Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, 1996.
- 15 Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, 2005.
- 16 Peter E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, pages 515–516, 1968.
- 17 Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.
- 18 Gabrielle Johnson. Algorithmic bias: On the implicit biases of social technology, May 2020. URL: <http://philsci-archive.pitt.edu/17169/>.
- 19 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.
- 20 David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. *CoRR*, abs/1802.06309, 2018. [arXiv:1802.06309](https://arxiv.org/abs/1802.06309).
- 21 Inderjeet Mani and Jianping Zhang. knn approach to unbalanced data distributions: A case study involving information extraction. *Workshop on Learning from Imbalanced Datasets II, ICML*, 126:1–7, 2003.
- 22 Daniel McCaffrey, Greg Ridgeway, and Andrew Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9:403–25, January 2005. doi:10.1037/1082-989X.9.4.403.

- 23 Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28:92–122, 2012.
- 24 Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105 4:987–993, 2016. URL: <https://api.semanticscholar.org/CorpusID:88521475>.
- 25 Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. doi:10.1016/j.dss.2014.03.001.
- 26 Hien M. Nguyen, Eric W. Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigms*, 3:4–21, 2009.
- 27 Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 854–863, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3461702.3462629.
- 28 Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *ArXiv*, abs/1910.11779, 2019. URL: <https://api.semanticscholar.org/CorpusID:204900934>.
- 29 Hongxiang Qiu, Xu Shi, Wang Miao, Edgar Dobriban, and Eric Tchetgen Tchetgen. Doubly robust proximal synthetic controls, 2023. arXiv:2210.02014.
- 30 M. A. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments, 2002.
- 31 Xu Shi, Kendrick Li, Wang Miao, Mengtong Hu, and Eric Tchetgen Tchetgen. Theory for identification and inference with synthetic controls: A proximal causal inference framework, 2023. arXiv:2108.13935.
- 32 Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning, 2020. arXiv:2009.10982.
- 33 I. Tomek. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(6):448–452, 1976. doi:10.1109/TSMC.1976.4309523.
- 34 I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:769–772, 1976.
- 35 Bureau of the Census U. S. Department of Commerce. Census of population and housing 1990 united states: Summary tape file 1a & 3a (computer files).
- 36 U.S. Students for fair admissions, inc. v. president and fellows of harvard college, 2023.
- 37 Bureau Of The Census Producer U.S. Department Of Commerce, 1992.
- 38 Bureau Of The Census Producer U.S. Department Of Commerce. U.s. department of justice, bureau of justice statistics, law enforcement management and administrative statistics (computer file), 1992.
- 39 Federal Bureau of Investigation U.S. Department of Justice. Crime in the united states (computer file), 1995.
- 40 Ioan Voicu. Using first name information to improve race and ethnicity classification. *Statistics and Public Policy*, 5:1–13, 2016.
- 41 Michael R. Wickens. A note on the use of proxy variables. *Econometrica*, 40(4):759–761, 1972. URL: <http://www.jstor.org/stable/1912971>.
- 42 Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.*, 2:408–421, 1972.
- 43 Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. arXiv:1710.09412.
- 44 Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Comparative Political Economy: Regulation eJournal*, 2016.
- 45 Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*. Washington, DC, 2003.

A Omitted Proofs

► **Lemma 20** (Progress via Vertex Split). *When a vertex V is split, forming new vertices $V0$ and $V1$, the distance from the convex hull to U decreases by at least a factor of $1 - \gamma$ if*

$$\begin{aligned} \langle R(V1) - R(V), U - U' \rangle / \|U - U'\|_2 &\geq f(\gamma) \text{ and} \\ \|R(V1) - R(V)\|_2 &\geq (1 - \gamma)^{-1} \sqrt{2\gamma - \gamma^2} \|U - U'\|_2, \quad R(V1) - U' \perp U - U'' \text{ where} \end{aligned} \quad (1)$$

$$f(\gamma) := \sqrt{(2\gamma - \gamma^2) \left(2 - \|R(V) - U'\|_2^2 + 2\|R(V) - U'\|_2 (1 - \gamma) \sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|_2^2 + 2} \right)}$$

Proof. We want to find sufficient conditions for $\|U - U''\| \leq (1 - \gamma)\|U - U'\|$. Let ϕ be the angle between the vectors $U - U''$ and $U - U'$. Then $\|U - U''\| = \|U - U'\| \cos \phi$. So, we would like to find conditions for which $\cos \phi \leq (1 - \gamma)$. By the law of cosines, $\|V1 - U'\|^2 = \|V - U'\|^2 + \|V1 - V\|^2 - 2\|V - U'\|\|V1 - V\| \cos(90 + \theta)$ and

$$\begin{aligned} \cos(\phi) &= \frac{\|V - U'\|^2 + \|V1 - U'\|^2 - \|R(V1) - R(V)\|^2}{2\|V - U'\|\|V1 - U'\|} \\ &= \frac{\|V - U'\| - \|R(V1) - R(V)\| \cos(90 + \theta)}{\sqrt{\|V - U'\|^2 + \|R(V1) - R(V)\|^2 - 2\|V - U'\|\|R(V1) - R(V)\| \cos(90 + \theta)}} \\ &= \frac{\|V - U'\| + \|R(V1) - R(V)\| \sin \theta}{\sqrt{\|V - U'\|^2 + \|R(V1) - R(V)\|^2 + 2\|V - U'\|\|R(V1) - R(V)\| \sin \theta}} \end{aligned}$$

Setting $-(1 - \gamma) \leq \cos \phi \leq 1 - \gamma$ and solving for $\sin \theta$, we see that this is satisfied by

$$\sin \theta \in \frac{(\gamma^2 - 2\gamma)\|R(V) - U'\|}{\|R(V1) - R(V)\|} \pm \frac{(1 - \gamma)\sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|^2 + \|R(V1) - R(V)\|^2}}{\|R(V1) - R(V)\|}$$

To find a set of values for $\cos \theta$ that make the above expression always true, we will consider only γ for which the set of values of $\sin \theta$ includes the origin. This is true for $\gamma \in \left[0, 1 - \sqrt{\frac{\|V - U'\|^2}{\|V - U'\|^2 + \|R(V1) - R(V)\|^2}}\right]$. Then,

$$\cos^2 \theta \geq 1 - \left(\frac{(\gamma^2 - 2\gamma)\|R(V) - U'\|}{\|R(V1) - R(V)\|} + \frac{(1 - \gamma)\sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|^2 + \|R(V1) - R(V)\|^2}}{\|R(V1) - R(V)\|} \right)^2$$

Rearranging, we have

$$\begin{aligned} \|R(V1) - R(V)\|^2 \cos^2 \theta &\geq \\ \|R(V1) - R(V)\|^2 - \left((\gamma^2 - 2\gamma)\|R(V) - U'\| + (1 - \gamma)\sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|^2 + \|R(V1) - R(V)\|^2} \right)^2 & \\ = (2\gamma - \gamma^2) \cdot (\|R(V1) - R(V)\|^2 - \|R(V) - U'\|^2 + & \\ 2\|R(V) - U'\| (1 - \gamma) \sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|^2 + \|R(V1) - R(V)\|^2}) & \end{aligned}$$

Using the fact that $\|R(V1) - R(V)\|^2 \leq 2$, we upper bound the right-hand side to say that a split satisfying the following condition will guarantee that we decrease the distance from U to the convex hull by $(1 - \gamma)$:

$$\begin{aligned} \|R(V1) - R(V)\| \cos \theta &\geq \\ (2\gamma - \gamma^2)^{\frac{1}{2}} \left(2 - \|R(V) - U'\|^2 + 2\|R(V) - U'\| (1 - \gamma) \sqrt{(\gamma^2 - 2\gamma)\|R(V) - U'\|^2 + 2} \right)^{\frac{1}{2}} &:= f(\gamma) \blacktriangleleft \end{aligned}$$

► **Theorem 2** (Learning an $(\alpha + \epsilon)$ -Disclosive Proxy). *Fix α, ϵ , suppose \mathcal{H} has finite VC dimension, and suppose $\exists \tilde{h}_V^* \in \Delta\mathcal{H}$ that is a feasible solution to Program (4). Then, Algorithm 3 returns a distribution \tilde{h}_V that is an ϵ -optimal solution to Program (4).*

Proof. We begin by upper bounding the L_2 norm of the gradient:

$$\begin{aligned} \|\nabla_\lambda L(\lambda, \tilde{h}_V)\|^2 &= \left(\sum_{k=1}^K \sum_{i=1}^m \mathbb{E}_{h_V \sim \tilde{h}_V} w(x_i) h_V(x_i) (\mathbb{1}_{z_i=k} - r_k - \alpha) \right)^2 + \\ &\quad \left(\sum_{k=1}^K \sum_{i=1}^m \mathbb{E}_{h_V \sim \tilde{h}_V} w(x_i) (1 - h_V(x_i)) (\mathbb{1}_{z_i=k} - r_k - \alpha) \right)^2 + \\ &\quad \left(\sum_{k=1}^K \sum_{i=1}^m \mathbb{E}_{h_V \sim \tilde{h}_V} w(x_i) h_V(x_i) (r_k - \mathbb{1}_{z_i=k} - \alpha) \right)^2 + \\ &\quad \left(\sum_{k=1}^K \sum_{i=1}^m \mathbb{E}_{h_V \sim \tilde{h}_V} w(x_i) (1 - h_V(x_i)) (r_k - \mathbb{1}_{z_i=k} - \alpha) \right)^2 \\ &\leq 4K^2 m^2 (1 + \alpha)^2 \end{aligned}$$

We now apply the regret bound for Online Gradient Descent from [45]. With an appropriate choice of η (derived below), we bound the Auditor's average regret over T rounds:

$$\frac{R_T}{T} \leq \frac{\sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\| \|\nabla_\lambda L(\tilde{h}, \lambda)\| \sqrt{T}}{T} \leq \frac{2Km(1+\alpha) \sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\|}{\sqrt{T}}$$

Setting $T \geq \left(\frac{2Km(1+\alpha) \sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\|}{\epsilon} \right)^2$ and $\eta = \frac{\sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\|}{2Km(1+\alpha)\sqrt{T}}$, we have that $\frac{R_T}{T} \leq \epsilon$. Because the Learner plays a no-regret strategy, we can apply Theorem (1) to assert that the mixed strategy of the Auditor and Learner together form an ϵ -approximate equilibrium. Next, we must show that an approximate solution to the game corresponds to an approximate solution to Program (4). We will show this using two cases. In the first case, we consider some \tilde{h}_V^* that is a feasible solution to Program (4) at vertex V and a $\hat{\lambda}$ that is an ϵ -approximate minimax solution to the Lagrangian game specified in the Lagrangian above. Now we will analyze the case in which we have a solution \tilde{h}_V that is an ϵ -approximate solution to the Lagrangian game but is not a feasible solution for Program (4) – we will show that this is impossible. To illustrate this, assume that we *do* have such a \tilde{h}_V . Because it is not a feasible solution for Program (4), some constraints must be violated. Let ξ be the magnitude of the violated constraint, and let λ be such that the dual variable for the violated constraint is set to $\lambda_{max} := \sup_{\lambda, \lambda' \in \Lambda} \|\lambda - \lambda'\|$. By definition of an ϵ -approximate minimax solution, we know that $L(\hat{\lambda}, \tilde{h}_V) \geq L(\lambda, \tilde{h}_V) \geq \mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^m w_V h_V(x_i) \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k) + \lambda_{max} \xi - \epsilon$. Then,

$$\begin{aligned} &\mathbb{E}_{h_V \sim \tilde{h}_V} \sum_{i=1}^m w_V h_V(x_i) \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k) + \lambda_{max} \xi \\ &\leq L(\tilde{h}_V, \hat{\lambda}) + \epsilon \leq L(\tilde{h}_V^*, \hat{\lambda}) + 2\epsilon \leq \mathbb{E}_{h_V \sim \tilde{h}_V^*} \sum_{i=1}^m w_V h_V(x_i) \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k) + 2\epsilon \end{aligned}$$

Finally, because $\mathbb{E}_{h_V \sim \tilde{h}_V^*} \sum_{i=1}^m w_V h_V(x_i) \sum_{k=1}^K \mathbb{1}_{z_i=k} (U'_k - U_k) \leq \frac{m(K-1)}{K}$, we have that $\lambda_{max} \xi \geq \frac{m(K-1)}{K} + 2\epsilon$. Therefore, the maximum constraint violation is no more than $\frac{\frac{m(K-1)}{K} + 2\epsilon}{\lambda_{max}}$. Setting $\lambda_{max} = \frac{m(K-1)}{K\epsilon_\alpha} + 2$, \tilde{h}_V does not violate any constraint by more than ϵ . ◀

► **Theorem 4** (Generalization). *Let $\epsilon, \delta, \gamma > 0$ and G be the proxy class. Let there be K sensitive groups. If each proxy group has at least $\frac{1}{2\epsilon^2} \ln \frac{8K \cdot VC(\mathcal{G})(\ln \beta - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$ samples, with probability $1 - \delta$, an (α, β) proxy in-sample will be an $(\alpha + 2\epsilon, \beta + K\epsilon\sqrt{\frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}})$ proxy out-of-sample.*

Proof. Let $\tilde{A}_{k,j} = \Pr_{(x,y,z) \sim \mathcal{P}}[z = k, g(x) = j]$ and $A_{i,j} = \frac{1}{n} \sum_{i=1}^n w_j(x_i) \mathbb{1}_{z_i=k, g(x_i)=j}$. Then, Hoeffding's inequality gives us that, for fixed k, j, g

$$\Pr_{D \sim \Omega} \left[\left| \frac{1}{n} \sum_{i=1}^n w_j(x_i) \mathbb{1}_{z_i=k, g(x_i)=j} - \mathbb{E}_{(x,y,z) \sim \mathcal{P}}[\mathbb{1}_{z_i=k, g(x_i)=j}] \right| > \epsilon \right] \leq 2e^{-2\epsilon^2 n}$$

Recall from Theorem 3, our decision tree proxy will contain at most $\frac{\ln \gamma - \ln \sqrt{2}}{\ln(1-\gamma)}$ splits, and therefore there will be at most $\frac{\ln \gamma - \ln \sqrt{2}}{\ln(1-\gamma)}$ unique proxy groups. Applying a union bound over all k, j pairs and fixed g , we see that

$$\Pr_{D \sim \Omega} \left[\left| \bigcap_{k,j} \mathbb{1}_{\left| \frac{1}{n} \sum_{i=1}^n w_j(x_i) \mathbb{1}_{z_i=k, g(x_i)=j} - \mathbb{E}_{(x,y,z) \sim \mathcal{P}}[\mathbb{1}_{z_i=k, g(x_i)=j}] \right|} > \epsilon \right] \leq 2K \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)} e^{-2\epsilon^2 n}$$

Again applying a union bound, this time over the model class g – with VC dimension d – as well as k, j pairs, we see that for all k, j, g ,

$$\Pr_{D \sim \Omega} \left[\left| \bigcap_{k,j} \mathbb{1}_{\left| \frac{1}{n} \sum_{i=1}^n w_j(x_i) \mathbb{1}_{z_i=k, g(x_i)=j} - \mathbb{E}_{(x,y,z) \sim \mathcal{P}}[\mathbb{1}_{z_i=k, g(x_i)=j}] \right|} > \epsilon \right] \leq 2dK \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)} e^{-2\epsilon^2 n}$$

Setting this to be less than $\frac{\delta}{3}$, we obtain $2dK \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)} e^{-2\epsilon^2 n} \leq \frac{\delta}{3}$, which implies $n \geq \frac{1}{2\epsilon^2} \ln \frac{6dk(\ln \beta - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$. Then, with probability $1 - \frac{\delta}{3}$

$$\|(A - \tilde{A})\rho\|_2 \leq \sqrt{\sum_{j=1}^J \left(\sum_{k=1}^K (A_{k,j} - \tilde{A}_{k,j}) \cdot \rho_j \right)^2} < \sqrt{\sum_{j=1}^J (K\epsilon\rho_j)^2} \leq K\epsilon \sqrt{\frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)}}$$

This bounds the degradation we expect in balance when we apply the proxy out of sample. Next, we consider the degradation in disclosiveness, which will depend on our estimates of $\Pr_{z \sim \mathcal{P}_z}[z = k]$ and $\Pr_{z|x \sim \mathcal{P}_{z|x}}[z = k|g(x) = j]$. First, we bound the empirical estimate of $\Pr_{z \sim \mathcal{P}_z}[z = k]$. Applying Hoeffding's inequality gives $\Pr_{D \sim \Omega} \left[\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{z_i=k} - \mathbb{E}_{z \sim \mathcal{P}_z}[\mathbb{1}_{z_i=k}] \right| > \epsilon \right] \leq 2e^{-2\epsilon^2 n}$. Applying a union bound over the range of Z gives $\Pr_{D \sim \Omega} \left[\bigcap_{i=1}^K \left| \frac{1}{n} \sum_{l=1}^n \mathbb{1}_{z_i=k} - \mathbb{E}_{z \sim \mathcal{P}_z}[\mathbb{1}_{z_i=k}] \right| > \epsilon \right] \leq 2Ke^{-2\epsilon^2 n}$. Setting this to be less than $\frac{\delta}{3}$ gives us: $2Ke^{-2\epsilon^2 n} \leq \frac{\delta}{3} \implies n \geq \frac{1}{2\epsilon^2} \ln \frac{6K}{\delta}$.

Finally, repeating the exercise for $\Pr_{z|x \sim \mathcal{P}_{z|x}}[z|g(x)]$, we have that for fixed $g \in \mathcal{G}$ and $z \in Z$,

$$\Pr_{D \sim \Omega} \left[\left| \sum_{i=1}^n \frac{w_j(x_i)}{\sum_{i=1}^n w_j(x_i)} \mathbb{1}_{z_i=k|g(x_i)=j} - \sum_{i=1}^n w_j(x_i) \mathbb{E}_{z|x \sim \mathcal{P}_{z|x}}[\mathbb{1}_{z_i=k|g(x_i)=j}] \right| > \epsilon \right] \leq 2e^{-2\epsilon^2 \sum_{i=1}^n w_j(x_i)}$$

Applying a union bound over z, g , and the VC dimension of \mathcal{G} , and setting the probability to be less than $\frac{\delta}{3}$ gives us $2dK \frac{\ln \beta - \ln \sqrt{2}}{\ln(1-\gamma)} e^{-2\epsilon^2 \sum_{i=1}^n w_j(x_i)} \leq \frac{\delta}{3}$, which implies $\sum_{i=1}^n w_j(x_i) \geq \frac{1}{2\epsilon^2} \ln \frac{6dK(\ln \beta - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$. Then, with probability $1 - \delta$ both of our estimates for $\Pr_{z|x \sim \mathcal{P}_{z|x}}[z|g(x)]$ and $\Pr_{z \sim \mathcal{P}_z}[z]$ must be within ϵ of the true parameters if we have sample

count $n \geq \frac{1}{2\epsilon^2} \max\{\ln \frac{6K}{\delta}, \ln \frac{6dK(\ln \gamma - \ln \sqrt{2})}{\delta \ln(1-\gamma)}\}$. Note that $\max\{\ln \frac{6K}{\delta}, \ln \frac{6dK(\ln \gamma - \ln \sqrt{2})}{\delta \ln(1-\gamma)}\} \leq \ln \frac{6dK(\ln \gamma - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$. Then, taking $n \geq \frac{1}{2\epsilon^2} \ln \frac{6dK(\ln \gamma - \ln \sqrt{2})}{\delta \ln(1-\gamma)}$ suffices. Finally, we can apply our concentration bounds to the expression for disclosure level. If we obtain an α -disclosive proxy in-sample, this is equivalent to satisfying, for all $z \in Z$ and $g \in \mathcal{G}$, $|\Pr_{z|x \sim D}[z|g(x)] - \Pr_{z \sim D}[z]| \leq \alpha \implies |\Pr_{z|x \sim \mathcal{P}_{z|x}}[z|g(x)] - \Pr_{z \sim \mathcal{P}_z}[z]| \leq \alpha + 2\epsilon$ \blacktriangleleft

B Additional Experimental Details

In Figure 2, we show trade-off curves for balance and disclosure when using XGB as the base model. In these plots we also explore a slight relaxation of our (α, β) Proxy, in which we remove the constraint $\sum_i q_i = 1$ when solving $\min_q \|qA - U\|_2$ subject to $q_i \geq 0 \forall i$. We find that both the original and relaxed version perform similarly. On Communities, there is less stability displayed by the proxies trained with the XGB base model compared to those trained with the PRC base model. On Adult, the proxies trained with XGB as a base model exhibit a smoother trade-off curve. On the Marketing data set, our proxy approach dominates in sample for smaller levels of α but struggles to generalize.

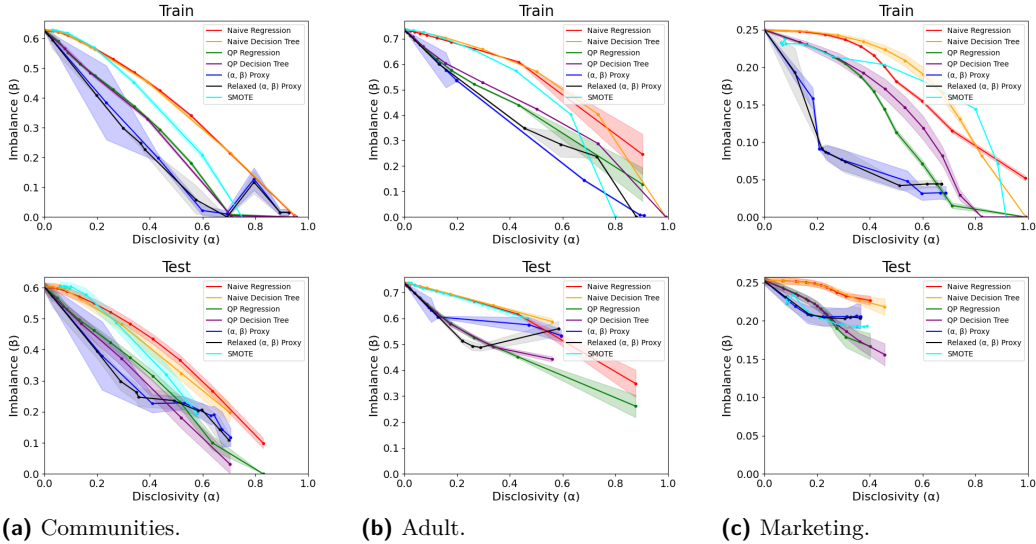
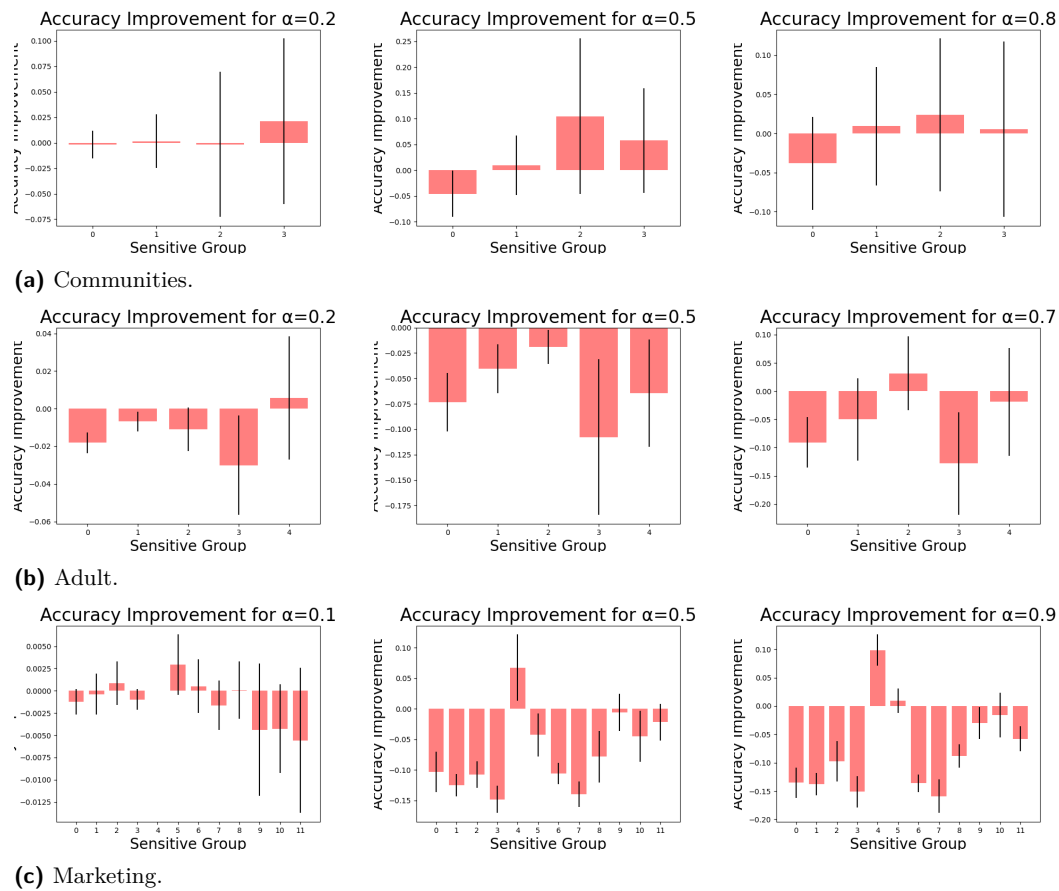


Figure 2 Trade-off of Disclosure and Balance for Proxy Models on the Communities, Adult, and Marketing data sets with XGBoost Base Model.

Next, we analyze the downstream fairness impact resulting from using our proxy filtering approach to prepare machine learning training datasets. Here, we train a model for the data-specific classification or regression task on an unfiltered sample and a filtered sample of the same size, and we compare the differences in group-wise accuracy obtained by each model. We consider the downstream fairness impact of training a model on data that has been filtered by our (α, β) proxy function but find our results are inconclusive. While we are able to theoretically guarantee a certain level of balance in the filtered data set, we cannot guarantee that the distribution over features and labels will not be skewed in the filtered set, nor can we guarantee that the distribution over features and labels given sensitive attributes will not be distorted. To test this, we first use an (α, β) proxy with a specified α budget to filter the Test set into a balanced sub-sample. Then, we train two model for the dataset specific classification task, one on the filtered data, and the other on a down-sampled version



■ **Figure 3** Difference in accuracy between models trained on filtered and unfiltered data on the Communities, Adult, and Marketing data sets with PRC base model.

of the original Test set of the same size. We calculate the accuracy of the models on each sensitive group and then plot the *difference* in accuracy between the two models, calculated as the group accuracy on the filtered data minus the group accuracy on the unfiltered data. Thus, positive values indicate an improvement in group accuracy from training on the filtered data, while negative values indicate a decrease. Between the three data sets, we see mixed results, displayed in Figure 3. On the Communities data set, we broadly see improvement on lower accuracy groups when using the model trained on the filtered data. However, results from the Adult data set in show a decrease in performance across all groups, and results from the Marketing data set show improvement for one of the least represented groups, but a decrease in performance for most others.

Distribution-Specific Auditing for Subgroup Fairness

Daniel Hsu ✉ 

Columbia University, New York, NY, USA

Jizhou Huang ✉ 

Washington University in St. Louis, MO, USA

Brendan Juba ✉ 

Washington University in St. Louis, MO, USA

Abstract

We study the problem of auditing classifiers for statistical subgroup fairness. Kearns et al. [20] showed that the problem of auditing combinatorial subgroups fairness is as hard as agnostic learning. Essentially all work on remedying statistical measures of discrimination against subgroups assumes access to an oracle for this problem, despite the fact that no efficient algorithms are known for it. If we assume the data distribution is Gaussian, or even merely log-concave, then a recent line of work has discovered efficient agnostic learning algorithms for halfspaces. Unfortunately, the reduction of Kearns et al. was formulated in terms of weak, “distribution-free” learning, and thus did not establish a connection for families such as log-concave distributions. In this work, we give positive and negative results on auditing for Gaussian distributions: On the positive side, we present an alternative approach to leverage these advances in agnostic learning and thereby obtain the first polynomial-time approximation scheme (PTAS) for auditing nontrivial combinatorial subgroup fairness: we show how to audit statistical notions of fairness over homogeneous halfspace subgroups when the features are Gaussian. On the negative side, we find that under cryptographic assumptions, no polynomial-time algorithm can guarantee any nontrivial auditing, even under Gaussian feature distributions, for general halfspace subgroups.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Fairness auditing, agnostic learning, intractability

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.5

Related Version *Previous Version:* <https://arxiv.org/abs/2401.16439>

Funding This work was partially supported by NSF awards IIS-2040971, IIS-1908287, and IIS-1942336, and NSF-Amazon award IIS-1939677.

1 Introduction

The deployment of decision rules obtained using machine learning has raised the risk that the rules may exhibit biases against historically marginalized communities. In particular, Kearns et al. [20] raised the concern that these decision rules may be biased against subgroups characterized by a combination of “protected” attributes. Since there are an exponential number of such subgroups, even detecting such statistical patterns of discrimination is a nontrivial computational problem; indeed, Kearns et al. [20] showed that the problem of finding disadvantaged subgroups is equivalent to the problem of agnostic learning, which is believed to be intractable in general for all but the simplest classes of sets. Essentially all work [20, 23, 18] on remedying statistical measures of discrimination against subgroups assumes access to an oracle for this problem, despite the fact that no efficient algorithms are known for it. In this work we are proposing a solution for a variant of the fairness



© Daniel Hsu, Jizhou Huang, and Brendan Juba;
licensed under Creative Commons License CC-BY 4.0
5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 5; pp. 5:1–5:20



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

auditing problem with provable guarantees of efficiency and correctness, as well as some strong limitations on the extent to which these solutions can be extended to richer families of subgroups.

1.1 Background and Motivation

Fairness learning has received massive attention in recent years. It turns out learning a fair classifier, in most cases, is equivalent to auditing [20, 23, 18]. In particular, if auditing is possible, learning a fair classifier is easy. There are many successful examples of fairness learning with auditing over a relatively small number of predetermined subgroups [1, 29]. However, a small number of predetermined subgroups, in many cases, is not enough to cover all the natural subgroups.

► **Example 1.** In the court case “DeGraffenreid v General Motors” [6], five Black women brought suit against General Motors for its discrimination against the group of Black women. Although no sex discrimination was revealed, the evidence showed that Black women hired after 1970 were discriminated against by the company’s seniority system. Such discrimination can be better illustrated by an example shown in Table 1. In particular, the hiring rate of a company could seemingly be fair in terms of gender or race alone, but clearly discriminates against the subgroups of white men and black women. The court rejected the plaintiffs’ attempt to bring a suit not on behalf of Blacks or women, but specifically on behalf of Black women. In the ruling, in favor of the defendant, the judge was specifically concerned about the proliferation of protected classes.

■ **Table 1** an example of discrimination against subgroups.

	men	women	total
black	50	0	50
white	0	50	50
total	50	50	100

More generally, a classifier may appear to be fair on each individual attribute, e.g., gender, race, age, incomes, etc., and yet perform unfairly on subgroups defined on multiple attributes, i.e., the conjunction of such attributes. In the case of *DeGraffenreid v General Motors*, it is the conjunction of race and gender being discriminated against. The possible number of the conjunctions grows exponentially as the number of the “protected” attributes increases.

Thereafter, [20] proposed more general notions of statistical fairness that require auditing over subgroups defined on simple combinations of data features. Specifically, such combinations of features can be any simple representations, such as conjunctions and halfspaces, which, however, can generate exponentially many subgroups. They also showed that the problem of auditing subgroups defined by such simple representation is as hard as “weak agnostic learning” in the standard “distribution-free” setting [17, 22]. While the problem of distribution-free weak agnostic learning is widely believed to be computationally intractable [22, 12], its hardness does not necessarily hold for specific distribution families. Thus, it is natural to consider auditing using distribution-specific agnostic learning approaches as agnostic learning is a much more extensively studied problem. However, it turns out there are still obstacles remaining for doing so.

1.2 Challenges of Auditing through Agnostic Learning

The main challenge that prevents us from applying existing agnostic learning techniques to perform auditing based on the reduction by [20] is that it is formulated in terms of weak agnostic learning, that is, finding classifiers with error rates that are nonnegligibly better than guessing, and correspondingly weak auditing guarantees. In particular, the approximation guarantees we obtain for distribution-specific agnostic learning yield vacuous guarantees for weak learning. When we have guarantees for arbitrary distributions, “boosting” [28] enables us to obtain high accuracy from such weak learners. Unfortunately, these techniques require re-weighting the data examples after which the distribution-specific properties may no longer hold.

One might hope to dodge this issue by casting the problem of finding a harmed subgroup as a Mixed-Integer Program and using solvers that, though they lack polynomial-time guarantees, obtain adequate performance in practice. In such an approach, the failure of the solver to find a feasible solution to the optimization problem is taken as the proof that the classifier is fair. Unfortunately, these solvers owe their speed in part to a lack of soundness, both due to numerical issues [5] and the complexity of the heuristics used to prune the search [2, 14], and it remains a current research challenge to obtain acceptable performance (using the various advanced techniques employed by commercial solvers) while retaining the guarantee that the solver correctly reports infeasibility [4]. In any case, the works by [20, 21] and [24] that empirically studied these approaches to obtaining fair classifiers used linear regression as a proxy for the agnostic learning or cost-sensitive classification subroutines. Unfortunately, these heuristics do not even provide in-principle guarantees.

In this paper, we will show auditing general halfspace subgroups is hard even for data with a Gaussian distribution, and present an alternative auditing approach for subgroups determined by homogeneous halfspaces with provable guarantees.

1.3 Our Contribution

Our first contribution is a more careful analysis of the relationship between auditing and agnostic learning: Given a fixed positive classification rate, the harm (w.r.t. statistical parity) suffered by a subgroup is affinely related to the error rate of the subgroup indicator. Thus, a solution to the agnostic learning problem directly gives a harmed subgroup. Note that whereas the fairness objective refers to conditioning on a group, which generally doesn’t preserve a distributional assumption, agnostic learning instead refers to the accuracy under that “nice” distribution, and hence is easier to analyze. Also note that under a standard normal distribution, the subclass of halfspaces with a fixed positive classification rate is given by the halfspaces with unit normal vectors and the same threshold.

► **Remark 2.** Our reduction to learning halfspaces with fixed positive classification rates can achieve arbitrarily high precision auditing and does not rely on re-weighting data examples or make any assumptions on the potentially unfair classifiers. This enables the use of the existing distribution-specific agnostic learning methods for auditing.

Based on the reduction and a inspiration from Diakonikolas et al. [7], our second major contribution is a lower bound on the unfairness detectable when auditing for halfspace subgroups under Gaussian distributions by reducing the problem of continuous Learning With Errors (cLWE) to auditing. Our hardness results include both multiplicative and additive forms. More interestingly, we can further show that even “nonconstructive auditing” is hard, where we do not need to exhibit a discriminated subgroup for a failed audit.

For our algorithmic results, we will present a general auditing framework given an oracle for (distribution-specific) agnostic learning. Also, we give a randomized PTAS auditing algorithm for subgroups determined by homogeneous halfspaces under Gaussian data by applying the method from Diakonikolas et al. [8].

► **Remark 3.** We stress that a PTAS for auditing subgroups defined by homogeneous halfspaces for Gaussian distributions is, in fact, the best guarantee we know so far, hence, not trivial.

At first blush, the reliance on a (prima facie unverifiable) distributional assumption for the analysis of our auditing algorithm may seem to be at odds with our desire to certify the fairness of a classifier. Nevertheless, a line of recent works by Rubinfeld and Vasilyan [27] and Gollakota et al. [15] have shown that the properties of the data that are crucial to these algorithms for distribution-specific learning of halfspaces *can be verified*. Thus, these methods give a way of certifying fairness for families of nice distributions: so long as the data passes these tests and the audit reveals no subgroup that is significantly harmed, we may *guarantee that the classifier is fair*.

This paper will be organized as follows. Some necessary background for our arguments are given in Section 2. We will present the main reduction from auditing to agnostic learning in Section 3. Then, we will show the hardness results in Section 4. Section 5 will present our auditing framework as well as the distribution-specific PTAS algorithm. Finally, we will discuss the limitations of our approach and suggest directions for future work.

1.4 Related Work

Many authors have considered the problem of ensuring fairness in classification, and Barocas et al. [3] give a good overview of the broader area. In particular, there are alternatives to the statistical, group-fairness notions we are considering, for example individual-level fairness as proposed by Dwork et al. [11], or based on causal modeling, such as the “counterfactual” fairness notion proposed by Kusner et al. [25]. We cannot do justice to the breadth of literature and philosophical issues here, and we strongly encourage the interested reader to consult Barocas et al. The group-fairness notions we consider have their roots in the game-theory-based approach of Kearns et al. [20] for learning representations with subgroup fairness by assuming there exists an efficient oracle for auditing. A follow-up study [21] evaluated their algorithm on real-world datasets. Hébert-Johnson et al. [18] showed a method of obtaining “multi-accurate” representations by assuming the existence of an efficient auditing oracle. Further, Kim et al. [23] proposed a variant of statistical fairness called “multi-fairness,” which allows them to efficiently learn a multi-fair classifier with querying “relative fairness” of data pairs. As we discussed previously, the auditing oracles in these works were provided by using linear regression as a heuristic for the optimal halfspace, which does not provide guarantees. They also did not consider auditing for specific families of distributions. On the other hand, the works on agnostic learning for specific families of distributions, e.g., [19, 9, 8, 10, 13] do not consider how their techniques may be applied to the subgroup fairness auditing problem.

2 Preliminaries

We use lowercase bold font characters to represent real vectors and subscripts to index the coordinates of each vector, e.g., \mathbf{x}_i represents the i -th coordinate of vector \mathbf{x} . We denote the l_p -norm by $\|\mathbf{x}\|_p = (\sum_i \mathbf{x}_i^p)^{1/p}$, and $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$. We model each individual as a vector of protected attributes, i.e., $\mathbf{x} \in \mathcal{X}$.

Further, the probability of an event under a distribution \mathcal{D} is denoted by $\Pr_{\mathbf{x} \sim \mathcal{D}}\{\cdot\}$. $\mathcal{N}(0, \mathbf{I})$ denotes a standard normal distribution, where \mathbf{I} represents the identity matrix. For simplicity of notation, we may use $\mathcal{N}, \mathcal{N}_\sigma$ instead of $\mathcal{N}(0, \mathbf{I}), \mathcal{N}(0, \sigma^2 \mathbf{I})$ or even drop \mathcal{D} and \mathcal{N} from the subscript when it is clear from the context.

► **Fact 4** (Rotational Invariance). *For any real vector \mathbf{u} , if $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$, then $\bar{\mathbf{u}}^\top \mathbf{x} \sim \mathcal{N}(0, 1)$.*

To understand the problem of fairness auditing, it is necessary to define fairness or unfairness precisely. In this work, we focus on the notion of Statistical Parity Subgroup Fairness (SPSF). Formally, we have the following definition.

► **Definition 5** (Statistical Parity Subgroup Fairness). *Fix any binary classifier $c \in \mathcal{C}$ such that $c: \mathbb{R}^d \rightarrow \{-1, +1\}$, data distribution \mathcal{D} , collection of subgroups \mathcal{G} , and parameter $\gamma \in [0, 1]$. Define*

$$d_{\mathcal{D}}(c, g) = \Pr_{\mathbf{x} \sim \mathcal{D}}\{c(\mathbf{x}) = 1\} - \Pr_{\mathbf{x} \sim \mathcal{D}}\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in g\} \quad (1)$$

We say that c does not satisfy γ -statistical parity fairness (or is γ -unfair) with respect to \mathcal{D} and \mathcal{G} , if $\exists g \in \mathcal{G}$ such that

$$\Pr_{\mathbf{x} \sim \mathcal{D}}\{\mathbf{x} \in g\} |d_{\mathcal{D}}(c, g)| \geq \gamma \quad (2)$$

Equation (1) is a straightforward way to quantify how much the positive classification rate within a subgroup deviates from that of the overall population. The weighting by the size of the group (i.e., $\Pr_{\mathbf{x} \sim \mathcal{D}}\{\mathbf{x} \in g\}$) is a concession to address the statistical issues that arise with estimating d on small groups: we cannot escape that our empirical estimates are less accurate as the size shrinks. Our approach makes no assumptions on the form of the function c ; note therefore, that by replacing c with other functions of \mathbf{x} , such as whether a given classifier agrees with a given label, or whether the classifier makes a false-positive error, our results will immediately extend to other standard notions of statistical subgroup fairness. The goal of fairness auditing is to develop an “auditing algorithm” to efficiently find such a certificate $g \in \mathcal{G}$ for any $c \in \mathcal{C}$ with sample access to \mathcal{D} , formalized as follows.

► **Definition 6** (Constructive Auditing [20]). *Fix a collection of group indicators \mathcal{G} over the protected features, and any $\delta, \gamma, \gamma' \in (0, 1)$ such that $\gamma' \leq \gamma$. A constructive (γ, γ') -auditing algorithm for \mathcal{G} with respect to distribution \mathcal{D} is an algorithm \mathcal{A} such that for any classifier h , when given access the joint distribution $(\mathcal{D}, h(\mathcal{D}))$, \mathcal{A} runs in time $\text{poly}(1/\gamma', \log(1/\delta))$, and with probability $1 - \delta$, outputs a γ' -unfair certificate for h whenever h is γ -unfair with respect to \mathcal{D} and \mathcal{G} . If h is γ' -fair, \mathcal{A} will output “fair”.*

Moreover, we will consider a more general type of auditing task, called “non-constructive auditing”, where the algorithms are only required to tell if a discriminated subgroup exists.

► **Definition 7** (Non-constructive Auditing). *Under the same setting as Definition 6, a non-constructive (γ, γ') -auditing algorithm for \mathcal{G} with respect to distribution \mathcal{D} is an algorithm \mathcal{A} such that for any classifier h , when given access the joint distribution $(\mathcal{D}, h(\mathcal{D}))$, \mathcal{A} runs in time $\text{poly}(1/\gamma', \log(1/\delta))$, and with probability $1 - \delta$, claims h is γ' -unfair whenever h is γ -unfair with respect to \mathcal{D} and \mathcal{G} . If h is γ' -fair, \mathcal{A} will output “fair”.*

In this work, we will mainly focus on subgroups defined on halfspaces, a.k.a. linear threshold functions (LTF) over a d -dimensional real domain. Formally:

5:6 Distribution-Specific Auditing for Subgroup Fairness

► **Definition 8** (Halfspaces). *The class of halfspaces over \mathbb{R}^d is defined as $\mathcal{H}^d := \{\mathbf{x} \mapsto \text{sgn}(\mathbf{v}^\top \mathbf{x} - t) \mid \mathbf{x}, \mathbf{v} \in \mathbb{R}^d, t \in \mathbb{R}\}$ where $\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & \text{otherwise} \end{cases}$. In particular, the class of homogeneous halfspaces can be defined as $\{\mathbf{x} \mapsto \text{sgn}(\mathbf{v}^\top \mathbf{x}) \mid \mathbf{x}, \mathbf{v} \in \mathbb{R}^d\}$.*

Since our reduction involves the subclass of halfspace subgroups of a fixed size, we give the formal definition of it as follows.

► **Definition 9** (Fixed-size Halfspaces). *We use \mathcal{H}^d to represent the collection of all halfspaces in \mathbb{R}^d . Then, for any arbitrary distribution \mathcal{D} over \mathbb{R}^d , we define the collection of all halfspaces with the same (relative) density μ as*

$$\mathcal{H}_\mu^{\mathcal{D}} := \{h \in \mathcal{H}^d \mid \Pr_{\mathbf{x} \in \mathcal{D}}\{h(\mathbf{x}) = 1\} = \mu\} \quad (3)$$

In particular, the class of homogeneous halfspaces for a mean-0 Gaussian distribution is $\mathcal{H}_{1/2}^{\mathcal{N}(0, \Sigma)}$.

For conciseness, we may abbreviate $\Pr\{f(\mathbf{x}) = 1\}$ and $\Pr\{f(\mathbf{x}) = -1\}$ to simply $\Pr\{f\}$ and $\Pr\{\neg f\}$ for any binary output functions $f: \mathcal{X} \rightarrow \{-1, +1\}$ in the rest of the paper.

To state the hardness results, we denote $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$, $\mathbb{Z}_q := \{0, 1, \dots, q-1\}$, $\mathbb{R}_q := [0, q)$, and $\text{mod}_q: \mathbb{R}^d \rightarrow \mathbb{R}_q$ for the unique translation of the input by $q\mathbb{Z}^d$ to \mathbb{R}_q for $q \in \mathbb{N}$. The hardness of distribution-specific auditing is based on the assumption that the problem of ‘‘Learning With Errors’’ (LWE) is computationally intractable. Informally speaking, in the problem of LWE, we are given labelled examples from two hypothesis cases. In one case, the labels are biased by some secret vector, while, in another case, the labels are generated uniformly at random. We wish to distinguish between these cases. We formally define the problem of LWE [26], following [7]:

► **Definition 10** (Learning With Errors). *For $m, d \in \mathbb{N}$, $q \in \mathbb{R}_+$, let $\mathcal{D}_{\text{sample}}, \mathcal{D}_{\text{secret}}, \mathcal{D}_{\text{noise}}$ be distributions on $\mathbb{R}^d, \mathbb{R}^d, \mathbb{R}$ respectively. In the $\text{LWE}(m, \mathcal{D}_{\text{sample}}, \mathcal{D}_{\text{secret}}, \mathcal{D}_{\text{noise}}, \text{mod}_q)$ problem, with m independent samples (\mathbf{x}, y) , we want to distinguish between the following two cases:*

- **Alternative hypothesis:** (\mathbf{x}, y) is generated as $y = \text{mod}_q(\mathbf{s}^\top \mathbf{x} + z)$, where $\mathbf{x} \sim \mathcal{D}_{\text{sample}}, \mathbf{s} \sim \mathcal{D}_{\text{secret}}, z \sim \mathcal{D}_{\text{noise}}$.
- **Null hypothesis:** y is sampled uniformly at random on the support of its marginal distribution in the alternative hypothesis, independent of $\mathbf{x} \sim \mathcal{D}_{\text{sample}}$.

An algorithm is said to be able to solve the LWE problem with Δ advantage if the probability that the algorithm outputs ‘‘alternative hypothesis’’ is Δ larger than the probability that it outputs ‘‘null hypothesis’’ when the given data is sampled from the alternative hypothesis distribution.

This problem is widely believed to be computationally hard, formalized as follows.

► **Assumption 11** (Sub-exponential LWE Assumption). *For $q, \kappa \in \mathbb{N}, \alpha \in (0, 1)$ and $C > 0$ being a sufficiently large constant, the problem $\text{LWE}(2^{O(n^\alpha)}, \mathbb{Z}_q^d, \mathbb{Z}_q^d, \mathcal{N}_\sigma, \text{mod}_q)$ with $q \leq d^\kappa$ and $\sigma = C\sqrt{d}$ cannot be solved in $2^{O(d^\alpha)}$ time with $2^{O(-d^\alpha)}$ advantage.*

3 From Auditing To Agnostic Learning

In this section, we describe our reduction from auditing to agnostic learning. In addition, we give a lower bound for fairness auditing under Gaussian distributions.

We are considering the auditing problem w.r.t. SPSF as in Definition 5, which naturally rules out the statistically small subgroups. Indeed, if the probability of accessing the data of certain sub-population is exponentially small, it is statistically hard to even estimate their deviation. Therefore, it makes sense to just consider the collection of subgroups \mathcal{G} that are statistically large enough, e.g., $\Pr\{\mathbf{x} \in g\} = \Theta(1)$ for $\mathbf{x} \in \mathbb{R}^d$.

Based on the observation, the following optimization program, $\mathcal{P}_{a,b}^{\mathcal{D}}(\mathcal{G})$, can capture the most unfair subgroup which is also statistically significant enough. That is

$$\begin{aligned} \max_{g \in \mathcal{G}} \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{\mathbf{x} \in g\} |d_{\mathcal{D}}(c, g)| \\ \text{s.t.} \quad & a \leq \Pr_{\mathbf{x} \in \mathcal{D}} \{\mathbf{x} \in g\} \leq b \end{aligned} \quad (4)$$

for some constants $0 < a \leq b < 1$.

Furthermore, if we only consider the subgroups represented by halfspaces, i.e., $\mathcal{G} \equiv \mathcal{H}^d$, there exists a simple reduction from $\mathcal{P}_{a,b}^{\mathcal{D}}(\mathcal{H}^d)$ to agnostic learning that, in particular, preserves the properties of the data distribution. We show our reduction as the following theorem.

► **Theorem 12 (Main Reduction).** *Given any binary classifier $c : \mathbb{R}^d \rightarrow \{-1, +1\}$, and a data distribution \mathcal{D} over \mathbb{R}^d whose 1-dimensional marginals have continuous cumulative distribution functions, if there exists an efficient algorithm for learning $\mathcal{H}_{\mu}^{\mathcal{D}}$ in the agnostic model on distribution \mathcal{D} , then there is an efficient auditing algorithm for c on subgroups represented by \mathcal{H}^d over distribution \mathcal{D} .*

We delay the proof of the above theorem to the end of this section, and show two fundamental hurdles we need to overcome in order to prove Theorem 12.

► **Remark 13.** While learning from a representation class like $\mathcal{H}_{\mu}^{\mathcal{D}}$ may seem to be hard at a first glance, there are actually examples [10] of learning $\mathcal{H}_{\mu}^{\mathcal{D}}$ in an agnostic setting under Gaussian data.

Instead of starting from the optimization problem (4), it turns out that solving a sequence of simpler optimization problems suffices to certify the γ -unfairness as stated in Definition 5. We state the equivalence in the following proposition. Its proof is deferred to the appendix.

► **Proposition 14.** *Consider any binary classifier $c : \mathbb{R}^d \rightarrow \{-1, +1\}$, any data distribution \mathcal{D} over \mathbb{R}^d whose 1-dimensional marginals have continuous cumulative distribution functions, and any $0 < a \leq b < 1$. For each pair of non-negative integers $k < n$, let $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$ denote the optimization program*

$$\begin{aligned} \max_{h \in \mathcal{H}^d} \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| \\ \text{s.t.} \quad & \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} = a + \frac{k(b-a)}{n}. \end{aligned}$$

Let h^* be a global optimizer of $\mathcal{P}_{a,b}^{\mathcal{D}}(\mathcal{H}^d)$, as defined in (4), and let $\gamma^* = \Pr\{h^*\} |d_{\mathcal{D}}(c, h^*)|$. For each $k = 0, \dots, n$, let h_k^* be a global optimizer of $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$. Then

$$\max_k \Pr\{h_k^*\} |d_{\mathcal{D}}(c, h_k^*)| \geq \gamma^* - \frac{2(b-a)}{n}.$$

The reason why this proposition is so crucial is that it allows us to solve a simpler optimization problem without compromising the guarantee. Being able to fix $\Pr\{h(\mathbf{x}) = 1\}$ as a constant will significantly simplify the overall optimization as it reduces the degree of

the optimization objective. In fact, it is because we can optimize $\Pr\{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)|$ over $\mathcal{H}_{\mu}^{\mathcal{D}}$ instead of \mathcal{H}^d that we can conduct the reduction from auditing to agnostic learning.

The following lemma shows a direct relationship between the unfairness level and the classification error.

► **Lemma 15.** *Given any binary classifier $c : \mathcal{X} \rightarrow \{-1, +1\}$, a data distribution \mathcal{D} over \mathcal{X} and a collection of subgroups $g \in \mathcal{G}$ such that $g : \mathcal{X} \rightarrow \{-1, +1\}$, we have*

$$2 \Pr\{g\} d_{\mathcal{D}}(c, g) = \Pr\{\neg c\} \Pr\{\neg g\} + \Pr\{c\} \Pr\{g\} - \Pr\{c(\mathbf{x}) = g(\mathbf{x})\}$$

for $\mathbf{x} \sim \mathcal{D}$.

Proof. By the law of total probability, we have

$$\Pr\{c \cap g\} = \Pr\{g\} - (\Pr\{\neg c\} - \Pr\{\neg c \cap \neg g\}).$$

which along with Definition 5 gives

$$\begin{aligned} d_{\mathcal{D}}(c, g) &= \Pr\{c\} - \Pr\{c \mid g\} \\ &= \frac{\Pr\{c\} \Pr\{g\} - \Pr\{c \cap g\}}{\Pr\{g\}} \\ &= \frac{\Pr\{\neg c\} \Pr\{\neg g\} - \Pr\{\neg c \cap \neg g\}}{\Pr\{g\}}. \end{aligned} \tag{5}$$

Summing up the two different forms of $d_{\mathcal{D}}(c, g)$ results to

$$\begin{aligned} 2d_{\mathcal{D}}(c, g) &= \frac{\Pr\{\neg c\} \Pr\{\neg g\} - \Pr\{\neg c \cap \neg g\}}{\Pr\{g\}} + \frac{\Pr\{c\} \Pr\{g\} - \Pr\{c \cap g\}}{\Pr\{g\}} \\ &= \frac{\Pr\{\neg c\} \Pr\{\neg g\} + \Pr\{c\} \Pr\{g\} - (\Pr\{\neg c \cap \neg g\} + \Pr\{c \cap g\})}{\Pr\{g\}} \end{aligned} \tag{6}$$

Notice that, because $c \cap g$ and $\neg c \cap \neg g$ are two disjoint events, we have

$$\begin{aligned} \Pr\{c(\mathbf{x}) = g(\mathbf{x})\} &= \Pr\{(c \cap g) \cup (\neg c \cap \neg g)\} \\ &= \Pr\{c \cap g\} + \Pr\{\neg c \cap \neg g\} \end{aligned}$$

Plugging it back in to Equation (6) produces the desired result. ◀

This immediately implies a duality between SPSF auditing and agnostic learning as follows.

► **Corollary 16.** *Given any binary classifier $c : \mathbb{R}^d \rightarrow \{-1, +1\}$, a data distribution \mathcal{D} and a collection of halfspaces $\mathcal{H}_{\mu}^{\mathcal{D}}$ over \mathbb{R}^d , we have the following two properties*

- (1) $d_{\mathcal{D}}(c, h^*) \geq d_{\mathcal{D}}(c, h), \forall h \in \mathcal{H}_{\mu}^{\mathcal{D}}$ if and only if $h^* = \operatorname{argmin}_{h \in \mathcal{H}_{\mu}^{\mathcal{D}}} \Pr_{\mathbf{x} \sim \mathcal{D}}\{c(\mathbf{x}) = h(\mathbf{x})\}$
- (2) $d_{\mathcal{D}}(c, h^*) \leq d_{\mathcal{D}}(c, h), \forall h \in \mathcal{H}_{\mu}^{\mathcal{D}}$ if and only if $h^* = \operatorname{argmax}_{h \in \mathcal{H}_{\mu}^{\mathcal{D}}} \Pr_{\mathbf{x} \sim \mathcal{D}}\{c(\mathbf{x}) = h(\mathbf{x})\}$

Proof. Because $\Pr\{c\}$ is a constant and $\Pr\{h\} = \mu, \forall h \in \mathcal{H}_{\mu}^{\mathcal{D}}$ by Definition 9, $d_{\mathcal{D}}(c, h)$ is simply an affine transformation of $\Pr\{c(\mathbf{x}) = h(\mathbf{x})\}$ for a fixed μ by Lemma 15, which implies the desired results. ◀

Proposition 14 tells us that solving $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$ for $k = 0, \dots, n$ would give us a good enough approximation to the maximum unfairness level, of course, with a large enough n . Therefore, we just need to further show that solving each $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$ is equivalent to learning $\mathcal{H}_{\mu}^{\mathcal{D}}$ to complete the reduction.

Formally, because $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$ can be equivalently written as

$$\max_{h \in \mathcal{H}_{\mu}^{\mathcal{D}}} \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| \quad (7)$$

for some $\mu = a + k(b - a)/n$, it suffices to prove the following theorem.

► **Lemma 17.** *Given any binary classifier $c: \mathbb{R}^d \rightarrow \{-1, +1\}$, a data distribution \mathcal{D} and a collection of halfspaces $\mathcal{H}_{\mu}^{\mathcal{D}}$ over \mathbb{R}^d such that*

$$\text{opt}_{\min} \leq \Pr_{\mathbf{x} \sim \mathcal{D}} \{c(\mathbf{x}) = h(\mathbf{x})\} \leq \text{opt}_{\max}$$

for all $h \in \mathcal{H}_{\mu}^{\mathcal{D}}$, if $h_{\mathbf{v}}, h_{\mathbf{u}} \in \mathcal{H}_{\mu}^{\mathcal{D}}$ satisfy that $\Pr\{c(\mathbf{x}) = h_{\mathbf{v}}(\mathbf{x})\} \leq \text{opt}_{\min} + 2\epsilon$ as well as $\Pr\{c(\mathbf{x}) = h_{\mathbf{u}}(\mathbf{x})\} \geq \text{opt}_{\max} - 2\epsilon$, we have either

$$\Pr_{\mathbf{x} \sim \mathcal{D}} \{h_{\mathbf{v}}(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h_{\mathbf{v}})| \geq \gamma^* - \epsilon \quad (8)$$

or

$$\Pr_{\mathbf{x} \sim \mathcal{D}} \{h_{\mathbf{u}}(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h_{\mathbf{u}})| \geq \gamma^* - \epsilon \quad (9)$$

where $\gamma^* = \max_{h \in \mathcal{H}_{\mu}^{\mathcal{D}}} \Pr_{\mathbf{x} \sim \mathcal{D}} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)|$.

Proof. By the proof of Lemma 15, we have

$$2 \Pr\{h\} |d_{\mathcal{D}}(c, h)| = \underbrace{|\Pr\{\neg c\} \Pr\{\neg h\} - \Pr\{\neg c \cap \neg h\}|}_{I_1} + \underbrace{|\Pr\{c\} \Pr\{h\} - \Pr\{c \cap h\}|}_{I_2}$$

Let $h^* \in \mathcal{H}_{\mu}^{\mathcal{D}}$ be such that $\Pr\{h^*\} |d_{\mathcal{D}}(c, h^*)| = \gamma^*$. Then for I_2 , we have

$$\begin{aligned} I_2 &= (\Pr\{c\} - \Pr\{c | h^*\} + \Pr\{c | h^*\}) \Pr\{h\} - \Pr\{c \cap h\} \\ &= \Pr\{h^*\} d_{\mathcal{D}}(c, h^*) + \Pr\{c \cap h^*\} - \Pr\{c \cap h\} \end{aligned}$$

where the last equation is because $h^* \in \mathcal{H}_{\mu}^{\mathcal{D}}$, then $\Pr\{h\} = \Pr\{h^*\} = \mu$ by Definition 9.

Similarly, for I_1 , we can write

$$\begin{aligned} I_1 &= \Pr\{\neg h^*\} (\Pr\{\neg c\} - \Pr\{\neg c | \neg h^*\}) + \Pr\{\neg c \cap \neg h^*\} - \Pr\{\neg c \cap \neg h\} \\ &= \Pr\{h^*\} d_{\mathcal{D}}(c, h^*) + \Pr\{\neg c \cap \neg h^*\} - \Pr\{\neg c \cap \neg h\} \end{aligned}$$

where the last equation follows because we have shown in the proof of Lemma 15 that $d_{\mathcal{D}}(c, h^*) = \Pr\{\neg h^*\} (\Pr\{\neg c\} - \Pr\{\neg c | \neg h^*\}) / \Pr\{h^*\}$.

Combining I_1 and I_2 will result to

$$\begin{aligned} \Pr\{h\} |d_{\mathcal{D}}(c, h)| &= |\Pr\{h^*\} d_{\mathcal{D}}(c, h^*) + \frac{\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h(\mathbf{x})\}}{2}| \\ &\geq \gamma^* - \frac{|\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h(\mathbf{x})\}|}{2} \end{aligned}$$

by triangle inequality. Further, since h^* maximizes $|d_{\mathcal{D}}(c, h)|$, it either maximizes or minimizes $d_{\mathcal{D}}(c, h)$. Then, by Corollary 16, we know

$$\Pr_{\mathbf{x} \sim \mathcal{D}} \{c(\mathbf{x}) = h^*(\mathbf{x})\} \in \{\text{opt}_{\min}, \text{opt}_{\max}\}$$

5:10 Distribution-Specific Auditing for Subgroup Fairness

which implies either

$$|\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h_{\mathbf{v}}(\mathbf{x})\}| \leq 2\epsilon$$

or

$$|\Pr\{c(\mathbf{x}) = h^*(\mathbf{x})\} - \Pr\{c(\mathbf{x}) = h_{\mathbf{u}}(\mathbf{x})\}| \leq 2\epsilon$$

Therefore, the proof is completed. \blacktriangleleft

► **Remark 18.** We emphasize that it is necessary for us to consider the guarantee of agnostic learning in an additive form rather than multiplicative form. Although Corollary 16 shows that the classification error, $\Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$, and the unfairness level, $\Pr\{h\} |d_{\mathcal{D}}(c, h)|$, are dual to each other over $\mathcal{H}_{\mu}^{\mathcal{D}}$, the affine relationship between them prohibits obtaining a guarantee on the unfairness from a multiplicative error. This also explains why the guarantee provided by [10] does not fit in our analysis.

Now we are ready to prove Theorem 12.

Proof of Theorem 12. To solve the auditing problem, we just need to solve the sequence of optimization problems, $\{\mathcal{P}_{a,b}^{\mathcal{D}}(k, n) \mid k = 0, \dots, n\}$ as described in Proposition 14. We can solve each $\mathcal{P}_{a,b}^{\mathcal{D}}(k, n)$ with an additive error ϵ by calling the given oracle of learning halfspaces with the same strategy specified in Lemma 17. Eventually, we solve all of these optimization problems with an $2(b-a)/n + \epsilon$ additive error and a running time of $O(n)$ factor overhead compared with that of the oracle. \blacktriangleleft

4 Intractability Of Auditing Under Gaussian Data

In this section, we will show that the problem of auditing halfspaces subgroups under a Gaussian distribution is computationally hard in two forms: the multiplicative form and additive form. To do so, we first show that distinguishing between fair and unfair cases with respect to halfspace subgroups for Gaussian data is hard. Then, the hardness of auditing will follow as corollaries.

4.1 Indistinguishability Of Unfairness

We claim it is computationally hard to distinguish between halfspace subgroups that are evenly fair and halfspace subgroups among which there exists a slightly unfair subgroup with significant advantage.

► **Theorem 19.** *Under Assumption 11, for any $d \in \mathbb{N}$, any constants $\alpha \in (0, 1)$, $\beta \in \mathbb{R}_+$, and any $\log^{\beta} d \leq k \leq cd$ where c is a sufficiently small constant, there is no algorithm that runs in time $d^{O(k^{\alpha})}$ and distinguishes between the following two cases of a joint distribution \mathcal{D} of $(\mathbf{x}, c(\mathbf{x}))$ supported on $\mathbb{R}^d \times \{-1, +1\}$ with marginal $\mathcal{D}_{\mathbf{x}} = \mathcal{N}(0, \mathbf{I})$, with $d^{-O(k^{\alpha})}$ advantage:*

- (i) **Alternative Hypothesis:** *There exist non-negligibly unfair halfspace subgroups, specifically $\exists h \in \mathcal{H}^d, \Pr_{\mathcal{D}}\{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| = \Omega(1/\sqrt{k \log d})$.*
- (ii) **Null Hypothesis:** *All halfspace subgroups are perfectly fair, i.e., $\Pr_{\mathcal{D}}\{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| = 0, \forall h \in \mathcal{H}^d$.*

The above theorem simply states that the closer the unfairness level of the alternative hypothesis is to zero ($k \log d$ is large), the harder it is to distinguish between these two cases, where the hardness is reflected on the running time $d^{O(k^{\alpha})}$. Hence, if we restrict the running

time to a certain order, there is a limitation on how large $k \log d$ can be for someone to be able to distinguish between them with a significant enough advantage. It is this observation that allows us to prove the hardness of auditing in the next section.

The idea behind the proof of this theorem is to observe that the data generated in the two hypotheses in certain LWE instances can be reduced to binary labelled ones through rounding. With such a reduction, the distribution from the null hypothesis case of LWE will produce perfectly fair data, while the distribution from alternative hypothesis will yield slightly biased labels where a unfair halfspace subgroup therefore exists. Thus, if we can distinguish between the fair case from the unfair case with some marginal error, we can solve the LWE problem. We defer the formal proof to the appendix.

4.2 Auditing With Small Error Is Hard

We now show that the hardness of distinguishability implies the hardness of auditing with both multiplicative error and additive error.

Suppose an auditing algorithm is guaranteed to return us a γ' -unfair certificate (a halfspace) given a γ -unfair classifier c , where $\gamma' \leq \gamma \leq 1$. The following corollaries show that γ' can never be close to γ .

► **Corollary 20** (multiplicative form). *Given Assumption 11, there is no polynomial-time $1/\text{poly}(d)$ -approximation algorithm for constructive auditing for halfspace subgroups under Gaussian marginals in \mathbb{R}^d .*

Proof. Suppose there exists an auditing algorithm that guarantees to return a $\delta\gamma$ -unfair certificate given a γ -unfair collection of halfspace subgroup and access to data with a Gaussian marginal, where $\delta \in (0, 1)$.

For the alternative hypothesis case as described in Theorem 19, given a $1/\sqrt{k \log d}$ -unfair collection of halfspace subgroups, we run such an algorithm to obtain a $\delta/\sqrt{k \log d}$ -unfair certificate, i.e., a halfspace h such that $\Pr_{\mathbf{x} \sim \mathcal{N}}\{h(\mathbf{x}) = 1\} | d_{\mathcal{N}}(c, h) | \geq \delta/\sqrt{k \log d}$. By the Hoeffding Bound, we can verify that the empirical estimation of $\Pr_{\mathbf{x} \sim \mathcal{N}}\{h(\mathbf{x}) = 1\} | d_{\mathcal{N}}(c, h) |$ is ε_1 -close to $\delta/\sqrt{k \log d}$ with high probability by drawing $O(1/\varepsilon_1^2)$ examples from the distribution constructed in the alternative hypothesis case.

For the null hypothesis case, with the same argument, we can verify there is no ε_2 -unfair subgroup with high probability given $O(1/\varepsilon_2^2)$ examples from the distribution in the null hypothesis case.

Suppose $\delta = \Omega(1/\text{poly}(d))$, notice that we only need $\varepsilon_1, \varepsilon_2$ to be $O(1/\text{poly}(d))$ to ensure $\delta/\sqrt{k \log d} - \varepsilon_1 > \varepsilon_2$. However, this implies that our auditing algorithm can distinguish between the two cases in Theorem 19 with high probability and only runs in polynomial time, which contradicts to the hardness assumption. ◀

► **Corollary 21** (additive form). *Given Assumption 11, for any constants $\alpha \in (0, 1), \beta \in \mathbb{R}_+$, and any $C/\sqrt{d \log d} \leq \epsilon \leq c'/\log^{(1+\beta)/2} d$ where C is a sufficiently large constant and c' is a sufficiently small constant, no auditing algorithm can return a unfair certificate for halfspace subgroups in \mathbb{R}^d with an additive error ϵ under Gaussian marginals and runs in time $d^{O(1/(\epsilon^2 \log d)^\alpha)}$.*

Proof. Suppose there exists an auditing algorithm that guarantees to return a $\gamma - \epsilon$ -unfair certificate given a γ -unfair collection of halfspace subgroups and access to data with a Gaussian marginal, where $\epsilon \in (0, 1)$.

Similar to the proof of Corollary 20, given a $1/\sqrt{k \log d}$ -unfair collection of halfspace subgroups, we run such an algorithm to obtain a $(1/\sqrt{k \log d} - \epsilon)$ -unfair certificate. Observe that, if $\epsilon = c'/\sqrt{k \log d}$ for some sufficiently small constant c' , we can solve the testing problem

in Theorem 19 within time $d^{O(k^\alpha)}$ by running this algorithm as well as drawing enough examples to estimate the unfairness of the returned certificates from the two cases respectively. On the other hand, given $\epsilon = c'/\sqrt{k \log d}$, we can rewrite $d^{O(k^\alpha)} = d^{O(1/(\epsilon^2 \log d)^\alpha)}$.

However, Theorem 19 tells that the above case is impossible for any $C/\sqrt{d \log d} \leq \epsilon \leq c'/\log^{(1+\beta)/2} d$, where C is a sufficiently large constant. \blacktriangleleft

Besides the general auditing problem, we also consider the “non-constructive auditing” problem as in Definition 7, where the algorithm is only required to tell if there exists an unfair subgroup without returning the unfair certificate. Actually, it turns out any non-constructive auditing algorithm can distinguish the two cases in Theorem 19.

► **Corollary 22** (non-constructive auditing is hard). *Given Assumption 11, for any constants $\alpha \in (0, 1), \beta \in \mathbb{R}_+$, and any $C/\sqrt{d \log d} \leq \epsilon \leq c'/\log^{(1+\beta)/2} d$ where C is a sufficiently large constant and c' is a sufficiently small constant, no auditing algorithm can tell if there exists a unfair certificate for halfspace subgroups in \mathbb{R}^d with*

- *an additive error ϵ under Gaussian marginals and running in time $d^{O(1/(\epsilon^2 \log d)^\alpha)}$.*
- *or a multiplicative approximation factor of $1/\text{poly}(d)$ and running in polynomial time.*

Proof. Suppose there exists an auditing algorithm that can either tell if a $\delta\gamma$ -unfair certificate or a $\gamma - \epsilon$ -unfair certificate exists given a γ -unfair collection of halfspace subgroup and access to data with a Gaussian marginal, where $\delta, \epsilon \in (0, 1)$. With the same argument as that of Corollary 20 and 21, we can achieve the desired results. \blacktriangleleft

To the best of our knowledge, there does not exist any PTAS for properly learning general halfspaces in the agnostic model with guarantees of additive error close to $O(1/\sqrt{\log d})$. However, in the next section, we will show that if we restrict our attention to just homogeneous halfspaces under a standard normal distribution, it is possible to achieve additive error of $O(1/\log^{1/C} d)$ for some constant $C > 2$.

5 Auditing Via Agnostic Learning Under Gaussian Distribution

In this section, we present our algorithmic results. Our approach is based on Theorem 12: auditing over subgroups determined by halfspaces can be accomplished by solving a sequence of simpler tasks of learning halfspaces. As a result, we are able to take advantage of existing agnostic learning methods to solve the auditing problem.

Meanwhile, we will discuss the testability of Gaussian distributions and show that existing distribution testing methods [15, 27] for learning halfspaces will not increase the running time significantly for our task. In fact, the running time of the testing method is asymptotically no greater than that of our auditing algorithm.

5.1 Auditing Algorithm for Homogeneous Halfspaces

Assuming there exists an efficient oracle for agnostic learning, Algorithm 1 will eventually return a halfspace h' as a certificate of the subgroup that has the highest unfairness level.

Notice, we create a negatively labelled data sets at Line 3 because maximizing (minimizing) the unfairness $\Pr\{h\} |d_{\mathcal{D}}(c, h)|$ for the $c(\mathbf{x}) = 1$ labelling is equivalent to minimizing (maximizing) $\Pr\{h\} |d_{\mathcal{D}}(c, h)|$ for $c(\mathbf{x}) = -1$. Thus, by reversing the labels, we can use the oracle to solve both the maximization and minimization directions.

In the loop, we simply follow our previous reduction by dividing the population constraint into multiple approximately-fixed-size constraints at Line 11. Then, we solve each sub-task with a fixed population size by calling the oracle on both data sets at Lines 7 and 8.

■ **Algorithm 1** Fairness Auditing.

Input: $n, a, b, \epsilon, \delta, \mathcal{D}$, classifier c , oracle \mathcal{O}
Result: μ', h'

- 1 $\hat{\mathcal{X}} \leftarrow$ draw $N(d, \epsilon, \delta)$ i.i.d. samples from \mathcal{D} ;
- 2 $\hat{\mathcal{D}}^+ \leftarrow \{\hat{\mathcal{X}}, c(\hat{\mathcal{X}})\}$;
- 3 $\hat{\mathcal{D}}^- \leftarrow \{\hat{\mathcal{X}}, -c(\hat{\mathcal{X}})\}$;
- 4 $\mu \leftarrow a$;
- 5 $(\mu', h') \leftarrow (1, c)$;
- 6 **while** $\mu \leq b$ **do**
- 7 $h_\mu^+ \leftarrow \mathcal{O}(\epsilon, \delta/2n, \mu, \hat{\mathcal{D}}^+)$;
- 8 $h_\mu^- \leftarrow \mathcal{O}(\epsilon, \delta/2n, \mu, \hat{\mathcal{D}}^-)$;
- 9 **if** $|d_{\mathcal{D}}(c, h_\mu^+)| < |d_{\mathcal{D}}(c, h_\mu^-)|$ **then** $h_\mu^+ \leftarrow h_\mu^-$;
- 10 **if** $\mu' |d_{\mathcal{D}}(c, h')| \leq \mu |d_{\mathcal{D}}(c, h_\mu^+)|$ **then** $(\mu', h') \leftarrow (\mu, h_\mu^+)$;
- 11 $\mu \leftarrow \mu + (b - a)/n$;
- 12 **end**

We give the guarantees of our algorithm below and defer the proof to the appendix.

► **Theorem 23 (Auditing Framework).** *Given any binary classifier $c : \mathbb{R}^d \rightarrow \{-1, +1\}$, a data distribution \mathcal{D} whose 1-dimensional marginals have continuous cumulative distribution functions, and collections of halfspaces $\{\mathcal{H}_\mu^{\mathcal{D}} \mid \mu > 0\}$ over \mathbb{R}^d , if there exists an oracle \mathcal{O} that takes $\epsilon, \delta, \mu \in (0, 1)$ and $N(d, \epsilon, \delta)$ labelled i.i.d. samples from \mathcal{D} in the form of $(\mathbf{x}, c(\mathbf{x}))$, runs in time $T(d, \epsilon, \delta)$, and returns a halfspace h_μ such that, with at least $1 - \delta$ probability*

$$\Pr_{\mathbf{x} \sim \mathcal{D}} \{h_\mu(\mathbf{x}) \neq c(\mathbf{x})\} \leq \min_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr_{\mathbf{x} \sim \mathcal{D}} \{h(\mathbf{x}) \neq c(\mathbf{x})\} + \epsilon$$

then there exists an algorithm that takes $n \in \mathbb{Z}^+$, $0 < a \leq b < 1$, $\epsilon, \delta \in (0, 1)$ and $O(N(d, \epsilon, \delta/n))$ labeled i.i.d samples from \mathcal{D} , runs in time $O(nT(d, \epsilon, \delta/n))$ and returns a halfspace h' as a certificate such that $a \leq \Pr_{\mathbf{x} \sim \mathcal{D}}\{h'\} \leq b$ and

$$\Pr_{\mathbf{x} \sim \mathcal{D}} \{h' \mid |d_{\mathcal{D}}(c, h')| \geq \max_{h \in \mathcal{H}^d} \Pr_{\mathbf{x} \sim \mathcal{D}} \{h\} |d_{\mathcal{D}}(c, h)| - O(\epsilon)$$

with at least $1 - \delta$ probability.

While our framework heavily relies on the methods of agnostic learning with small additive error, unfortunately, there are no known methods for learning general halfspaces that can achieve additive error better than a constant, even under distributions as nice as standard normal ones.

However, if we restrict our audit to the class of homogeneous halfspaces, Diakonikolas et al. [8] proposed an agnostic learning PTAS for homogeneous halfspaces under Gaussian data. That is, we only audit for subgroups with probability mass $1/2$.

► **Lemma 24 (Learning Homogeneous Halfspaces [8]).** *Let \mathcal{D} be a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$ whose \mathbf{x} -marginal is $\mathcal{N}(0, \mathbf{I})$. There exists an algorithm that, given $\tau, \epsilon, \delta > 0$, and $N = d^{\text{poly}(1/\tau)} \text{poly}(1/\epsilon) \log(1/\delta)$ i.i.d. samples from \mathcal{D} , the algorithm runs in time $\text{poly}(N, d)$, and computes a halfspace $h_{\mathbf{v}}$ such that, with probability at least $1 - \delta$, it holds that $\Pr_{\mathcal{D}}\{y \neq h_{\mathbf{v}}(\mathbf{x})\} \leq (1 + \tau) \min_{h \in \mathcal{H}_{1/2}^{\mathcal{N}}} \Pr_{\mathcal{D}}\{y \neq h(\mathbf{x})\} + \epsilon$.*

Now, notice that Lemma 24 gives us an oracle for auditing halfspace subgroups with population size $1/2$ under Gaussian distributions, since by Lemma 15, we know that agnostic learning with fixed threshold will have constant population size under a Gaussian distribution and, hence, is equivalent to auditing with fixed population size. Therefore, we can use this oracle in Algorithm 1 to audit the subgroup class $\mathcal{H}_{1/2}^d$ for $\mathcal{D} = \mathcal{N}(0, \mathbf{I})$. We show our algorithmic guarantee of a PTAS in the following corollary.

► **Corollary 25** (Auditing Under Gaussian). *Given any binary classifier $c : \mathbb{R}^d \rightarrow \{-1, +1\}$, a data distribution $\mathcal{N}(0, \mathbf{I})$ and a collection of halfspaces $\mathcal{H}_{1/2}^N$ over \mathbb{R}^d , there exists an auditing algorithm that takes $\epsilon, \delta > 0$ and $N = d^{\text{poly}(1/\epsilon)} \text{poly}(1/\epsilon) \log(1/\delta)$ labeled i.i.d. examples from $\mathcal{N}(0, \mathbf{I})$ in the form of $(\mathbf{x}, c(\mathbf{x}))$, runs in time $\text{poly}(N, d)$, and returns a halfspace h' as a certificate such that $\Pr_{\mathbf{x} \sim \mathcal{D}}\{h'\} = 1/2$ and*

$$|d_{\mathcal{N}}(c, h')| \geq \max_{h \in \mathcal{H}_{1/2}^N} |d_{\mathcal{N}}(c, h)| - 2\epsilon$$

with at least $1 - \delta$ probability.

Proof. We can simply run Algorithm 1 for just one iteration with the same set of parameters except that $\mathcal{D} = \mathcal{N}(0, \mathbf{I})$, $n = 1$, $a = b = 1/2$ and the oracle being as described by Lemma 24 for $\tau = \epsilon$. Notice that Lemma 24 guarantees us that the requirement on the oracle in Theorem 23 is satisfied. Thus, we can refer to the proof of Theorem 23 to establish that running Algorithm 1 for just one iteration suffices. Also, since we only run the algorithm for one iteration, we have $T = 1$, hence, the running time is dominated by the running time of the oracle, which is $\text{poly}(N, d)$. ◀

5.2 Testability Of Gaussian Distribution

Given the assumption that our algorithm only works under Gaussian distributions, one might ask if a set of data examples can be tested to be Gaussian without increasing the running time guarantee in Corollary 25 asymptotically. We will show that this kind of testing can be accomplished within the same running time as our auditing algorithm.

A recent work by Rubinfeld and Vasilyan [27] has proposed a moment matching method for testing Gaussian assumptions specifically for agnostic learning. Their method is based on the observation that linear threshold functions have degree $\text{poly}(1/\epsilon)$ polynomial approximations with additive error of ϵ [19, 8]. Abstractly, this moment matching testing method estimates the moments of the data samples up to degree $O(1/\epsilon^4)$ and check if the element-wise difference between the estimated moments and the actual Gaussian moments are small. They proved that running their testing method along with the agnostic learning algorithm proposed by Kalai et al. [19] will not increase the running asymptotically, i.e., $d^{O(1/\epsilon^4)}$.

To see why the testing method in [27] will not increase the asymptotic running time of our auditing algorithm, we need to dig deeper into the algorithm described by Lemma 24 from [8]. First, they run the learning algorithm of Kalai et al. [19] to get an approximating polynomial of degree $O(1/\epsilon^4)$. Then, they estimate the moments of the outer product of the derivatives of the learned polynomial. Finally, they estimate the classification error of a collection of halfspaces in a subspace of degree $O(1/\epsilon^4)$. See [8] for further details.

The most important observation is that every step in the algorithm stated in Lemma 24 only requires estimating the moments of the data up to degree $O(1/\epsilon^4)$. Thus, running the moment matching testing method of [27] will only require an additional $d^{O(1/\epsilon^4)}$ running time, which will not increase the asymptotic running time of the agnostic learning algorithm in Lemma 24 or our auditing algorithm.

6 Future Work

The major drawback of our result is still the lack of approaches of learning halfspaces with a sub-constant error guarantee for more general distributions. Therefore, a major direction for fairness auditing remains to develop an agnostic learning method with additive error guarantees for broader classes, such as log-concave distributions – subject to the constraints of Corollary 21/Diakonikolas et al. [7]. Even a computationally efficient learning algorithm for general halfspaces that can achieve additive error close to $O(1/\sqrt{\log d})$ under Gaussian distributions would be an interesting improvement.

An alternative direction is to seek stronger guarantees for conjunctions on such families of distributions. Conjunctions are more natural in the context of auditing, and their relative lack of expressive power might enable a better guarantee.

References

- 1 Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.
- 2 Özgür Akgün, Ian P Gent, Christopher Jefferson, Ian Miguel, and Peter Nightingale. Metamorphic testing of constraint solvers. In *Principles and Practice of Constraint Programming: 24th International Conference, CP 2018, Lille, France, August 27-31, 2018, Proceedings 24*, pages 727–736. Springer, 2018.
- 3 Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- 4 Bart Bogaerts, Stephan Gocht, Ciaran McCreesh, and Jakob Nordström. Certified dominance and symmetry breaking for combinatorial optimisation. *Journal of Artificial Intelligence Research*, 77:1539–1589, 2023.
- 5 William Cook, Thorsten Koch, Daniel E Steffy, and Kati Wolter. A hybrid branch-and-bound approach for exact rational mixed-integer programming. *Mathematical Programming Computation*, 5(3):305–344, 2013.
- 6 Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 2013.
- 7 Ilias Diakonikolas, Daniel Kane, and Lisheng Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *International Conference on Machine Learning*, pages 7922–7938. PMLR, 2023.
- 8 Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Conference on Learning Theory*, pages 1522–1551. PMLR, 2021.
- 9 Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *Advances in Neural Information Processing Systems*, 33:18540–18549, 2020.
- 10 Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning*, pages 5118–5141. PMLR, 2022.
- 11 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, 2012.
- 12 Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

- 13 Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of halfspaces with gradient descent via soft margins. In *International Conference on Machine Learning*, pages 3417–3426. PMLR, 2021.
- 14 Xavier Gillard, Pierre Schaus, and Yves Deville. Solvercheck: Declarative testing of constraints. In *Principles and Practice of Constraint Programming: 25th International Conference, CP 2019, Stamford, CT, USA, September 30–October 4, 2019, Proceedings 25*, pages 565–582. Springer, 2019.
- 15 Aravind Gollakota, Adam R Klivans, and Pravesh K Kothari. A moment-matching approach to testable learning and a new characterization of rademacher complexity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1657–1670, 2023.
- 16 Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous lwe is as hard as lwe & applications to learning gaussian mixtures. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1162–1173. IEEE, 2022.
- 17 David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.
- 18 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.
- 19 Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.
- 20 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.
- 21 Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.
- 22 Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.
- 23 Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31, 2018.
- 24 Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.
- 25 Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.
- 26 Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM (JACM)*, 56(6):1–40, 2009.
- 27 Ronitt Rubinfeld and Arsen Vasilyan. Testing distributional assumptions of learning algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1643–1656, 2023.
- 28 Robert E Schapire. The strength of weak learnability. *Machine learning*, 5:197–227, 1990.
- 29 Ji Wang, Ding Lu, Ian Davidson, and Zhaojun Bai. Scalable spectral clustering with group fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 6613–6629. PMLR, 2023.

A **Analysis Of Reduction**

We formally prove that the auditing program (4) can be approximated by a sequence of simpler optimization problems with arbitrary precision.

Proof of Proposition 14. For conciseness of the proof, we define

$$\alpha(k) := a + \frac{k(b-a)}{n}$$

Since $a \leq \Pr\{h^*(\mathbf{x}) = 1\} \leq b$ by definition, there must exist a $k \in \{0, \dots, n-1\}$ such that

$$\alpha(k) < \Pr\{h^*(\mathbf{x}) = 1\} < \alpha(k+1)$$

Then, since we assumed that \mathcal{D} has a continuous CDF w.r.t. the normal of h^* , we can construct another halfspace h' by either increasing or decreasing the threshold of h^* until $\Pr\{\mathbf{x} \in h'\}$ hits either $\alpha(k)$ or $\alpha(k+1)$. We thus obtain

$$\begin{aligned} \Pr\{h'(\mathbf{x}) \neq h^*(\mathbf{x})\} &= |\Pr\{h^*\} - \Pr\{h'\}| \\ &\leq \alpha(k+1) - \alpha(k) \\ &= \frac{(b-a)}{n} \end{aligned} \tag{10}$$

Let $\mathbf{dom} := \{\mathbf{x} \mid h'(\mathbf{x}) \neq h^*(\mathbf{x})\}$. Then, by the triangle inequality and the fact that $\Pr\{c(\mathbf{x}) = 1\} \leq 1$, we have

$$\begin{aligned} |\Pr\{h^*\}d_{\mathcal{D}}(c, h^*)| - |\Pr\{h'\}d_{\mathcal{D}}(c, h')| &\leq |\Pr\{h^*\} - \Pr\{h'\}| + |\Pr\{h' \cap c\} - \Pr\{h^* \cap c\}| \\ &\leq \frac{(b-a)}{n} + |\Pr\{h' \cap c \cap \mathbf{dom}\} - \Pr\{h^* \cap c \cap \mathbf{dom}\}| \\ &\leq \frac{(b-a)}{n} + |\Pr\{\mathbf{x} \in \mathbf{dom}\}| \\ &\leq \frac{2(b-a)}{n} \end{aligned} \tag{11}$$

where the second inequality is obtained by expanding $\Pr\{h \cap c\}$ on the event $\mathbf{x} \in \mathbf{dom}$ using the law of total probability and exploiting the fact that h' always agrees with h^* on the complement of \mathbf{dom} , i.e., $\Pr\{h' \cap c \cap \mathbf{dom}^c\} = \Pr\{h^* \cap c \cap \mathbf{dom}^c\}$; the third inequality holds because at most one of $h^*(\mathbf{x}) = 1$ and $h'(\mathbf{x}) = 1$ holds for any $\mathbf{x} \in \mathbf{dom}$ by definition; and the last inequality is due to equation (10).

Finally, due to the optimality of h_k^* , we have

$$\begin{aligned} \Pr\{h_k^*\}d_{\mathcal{D}}(c, h_k^*) &\geq \Pr\{h'\}d_{\mathcal{D}}(c, h') - \gamma^* + \gamma^* \\ &\geq \gamma^* - \frac{2(b-a)}{n} \end{aligned}$$

by inequality (11) with $\Pr\{h^*(\mathbf{x}) = 1\}d_{\mathcal{D}}(c, h^*) = \gamma^*$. ◀

B Proof Of Hardness

We will need the following proposition from [16, 7] in the proof of theorem 19.

► **Proposition 26** ([16, 7] Hardness of cLWE). *Given Assumption 11, for any $d \in \mathbb{N}$, any constants $\kappa \in \mathbb{N}$, $\alpha \in (0, 1)$, $\beta \in \mathbb{R}_+$ and any $\log^\beta d \leq k \leq Cd$ where $C > 0$ is a sufficiently small universal constant, the problem $\text{LWE}(d^{O(k^\alpha)}, \mathcal{N}, \mathbb{S}^{d-1}, \mathcal{N}_\sigma, \text{mod}_T)$ over \mathbb{R}^d with $\sigma \geq k^{-\kappa}$ and $T = 1/C' \sqrt{k} \log d$, where $C' > 0$ is a sufficiently large universal constant, cannot be solved in time $d^{O(k^\alpha)}$ with $d^{-O(k^\alpha)}$ advantage*

The problem of continuous Learning With Error (cLWE) under Gaussian distribution is known to be as hard as LWE. Now we are ready to prove the main theorem.

5:18 Distribution-Specific Auditing for Subgroup Fairness

Proof of Theorem 19. We give an efficient method taking as input samples from a distribution \mathcal{D}' , that is either from the alternative hypothesis or the null hypothesis of $\text{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0, \mathbf{I}), \mathbb{S}^{d-1}, \mathcal{N}(0, \sigma), \text{mod}_T)$ from Proposition 26, and generate samples from another distribution \mathcal{D} with the following properties: if \mathcal{D}' is from the alternative (resp. null) hypothesis of the LWE problem, then the resulting distribution \mathcal{D} will satisfy the alternative (resp. null) hypothesis requirement of the theorem for the halfspace auditing problem.

The reduction process can be formulated as follow: for a sample (\mathbf{x}, y) from a instance \mathcal{D}' of the problem $\text{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0, \mathbf{I}), \mathbb{S}^{d-1}, \mathcal{N}(0, \sigma), \text{mod}_T)$ from Proposition 26, we simply output $(\mathbf{x}, c(\mathbf{x})) \sim \mathcal{D}$, where

$$c(\mathbf{x}) = \begin{cases} +1, & \text{if } y \leq T/2 \\ -1, & \text{otherwise} \end{cases}$$

We argue that \mathcal{D} satisfies the desired requirement stated above.

For the alternative hypothesis case, let \mathcal{D}' be from the alternative hypothesis case of the LWE. Let \mathbf{s} be the secret vector in the LWE problem. We consider the following two halfspaces:

$$\begin{aligned} h_1(\mathbf{x}) &= \text{sgn}(\mathbf{s}^\top \mathbf{x} - T/6) \\ h_2(\mathbf{x}) &= \text{sgn}(-\mathbf{s}^\top \mathbf{x} + T/3) \end{aligned}$$

If we can show $\left| \Pr_{\mathbf{x} \sim \mathcal{D}_x} \{h_1(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_1) + \Pr_{\mathbf{x} \sim \mathcal{D}_x} \{h_2(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_2) \right| = \Omega(T)$, then either $h = h_1$ or $h = h_2$ satisfies $\Pr_{\mathbf{x} \sim \mathcal{D}_x} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| = \Omega(T)$, which implies the desired property of the alternative hypothesis we would like to prove. By Lemma 15, we have

$$\begin{aligned} & 2 \Pr_{\mathbf{x} \sim \mathcal{D}_x} \{h_1(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_1) + 2 \Pr_{\mathbf{x} \sim \mathcal{D}_x} \{h_2(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_2) \\ &= \underbrace{\Pr\{-c\}(\Pr\{-h_1\} + \Pr\{-h_2\}) + \Pr\{c\}(\Pr\{h_1\} + \Pr\{h_2\})}_{I_1} \\ & \quad - \underbrace{(\Pr\{c(\mathbf{x}) = h_1(\mathbf{x})\} + \Pr\{c(\mathbf{x}) = h_2(\mathbf{x})\})}_{I_2} \end{aligned}$$

To bound I_1, I_2 , we first examine the subset of domain where h_1 and h_2 agree, namely

$$\begin{aligned} B &:= \{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = h_2(\mathbf{x})\} \\ &= \{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = 1 \cap h_2(\mathbf{x}) = 1\} \\ &= \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{s}^\top \mathbf{x} \in [T/6, T/3]\} \end{aligned}$$

Then, for I_1 , by the law of total probability, we have

$$\begin{aligned} I_1 &= \Pr\{c(\mathbf{x}) = -1\}(\Pr\{h_1(\mathbf{x}) = -1\} + \Pr\{h_2(\mathbf{x}) = -1\} + \Pr\{\mathbf{x} \in B\} - \Pr\{\mathbf{x} \in B\}) \\ & \quad + \Pr\{c(\mathbf{x}) = 1\}(\Pr\{h_1(\mathbf{x}) = 1\} + \Pr\{h_2(\mathbf{x}) = 1 \cap \mathbf{x} \notin B\} + \Pr\{h_2(\mathbf{x}) = 1 \cap \mathbf{x} \in B\}) \\ & \stackrel{(i)}{=} \Pr\{c(\mathbf{x}) = -1\}(1 - \Pr\{\mathbf{x} \in B\}) + \Pr\{c(\mathbf{x}) = 1\}(1 + \Pr\{\mathbf{x} \in B\}) \\ &= 1 + \Pr\{\mathbf{x} \in B\}(\Pr\{c(\mathbf{x}) = 1\} - \Pr\{c(\mathbf{x}) = -1\}) \\ &= 1 + \Pr\{\mathbf{x} \in B\}(2\Pr\{c(\mathbf{x}) = 1\} - 1) \end{aligned}$$

where (i) is because $\{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = -1\}, \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = -1\}, \{\mathbf{x} \in B\}$ are pairwise disjoint and their union equals to \mathbb{R}^d , $\{\mathbf{x} \in \mathbb{R}^d \mid h_1(\mathbf{x}) = 1\}, \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = 1 \cap \mathbf{x} \notin B\}$ are disjoint and their union equals to \mathbb{R}^d ; and since $\{\mathbf{x} \in B\} \subset \{\mathbf{x} \in \mathbb{R}^d \mid h_2(\mathbf{x}) = 1\}$ by definition, $\{\mathbf{x} \in B\} = \{\mathbf{x} \in B \mid h_2(\mathbf{x}) = 1\}$.

For I_2 , because for any $\mathbf{x} \in B$, $h_1(\mathbf{x}) = h_2(\mathbf{x}) = 1$ by construction, and by the law of total probability, we have

$$\begin{aligned} I_2 &= \Pr\{c(\mathbf{x}) = h_1(\mathbf{x}) \cap \mathbf{x} \notin B\} + \Pr\{c(\mathbf{x}) = h_2(\mathbf{x}) \cap \mathbf{x} \notin B\} + 2\Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} \\ &= \Pr\{\mathbf{x} \notin B\} + 2\Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} \\ &= 1 + \Pr\{c(\mathbf{x}) = 1 \cap \mathbf{x} \in B\} - \Pr\{c(\mathbf{x}) = -1 \cap \mathbf{x} \in B\} \\ &= 1 - \Pr\{\mathbf{x} \in B\}(1 - 2\Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}) \end{aligned}$$

By the definition of c as well as the Alternative case distribution of the LWE problem, $\{\mathbf{x} \in \mathbb{R}^d \mid c(\mathbf{x}) = 1\}$ is equivalent to $\{\mathbf{x} \in \mathbb{R}^d \mid \text{mod}_T(\mathbf{s}^\top \mathbf{x} + z) \leq T/2\}$ for some $z \sim \mathcal{N}(0, \sigma^2)$. Furthermore, we have

$$\{\mathbf{x} \in \mathbb{R}^d \mid \text{mod}_T(\mathbf{s}^\top \mathbf{x} + z) \leq T/2\} \equiv \bigcup_{k \in \mathbb{Z}} \{\mathbf{s}^\top \mathbf{x} + z \in (kT, kT + T/2]\}$$

Notice that $\mathbf{s}^\top \mathbf{x} + z$ is a one dimensional Gaussian random variable, which, by symmetry of Gaussian distribution, implies $\Pr\{c(\mathbf{x}) = 1\} = \Pr\{\cup_{k \in \mathbb{Z}} \{\mathbf{s}^\top \mathbf{x} + z \in (kT, kT + T/2]\}\} = 1/2$. Therefore, combining I_1 and I_2 gives

$$\begin{aligned} I_1 - I_2 &= 2\Pr\{\mathbf{x} \in B\}(\Pr\{c(\mathbf{x}) = 1\} - \Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}) \\ &= \Omega(T)(1/2 - \Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\}) \end{aligned} \tag{12}$$

where the last equation is because $\mathbf{s}^\top \mathbf{x} \sim \mathcal{N}(0, 1)$, hence, $\Pr\{\mathbf{x} \in B\} = \Pr\{\mathbf{s}^\top \mathbf{x} \in [T/6, T/3]\} = \Omega(T)$. Since we were only concerned with showing $|I_1 - I_2|$ is large, it suffices to show $\Pr\{c(\mathbf{x}) = 1 \mid \mathbf{x} \in B\} - 1/2 = \Omega(1)$.

For $\mathbf{x} \in B$, we have $\mathbf{s}^\top \mathbf{x} \in [T/6, T/3]$, therefore $c(\mathbf{x}) = -1$ only if $|z| \geq T/6$. Notice that $z \sim \mathcal{N}(0, \sigma^2)$ and Proposition 26 states that the LWE problem is hard for any fixed constant $\kappa \in \mathbb{N}$ and $\sigma \geq k^{-\kappa}$. Given the constant $\beta \in \mathbb{R}_+$ in this theorem, we can take $\kappa = \lceil 1/2\beta + 1/2 + 1 \rceil$, which is a fixed constant. Then, by Proposition 26, the LWE problem is hard for $\sigma = k^{-\kappa} \leq 1/(k^{3/2} \sqrt{\log d}) = o(T)$. Therefore, by a Gaussian tail bound, we have

$$\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \{c(\mathbf{x}) = -1 \mid \mathbf{x} \in B\} \leq \Pr_{z \sim \mathcal{N}(0, \sigma^2)} \{|z| \geq T/6\} = o(1)$$

Plugging the above back into Equation (12), we can conclude that

$$\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \{h_1(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_1) + \Pr_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \{h_2(\mathbf{x}) = 1\} d_{\mathcal{D}}(c, h_2) = \Omega(T)$$

Thus, either $h = h_1$ or $h = h_2$ must satisfy $\Pr_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| = \Omega(T)$, which completes the proof for the alternative hypothesis case.

For the null hypothesis, we can immediately see that $\Pr_{\mathbf{x} \in \mathcal{N}} \{h\} d_{\mathcal{N}}(c, h) = 0, \forall h \in \mathcal{H}^d$ because $c(\mathbf{x})$ is independent from each $h \in \mathcal{H}^d$.

It remains to verify the time lower bound and the distinguishing advantage for auditing halfspace subgroups. From Proposition 26, we know that under Assumption 11, for the problem $\text{LWE}(d^{O(k^\alpha)}, \mathcal{N}(0, \mathbf{I}), \mathbb{S}^{d-1}, \mathcal{N}(0, \sigma^2), \text{mod}_T)$ with any $\sigma \geq k^{-\kappa}$ (where $\kappa \in \mathbb{N}$ is a

5:20 Distribution-Specific Auditing for Subgroup Fairness

constant) and $T = 1/c' \sqrt{k \log d}$, where $c' > 0$ is a sufficiently large universal constant, the problem cannot be solved in $d^{O(k^\alpha)}$ time with $d^{-O(k^\alpha)}$ advantage. Therefore, under the same assumption, there is no algorithm that can solve the decision version of auditing problem w.r.t. halfspace subgroups in $d^{O(k^\alpha)}$ time with $d^{-O(k^\alpha)}$ advantage. ◀

C Analysis Of Algorithm

We prove the correctness, time and sample complexity of Algorithm 1.

Proof of Theorem 23. Let's notice that, although each iteration of the loop in Algorithm 1 solves $\min_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$ and $\max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} \Pr\{c(\mathbf{x}) \neq h(\mathbf{x})\}$, it is essentially equivalent to solving $\max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} |d_{\mathcal{D}}(c, h)|$ according to Lemma 17. As the oracle returns a halfspace with additive error smaller than ϵ with probability at least $1 - \delta$, we have that

$$\max(|d_{\mathcal{D}}(c, h_\mu^+)|, |d_{\mathcal{D}}(c, h_\mu^-)|) \geq \max_{h \in \mathcal{H}_\mu^{\mathcal{D}}} |d_{\mathcal{D}}(c, h_\mu^+)| - \frac{\epsilon}{\mu}$$


with probability at least $1 - \delta/n$ because of Lemma 17 as well as a union bound.

Across all iterations, the algorithm maximizes $\mu |d_{\mathcal{D}}(c, h_\mu^+)|$ over $\mathcal{H}_\mu^{\mathcal{D}}$ for μ increase from a to b with step size $(b - a)/n$. With a union bound over all n iterations, we obtain the same additive error ϵ in every iteration, with probability at least $1 - \delta$. As a result, the algorithm equivalently solves

$$\begin{aligned} \max_{h \in \mathcal{H}^{\mathcal{D}}} \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} |d_{\mathcal{D}}(c, h)| \\ \text{s.t. } a \leq \Pr_{\mathbf{x} \in \mathcal{D}} \{h(\mathbf{x}) = 1\} \leq b \end{aligned}$$

with probability at least $1 - \delta$ for an additive error at most $2(b - a)/n + \epsilon$ according to Proposition 14, which completes the proof. ◀

Modeling Diversity Dynamics in Time-Evolving Collaboration Networks

Christopher Archer ✉ 

University of California, Berkeley, CA, USA

Gireeja Ranade ✉

University of California, Berkeley, CA, USA

Abstract

Increasing diversity in a community or an organization requires paying attention to many different aspects, including recruitment, hiring, retention, climate, and more. In this paper, we focus on how climate, captured through network interactions, can affect the growth or decay of minority populations within that community. Building on previous work, we develop a dynamic stochastic block model that grows according to a weighted version of preferential attachment, while having some memory of previous edges as well. This models how interactions between nodes in the network can influence the recruitment of new nodes to the network. We derive a deterministic approximation of this random system and prove its convergence is determined by the network parameters. Additionally, we show how the memory of the network affects convergence under different parameter regimes, and we validate this model by assessing the growth of women scientists in the American Physics Society’s co-authorship network.

2012 ACM Subject Classification Applied computing → Sociology

Keywords and phrases Network Models, Diversity, Collaboration Networks, Stochastic Block Model

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.6

Supplementary Material *Software (Source Code):*

<https://github.com/chris-archer110/SBM-diversity-model-code/blob/main/README.md> [7]

archived at `swh:1:cnt:6b61695dfc862850edb54411a21115587c29c8c2`

Funding Thanks to the NSF for CAREER grant ECCS-2240031.

Acknowledgements Thanks to the NSF for CAREER grant ECCS-2240031. Additionally, the authors would like to thank user “Actually Fritz” From Mathematics StackExchange for assisting in the proof of Lemma 6. We also thank Moses Won, Ebonye Smith, and the reviewers whose comments improved the presentation of this paper. Thanks to The American Physics Society (APS) dataset for the use of their dataset which can be requested at <https://journals.aps.org/datasets>.

1 Introduction

While efforts have been made since the 1920s to desegregate and diversify the workplace, modern notions of diversity primarily originate from policies in the 1960s during the Civil Rights Era [21]. The Civil Rights Act of 1964 put an end to the “de facto” policies that discriminated against classes of workers [16]. The benefits of diverse organizations are well-documented; studies and computational experiments show that diverse organizations have increased feelings of belonging and satisfaction amongst workers, and increased problem-solving ability [17, 23].

There has been a large body of work in recent years that has focused on how hiring practices can be amended to be more inclusive and support diversity [33, 40, 45, 29, 37, 48]. However, real-world experience shows that many times so-called “inclusive hiring” programs can mask deeply entrenched social biases that still privilege the status quo [27, 11].



© Christopher Archer and Gireeja Ranade;

licensed under Creative Commons License CC-BY 4.0

5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 6; pp. 6:1–6:21

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In addition to hiring, growing diversity in an organization also requires focusing on what happens after someone is hired. Harvard Business Review reports that U.S. companies invest a total of 8 billion dollars on Diversity, Equity, and Inclusion (DEI) trainings per year [26]. However, studies show that there remains a large problem with diversity in these organizations [5, 38, 27].

Culture is a huge part of an individual’s experience in a community. An institution that does a good job of recruiting under-represented people but does not have a climate that encourages their retention will ultimately revert to the status quo [6]. In this paper, we explore the connection between the growth of diversity in a community and the connections and collaborations among the people in the community.

Network analysis allows a way to link “micro-scale” interactions to “macro-scale” organizational structure and dynamics [19, 35, 28]. In this sense, the key micro-scale insight about organizational networks is that in recruiting, people tend to favor candidates similar to them, or those who fit the dominant culture of the organization [43, 11]. The employee referral system, hiring bias, and the infamous “culture fit” qualification are all examples of this phenomenon [43, 44]. As an organization’s network grows in this way, it tends to reproduce the demographics already present in the network. This is known as *homophily*, a term developed to capture how interpersonal networks tend to self-cluster around shared characteristics [35]. In the context of networks, this can be captured through the notion of preferential attachment [9], a weighted version of which we use in this paper.

It has been shown that social ties and networks can impact employment [14, 49] and education outcomes [15]. This property is important to consider for diversity because networks with strong homophily tend to become less diverse over time [2, 44]. We wish to create a model that can capture this phenomenon, and demonstrate how under certain network conditions, it can be prevented. Analysis like this creates an opening for researchers to go beyond assessing inclusive policies from a purely qualitative standpoint and towards a mathematical characterization of organizational “climate”. Additionally, to validate this model in real-world settings we look towards the coauthorship network of the American Physical Society (APS) from 1980 to 2009, to assess how the collaboration structure has impacted the proportion of women authors in APS.

2 Background and Related Work

Many works try to mathematically assess the dynamics of organizations and their impact on diversity. Some works propose ecological models, particularly ecological theories of affiliation to understand how organizational networks grow over time [34]. Others use agent-based modeling approaches, where agents of different communities will have different access to information and algorithms to solve problems [23, 18]. We will focus on network-oriented approaches.

The classic preferential attachment algorithm for network growth was introduced in [9]. The growth of scientific collaboration networks has been specifically studied in [10, 24]. Strategic perspectives on these collaborations have also been studied in [25]. How network growth relates to the communities formed is studied in [8]. [41] provides a formulation of “social capital” in networks, which defines how value can be generated based on one’s position in a social network, through useful information, personal relationships, or the ability to organize groups. In heterogeneous networks, it has been shown that “broker” nodes with ties that connect clusters and span the “structural holes” of the network have greater access to social capital due to their unique access to diverse information [13]. To this end, measures

of “betweenness” in networks have been used as a way to identify which nodes tend to span structural holes [33]. Research has also been done on the strength of weak ties, or connections that are accessed less frequently, in networks. Due to the nature of strong ties being connections that are accessed frequently, much of the information shared in networks of strong ties becomes redundant, and through weak ties, more novel information can be accessed [19].

Recently, [12] utilized the Stochastic Block Model to show how diversity in a heterogeneous network evolves under homophily and preferential attachment. The model in [12] is very similar to the model in [47], which also tries to capture biased network growth through homophily. This model was able to mathematically prove threshold effects under which minority populations would proportionally vanish or reach parity in the network, showing that low cross-community collaboration rates will always lead to the minority vanishing [12]. However, one of the main oversights of this model was its lack of long-term memory. In particular, it was assumed that the network was renewed at every time step. In time-evolving networks, memory is crucial in the understanding of complex temporal systems and can have great influence on emergent properties of the network [42, 36]. Works such as [20] have also explored how memory of social connections influences network communities. Without any encoding of memory, the model dynamics in [12] were essentially a sequence of independent static graphs with new nodes added. This paper’s model seeks to build on [12] with the addition of a “memory” parameter which influences how long edges persist over time steps.

3 Main Contributions

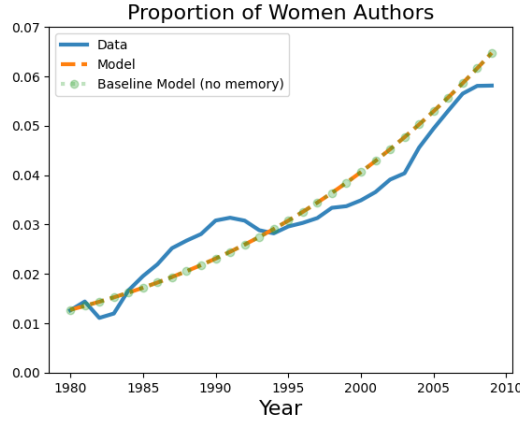
In our paper, we contribute the following:

- (1) Develop a model that builds on [12] that captures how homophily can affect diversity in collaboration networks while accounting for memory in the collaborations (edges) of the graph. We show that certain network conditions can lead to a decline in minority populations (Section 4).
- (2) Characterize the effect memory has on the rate of growth of the minority population and provide parameter regimes where memory can change the fixed point of the system, using a deterministic approximation to the stochastic system (Sections 5 and 6).
- (3) Validate this model on gender diversity in scientific collaborations using the American Physics Society’s citation network dataset [1]. In particular, we see in Figure 1 that the best-fit model we propose can roughly predict the growth of the minority population (women) in this dataset. We discuss this figure in more detail in Section 7.

4 Model Overview

Our basic model here replicates many of the features of [12]. We consider a Stochastic Block Model with two communities as the underlying structure. Nodes have weights that correspond to their “success” (which is the weighted sum of the collaborations (edges) of the node). Successful nodes are more influential in the recruitment of new nodes to the network (preferential attachment). Homophily plays a key role in recruitment since nodes only recruit members of their own community. The combination of homophily and preferential attachment leads to a rich-get-richer phenomenon, and we are interested in how the minority community evolves over time. We summarize key model features here:

- (i) *Community Structure*: We use a two-community (Red/Blue) Stochastic Block Model to account for differential interactions between different communities.



■ **Figure 1** Plot of the proportion of women in the APS co-authorship network from 1980-2009. The best-fit model identified, with memory parameter $q = 0.265$ is shown in orange, and the model from [12], with memory parameter $q = 0$ (i.e. no memory) is shown in green.

- (ii) *Collaborations*: In this network model, nodes represent people, and edges between nodes represent collaborations between two people. These edges will be weighted to account for how “successful” a collaboration is.
- (iii) *Node Influence*: The weight of a node is the sum of the weights of all its edges. Nodes with higher weight can recruit more new nodes.
- (iv) *Homophily*: A node will always recruit new nodes of the same color (sub-community). This leads to a rich-get-richer phenomenon since nodes with higher weight tend to recruit more nodes (weighted preferential attachment). The total number of nodes from a community (say red nodes) arriving at a particular time depends on the total weight of red nodes in the graph at the preceding time.
- (v) *Memory*: Given a collaboration network, people who have collaborated in one timestep are more likely to collaborate again in the next timestep. This is the key point of divergence from the model in [12]. While [12] assumed that brand-new collaborations were created at every time step, here we allow collaborations to persist over multiple time steps.

4.1 Weighted Stochastic Block Model

The Stochastic Block Model (SBM) [22] is a generalization of the Erdős-Rényi $G(n, p)$ random graph, which supports the interactions of multiple communities within the graph [31]. Weighted SBMs have been studied in [3, 4, 39]. In this paper, we consider two communities, Red and Blue. Without loss of generality, assume Red is the minority community. To maintain consistency with [12], we mirror much of the notation below.

► **Definition 1** (Weighted Stochastic Block Model). Let $[n] = \{1, \dots, n\}$ be a set of nodes, where each node $i \in [n]$ has color $c[i] \in \{R, B\}$. Define interaction matrix $\mathbf{P} = \begin{bmatrix} \mu_{RR} & \mu_{RB} \\ \mu_{BR} & \mu_{BB} \end{bmatrix}$ and weight matrix $\mathbf{W} = \begin{bmatrix} w_{RR} & w_{RB} \\ w_{BR} & w_{BB} \end{bmatrix}$ where $\mathbf{P}, \mathbf{W} > 0$. Generate a weighted random graph on $[n]$, so that for every pair of nodes $i, j \in [n]^2$, the edge’s weight $w_{i,j} = w_{c[i],c[j]}$ with probability $\mu_{c[i],c[j]}/n$, and otherwise $w_{i,j} = 0$.

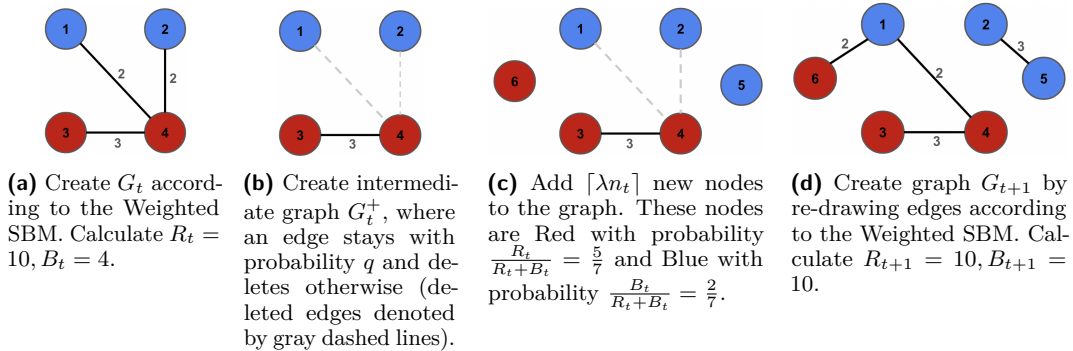
For this paper, we assume our matrices \mathbf{P}, \mathbf{W} are such that $\mu_{RR} = \mu_{BB}, w_{RR} = w_{BB}$ and $\mu_{RB} = \mu_{BR}, w_{RB} = w_{BR}$. Also note that the entries $\mu_{c[i],c[j]}$ of matrix \mathbf{P} are not themselves probabilities. The probability of an edge existing between two nodes is $\mu_{c[i],c[j]}/n$. One can think of $\mu_{c[i],c[j]}$ as the expected number of edges a node of color $c[i]$ will form with nodes of color $c[j]$ in the network without any memory.

4.2 Stochastic Block Model Dynamics

We assume that the recruitment of new nodes to the network happens in discrete time steps, and a constant fraction of new nodes join the network at each time step.

► **Definition 2** (Stochastic Block Model Dynamics). *Let $G_t = (V_t, E_t)$ be our graph at time t . Assume this graph contains n_t nodes, with n_t^R Red nodes and n_t^B Blue nodes. At each time step, $\lceil \lambda n_t \rceil, \lambda > 0$ new nodes are added to the network. Additionally, define edge holdover probability $q \in [0, 1]$. Our procedure for generating the subsequent graph G_{t+1} is as follows:*

- (1) Calculate the total weight of Red nodes and Blue nodes in G_t , defined as $R_t = \sum_{c[i]=R} \sum_{j \in [n_t]} w_{ij}$, and $B_t = \sum_{c[i]=B} \sum_{j \in [n_t]} w_{ij}$.
- (2) Define intermediate graph $G_t^+ = (V_t, E_t^+)$ where for every edge $(i, j) \in E_t$, with probability q let $(i, j) \in E_t^+$, otherwise remove it.
- (3) Add $m_{t+1} := \lceil \lambda n_t \rceil$ new nodes to graph G_t^+ . Each incoming node is Red with probability $\frac{R_t}{R_t+B_t}$, or Blue with probability $\frac{B_t}{R_t+B_t}$.
- (4) Initialize $G_{t+1} = G_t^+$, and for all potential edges $(i, j) \notin E_t^+$, generate them according to the Weighted Stochastic Block Model with parameters \mathbf{P} and \mathbf{W} .



■ **Figure 2** Visualization of the Stochastic Block Model Dynamics described in Definition 2 for $\mathbf{P} = \begin{bmatrix} 0.7 & 1 \\ 1 & 0.7 \end{bmatrix}, \mathbf{W} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}, q = 0.5, \lambda = 0.5$. Figure 2a shows initialization of G_t . Figure 2b shows the intermediate graph G_t^+ with some edges persisting according to the q parameter. Figure 2c shows new nodes being added to the network according to preferential attachment, and Figure 2d shows the creation of G_{t+1} according to the Weighted SBM from Definition 1.

The recruitment dynamics in step (3) of Definition 2 come from our assumption of homophily and preferential attachment. New Red nodes are recruited by existing Red nodes and arrive proportional to the total weight of Red nodes in the network. The same is true for Blue nodes. Let $w_i = \sum_{j \in [n_t]} w_{i,j}$ be the weight of a node i , and $w = \sum_{i \in [n_t]} w_i$ be the total weight of all nodes. Then our recruitment dynamics in step (3) are equivalent to every node i recruiting on average $m_{t+1} \cdot \frac{w_i}{w}$ nodes of the same color.

5 Deterministic Approximation

To understand the stochastic system, we will construct a deterministic approximation that follows the mean of the system. We will show that the stochastic system does not deviate too far from the deterministic system. For this, we first compute the probability that an edge exists at a given time. Because the graph has memory, the probability of an edge existing at time t is dependent on all previous graphs $\{G_k | k \leq t\}$. We define the event $\mathcal{E}_{ij}(t)$ on each edge $(i, j) \in E_t$, where $\mathcal{E}_{ij}(t) := \{\text{edge } (i, j) \text{ exists at time } t\}$.

► **Lemma 3.** *Let $\Pi_{RR}(t) := Pr(\mathcal{E}_{ij}(t) | c[i] = c[j] = R)$. Then:*

$$\Pi_{RR}(t) = \frac{\mu_{RR}}{n_t} + \sum_{k=0}^{t-1} \frac{\mu_{RR}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_j}\right). \quad (1)$$

Proof. Fix edge $(i, j) \in E_t$. Without loss of generality, assume $c[i] = c[j] = R$. We derive the probability of event $\mathcal{E}_{ij}(t)$ by conditioning on the previous timestep $\mathcal{E}_{ij}(t-1)$ with our edge holdover probability q .

$$Pr(\mathcal{E}_{ij}(t)) = Pr(\mathcal{E}_{ij}(t) | \mathcal{E}_{ij}(t-1)) Pr(\mathcal{E}_{ij}(t-1)) + Pr(\mathcal{E}_{ij}(t) | \mathcal{E}_{ij}^c(t-1)) Pr(\mathcal{E}_{ij}^c(t-1)) \quad (2)$$

$$= \left(q + (1-q) \frac{\mu_{RR}}{n_t} \right) Pr(\mathcal{E}_{ij}(t-1)) + \left(\frac{\mu_{RR}}{n_t} \right) (1 - Pr(\mathcal{E}_{ij}(t-1))) \quad (3)$$

$$= \frac{\mu_{RR}}{n_t} + q \left(1 - \frac{\mu_{RR}}{n_t} \right) Pr(\mathcal{E}_{ij}(t-1)). \quad (4)$$

This produces a recursive relationship with initial condition $Pr(\mathcal{E}_{ij}(0)) = \frac{\mu_{RR}}{n_0}$. We solve this inductively, yielding

$$\Pi_{RR}(t) := Pr(\mathcal{E}_{ij}(t) | c[i] = c[j] = R) = \frac{\mu_{RR}}{n_t} + \sum_{k=0}^{t-1} \frac{\mu_{RR}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_j}\right). \quad (5)$$

completing the proof. ◀

We can similarly define and compute probabilities $\Pi_{RB}(t)$ and $\Pi_{BB}(t)$. Define the minority fraction in the network as

$$(\phi_t)_{t \geq 0} := \frac{n_t^R}{n_t} = \frac{n_t^R}{n_t^R + n_t^B}. \quad (6)$$

Where n_t^R, n_t^B are the number of Red and Blue nodes in the network at time t , respectively.

► **Lemma 4.** *Assume $\exists \varepsilon \in (0, \frac{1}{2})$ such that $\left(\frac{1}{n_t}\right)^{\frac{1}{2}-\varepsilon} \leq \phi_t \leq \frac{1}{2}$. Let \mathcal{F}_t be a filtration until time t . Let*

$$\Gamma_q(x_t) := \frac{x_t^2 w_{RR} \Pi_{RR}(t) + x_t(1-x_t) w_{RB} \Pi_{RB}(t)}{(x_t^2 + (1-x_t)^2) w_{RR} \Pi_{RR}(t) + 2x_t(1-x_t) w_{RB} \Pi_{RB}(t)}. \quad (7)$$

Then, we can bound the conditional expectation of our process $\mathbb{E}[\phi_{t+1} | \mathcal{F}_t]$ as follows:

$$\frac{\phi_t + \frac{\lfloor \lambda n_t \rfloor}{n_t} \cdot \left(1 - \frac{1}{n_t^{\varepsilon/5}}\right) \Gamma_q(\phi_t)}{1 + \frac{\lfloor \lambda n_t \rfloor}{n_t}} \leq \mathbb{E}[\phi_{t+1} | \mathcal{F}_t] \leq \frac{\phi_t + \frac{\lfloor \lambda n_t \rfloor}{n_t} \cdot \left(1 + \frac{1}{n_t^{\varepsilon/5}}\right) \Gamma_q(\phi_t)}{1 + \frac{\lfloor \lambda n_t \rfloor}{n_t}}, \quad (8)$$

with probability at least $1 - \frac{8}{\exp(C_1 n_t^\varepsilon)}$, where $C_1 = \min \left\{ \frac{\mu_{RR}}{24}, \frac{\mu_{RB}}{12} \right\}$ and $\varepsilon \in (0, \frac{1}{2})$.

The proof of Lemma 4 is given in the Appendix A.1. We notice that the limit as $t \rightarrow \infty$ of the lower and upper bounds in Lemma 4 goes to $\frac{\phi_t + \lambda \cdot \Gamma_q(\phi_t)}{1 + \lambda}$. This motivates the consideration of the following deterministic system, Φ_t , to approximate the stochastic system behavior.

► **Definition 5** (Deterministic System). *We define the deterministic system Φ_t as:*

$$\Phi_{t+1} = \frac{\Phi_t + \lambda \cdot \Gamma_q(\Phi_t)}{1 + \lambda}, \Phi_0 = \frac{n_0^R}{n_0}, \quad (9)$$

where $\Gamma_q(x_t)$ is as before:

$$\Gamma_q(x_t) = \frac{x_t^2 w_{RR} \Pi_{RR}(t) + x_t(1-x_t) w_{RB} \Pi_{RB}(t)}{(x_t^2 + (1-x_t)^2) w_{RR} \Pi_{RR}(t) + 2x_t(1-x_t) w_{RB} \Pi_{RB}(t)}. \quad (10)$$

We derive this expression in more detail in Appendix Sec. A.2.

6 Analysis & Discussion of Deterministic System

In this section, we prove the existence of parameter regimes that will dictate the convergence of the deterministic system. Define $\rho_t := \frac{\Pi_{RR}(t) w_{RR}}{\Pi_{RB}(t) w_{RB}}$. This will be a key parameter in our analysis of the system's behavior.

► **Lemma 6.** *For $q > 0$, the limit $\lim_{t \rightarrow \infty} \rho_t := \rho = \frac{\mu_{RR} w_{RR} S_{RR}^q}{\mu_{RB} w_{RB} S_{RB}^q}$ exists, where we define S_{RR}^q as:*

$$S_{RR}^q = \lim_{t \rightarrow \infty} (1 + \lambda)^{-t} + \sum_{k=0}^{t-1} (1 + \lambda)^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} (1 + \lambda)^{-j} \right), \quad (11)$$

$$S_{RB}^q = \lim_{t \rightarrow \infty} (1 + \lambda)^{-t} + \sum_{k=0}^{t-1} (1 + \lambda)^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_0} (1 + \lambda)^{-j} \right). \quad (12)$$

For $q = 0$, we have $\rho_t = \rho_0 = \frac{\mu_{RR} w_{RR}}{\mu_{RB} w_{RB}}$.

The proof is in Appendix Section A.3. Now, consider the function $f_t : [0, 1] \rightarrow [0, 1]$, such that

$$f_t(x) = \frac{x + \lambda \Gamma_q(x)}{1 + \lambda} = \frac{x + \lambda \left(\frac{x^2 w_{RR} \Pi_{RR}(t) + x(1-x) w_{RB} \Pi_{RB}(t)}{(x^2 + (1-x)^2) w_{RR} \Pi_{RR}(t) + 2x(1-x) w_{RB} \Pi_{RB}(t)} \right)}{1 + \lambda} \quad (13)$$

$$= \frac{2x^3(\rho_t - 1) - x^2(\rho_t - 1)(2 - \lambda) + x(\rho_t + \lambda)}{(1 + \lambda)(2x^2(\rho_t - 1) - 2x(\rho_t - 1) + \rho_t)}. \quad (14)$$

The simplifications above follow from algebra. $f_t(x)$ captures the update function for the deterministic system in Def. 5. We establish the following properties of f_t .

► **Lemma 7.** *$f_t(x)$ converges uniformly to $f(x)$ as $t \rightarrow \infty$, where*

$$f(x) = \frac{2x^3(\rho - 1) - x^2(\rho - 1)(2 - \lambda) + x(\rho + \lambda)}{(1 + \lambda)(2x^2(\rho - 1) - 2x(\rho - 1) + \rho)}.$$

Proof. The denominator of $f_t(x)$ is $(1 + \lambda)(2x^2(\rho_t - 1) - 2x(\rho_t - 1) + \rho_t)$. This denominator is quadratic, with discriminant

$$D = 4(1 + \lambda)^2(\rho_t - 1)^2 - 4(2)(1 + \lambda)^2(\rho_t - 1)\rho_t = -4(1 + \lambda)^2(\rho_t^2 - 1). \quad (15)$$

In the case of $\rho_t > 1$, we have $D < 0$, meaning $f_t(x)$ is strictly bounded away from 0 for all $x \in \mathbb{R}$. For $\rho_t = 1$, our denominator is simply $(1 + \lambda)\rho_t$, and we have $\rho_t > 0$ by assumption. However, in the case of $\rho_t \in (0, 1)$, we see that the discriminant $D > 0$ and the denominator of $f_t(x)$ has real roots. We show that these roots must lie outside the domain of $f_t(x)$, and thus do not impact its convergence.

To prove this, assume $\rho_t \in (0, 1)$. By the quadratic formula the denominator of $f_t(x)$ has roots x_1, x_2 , defined as

$$x_1, x_2 = \frac{2(\rho_t - 1) \pm \sqrt{-4(\rho_t^2 - 1)}}{4(\rho_t - 1)}. \quad (16)$$

We first show that $x_1 := \frac{2(\rho_t - 1) + \sqrt{-4(\rho_t^2 - 1)}}{4(\rho_t - 1)} < 0$. This is equivalent to proving $2(\rho_t - 1) + \sqrt{-4(\rho_t^2 - 1)} > 0$. We construct the following chain of implications

$$2(\rho_t - 1) + \sqrt{-4(\rho_t^2 - 1)} > 0 \quad (17)$$

$$\iff -2(\rho_t - 1) < \sqrt{-4(\rho_t^2 - 1)} \quad (18)$$

$$\iff 2(1 - \rho_t) < 2\sqrt{1 - \rho_t^2}. \quad (19)$$

Since $\rho_t \in (0, 1)$, we know both $1 - \rho_t < 1 - \rho_t^2$, and $1 - \rho_t^2 < \sqrt{1 - \rho_t^2}$, thus proving the claim. To show that $x_2 := \frac{2(\rho_t - 1) - \sqrt{-4(\rho_t^2 - 1)}}{4(\rho_t - 1)} > 1$, we aim to prove the equivalent statement $2(\rho_t - 1) - \sqrt{-4(\rho_t^2 - 1)} < 4(\rho_t - 1)$. Again we construct a chain of implications.

$$2(\rho_t - 1) - \sqrt{-4(\rho_t^2 - 1)} < 4(\rho_t - 1) \quad (20)$$

$$\iff \sqrt{-4(\rho_t^2 - 1)} > -2(\rho_t - 1). \quad (21)$$

Inequality (21) is identical to (18), thus proving the claim. Since we have shown that no roots can exist in the denominator of $f_t(x)$ for $x \in [0, 1]$, we can express the limit as follows:

$$\lim_{t \rightarrow \infty} f_t(x) = \frac{\lim_{t \rightarrow \infty} 2x^3(\rho_t - 1) - x^2(\rho_t - 1)(2 - \lambda) + x(\rho_t + \lambda)}{\lim_{t \rightarrow \infty} (1 + \lambda)(2x^2(\rho_t - 1) - 2x(\rho_t - 1) + \rho_t)} \quad (22)$$

$$= \frac{2x^3(\rho - 1) - x^2(\rho - 1)(2 - \lambda) + x(\rho + \lambda)}{(1 + \lambda)(2x^2(\rho - 1) - 2x(\rho - 1) + \rho)}. \quad (23)$$

which gives the desired result. \blacktriangleleft

► **Lemma 8.** For any fixed t , if $0 < x < 1/2$:

(1) If $\rho_t > 1$, then $f_t(x) < x$.

(2) If $\rho_t < 1$, then $f_t(x) > x$.

(3) If $\rho_t = 1$, $f_t(x) = x$.

Similarly, if $0 < x < 1/2$

(1) If $\rho > 1$, then $f(x) < x$.

(2) If $\rho < 1$, then $f(x) > x$.

(3) If $\rho = 1$, $f(x) = x$.

The proof of this follows similarly to [12] and is omitted.

► **Lemma 9.** $f(x)$ has fixed points at $x = \{0, \frac{1}{2}, 1\}$ if $\rho \neq 1$. If $\rho = 1$, then $f(x) = x \forall x$. Furthermore, for $x \in (0, 1/2)$ the function $f(x)$ monotonically converges to 0 when $\rho > 1$, monotonically converges to $1/2$ when $\rho < 1$, and remains constant if $\rho = 1$.

The proof follows directly from [12] and is omitted.

► **Theorem 10.**

- If $\rho > 1$, the deterministic system Φ_t will converge to 0.
- If $\rho < 1$, the deterministic system Φ_t will converge to $\frac{1}{2}$.

Proof. If $\rho > 1$, then there exists T_0 where we have $\rho_t > 1$ for $t \geq T_0$. Thus for $t > T_0$ Lemma 8 implies that $\Phi_{t+1} < \Phi_t$. Since $\Phi_t \in [0, 1/2]$ is monotonically strictly decreasing, it must converge to a limit. We claim that this limit must be 0.

For contradiction, let Φ_t converge to $\alpha > 0$. Since Φ_t is decreasing it must converge from the right and there exists T_1 such that for $t > T_1$, we have $\Phi_t > \alpha$.

Let $f(\alpha) = \beta < \alpha$, by Lemma 8. By the continuity of f , we have a neighborhood of α such that for all x in that neighborhood, $f(x)$ is arbitrarily close to β . Hence, there exists δ such that

$$f(\alpha + \delta_1) < \alpha - \frac{\alpha - \beta}{2} \quad (24)$$

for all $\delta_1 \in (0, \delta)$.

Lemma 7 implies that f_t converges uniformly to f , and there exists T_2 such that for $t > T_2$, we have $|f_t(x) - f(x)| < \frac{\alpha - \beta}{4}$ for all x . Note that $f_t(\Phi_t) = \Phi_{t+1}$ by definition of the update function $f_t(x)$, so we have

$$|f_t(\Phi_t) - f(\Phi_t)| = |\Phi_{t+1} - f(\Phi_t)| < \frac{\alpha - \beta}{4}. \quad (25)$$

Consider $\Phi_t \in (\alpha, \alpha + \delta)$ for $t > \max\{T_0, T_1, T_2\}$. We combine equations (24) and (25) to get:

$$\Phi_{t+1} < f(\Phi_t) + \frac{\alpha - \beta}{4} < \alpha - \frac{\alpha - \beta}{2} + \frac{\alpha - \beta}{4} < \alpha. \quad (26)$$

But this is a contradiction, and therefore Φ_t must converge to 0. A similar argument holds in the case of $\rho < 1$. ◀

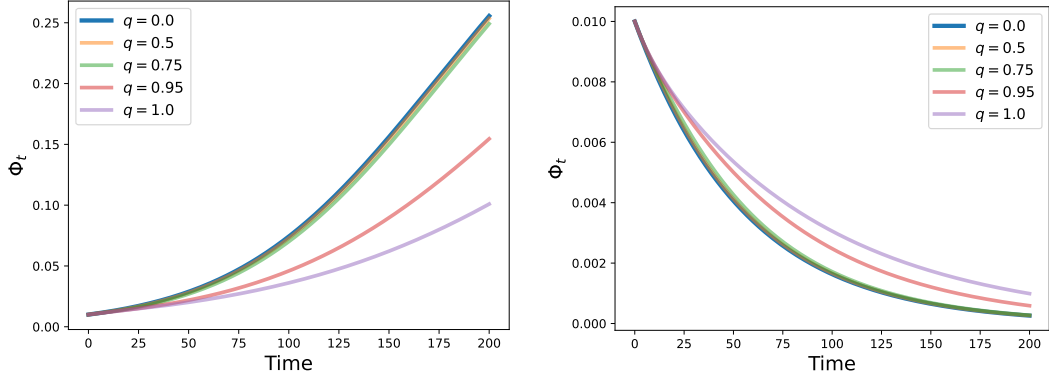
6.1 The Role of Memory in the Deterministic Approximation

This paper considers a memory parameter q , which is the probability that an edge from one time step persists to the next time step, and this is the primary extension to the model in [12]. Define $\rho_0 := \frac{\mu_{RR} w_{RR}}{\mu_{RB} w_{RB}}$, which is the threshold parameter of the model in [12] with $q = 0$, as well as the value of ρ_t at $t = 0$.

In the case that $\rho_0 \neq 1$, we empirically observe that the value of q only changes the convergence rate of the system to a particular fixed point, and does not change the fixed point itself (Figure 3). In Figure 3a, we see that since $\mu_{RR} < \mu_{RB}$ and $\rho_0 < 1$, increasing q causes our process to converge to $\frac{1}{2}$ at a slower rate. In Figure 3b, we see that $\mu_{RR} > \mu_{RB}$ and $\rho_0 > 1$, therefore increasing q causes the process to converge to 0 slower. Additionally, we see that for $q \in \{0, 0.5, 0.75\}$, the approximation trajectories are all roughly the same. It is only once $q \in \{0.95, 1.0\}$ that the approximation starts to noticeably differ in its trajectory.

In the case where $\rho_0 = 1$, in the absence of any memory the fraction of the minority will remain constant. However, we see that the inclusion of memory propels the system towards a fixed point (i.e. reaching parity or vanishing). In Figure 4, we see with $\mu_{RR} > \mu_{RB}$ and $\rho_0 = 1$, that increasing q causes the once stationary process to converge to $\frac{1}{2}$. Figure 4b plots the limit point of the deterministic process Φ_∞ over different values of our memory parameter q . Φ_{1000} is used as an approximation for Φ_∞ . We see that a phase transition happens soon after q crosses 0.9. We summarize these observations in Table 1.

6:10 Modeling Diversity Dynamics in Time-Evolving Collaboration Networks



(a) Deterministic Approximation Φ_t^1 where $\mathbf{P} = \begin{bmatrix} 8 & 10 \\ 10 & 8 \end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. (b) Deterministic Approximation Φ_t^2 where $\mathbf{P} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

■ **Figure 3** Plot of two different deterministic approximations Φ_t^1, Φ_t^2 with different \mathbf{P}, \mathbf{W} parameters over $q \in \{0, 0.5, 0.75, 0.95, 1\}$ with $\lambda = 0.1, N_0 = 100$. In Figure 3a, we see $\mu_{RR} < \mu_{RB}$ and $\rho_0 < 1$, so as q increases to 1, our process Φ_t^1 converges to $\frac{1}{2}$ at a slower rate. In Figure 3b, we have $\mu_{RR} > \mu_{RB}$, and $\rho_0 > 1$, so increasing q to 1 has the effect of Φ_t^2 decreasing the rate of convergence to 0.

■ **Table 1** The convergence of the system with memory in comparison with the memoryless baselines model from [12].

	Baseline model [12]	$\mu_{RR} < \mu_{RB}$	$\mu_{RR} > \mu_{RB}$	$\mu_{RR} = \mu_{RB}$
$\rho_0 > 1$	Converges to 0	Faster to 0	Slower to 0	No change
$\rho_0 < 1$	Converges to $\frac{1}{2}$	Slower to $\frac{1}{2}$	Faster to $\frac{1}{2}$	No change
$\rho_0 = 1$	Constant	Converges to 0	Converges to $\frac{1}{2}$	No change

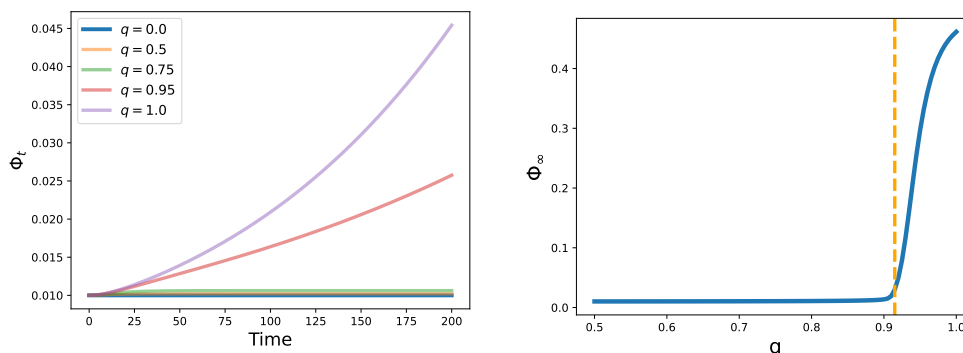
7 APS Dataset & Model Validation

To validate this model, we assess gender diversity in scientific collaboration networks. We use the co-authorship network from the American Physical Society (APS) [1], a database of Physics publications that has been used on a variety of meta-analyses of research collaborations [32, 46]. In particular, we use a filtered version of the data set as in [46]. We join this with the citation data from [1]. We use this network to find how often a paper is cited within the APS community, which serves as a proxy for edge weight in our model.

7.1 APS Dataset Information

7.1.1 Data Processing

We largely follow the procedure used in [32]. First, we used data provided in [46] which de-duplicates author names. We also restrict all publications to those that have been published between 1980 and 2009. The primary reason for this is because [46] provides a supplementary dataset of publications with the de-duplicated author names from 1893 until 2010, and to control for large political changes, we only selected the final three decades. We also exclude the year 2010 because the supplementary data in [46] did not provide data for the full

(a) Deterministic Approximation Φ_t .(b) Deterministic Approximation limit Φ_{1000} .

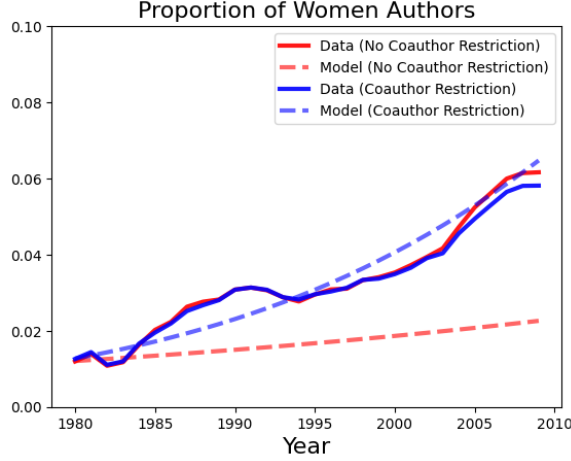
■ **Figure 4** Plot of our deterministic approximation Φ_t and threshold parameter ρ_t for $\mathbf{P} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix} 1 & 1.25 \\ 1.25 & 1 \end{bmatrix}$ over $q \in \{0, 0.5, 0.75, 0.95, 1\}$. Here we see $\mu_{RR} > \mu_{RB}$ and $\rho_0 = 1$, so as q increases, we see our process change from stationary to convergent towards $\frac{1}{2}$. Additionally we plot Φ_{1000} over different values of q . We empirically observe that there is a threshold $q^* \approx 0.915$ where the Φ_{1000} values quickly jump from 0 towards converging to $\frac{1}{2}$.

year of 2010. We also only include *active authors*, defined as authors who have, since their first publication, published at least once every five years until the most recent year 2009. Additionally, we filtered the data by removing all publications with 0 authors.

To predict the gender of authors, we used the pre-trained gender classifier GenderPerformer [50]. We let GenderPerformer predict gender for our authors in the dataset and only kept the authors whose gender could be predicted with $> 80\%$ confidence. We note that this model could potentially be biased towards Western-sounding names, which could reduce the relevance of our analysis as applied to diversity in Physics collaborations. We also note that this model classifies gender on a Male/Female binary, which does not accurately reflect all authors' gender identities.

This left us with a set of 14,793 authors, which formed our collaboration network. We consider our collaboration network as a sequence of graphs over time, $G_t = (V_t, E_t)$ where $t \in \{1980, \dots, 2009\}$. For time t , our nodes V_t are the set of authors whose first publication was *at or before* time t . This formulation ensures that our network will constantly grow over time. A pair of nodes is connected with an edge if the two nodes co-authored a paper together. In the case of a paper having more than two authors, we default to connecting the first and last authors of the paper, following the procedure in [32]. However, in the case that the first or last co-author is either a non-active author or their gender couldn't be predicted by GenderPerformer, we choose the authors nearest to the first/last position (i.e. if the first author isn't available, choose the second author, same with last and second to last authors and so on). This procedure alters the data by removing all author-publication pairs if the author was not nearest to the first or last (i.e. if Author A wrote one paper in 1980 where they were the third author out of seven, they would not appear in the graph for 1980, but instead would appear whenever they were nearest to first/last author of a paper). We allow for self-edges if the author published a paper with no co-authors. Each edge is given weight according to the number of citations received five years after the paper has been published. The best-fit model to the data according to this edge generation is shown in Figure 1.

An alternate strategy for creating a collaboration network is to connect all pairs of coauthors on a paper. However, this leads to an over-representation of papers with large numbers of co-authors in our weight computation, and reduces our model's predictive accuracy, as seen in Figure 5.



■ **Figure 5** Plot of the proportion of women in the APS co-authorship network from 1980-2009 for data created with no restriction on coauthorship (every pair of coauthors is connected in the graph) in red, and data created where only the first and last authors connected (approximately, see text for exact procedure) in blue. The deterministic model's predictions are overlaid in the dashed lines.

7.2 Estimating Model Parameters

To fit our model to the data, we perform the following procedure to estimate model parameters. We assume our probability and weight matrices at time t are of the form $\mathbf{P}_t = \begin{bmatrix} \mu_{MM}(t) & \mu_{MF}(t) \\ \mu_{MF}(t) & \mu_{MM}(t) \end{bmatrix}$, $\mathbf{W}_t = \begin{bmatrix} w_{MM}(t) & w_{MF}(t) \\ w_{MF}(t) & w_{MM}(t) \end{bmatrix}$, meaning subscripts MM and MF represent *in-community* and *cross-community* parameters respectively. To estimate our in-community probability parameter $\hat{\mu}_{MM}(t)$, we counted all in-community edges (both $M - M$ and $F - F$) and divided them by the total number of possible in-community edges in our Graph. Let $n_{MM}(t), n_{MF}(t), n_{FF}(t)$ be the number of Male-Male, Male-Female, and Female-Female edges present in our collaboration network G_t , respectively. Then our estimated parameter can be defined as

$$\hat{\mu}_{MM}(t) = n_t \left(\frac{n_{MM}(t) + n_{FF}(t)}{(n_t^M)^2 + (n_t^F)^2} \right). \quad (27)$$

$\hat{\mu}_{MF}(t)$ is computed similarly. To derive our in-community weight parameter $\hat{w}_{MM}(t)$ (or similarly $\hat{w}_{MF}(t)$), we averaged the weights over all in-community edges at time t .

$$\hat{w}_{MM}(t) = \frac{\sum_{c[i], c[j] \in (M, M), (F, F)} (w_{ij})_t}{n_{MM}(t) + n_{FF}(t)}. \quad (28)$$

To derive the probability of an edge persisting from time t to time $t + 1$, we consider the set $E_t \cap E_{t+1}$, which is the set of edges present in both G_t and G_{t+1} and find their ratio with respect to the number of edges in G_t as

$$\hat{q}_t = \frac{|E_t \cap E_{t+1}|}{|E_t|}. \quad (29)$$

Finally, to estimate the network growth parameters, we perform an exponential curve-fitting on the size of the network over time. Namely, choose approximation $f(t) := \hat{N}_0(1 + \hat{\lambda})^t$ which minimizes L_2 error over our network size $|V_t|$. Formally, we have

$$\hat{\lambda}, \hat{N}_0 = \arg \min_{\lambda, N_0} \sum_{t=1980}^{2009} (N_0(1 + \lambda)^t - |V_t|)^2. \quad (30)$$

In order to find the overall estimates $\hat{\mathbf{P}}, \hat{\mathbf{W}}, \hat{q}$, we simply average over times t , which results in $\hat{\mathbf{P}} = \begin{bmatrix} 0.359 & 0.468 \\ 0.468 & 0.359 \end{bmatrix}$, $\hat{\mathbf{W}} = \begin{bmatrix} 14.337 & 18.026 \\ 18.026 & 14.337 \end{bmatrix}$, $\hat{q} = 0.265$, $\hat{\lambda} = 0.115$, $\hat{N}_0 = 466.2$ being the estimated parameters for our model with coauthor restriction. For the model in Figure 5 with no coauthor restriction, these estimated parameters were $\hat{\mathbf{P}} = \begin{bmatrix} 0.511 & 0.718 \\ 0.718 & 0.511 \end{bmatrix}$, $\hat{\mathbf{W}} = \begin{bmatrix} 54.71 & 47.17 \\ 47.17 & 54.71 \end{bmatrix}$, $\hat{q} = 0.309$, $\hat{\lambda} = 0.116$, $\hat{N}_0 = 473.0$. Note that we do not use \hat{N}_0 in generating any of the plots. Instead, we use the true value of $N_0 = 251$.

7.3 Results & Discussion

With the optimal estimated parameter values from Section 7.2, we fit the model to our dataset and assessed how accurately the model predicts the growth of women authors in the scientific collaboration network, as in Figure 1. In the data, we see that from 1980-2009, the proportion of active women researchers grew from around 1% to 6%. This is roughly captured by our model when fitted to the dataset, predicting slightly over 6% active women researchers in 2009. We also observe that the addition of memory did not significantly change the model's predicted trajectory. The estimated parameters from Section 7.2 suggest that increasing q will cause a slower rate of convergence to $\frac{1}{2}$. However, we empirically observe that the estimated memory parameter $\hat{q} = 0.265$ does not meaningfully alter the trajectory when compared to the no memory case of $q = 0$. This was suggested by Figures 3 & 4, since in those figures large shifts away from the baseline trajectory occurred only after q was close to 1.

These results show that a combination of node influence (through preferential attachment) and homophily can partially explain the growth of the minority population in a scientific collaboration network. This result offers a powerful tool for understanding the impact of network dynamics on diversity in scientific communities, though some finer-tuned analysis is necessary to make it more accurate for prediction.

8 Conclusion

The above model is a simple abstraction that captures the effect of homophily in networks on long-term diversity, which matches sociological observations of workplace diversity and networks with preferential attachment [44, 28, 43]. Many other factors must be considered before we can have an end-to-end model, from pre-recruitment to promotion, and this is a clear limitation of our work. To expand the model, we could also consider more subtle versions of preferential attachment (e.g. a red node recruits α fraction red and $1 - \alpha$ fraction blue nodes). We could also look at node creation mechanisms beyond preferential attachment, such as a constant number of new nodes joining each timestep, or a time-inhomogenous rate parameter λ_t , to model seasonality trends in the recruitment step.

Mathematically, we have not derived an explicit definition for our threshold parameter ρ . With more knowledge of this parameter, we could precisely determine when memory significantly affects convergence. This model is also hindered by the fact that it only has fixed

points of $\{0, \frac{1}{2}, 1\}$ when in actuality, diversity isn't usually judged by equivalent parity, but rather representative parity. Finding ways to mathematically extend the model to arbitrary fixed points would be an important future step. Additionally, in our model, the network grows infinitely large as $t \rightarrow \infty$. Allowing for node departures and finite population size may make the model more applicable to real-world collaboration networks.

With this in mind, our model and our findings about its threshold property could point towards a network analysis of institutions to establish whether cross-community collaborations are frequent enough or weighted highly enough to encourage lasting diversity. This is a powerful tool because an organization can look at how a network is at one point in time, and use it to extrapolate into the future, as well as develop interventions for the present (e.g. incentives to encourage more cross-community collaborations).

References

- 1 APS Data Sets for Research — journals.aps.org. <https://journals.aps.org/datasets>. [Accessed 24-02-2024].
- 2 Joan Acker. Inequality regimes: Gender, class, and race in organizations. *Gender and Society*, 20(4):441–464, 2006.
- 3 Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Adapting the stochastic block model to edge-weighted networks. *arXiv preprint arXiv:1305.5782*, 2013.
- 4 Christopher Aicher, Abigail Z. Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, June 2014.
- 5 Lauren M. Alfrey. Diversity, Disrupted: A Critique of Neoliberal Difference in Tech Organizations. *Sociological Perspectives*, 65(6):1081–1098, December 2022. Publisher: SAGE Publications Inc.
- 6 Mackenzie Alston. Eliminating discrimination in hiring isn't enough. *IZA World of Labor*, 2023.
- 7 Christopher Archer. Modeling Diversity Dynamics in Time-Evolving Collaboration Networks. Software (visited on 13/05/2024). URL: <https://github.com/chris-archer110/SBM-diversity-model-code/blob/main/README.md>.
- 8 Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.
- 9 Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
- 10 Albert-Laszlo Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.
- 11 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, September 2004. URL: <https://www.aeaweb.org/articles?id=10.1257%2F0002828042002561&ref=exo-insight>.
- 12 Simina Brânzei, Nithish Kumar, and Gireeja Ranade. Phase transitions of diversity in stochastic block model dynamics. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2023.
- 13 Ronald S. Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, 2004. Publisher: The University of Chicago Press. URL: <https://www.jstor.org/stable/10.1086/421787>.
- 14 Antoni Calvo-Armengol and Yannis M. Ioannides. Social Networks in Labor Markets. Discussion Papers Series, Department of Economics, Tufts University 0517, Department of Economics, Tufts University, 2005. URL: <https://ideas.repec.org/p/tuf/tuftec/0517.html>.

- 15 Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267, 2009. URL: <http://www.jstor.org/stable/40247641>.
- 16 Kenneth A. Couch, Joni Hersch, and Jennifer Bennett Shinall. Fifty years later: The legacy of the civil rights act of 1964. *Journal of Policy Analysis and Management*, 34(2):424–456, 2015. URL: <http://www.jstor.org/stable/43866378>.
- 17 T. Cox. *Cultural Diversity in Organizations: Theory, Research and Practice*. Berrett-Koehler Publishers, 1993. URL: <https://www.semanticscholar.org/paper/Cultural-diversity-in-organizations-%3A-theory%2C-and-Cox/8fbae390ceab816f036c2a5835a79f6bc7002d8e>.
- 18 Matthew Eichhorn, Siddhartha Banerjee, and David Kempe. Online team formation under different synergies. In *Workshop on Internet and Network Economics*, 2022. URL: <https://api.semanticscholar.org/CorpusID:252846784>.
- 19 Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. Publisher: University of Chicago Press. URL: <https://www.jstor.org/stable/2776392>.
- 20 Peter Grindrod and Mark Parsons. Social networks: Evolving graphs with memory dependent edges. *Physica A: Statistical Mechanics and its Applications*, 390(21-22):3970–3981, 2011.
- 21 Cedric Herring and Loren Henderson. From affirmative action to diversity: Toward a critical diversity perspective. *Critical Sociology*, 38(5):629–643, 2012.
- 22 Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
- 23 Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.
- 24 Jian Huang, Ziming Zhuang, Jia Li, and C Lee Giles. Collaboration over time: characterizing and modeling network evolution. In *WSDM*, pages 107–116, 2008.
- 25 Matthew O Jackson and Asher Wolinsky. A strategic model of social and economic networks. In *Networks and groups*, pages 23–49. Springer, 2003.
- 26 Jamie Dolkas Joan C. Williams. Data-Driven Diversity — hbr.org. <https://hbr.org/2022/03/data-driven-diversity>. [Accessed 26-01-2024].
- 27 Judd Kessler and Corinne Low. Research: How Companies Committed to Diverse Hiring Still Fail — hbr.org. <https://hbr.org/2021/02/research-how-companies-committed-to-diverse-hiring-still-fail>. [Accessed 16-01-2024].
- 28 Kibae Kim and Jörn Altmann. Effect of homophily on network formation. *Communications in Nonlinear Science and Numerical Simulation*, 44:482–494, March 2017. URL: <https://www.sciencedirect.com/science/article/pii/S1007570416302805>.
- 29 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2016.
- 30 Oliver Knill. Probability theory and stochastic processes with applications. <https://people.math.harvard.edu/~knill/books/KnillProbability.pdf>, 2009. [Accessed 23-02-2024].
- 31 Clement Lee and Darren J. Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1), December 2019.
- 32 Weihua Li, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. Early coauthorship with top scientists predicts success in academic careers. *Nature Communications*, 10:5170, November 2019.
- 33 Stephanie Lunn and Monique Ross. Cracks in the foundation: Issues with diversity and the hiring process in computing fields, 2021.
- 34 Miller McPherson. An ecology of affiliation. *American Sociological Review*, 48(4):519–532, 1983. URL: <http://www.jstor.org/stable/2117719>.

6:16 Modeling Diversity Dynamics in Time-Evolving Collaboration Networks

- 35 Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. URL: <http://www.jstor.org/stable/2678628>.
- 36 Francisco Bauzá Mingueza, Mario Floría, Jesús Gómez-Gardeñes, Alex Arenas, and Alessio Cardillo. Characterization of interactions' persistence in time-varying networks - Scientific Reports — nature.com. <https://www.nature.com/articles/s41598-022-25907-7#citeas>. [Accessed 21-01-2024].
- 37 Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.
- 38 Pamela Newkirk. *Diversity, Inc.: The Failed Promise of a Billion-Dollar Business*. Bold Type Books, New York, October 2019.
- 39 Tiago P. Peixoto. Nonparametric weighted stochastic block models. *Phys. Rev. E*, 97:012306, January 2018.
- 40 Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:125–134, October 2019.
- 41 S. S. Phulari, S. D. Khamitkar, N. K. Deshmukh, P. U. Bhalchandra, S. N. Lokhande, and A. R. Shinde. Understanding Formulation of Social Capital in Online Social Network Sites (SNS), February 2010. URL: <https://arxiv.org/abs/1002.1201v1>.
- 42 Fereshteh Rabbani, Tamer Khraisha, Fatemeh Abbasi, and Gholam Reza Jafari. Memory effects on link formation in temporal networks: A fractional calculus approach. *Physica A: Statistical Mechanics and its Applications*, 564:125502, 2021. URL: <https://www.sciencedirect.com/science/article/pii/S0378437120308001>.
- 43 Lauren A. Rivera. Hiring as cultural matching: The case of elite professional service firms. *American Sociological Review*, 77(6):999–1022, 2012.
- 44 Roberto M. Rubineau, Brian; Fernandez. Tipping Points: Referral Homophily and Job Segregation — dspace.mit.edu. <https://dspace.mit.edu/handle/1721.1/66931>. [Accessed 20-01-2024].
- 45 Jad Salem, Deven Desai, and Swati Gupta. Don't let Ricci v. DeStefano hold you back: A bias-aware legal solution to the hiring paradox. In *Facct*, pages 651–666, 2022.
- 46 Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.
- 47 Ana-Andreea Stoica, Jessy Xinyi Han, and Augustin Chaintreau. Seeding network influence in biased networks and the benefits of diversity. In *WWW*, pages 2089–2098. ACM / IW3C2, 2020.
- 48 Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. In *AIES*, pages 989–999, 2021.
- 49 Phyllis Tharenou. Explanations of managerial career advancement. *Australian Psychologist*, 32(1):19–28, 1997.
- 50 Zijian Wang and David Jurgens. It's going to be okay: Measuring access to support in online communities. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Association for Computational Linguistics.

A Appendix

A.1 Proof of Lemma 4:

Before we prove the bounds on the ϕ_t process itself, we first must prove bounds on $\mathbb{E}\left[\frac{R_t}{R_t+B_t}\right]$, as defined in the following Lemma.

► **Lemma 11.** *Assume $\exists \varepsilon \in (0, \frac{1}{2})$ such that $\left(\frac{1}{n_t}\right)^{\frac{1}{2}-\varepsilon} \leq \phi_t \leq \frac{1}{2}$. Then there exists N_0 such that for $n_t \geq N_0$ the following holds*

$$\left(1 - \frac{1}{n_{t-1}^{\varepsilon/4}}\right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t+B_t]} < \frac{R_t}{R_t+B_t} < \left(1 + \frac{1}{n_{t-1}^{\varepsilon/4}}\right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t+B_t]}. \quad (31)$$

With probability at least $1 - \frac{8}{\exp(C_1 n_{t-1}^\varepsilon)}$, where $C_1 = \min\left\{\frac{\mu_{RR}}{24}, \frac{\mu_{RB}}{12}\right\}$ and $\varepsilon \in (0, \frac{1}{2})$. Additionally, we have

$$\left(1 - \frac{1}{n_{t-1}^{\varepsilon/5}}\right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t+B_t]} \leq \mathbb{E}\left[\frac{R_t}{R_t+B_t}\right] \leq \left(1 + \frac{1}{n_{t-1}^{\varepsilon/5}}\right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t+B_t]}. \quad (32)$$

Prerequisite (Chernoff Bound). We define a Chernoff Bound [30] on a random variable X :

$$\Pr(X \geq \delta) \leq \mathbb{E}[e^{sX}]e^{-s\delta}. \quad (33)$$

For all $s > 0$. A corollary of this is also

$$\Pr(|X - \mathbb{E}[X]| \geq \delta \mathbb{E}[X]) \leq 2e^{-\delta^2 \mathbb{E}[X]/3}. \quad (34)$$

◀

Proof of Lemma 11. Consider random variables RR, RB, BB , representing the total weight of Red-Red, Red-Blue, and Blue-Blue edges respectively. We can bound the number of these edges (RR/w_{RR}), (RB/w_{RB}) using the Chernoff Bound

$$\begin{cases} \Pr\left(\left|\frac{RR}{w_{RR}} - \frac{\mathbb{E}[RR]}{w_{RR}}\right| \geq \delta \frac{\mathbb{E}[RR]}{w_{RR}}\right) \leq 2 \exp\left(-\frac{\delta^2}{6} (n_t^R)^2 \Pi_{RR}(t)\right) \\ \Pr\left(\left|\frac{RB}{w_{RB}} - \frac{\mathbb{E}[RB]}{w_{RB}}\right| \geq \delta \frac{\mathbb{E}[RB]}{w_{RB}}\right) \leq 2 \exp\left(-\frac{\delta^2}{6} (n_t^R n_t^B) \Pi_{RB}(t)\right) \end{cases}. \quad (35)$$

From our assumption, we find that $n_t^R \geq n_t^{\frac{1}{2}+\varepsilon}$ where $\varepsilon \in (0, \frac{1}{2})$, and $n_t^B \geq \frac{n_t}{2}$. Define constant $C_1 = \min\left\{\frac{\mu_{RR}}{24}, \frac{\mu_{RB}}{12}\right\}$. Letting $\delta = \frac{1}{n_t^{\varepsilon/2}}$, we can establish the following probability bounds on edge weights RR, RB, BB :

$$\Pr\left(\left|RR - \mathbb{E}[RR]\right| \geq \frac{\mathbb{E}[RR]}{n_t^{\varepsilon/2}}\right) \leq 2 \exp\left(-\frac{n_t \Pi_{RR}(t)}{6} n_t^\varepsilon\right) \leq 2 \exp(-C_1 n_t^\varepsilon). \quad (36)$$

$$\Pr\left(\left|RB - \mathbb{E}[RB]\right| \geq \frac{\mathbb{E}[RB]}{n_t^{\varepsilon/2}}\right) \leq 2 \exp\left(-\frac{n_t \Pi_{RB}(t)}{12} n_t^{\frac{1}{2}}\right) \leq 2 \exp(-C_1 n_t^\varepsilon). \quad (37)$$

$$\Pr\left(\left|BB - \mathbb{E}[BB]\right| \geq \frac{\mathbb{E}[BB]}{n_t^{\varepsilon/2}}\right) \leq 2 \exp\left(-\frac{n_t \Pi_{BB}(t)}{24} n_t^{1-\varepsilon}\right) \leq 2 \exp(-C_1 n_t^\varepsilon). \quad (38)$$

6:18 Modeling Diversity Dynamics in Time-Evolving Collaboration Networks

See that $R_t = RR_t + RB_t$, so we can establish bounds on the total weight of red R by union bound

$$\Pr\left(|R - \mathbb{E}[R]| \geq \frac{\mathbb{E}[R]}{n_t^{\varepsilon/2}}\right) \leq 4 \exp(-C_1 n_t^\varepsilon). \quad (39)$$

Likewise for the total blue weight B

$$\Pr\left(|B - \mathbb{E}[B]| \geq \frac{\mathbb{E}[B]}{n_t^{\varepsilon/2}}\right) \leq 4 \exp(-C_1 n_t^\varepsilon). \quad (40)$$

We see that the event $|R - \mathbb{E}[R]| < \frac{\mathbb{E}[R]}{n_t^{\varepsilon/2}}$ implies $\left(1 - \frac{1}{n_t^{\varepsilon/2}}\right) \mathbb{E}[R] < R < \left(1 + \frac{1}{n_t^{\varepsilon/2}}\right) \mathbb{E}[R]$, so we use this to bound our ratio $\frac{R}{R+B}$

$$\left(\frac{\left(1 - \frac{1}{n_t^{\varepsilon/2}}\right)}{\left(1 + \frac{1}{n_t^{\varepsilon/2}}\right)}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]} < \frac{R}{R+B} < \left(\frac{\left(1 + \frac{1}{n_t^{\varepsilon/2}}\right)}{\left(1 - \frac{1}{n_t^{\varepsilon/2}}\right)}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]} \quad (41)$$

$$\left(1 - \frac{1}{n_t^{\varepsilon/4}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]} < \frac{R}{R+B} < \left(1 + \frac{1}{n_t^{\varepsilon/4}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]}. \quad (42)$$

Define event \mathcal{G} as the event of this inequality holding. By union bound on the values from (39) and (40), we see that $\Pr(\mathcal{G}^C) \leq 8 \exp(-C_1 n_t^\varepsilon)$. Therefore our inequality holds with probability at least $1 - \frac{8}{\exp(C_1 n_t^\varepsilon)}$, concluding the proof for inequality (31).

To prove the second inequality (32), we extend our bound to the expected value $\mathbb{E}\left[\frac{R_t}{R_t+B_t}\right]$. Notice our ratio $0 < \frac{R_t}{R_t+B_t} < 1$, so to upper bound our expectation we condition on our inequality \mathcal{G} from (42).

$$\mathbb{E}\left[\frac{R_t}{R_t+B_t}\right] \leq \left(\left(1 + \frac{1}{n_t^{\varepsilon/4}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]}\right) \Pr(\mathcal{G}) + 1 \cdot (1 - \Pr(\mathcal{G})) \quad (43)$$

$$\leq \left(1 - \frac{8}{\exp(C_1 n_t^\varepsilon)}\right) \left(1 + \frac{1}{n_t^{\varepsilon/4}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]} + \frac{8}{\exp(C_1 n_t^\varepsilon)} \quad (44)$$

$$\leq \left(1 + \frac{1}{n_t^{\varepsilon/4}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]} + \frac{8}{\exp(C_1 n_t^\varepsilon)} \quad (45)$$

$$\leq \left(1 + \frac{1}{n_t^{\varepsilon/5}}\right) \frac{\mathbb{E}[R]}{\mathbb{E}[R+B]}. \quad (46)$$

This above inequality only holds when $n_t \geq N_0$. Therefore, defining constants N_1, N_2

N_1 : When $n_t \geq N_1 = \exp\left(\frac{20 \ln 2}{\varepsilon}\right)$ then $\frac{1}{n_t^{\varepsilon/5}} - \frac{1}{n_t^{\varepsilon/4}} \geq \frac{1}{n_t^{\varepsilon/4}}$

N_2 : $n_t^\varepsilon \geq \ln(n_t) \left(\frac{2-3\varepsilon}{4C_1}\right) + \frac{\ln 2}{4C_1} - \frac{1}{C} \ln\left(\frac{\mu_{RR} \mu_{WB}}{5\mu_{RR} \mu_{RR} + 4\mu_{RB} \mu_{RB}}\right)$

Thus, let $N_0 = \max\{N_1, N_2\}$ so that for $n_t \geq N_0$ both bounds hold, completing the proof for the upper bound. Now to prove the lower bound, we condition again

$$\mathbb{E} \left[\frac{R_t}{R_t + B_t} \right] \geq \left(\left(1 - \frac{1}{n_t^{\varepsilon/4}} \right) \frac{\mathbb{E}[R]}{\mathbb{E}[R + B]} \right) Pr(\mathcal{G}) + 0 \cdot (1 - Pr(\mathcal{G})) \quad (47)$$

$$\geq \left(1 - \frac{8}{\exp(C_1 n_t^\varepsilon)} \right) \left(1 - \frac{1}{n_t^{\varepsilon/4}} \right) \frac{\mathbb{E}[R]}{\mathbb{E}[R + B]} \quad (48)$$

$$\geq_{N_3} \left(1 - \frac{1}{n_t^{\varepsilon/5}} \right) \frac{\mathbb{E}[R]}{\mathbb{E}[R + B]}. \quad (49)$$

There exists N_3 such that inequality (49) holds for $n_t \geq N_3$

$$\left(1 - \frac{8}{\exp(8C_1 n_t^\varepsilon)} \right) \left(1 - \frac{1}{n_t^{\varepsilon/4}} \right) \geq 1 - \frac{1}{n_t^{\varepsilon/5}}. \quad (50)$$

Thus let $N_0 \geq \max\{N_1, N_2, N_3\}$ such that all our bounds hold for $n_t \geq N_0$ and we have

$$\left(1 - \frac{1}{n_t^{\varepsilon/5}} \right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]} \leq \mathbb{E} \left[\frac{R_t}{R_t + B_t} \right] \leq \left(1 + \frac{1}{n_t^{\varepsilon/5}} \right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]}. \quad (51)$$

Concluding the proof of (32). \blacktriangleleft

Now with Lemma 11, we can prove Lemma 4.

Proof of Lemma 4. Consider the expected number of new red nodes arriving at time $t + 1$, denoted $\mathbb{E}[m_{t+1}^R]$, we can bound this quantity by rounding our total expected new nodes λn_t

$$\lceil \lambda n_t \rceil \mathbb{E} \left[\frac{R_t}{R_t + B_t} \right] \leq \mathbb{E}[m_{t+1}^R | \mathcal{F}_t] \leq \lceil \lambda n_t \rceil \mathbb{E} \left[\frac{R_t}{R_t + B_t} \right]. \quad (52)$$

Additionally, we can use inequality (32) from Theorem 1 to bound $\mathbb{E} \left[\frac{R_t}{R_t + B_t} \right]$

$$\lceil \lambda n_t \rceil \left(1 - \frac{1}{n_t^{\varepsilon/5}} \right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]} \leq \mathbb{E}[m_{t+1}^R | \mathcal{F}_t] \leq \lceil \lambda n_t \rceil \left(1 + \frac{1}{n_t^{\varepsilon/5}} \right) \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]}. \quad (53)$$

Also note that we define $\Gamma_q(\phi_t) := \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]}$

$$\frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]} = \frac{(n_t^R)^2 w_{RR} \Pi_{RR}(t) + (n_t^R n_t^B) w_{RB} \Pi_{RB}(t)}{((n_t^R)^2 + (n_t^B)^2) w_{RR} \Pi_{RR}(t) + 2(n_t^R n_t^B) w_{RB} \Pi_{RB}(t)} \quad (54)$$

$$= \frac{\phi_t w_{RR} \Pi_{RR}(t) + \phi_t (1 - \phi_t) w_{RB} \Pi_{RB}(t)}{(\phi_t^2 + (1 - \phi_t)^2) w_{RR} \Pi_{RR}(t) + 2\phi_t (1 - \phi_t) w_{RB} \Pi_{RB}(t)} \quad (55)$$

$$= \Gamma_q(\phi_t). \quad (56)$$

Using this to bound our original equation for $\mathbb{E}[\phi_{t+1} | \mathcal{F}_t] = \frac{n_t^R + \mathbb{E}[m_{t+1}^R | \mathcal{F}_t]}{(1 + \lambda) n_t}$

$$\frac{\phi_t + \frac{\lceil \lambda n_t \rceil}{n_t} \left(1 - \frac{1}{n_t^{\varepsilon/5}} \right) \cdot \Gamma_q(\phi_t)}{1 + \frac{\lceil \lambda n_t \rceil}{n_t}} \leq \mathbb{E}[\phi_{t+1} | \mathcal{F}_t] \leq \frac{\phi_t + \frac{\lceil \lambda n_t \rceil}{n_t} \left(1 + \frac{1}{n_t^{\varepsilon/5}} \right) \cdot \Gamma_q(\phi_t)}{1 + \frac{\lceil \lambda n_t \rceil}{n_t}}. \quad (57)$$

As $t \rightarrow \infty$, our bounds converge to the following expression

$$\mathbb{E}[\phi_{t+1} | \mathcal{F}_t] = \frac{\phi_t + \lambda \cdot \Gamma_q(\phi_t)}{1 + \lambda}. \quad (58)$$

Which is our deterministic approximation, concluding the proof. \blacktriangleleft

A.2 Derivation of Deterministic Approximation

Using Theorem 3, we define $\Pi_{RR}(t), \Pi_{RB}(t), \Pi_{BB}(t)$, where for Red-Blue edges:

$$\Pi_{RB}(t) = \frac{\mu_{RB}}{n_t} + \sum_{k=0}^{t-1} \frac{\mu_{RB}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_j} \right). \quad (59)$$

And for Blue-Blue edges:

$$\Pi_{BB}(t) = \frac{\mu_{BB}}{n_t} + \sum_{k=0}^{t-1} \frac{\mu_{BB}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{BB}}{n_j} \right). \quad (60)$$

In our graph, the existence of an edge is not dependent on the existence of any other edges, so our probability is independent across edges. As such, we define the expected weight of Red and Blue nodes at time t as the expected weight of all Red-Red, Red-Blue, and Blue-Blue edges in our graph.

$$\mathbb{E}[R_t] = n_t^R \Pi_{RR}(t) w_{RR} + \frac{(n_t^R)(n_t^R - 1)}{2} 2w_{RR} \Pi_{RR}(t) + n_t^R n_t^B w_{RB} \Pi_{RB}(t) \quad (61)$$

$$= (n_t^R)^2 \Pi_{RR}(t) w_{RR} + (n_t^R + n_t^B) \Pi_{RB}(t) w_{RB} \quad (62)$$

$$\mathbb{E}[B_t] = n_t^B \Pi_{BB}(t) w_{BB} + \frac{(n_t^B)(n_t^B - 1)}{2} 2w_{BB} \Pi_{BB}(t) + n_t^R n_t^B w_{RB} \Pi_{RB}(t) \quad (63)$$

$$= (n_t^B)^2 w_{BB} \Pi_{BB}(t) + n_t^R n_t^B w_{RB} \Pi_{RB}(t). \quad (64)$$

Assume $\mathbf{P} = \begin{bmatrix} \mu_{RR} & \mu_{RB} \\ \mu_{RB} & \mu_{RR} \end{bmatrix}$, $\mathbf{W} = \begin{bmatrix} w_{RR} & w_{RB} \\ w_{RB} & w_{RR} \end{bmatrix}$. With this, we calculate $\frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]}$

$$\frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]} = \frac{(n_t^R)^2 w_{RR} \Pi_{RR}(t) + (n_t^R n_t^B) w_{RB} \Pi_{RB}(t)}{((n_t^R)^2 + (n_t^B)^2) w_{RR} \Pi_{RR}(t) + 2(n_t^R n_t^B) w_{RB} \Pi_{RB}(t)}. \quad (65)$$

Let ϕ_t be the fraction of red nodes and m_t be the new nodes joining the graph at time t . From our growth dynamics we know $\phi_t = \frac{n_{t-1}^R + m_t^R}{n_{t-1} + m_t}$. Note that while n_t^R and m_t^R are random variables, n_t and m_t are deterministic. We can approximate this expectation

$$\mathbb{E}[\phi_{t+1} | \mathcal{F}_t] = \frac{n_t^R + \lambda n_t \mathbb{E} \left[\frac{R_t}{R_t + B_t} \right]}{(1 + \lambda) n_t} \approx \frac{n_t^R + \lambda n_t \frac{\mathbb{E}[R_t]}{\mathbb{E}[R_t + B_t]}}{(1 + \lambda) n_t} \approx \frac{\phi_t + \lambda \cdot \Gamma_q(\phi_t)}{1 + \lambda}. \quad (66)$$

Where

$$\Gamma_q(x_t) = \frac{x_t^2 w_{RR} \Pi_{RR}(t) + x_t(1 - x_t) w_{RB} \Pi_{RB}(t)}{(x_t^2 + (1 - x_t)^2) w_{RR} \Pi_{RR}(t) + 2x_t(1 - x_t) w_{RB} \Pi_{RB}(t)}. \quad (67)$$

A.3 Proof of Lemma 6:

Proof. Consider $\rho_t = \frac{w_{RR} \Pi_{RR}(t)}{w_{RB} \Pi_{RB}(t)}$. Note that if $q = 0$, then $\rho_t = \rho_0 = \frac{w_{RR} \mu_{RR}}{w_{RB} \mu_{RB}}$ for all t . Hence we restrict our attention to the case where $q > 0$. We expand the definition of ρ_t using the explicit definitions of $\Pi_{RR}(t), \Pi_{RB}(t)$ from (5),(59), yielding

$$\rho_t = \frac{w_{RR} n_t \Pi_{RR}(t)}{w_{RB} n_t \Pi_{RB}(t)} = \frac{w_{RR} \left(\mu_{RR} + n_t \sum_{k=0}^{t-1} \frac{\mu_{RR}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_j} \right) \right)}{w_{RB} \left(\mu_{RB} + n_t \sum_{k=0}^{t-1} \frac{\mu_{RB}}{n_k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_j} \right) \right)}. \quad (68)$$

Additionally, let $r := 1 + \lambda$, we have

$$\rho_t = \frac{w_{RR} \left(\mu_{RR} + \mu_{RR} r^t \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} r^{-j} \right) \right)}{w_{RB} \left(\mu_{RB} + \mu_{RB} r^t \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_0} r^{-j} \right) \right)} \quad (69)$$

$$= \frac{w_{RR} \left(\mu_{RR} r^t \sum_{k=0}^t r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} r^{-j} \right) \right)}{w_{RB} \left(\mu_{RB} r^t \sum_{k=0}^t r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_0} r^{-j} \right) \right)} \quad (70)$$

$$= \frac{w_{RR} \mu_{RR} \left(r^{-t} + \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} r^{-j} \right) \right)}{w_{RB} \mu_{RB} \left(r^{-t} + \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_0} r^{-j} \right) \right)}. \quad (71)$$

Define the series $S_{RR}^q(t) := r^{-t} + \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} r^{-j} \right)$ and define $S_{RB}^q(t)$ similarly. Additionally, note both $S_{RR}^q(t)$ and $S_{RB}^q(t)$ must converge because they are the partial series of a geometric series. Namely,

$$S_{RR}^q(t) = r^{-t} + \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RR}}{n_0} r^{-j} \right) \leq \sum_{k=0}^t r^{-k}, \quad (72)$$

$$S_{RB}^q(t) = r^{-t} + \sum_{k=0}^{t-1} r^{-k} \prod_{j=k+1}^t q \left(1 - \frac{\mu_{RB}}{n_0} r^{-j} \right) \leq \sum_{k=0}^t r^{-k}. \quad (73)$$

Therefore we have $\lim_{t \rightarrow \infty} S_{RR}^q(t) \leq \frac{1}{1-r^{-1}} = \frac{1+\lambda}{\lambda}$ and $\lim_{t \rightarrow \infty} S_{RB}^q(t) \leq \frac{1+\lambda}{\lambda}$. Define these limits as S_{RR}^q and S_{RB}^q respectively. Also, we observe that these quantities are bounded away from 0 for $q > 0$. Observing $S_{RR}^q(0) = S_{RB}^q(0) = 1 > 0$. Since $S_{RR}^q(t), S_{RB}^q(t)$ are both positive series, this inequality then holds for all $t \geq 0$ and thus holds for the limits S_{RR}^q, S_{RB}^q as well. This allows us to express the limit of the overall ratio ρ_t :

$$\rho = \lim_{t \rightarrow \infty} \rho_t = \lim_{t \rightarrow \infty} \frac{w_{RR} \mu_{RR} S_{RR}^q(t)}{w_{RB} \mu_{RB} S_{RB}^q(t)} \quad (74)$$

$$= \frac{\lim_{t \rightarrow \infty} w_{RR} \mu_{RR} S_{RR}^q(t)}{\lim_{t \rightarrow \infty} w_{RB} \mu_{RB} S_{RB}^q(t)} \quad (75)$$

$$= \frac{w_{RR} \mu_{RR} S_{RR}^q}{w_{RB} \mu_{RB} S_{RB}^q}. \quad (76)$$

Note that because $S_{RR}^q(0) = S_{RB}^q(0) = 1$, and also for $q = 0$ we have $S_{RR}^0(t) = S_{RB}^0(t) = r^{-t}$ for all $t \geq 0$, we know that the baseline threshold parameter $\rho_0 = \frac{w_{RR} \mu_{RR}}{w_{RB} \mu_{RB}}$ is consistent with the threshold parameter in [12]. As far as we know, there is no explicit expression for S_{RR}^q or S_{RB}^q . ◀

Drawing Competitive Districts in Redistricting

Gabriel Chuang ✉

Computer Science, Columbia University, New York, NY, USA

Oussama Hanguir ✉

Lyft, Inc., New York, NY, USA

Columbia University, New York, NY, USA

Clifford Stein ✉

Industrial Engineering and Operations Research, Columbia University, New York, NY, USA

Abstract

In the process of redistricting, one important metric is the number of *competitive districts*, that is, districts where both parties have a reasonable chance of winning a majority of votes. Competitive districts are important for achieving proportionality, responsiveness, and other desirable qualities; some states even directly list competitiveness in their legally-codified districting requirements. In this work, we discuss the problem of drawing plans with at least a fixed number of competitive districts. In addition to the standard, “vote-band” measure of competitiveness (i.e., how close was the last election?), we propose a measure that explicitly considers “swing voters” - the segment of the population that may choose to vote either way, or not vote at all, in a given election. We present two main, contrasting results. First, from a computational complexity perspective, we show that the task of drawing plans with competitive districts is NP-hard, even on very natural instances where the districting task itself is easy (e.g., small rectangular grids of population-balanced cells). Second, however, we show that a simple hill-climbing procedure can in practice find districtings on real states in which *all* the districts are competitive. We present the results of the latter on the precinct-level graphs of the U.S. states of North Carolina and Arizona, and discuss trade-offs between competitiveness and other desirable qualities.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases Redistricting, Computational Complexity, Algorithms

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.7

Related Version *Full Version*: <https://arxiv.org/abs/2404.10964>

Funding *Gabriel Chuang*: NSF Graduate Research Fellowship.

Clifford Stein: Research supported in part by NSF grant CCF-2218677 and ONR grant ONR-13533312, and by the Wai T. Chang Chair in Industrial Engineering and Operations Research.

1 Introduction

In the United States, *redistricting* is the task of geographically dividing a state into a fixed number of regions called *districts*, each of which elects one representative to a legislative body (such as the U.S. House of Representatives or a state legislature). The process is prone to various types of manipulation, collectively known as *gerrymandering*, in which parties draw districting maps that are optimized for particular outcomes. For example, a party may wish to maximize seats in which their preferred voters constitute a majority, protect their party’s incumbents, or force opposition-party incumbents to run against each other. This process often results in many districts that are uncompetitive, i.e., districts in which one party’s voters constitute such a large majority that voters are denied any meaningful choice and the winning party is effectively pre-determined.



© Gabriel Chuang, Oussama Hanguir, and Clifford Stein;
licensed under Creative Commons License CC-BY 4.0

5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 7; pp. 7:1–7:22

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

To combat gerrymandering, many quantitative and qualitative measures capturing various normative criteria have been proposed. These include notions of proportionality [18], responsiveness [26], partisan symmetry [30], typicality [15], stability under perturbation [17], and many others (e.g., [9], [10]; see [31] for a comparison of several partisan-based measures).

In this work, we focus on *competitive districts*: those where one expects elections to be close and highly contested, i.e., where the outcomes of future elections are not pre-determined by the geography of the district. There are three reasons we consider competitiveness to be of particular importance:

1. Several jurisdictions in the United States explicitly require competitiveness as a quality that their districting plans must satisfy. Colorado, for example, has a requirement that plans must “maximize the number of politically competitive districts,” where competitive is defined as “having a reasonable potential for the party affiliation of the district’s representative to change at least once between federal decennial censuses” [14]. See [15] for an overview.
2. Having competitive districts is key for the responsiveness of the plan. Responsiveness is a measure of how much a given change in popular vote translates into a change in the proportion of seats held by a particular party. For example, the 14-district Congressional map enacted in North Carolina in 2023 is highly non-responsive: the state could swing 5 points more Democratic and result in zero more Democratic-held seats, or 17 points more Republican and result in only one Republican pickup. This lack of responsiveness is a direct consequence of there being only one competitive seat of the fourteen.
3. Competitive elections are generally seen as promoting various positive civic qualities, such as voter engagement, high turnout, close attention to local issues, and others.

The number of competitive districts in the country has been shrinking rapidly. In 2020, only 45 of 435 districts have a Cook Partisan Voting Index between R+3 and D+3, down from 107 in 1999 [32]; Wasserman estimates that around half of the lost swing seats are due to changes to district boundaries, rather than “true” changes in electoral behavior (e.g. changing voter preferences, geographic polarization, etc.).

1.1 Contributions

Motivated by the desirable qualities of competitive districts, we study a version of the districting problem where the goal is to draw maps with at least some fixed number of competitive districts. In addition to the intuitive notion of competitiveness, where a district is considered competitive if recent elections have been decided by close margins (e.g., by 5% or less), we consider a characterization that relies on separately counting “swing voters” – that is, voters who have a reasonable chance of voting for either party, a formulation that aligns with the Colorado requirement of having districts with a “reasonable chance of being won by either party”. We show that the problem of maximizing swing districts is NP-hard for both our characterization and the standard model, even on instances where the underlying districting problem is polynomial-time solvable: that is, the hardness comes from the competitive-districts requirement, rather than the inherent hardness of balanced graph partitioning. Despite this, we show that a simple hill-climbing procedure can achieve a very high number of competitive districts without significantly sacrificing compactness, equal population, and other desirable qualities. We demonstrate the results on data from North Carolina and Arizona, showing that it is possible to make every district competitive (although we do not necessarily advocate for doing so). We also complement the NP-completeness results by giving restrictions that make the problem of maximizing swing districts more tractable.

2 Related Work

2.1 Competitiveness of districting plans

Among the plethora of proposed evaluation metrics for districting plans, Deford et al. [15] take a comprehensive look at various criteria that aim to operationalize competitiveness, including “evenness” (how close is the vote share to 50%?), “typicality” (how close is the vote share to the national or statewide average?), and “vote-band” metrics (does the vote fall within a fixed percentage of 50%, or the statewide average?). They observe that “there is no guarantee that it is even possible to construct a plan with a large number of ... districts [that fall within a given vote-band] while adhering to reasonable compactness and boundary preservation norms,” an idea we will expand on in this work. We will adopt the vote-band metric, because its binary nature allows us to easily express the problem of drawing competitive districts as a decision problem. In addition, we will propose another metric based on swing voters.

In the same work, Deford et al. [15] also conduct an extensive ensemble analysis of how many competitive districts arise in “typical” districtings, and present two hill-climbing algorithms for optimizing directly for competitive districts. We conduct similar experiments for both vote-band and swing-voter metrics, but use a randomized weighted scheme that incorporates compactness directly via the isoperimetric score.

Other works that explore the competitiveness of enacted and proposed plans largely use margins in recent elections as their measurement for competitiveness, including [14], which investigates plans for Colorado, and [20], which studies whether independent commissions tend to draw more competitive plans.

2.2 Computational complexity of redistricting

There has also been work exploring hardness of the districting problem in a computational complexity sense, including work on auditing for local deviating groups [24], on minimizing the margin of victory in non-geographically-bounded districts [29], and on describing classes of graphs for which various redistricting tasks are NP-hard [21]. Notably, even the balanced graph partition problem (i.e., drawing population-balanced districts, without any other restrictions) is NP-hard, even for planar graphs [5]. Given this, [25] show that the problem of drawing districts where each party wins at least c seats is NP-hard *even on instances where valid (contiguous and population-balanced) maps can be found in polynomial time*, by showing a reduction from Var-Linked Planar 3-SAT. We will adopt this structure for our hardness results: we will show reductions that create instances where population-balanced districts can be drawn in polynomial time but drawing competitive districts (both vote-band competitive and swing) is hard.

3 Preliminaries and Problem Formulation

3.1 Voting setting

Our setting consists of a set of voters distributed over a given geographic area (such as a U.S. state), represented by a (typically planar) graph G with n cells (nodes). The cells are some fixed geographic units (such as counties, precincts, etc.) and edges represent adjacency¹.

¹ Depending on the jurisdiction, legal requirements may mandate either rook adjacency or queen adjacency.

7:4 Drawing Competitive Districts in Redistricting

We assume that there are only two parties in consideration, Party **A** and Party **B**; every voter may be either a partisan voter or a swing voter. Specifically, each cell $c_i : i \in \{1, \dots, n\}$ has the following four quantities, all non-negative integers:

- $Pop_i \geq 0$ indicating the total population of the cell, which we abbreviate Pop_i ;
- a_i , the number of voters that vote for Party **A**;
- b_i , the number of voters that vote for Party **B**;
- optionally, s_i , the number of swing voters (who may vote either way).

3.2 Districtings and Competitive Districts

For a fixed $d \in \{2, \dots, n-1\}$, a d -*districting* is a partition of the cells of G into d disjoint subgraphs D_1, \dots, D_d , which we call *districts*. For any district D_j , we have:

$$Party_A(D_j) = A_j = \sum_{c_i \in D_j} a_i, \quad Party_B(D_j) = B_j = \sum_{c_i \in D_j} b_i, \quad Swing(D_j) = S_j = \sum_{c_i \in D_j} s_i.$$

Since every voter is either A, B, or swing, we have $Pop(D_j) = Pop_j = A_j + B_j + S_j$.

In general, a d -districting of a graph G is ε -valid if it satisfies the following constraints:

1. *Contiguity*, i.e., the subgraph induced by D_j must be connected for each $j = 1, \dots, d$. Thus, each district is contiguous, which is typically required by law.
2. ε -*Population-balance*, i.e. $(1 - \varepsilon) \left(\frac{Pop(G)}{d} \right) \leq Pop(D_j) \leq (1 + \varepsilon) \left(\frac{Pop(G)}{d} \right)$ for each $j = 1, \dots, d$, which ensures that each district has approximately the same population, which is required under the ‘‘One Person, One Vote’’ rule [28]. We only consider $\varepsilon < \frac{1}{6}$, as $\varepsilon \geq \frac{1}{6}$ allows one district to have *double* the population of another district. Different jurisdictions may require different values of ε .

3.2.1 Competitiveness

We consider two notions that aim to capture the competitiveness of a given district.

First, we consider the intuitive notion that a district is competitive if the most recent election was decided by a very close margin - for example, 51% of votes cast for Party A and 49% cast for Party B. Deford et al. [15] call this ‘‘vote-band’’ competitiveness, named for the ‘‘band’’ of outcomes (for example, 45-55%) that the vote share should fall into in order for the election to be considered competitive². This notion does *not* depend on counting swing voters separately, so, in this context, $s_i = 0$ for all cells c_i .

► **Definition 1.** A district D_j is δ -Vote-Band Competitive (δ -VBC) iff $\frac{A_j}{Pop_j}, \frac{B_j}{Pop_j} \in \left[\frac{1}{2} - \delta, \frac{1}{2} + \delta \right]$.

An alternative notion is that a district is competitive if the outcome depends on how the swing voters vote. In particular, elections can be both close and uncompetitive. For example, a district comprised of 40% Party A voters, 40% Party B voters, and 20% swing voters (who may vote either way) is likely to be highly competitive, whereas a district with 52% Party A voters, 47% Party B voters, and 1% swing voters is less likely to be competitive.

► **Definition 2.** A district D_j is Swing if $S_j \geq |A_j - B_j|$.

² Deford et al. consider both centering the vote band around 50% and centering it around the statewide or nationwide average. For simplicity, we only discuss the former.

This formulation more directly captures the notion that districts are competitive if there is a reasonable chance for either party to win. However, evaluation of this metric depends on having a reasonable estimate of the number of such voters; we discuss using statistical methods of Ecological Inference for this task in Section 5.1.2. In particular, while polling data is valuable for estimating voter transitions, we need this information at a precinct-by-precinct level to do redistricting analysis, and polls are mostly conducted at a national or statewide level.

This formulation also allows us to capture the fact that, due to partisan polarization, elections in the US are often determined by turnout, rather than vote-switching [19]. For example, we can count “unreliable partisan voters” (those who will either vote for Party A or stay home) as half of a “true” swing voter (those who may vote for Party A or Party B), as their decision will affect the eventual vote margin in their precinct, district, or state by half as much.

3.3 Maximizing Competitive Districts

► **Definition 3** (δ -VBC-Max). *Given a graph G with n cells c_1, \dots, c_n , $d \in [2, n - 1]$, $\varepsilon > 0$, and $\delta > 0$, the δ -VBC District Maximization Problem is to compute a ε -valid d -districting that maximizes the number of δ -VBC districts.*

► **Definition 4** (Swing-Max). *Given a graph G with n cells c_1, \dots, c_n , $d \in [2, n - 1]$ and $\varepsilon > 0$, the Swing District Maximization Problem is to compute a ε -valid d -districting that maximizes the number of swing districts.*

4 Hardness results

Our main results in this section are to show that the δ -VBC District Maximization Problem and the Swing District Maximization Problem are both NP-hard. On its own, deciding the existence of any ε -valid d -districting for a graph G is NP-hard, regardless of competitive districts. Therefore, we show that both δ -VBC-Max and Swing-Max are hard, even on instances where:

- the underlying graph G is a grid,
- the number of districts d is 2, and
- there exists a polynomial-time computable ε -valid d -districting.

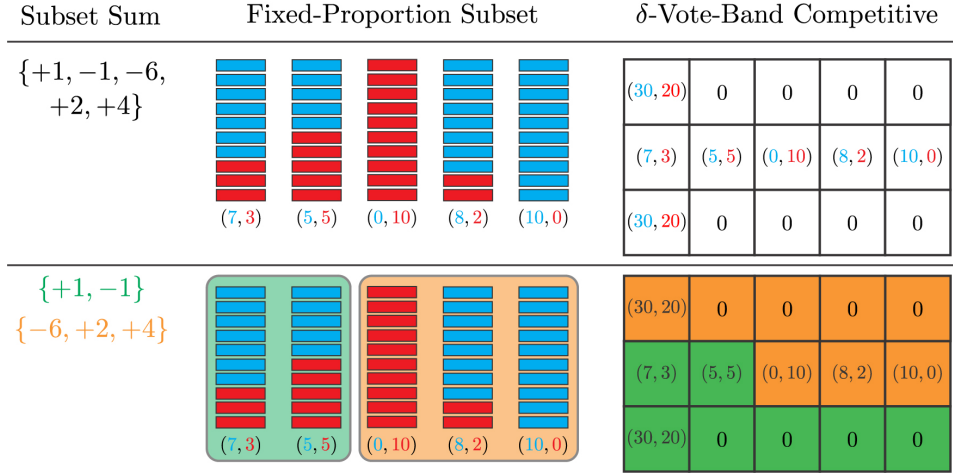
This setup is similar that of Kueng et al. [25], who show that drawing districtings where both parties have at least c seats is hard even on instances where balanced-population districting is easy.

► **Theorem 1** (Vote-Band Hardness). *For all positive $\varepsilon < \frac{1}{6}$, $\delta < \frac{1}{2}$, the δ -VBC ε -valid District Maximization Problem (δ -VBC-Max(ε)) is NP-hard, even on instances where G is a grid, $d = 2$, and ε -valid districtings (that are not δ -VBC-maximal) are poly-time computable.*

► **Theorem 2** (Swing Hardness). *For all positive $\varepsilon < \frac{1}{6}$, the Swing ε -valid District Maximization Problem (Swing-Max(ε)) is NP-hard, even on instances where G is a grid, $d = 2$, and ε -valid districtings (that are not Swing-maximal) are poly-time computable.³*

³ This constraint of $\varepsilon < \frac{1}{6}$ is extremely lenient: in practice, most districtings have population balance under 1%, and $\varepsilon = \frac{1}{6}$ would allow one district to have *double* the population of another - which is certainly illegal.

7:6 Drawing Competitive Districts in Redistricting



■ **Figure 1** The chain of reductions from Subset Sum to FPSP to δ -VBC-Max, for $\delta = 0.1, \varepsilon = \frac{1}{6}$. Given the Subset Sum instance T (left), we construct an instance of FPSP (center) where bin i has $6 + t_i$ type-A units and $4 - t_i$ type-B units. This induces a districting instance (right), where the central row has Party A and Party B voters corresponding to type-A and type-B units. The green and orange districts (where each district has population in $(\frac{1}{2} \pm \varepsilon)pop_{total}$ and exactly 60% Party A voters) correspond to a FPSP partition (where each partition has exactly 60% type-A units), which in turn corresponds to a solution to the original Subset Sum instance.

We will prove both of these via a reduction from Subset Sum [22]. Specifically, we will reduce through an intermediate problem, the Fixed-Proportion Subset Problem.

4.1 The Fixed-Proportion Subset Problem

► **Definition 5** (FPSP(δ)). For any fixed positive $\delta < \frac{1}{2}$, the Fixed-Proportion Subset Problem (FPSP) takes a list of tuples $S = [(a_1, b_1) \cdots (a_n, b_n)]$, where:

$$\sum_{i=1}^n a_i = \left(\frac{1}{2} + \delta\right) \sum_{i=1}^n (a_i + b_i) \quad \text{and} \quad a_1 + b_1 = a_2 + b_2 = \cdots = a_n + b_n$$

The task is to find nonempty proper subsets $S_1, S_2 \subsetneq [n]$ that partition $[n]$, such that

$$\sum_{i \in S_1} a_i = \left(\frac{1}{2} + \delta\right) \sum_{i \in S_1} (a_i + b_i) \quad \text{and} \quad \sum_{i \in S_2} a_i = \left(\frac{1}{2} + \delta\right) \sum_{i \in S_2} (a_i + b_i).^4$$

Intuitively, the Fixed-Proportion Subset Problem asks to partition a population of units with $(\frac{1}{2} + \delta)$ type- a units and $(\frac{1}{2} - \delta)$ type- b units, which are grouped into equal-size bins, into two subsets, where each subset is also exactly $(\frac{1}{2} + \delta)$ type- a and $(\frac{1}{2} - \delta)$ type- b .

► **Theorem 3** (FPSP Hardness). The Fixed-Proportion Subset Problem is NP-hard for all $\delta \in [0, \frac{1}{2})$.

Proof. Fix any $\delta \in [0, \frac{1}{2})$. Given an instance of Subset Sum $T = [t_1, \cdots, t_{n-1}]$ with desired sum 0, let $T' = [t_1, \cdots, t_{n-1}, t_n]$, where $t_n = -\sum_{i=1}^{n-1} t_i$, so that $\sum T' = 0$.

⁴ The existence of any S_1 satisfying the first condition is sufficient to have $S_2 = [n] \setminus S_1$ satisfy the second condition, but we explicitly write both for clarity.

We construct an instance of FPSP(δ) as follows. Let $c = -\frac{\min t_i}{\frac{1}{2} + \delta}$. Our instance of FPSP will be

$$S = [(a_1, b_1) \cdots (a_{n+1}, b_{n+1})] \text{ where } a_i = t_i + \left(\frac{1}{2} + \delta\right) c \text{ and } b_i = c - a_i .$$

Each “bin” will have c units total, satisfying the second condition of FPSP. We can verify the first condition holds, using the fact that $\sum t_i = 0$:

$$\sum_{i=1}^n a_i = \sum_{i=1}^n \left(t_i + \left(\frac{1}{2} + \delta\right) c \right) = \sum_{i=1}^n t_i + \left(\frac{1}{2} + \delta\right) \sum_{i=1}^n c = 0 + \left(\frac{1}{2} + \delta\right) \sum_{i=1}^n (a_i + b_i) .$$

Now, suppose that $S_1, S_2 \subsetneq [n]$ is a solution to FPSP(δ) for this instance S . Without loss of generality, suppose $n \notin S_1$. We claim that $\{t_i : i \in S_1\}$ is a solution to the original Subset Sum instance T :

$$\sum_{i \in S_1} a_i = \left(\frac{1}{2} + \delta\right) \sum_{i \in S_1} (a_i + b_i) \implies \sum_{i \in S_1} \left(t_i + \left(\frac{1}{2} + \delta\right) c \right) = \left(\frac{1}{2} + \delta\right) \sum_{i \in S_1} c \implies \sum_{i \in S_1} t_i = 0$$

◀

4.2 Reduction from FPSP to Competitive Districts

We prove Theorem 1 via a reduction from FPSP(δ) to δ -VBC-Max.

Proof. Fix ε and δ . Let $S = [(a_1, b_1) \cdots (a_n, b_n)]$ be an instance of FPSP(δ), and let $Z = \sum_{i=1}^n (a_i + b_i)$. Our redistricting instance (“state”) for δ -VBC-Max(ε) will be a $3 \times n$ grid of cells $C_{i,j}$, where:

- The top left and bottom left cells $C \in \{C_{1,1}, C_{3,1}\}$ will each have

$$\text{pop}(C) = \frac{1}{2}P, \quad \text{Party}_A(C) = \left(\frac{1}{2} + \delta\right) \frac{1}{2}P, \quad \text{Party}_B(C) = \left(\frac{1}{2} - \delta\right) \frac{1}{2}P ,$$

where $P = \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} Z$.

- All other cells in the first and third row have zero population⁵: $\text{pop}(C_{1,\cdot}) = \text{pop}(C_{3,\cdot}) = 0$.
- For each cell $C_{2,i}$ in the second row (for $i = 1, 2, \dots, n$),

$$\text{pop}(C_{2,i}) = (a_i + b_i) \quad \text{Party}_A(C_{2,i}) = a_i \quad \text{Party}_B(C_{2,i}) = b_i .$$

The total population of the state is $\text{pop}_{total} = Z + P = \frac{\frac{1}{2} - \varepsilon + \frac{1}{2} + \varepsilon}{\left(\frac{1}{2} - \varepsilon\right)} Z = \frac{1}{\frac{1}{2} - \varepsilon} Z$.

Observation 0. Since all cells $C_{2,1 \dots n}$ have the same population, it is trivial to draw an ε -valid districting: one can assign the first row and half of the second row to D_1 , and the rest to D_2 . That is, one can district this instance in polynomial time, if competitiveness is not considered.

⁵ One can easily modify the reduction to have all cells have nonzero population by multiplying all other cells by $2(n+2)+1$ and having the first- and third-row cells have $\text{pop}(C) = 2, \text{Party}_A(C) = 1, \text{Party}_B(C) = 1$.

7:8 Drawing Competitive Districts in Redistricting

Observation 1. The two left “corner” cells $C_{1,1}$ and $C_{3,1}$ must be assigned to different districts. If they were assigned the same district D_i (along with at least one center-row cell c_j , required for connectivity), that district’s population would exceed $(\frac{1}{2} + \varepsilon) \text{pop}_{total}$:

$$\text{pop}(C_{1,1}) + \text{pop}(C_{3,1}) + \text{pop}(C_{2,j}) = P + (a_j + b_j) > P = \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} Z = \left(\frac{1}{2} + \varepsilon\right) \text{pop}_{total}.$$

Observation 2. Each district must contain at least one of the center-row cells $C_{2,i}$; a “corner” cell alone is under the population bounds.

Besides these two cells, the structure of the state allows all other second-row cells to be assigned to either district while respecting contiguity. For example, one can assign all cells in the first row to district 1 and all cells in the third row to district 2.

Observation 3. $\frac{1}{2} + \delta$ of the overall population is Party A voters:

$$\text{total Party}_A = \sum_{i=1}^n a_i + \left(\frac{1}{2} + \delta\right) \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} Z = \left(\frac{1}{2} + \delta\right) Z + \left(\frac{1}{2} + \delta\right) \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} Z = \left(\frac{1}{2} + \delta\right) \text{pop}_{total},$$

where the last equality holds because $\text{pop}_{total} = \frac{1}{\frac{1}{2} - \varepsilon} Z$.

Let $\{D_1, D_2\}$ be a ε -valid δ -Vote-Band-Competitive districting on this state. We claim that $S_1 = \{i : C_{2,i} \in D_1\}, S_2 = \{i : C_{2,i} \in D_2\}$ is a valid solution to the FPSP(ε) instance.

For S_1, S_2 to be a solution to FPSP(ε), we must show (a) both are nonempty (which follows from Observation 2 above), and (b) $\sum_{i \in S_1} a_i = (\frac{1}{2} + \delta) \sum_{i \in S_1} (a_i + b_i)$.

Since both districts’ margins fall in the $\frac{1}{2} \pm \delta$ vote band, each one must have *exactly* $\frac{1}{2} + \delta$ Party A voters; if (for example) D_1 had $< \frac{1}{2} + \delta$ Party A voters, then D_2 would end up with $> \frac{1}{2} + \delta$ Party A voters, falling outside the vote band. So,

$$\frac{1}{2} + \delta = \frac{\text{Party}_A(D_1)}{\text{Party}_A(D_1) + \text{Party}_B(D_1)} = \frac{(\frac{1}{2} + \delta) \frac{1}{2} P + \sum_{i \in S_1} a_i}{\frac{1}{2} P + \sum_{i \in S_1} (a_i + b_i)}.$$

Solving this equation for $\sum_{i \in S_1} a_i$, yields $\sum_{i \in S_1} a_i = (\frac{1}{2} + \delta) \sum_{i \in S_1} (a_i + b_i)$ as required. Thus, S_1, S_2 are a valid solution to the FPSP instance. ◀

The chain of reductions, using the Subset Sum instance $T = \{1, -1, -6, 2, 4\}$ as an example, is shown in Fig. 1.

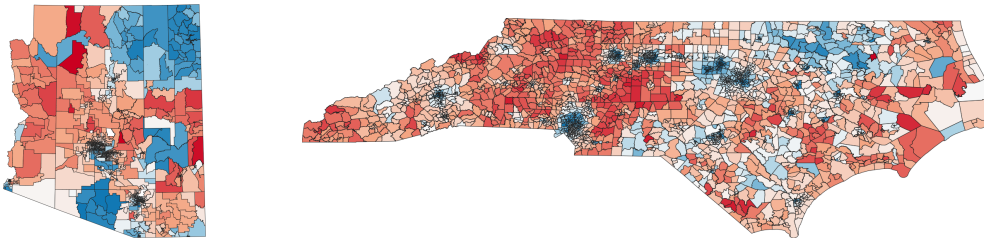
The proof for Swing district maximization is nearly identical. We present it in Appendix A.1.

► **Corollary 4** (Hardness for more than two districts). *For any k, d with $2 \leq k \leq d$, it is NP-hard to generate an ε -valid d -districting plan where at least k districts are competitive (for either δ -VBC or SWING).*

The proof for Corollary 4 is given in Appendix A.2.

4.3 Algorithms for Special Cases

In Appendix B, we present some special cases of graphs which admit polynomial time algorithms. Although these cases are somewhat artificial if expressed as standard redistricting instances, they may be of interest to those interested in districting-flavored problems on constrained structures, such as road networks or low-population grids.



■ **Figure 2** Precinct-level voting data for Arizona and North Carolina. Nodes are colored according to the margin in the 2020 presidential election.

4.4 Discussion of Complexity Results

First, we re-emphasize that we have shown that the hardness is a result of the competitive district maximization, not the intrinsic hardness of the balanced partition problem, because the instances we construct can be partitioned to population balance $\varepsilon = 0$ trivially.

There are reasons to see this reduction as somewhat more “natural” than previous hardness results, which often construct instances with strange features (for example, [25] constructs instances with that have long, tendril-like geography with many holes; [29] deals with a setting without geographic contiguity requirements; [24] creates instances where half of the nodes are connected to only one other node (i.e., there are many “donut” precincts, where one precinct entirely surrounds another)). In contrast, we construct instances where the underlying graph is a rectangular grid; many states (especially in the Midwest) have precinct graphs that have this approximate structure.

On the other hand, our reduction relies heavily on the hard cutoff at the edge of a desired “vote band” – that is, the hardness comes from distinguishing, in a binary way, between a district that is 59.99% Party A and one that is 60.01% Party A. In practice, the difference is likely to be negligible. Therefore, one may hope that we can, in practice, draw maps that significantly increase the number of competitive districts. We explain our approach to doing so in the subsequent section.

5 Algorithms, Heuristics, and Experiment Setup

Given the hardness of drawing competitive districting plans (as measured both in vote-band and swing-voter flavors), even on fairly constrained and natural varieties of graphs, one may wonder whether it is tractable to draw such districtings on real-world graphs. In this section, we argue that the answer is a definitive *yes*. In particular, we find that very simple hill-climbing procedure can yield reasonable districting plans that are highly competitive.

We propose a heuristic hill-climbing procedure based on making local moves called “single node flips”, and run several experiments on the U.S. states of North Carolina and Arizona: one set seeking to maximize δ -Vote-Band-Competitive districts for $\delta = 0.1, 0.05$ (corresponding to the thresholds used by the Center for Voting and Democracy for “landslides” (margin difference above 20%) and “competitive” (margin under 10%) [4]), and one seeking to maximize swing districts. In all cases, we fixed the allowable population deviation at $\varepsilon = 5\%$ and required all districts to be contiguous.

5.1 Data

We used precinct-level shapefiles and election data of the U.S. states of Arizona and North Carolina, two medium-sized states that have been highly competitive in recent elections. The

7:10 Drawing Competitive Districts in Redistricting

geographic data was collected and processed by the Metric Geometry and Gerrymandering Group [2], and the election data was obtained from the Redistricting Data Hub [3]. Specifically, the data includes the area, perimeter, and population of each precinct; voting history for the last several elections, and adjacency information. North Carolina has 2,650 precincts and fourteen Congressional districts. Arizona has 1489 precincts and nine Congressional districts. We chose these two states due to their extremely close statewide margins in recent elections, and their moderate size (for reasonable computational burden). They are shown in Figure 2.

5.1.1 Previous Election Votes for Vote-Band Competitiveness

Evaluating the vote-band competitiveness of a given districting plan depends on the votes cast in a past election. For these experiments, we used the results of the 2020 Presidential election, shown in Fig. 2.

5.1.2 Ecological Inference for Estimating Swing Voters

Evaluating whether a given district is swing or not depends on having estimates for the number of reliable voters of each party, and the number of swing voters. The task of estimating voter transitions from election data is a well-studied problem in Ecological Inference (EI). Given only top-level voting information (i.e., the number of votes cast for each candidate in two subsequent elections), the task of finding the number of voters who switched their vote from one election to the next (or who voted in one election and not the other) is highly underdetermined. As a result, EI techniques have no worst-case guarantees; however, they have been shown to perform well in practice. For an overview, see [23].

The task is made easier by the fact that we have the marginals (i.e., total votes cast) for each precinct in the state. `nsIphom` is a multi-iteration Linear Programming technique that takes advantage of this fact, using statewide homogeneity assumptions, developed by Pavia et al. [27]. It is available as an R package, which we used to estimate the inner cell values of the 3×3 tables of the form shown in Fig. 3 (one table per precinct). For example, 71.30 is the estimate for the number of voters who voted for the Democratic candidate in 2012 but did not vote in 2016.

		2016 votes			Total
		Democratic	Republican	Nonvote/Other	
2012 votes	Democratic	407.45	9.24	71.30	488
	Republican	3.55	1583.69	73.76	1660
	Nonvote/Other	0.00	272.07	2713.93	2986
Total		410	1864	2859	

■ **Figure 3** The EI table for votes cast in Alamance County, North Carolina, Precinct 01. The inner cell values are estimated using the R package `nsIphom`, using the known marginal values. In this case, the estimated number of swing voters is $3.55 + 9.24 + \frac{1}{2}(71.30 + 73.76 + 0.0 + 272.07) = 221.36$.

We use the election results of the 2012 and 2016 elections to estimate swing voter counts and the results of the 2020 election to evaluate the final redistricting maps in order to keep estimation and evaluation metrics separate. For each precinct, we compute the final estimate for “swing voters” by summing the off-diagonal elements of the EI table, with the entries in a “Nonvote/Other” row or column halved. Intuitively, compared to a voter who switches their vote from Party A to Party B, a voter who merely goes from abstaining to voting (or from voting to abstaining) only changes the top-level margin by half as much.

5.2 Heuristic-based optimization procedure

In order to find districtings that maximize the number of competitive districts, we use a simple randomized greedy hill-climbing procedure based on repeatedly making “single node flips” (also called spin flips) that incrementally improve the plan when measured against some objective(s). These methods are extensively used in ensemble-based analyses of redistricting; see, for example, [13, 12].⁶

A *single-node-flip* is a step that flips a node that is on the boundary of two districts from one district to another, subject to some constraints. Specifically, a node $u \in D_j$ is eligible to be flipped to district D_i (where $i \neq j$) if (a) u is adjacent to some $v \in D_i$, (b) $D_j \setminus \{u\}$ remains connected and within population bounds; and (c) $D_i \cup \{u\}$ is within population bounds.

In each step, we randomly choose a single-node-flip (u, D_1, D_2) from the set of all valid flips, proportional to a score function J_i , which quantifies the “desirability” of the resulting districts $(D_1 \setminus \{u\}, D_2 \cup \{u\})$, compared to the original districts (D_1, D_2) , for particular objectives of interest:

$$J(u, D_1, D_2) = \exp \left(\sum_{\text{objectives } i} -w_i \left(J_i(D_1 \setminus \{u\}) + J_i(D_2 \cup \{u\}) - (J_i(D_1) + J_i(D_2)) \right) \right).$$

Here, low values for J_i are preferred; this weight function weights a flip more heavily the more it decreases the J_i values.⁷ In this work, our weight includes two score functions: compactness and competitiveness (via swing districts or δ -VBC districts).

We employ the widely-used measure of compactness, the *isoperimetric ratio* [8], defined as the perimeter squared divided by area. The more compact a district, the lower its isoperimetric ratio, and we define $J_{iso}(D) = \frac{\text{perimeter}(D)^2}{\text{area}(D)}$.

The vote-band competitiveness term in the score function prioritizes districts that are close to even (i.e., 50-50 margin), and further prioritizes districts where the margin is in the range $\frac{1}{2} \pm \delta$:

$$J_{VBC}(D) = \begin{cases} \left(\frac{\text{PartyA}(D)}{\text{Pop}(D)} - \frac{1}{2} \right)^2 & \text{if } \frac{\text{PartyA}(D)}{\text{Pop}(D)} \notin \frac{1}{2} \pm \delta, \\ \frac{1}{16} \left(\frac{\text{PartyA}(D)}{\text{Pop}(D)} - \frac{1}{2} \right)^2 & \text{if } \frac{\text{PartyA}(D)}{\text{Pop}(D)} \in \frac{1}{2} \pm \delta. \end{cases}$$

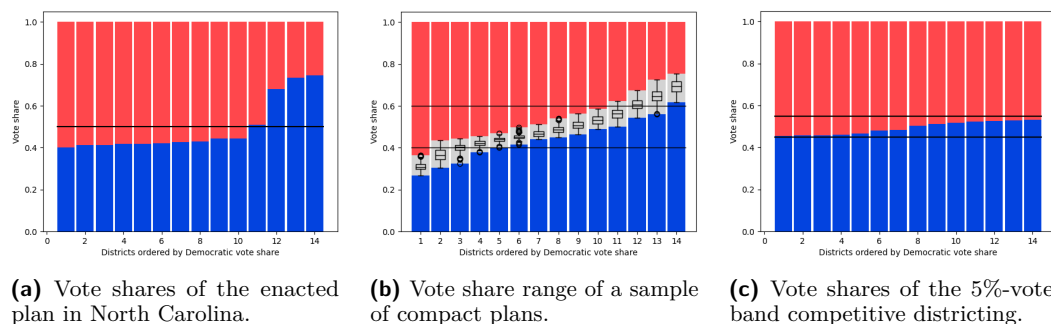
The swing term in the score function prioritizes having districts where the number of Party A voters plus half the swing voters are close to half of the total population, and further prioritizes districts where neither party’s reliable voters comprise more than half of the overall population:

$$J_{sw}(D) = \begin{cases} \left(\frac{1}{2} - \left(\frac{1}{2} \frac{\text{Swing}(D)}{\text{Pop}(D)} + \frac{\text{PartyA}(D)}{\text{Pop}(D)} \right) \right)^2 & \text{if } \frac{\text{PartyA}(D)}{\text{Pop}(D)} > \frac{1}{2} \text{ or } \frac{\text{PartyB}(D)}{\text{Pop}(D)} > \frac{1}{2}, \\ \left(\frac{1}{2} - \left(\frac{1}{2} \frac{\text{Swing}(D)}{\text{Pop}(D)} + \frac{\text{PartyA}(D)}{\text{Pop}(D)} \right) \right)^2 \cdot 0.8^2 & \text{if } \frac{\text{PartyA}(D)}{\text{Pop}(D)} < \frac{1}{2} \text{ and } \frac{\text{PartyB}(D)}{\text{Pop}(D)} < \frac{1}{2}. \end{cases}$$

⁶ However, we are *not* doing an ensemble analysis; we will not attempt to sample from a measure or use the Metropolis-Hastings algorithm, as these works do. Rather, we are simply investigating the degree to which a direct optimization can achieve competitive districtings.

⁷ This type of score function can be considered to be a “tempered choice” according to the measure $\pi(D) = \exp \left(\sum_i -w_i J(D_i) \right)$. This form of measure is commonly used in ensemble methods, e.g., in [7, 6].

7:12 Drawing Competitive Districts in Redistricting



■ **Figure 4** Explicitly considering competitiveness allows us to draw plans with significantly fewer safe districts than both the enacted plan and a sample of compact plans.

5.3 Run Parameters

We ran the hill-climbing procedure for 36,000 steps, restarting from a random initial state every 3,000 steps and taking the most compact plan with the maximal number of competitive districts. We used weights of $w_{iso} = 3$, $w_{VBC} = 10^5$, for VBC runs and $w_{iso} = 3$, $w_{sw} = 10^5$ for the Swing runs. We implemented the procedure in Python using the `gerrychain` package [1]. With unoptimized, single-threaded Python code, the procedure takes about three hours to run on Intel Xeon Gold 6226 2.9 Ghz machines.

As a basis for comparison, we also ran the hill-climbing procedure for 40,000 steps, restarting every 200 steps, with $w_{iso} = 3$, $w_{VBC} = 0$, $w_{sw} = 0$, i.e., prioritizing only compactness. We logged the most-compact plan found in each 200-step interval. We will refer to this set of plans as the *compact sample* below.

6 Experimental Results

We find that the heuristics are extremely effective for constructing districting plans with a significant proportion of swing districts. In fact, we obtain districtings on North Carolina and Arizona where *every single* district is competitive (for 10%-vote-band, 5%-vote-band, and swing metrics).

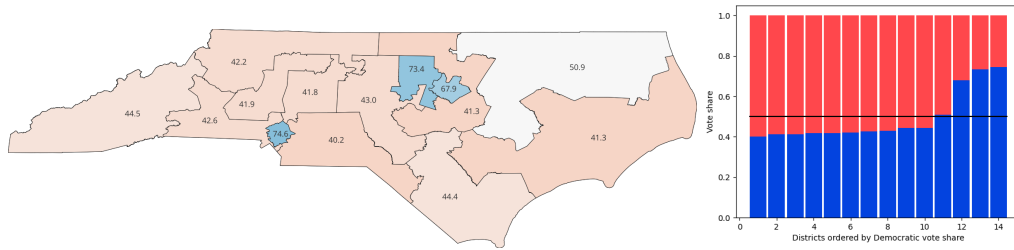
In Fig. 5, we display the most compact plan for each of $\delta = 0.1, 0.05$, and Swing, alongside the current enacted plan, for North Carolina. We also display the vote share of each district: that is, the percentage of Democratic and Republican (and Swing, for the last plot) voters in each district, based on votes cast in the 2020 presidential election. In Fig. 4, we compare the vote share distribution of the 5%-VBC plan to that of the enacted plan and of the compact sample.

We display the same results for Arizona in Fig. 6.

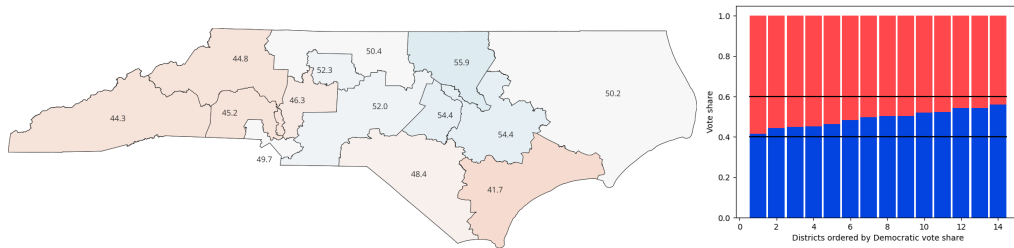
Notably, the VBC plans exhibit significantly improved responsiveness: the median district is extremely competitive (49.9% D-voting) and a small statewide swing in vote share would correspond to a larger change in number of seats won.

7 Discussion

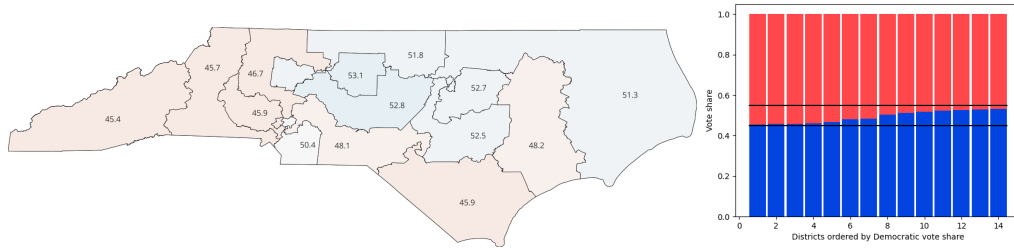
We explicitly do *not* present these plans as examples of *ideal* districting plans. Rather, we present them in contrast to the hardness results presented above, and to explore the consequences of fully prioritizing competitiveness as an objective. Although both swing district and vote-band-competitive district maximization is NP-hard (even if only seeking to make two out of d districts competitive), we explicitly show that it is tractable in practice,



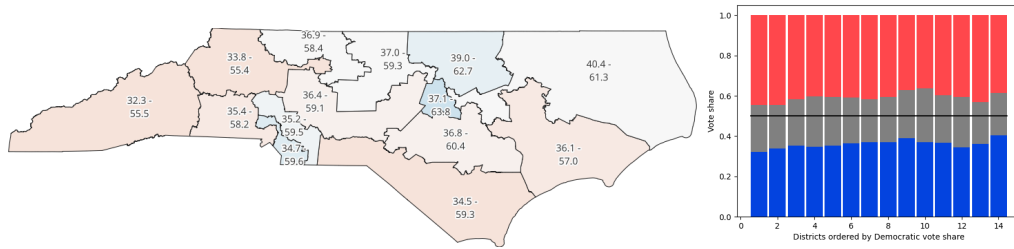
(a) The currently-enacted plan. Despite winning just over half of the two-party vote, Trump would have carried ten of the fourteen Congressional districts.



(b) In this plan, all districts are δ -Vote-Band competitive for $\delta = 10\%$. That is, all districts have a Biden vote share between 40% and 60%. Note that Biden and Trump would have each carried seven of the fourteen districts: a result that is significantly more reflective of the fact that they won nearly the same number of votes.



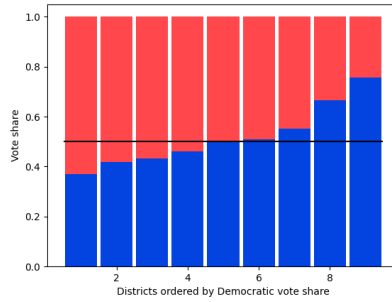
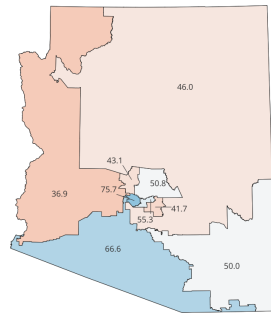
(c) In this plan, all districts are δ -Vote-Band competitive for $\delta = 5\%$: all districts have a Biden vote share between 45% and 55%. Again, Biden and Trump would have each carried seven districts. However, the districts are noticeably less compact, with significantly contorted boundary shapes; the heavily Democratic areas of Charlotte and Raleigh-Durham are visibly “cracked” among many districts.



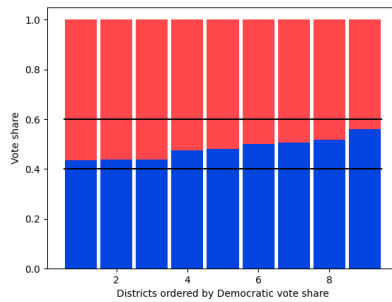
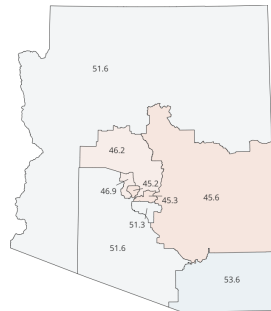
(d) In this plan, all districts are swing. The range of outcomes (ranging from all swing voters voting for Trump to all swing voters voting for Biden) is shown for each district.

■ **Figure 5** Our simple hill-climbing procedure successfully finds plans for North Carolina where all fourteen districts are competitive.

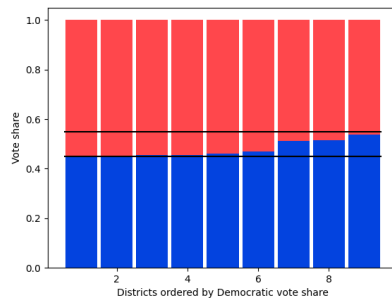
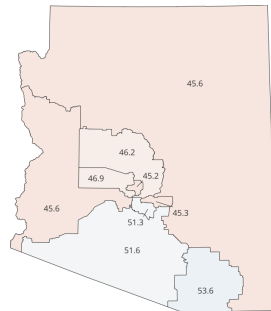
7:14 Drawing Competitive Districts in Redistricting



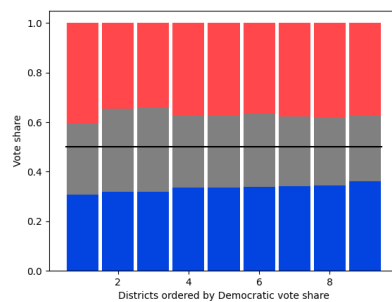
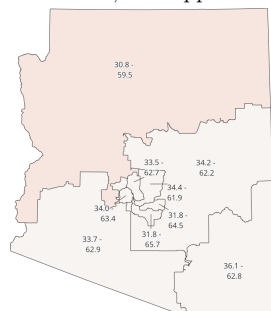
(a) The current enacted plan.



(b) A 0.1-Vote-Band-Competitive districting: all districts have a Biden vote share between 40% and 60%.



(c) A 0.05-Vote-Band-Competitive districting: all districts have a Biden vote share between 45% and 55%. Unlike in North Carolina, this appears to be achievable without significant loss of compactness.



(d) In this plan, all districts are swing. The range of outcomes (ranging from all swing voters voting for Trump to all swing voters voting for Biden) is shown for each district.

■ **Figure 6** Our hill-climbing procedure successfully finds plans for Arizona where all nine districts are competitive. Unlike in North Carolina, there is no significant sacrifice in compactness to achieve a $\delta = 5\%$ -vote-band-competitive districting.

even achieving plans that make all d districts competitive. Thus, the point is to dissuade the reader from drawing the conclusion that “because drawing swing districts is NP-hard, attempting to do so is a lost cause”, or that “policymakers are exempt from drawing competitive districts because of computational intractability.” Instead, drawing competitive districts is very tractable in practice.

Notably, whereas the enacted plan “packs” the heavily-Democratic areas of Charlotte and Raleigh-Durham into as few districts as possible, maximizing competitive districts entails “cracking” those areas up into multiple districts. We observe this among all plans that achieved 14 competitive districts; in some sense, this is likely unavoidable given the political geography of the state (in which Democratic voters are heavily concentrated in small geographic areas).

Indeed, the *ideal* number of competitive districts is almost certainly not the *maximal* number. We leave the question of exactly *how many* districts should be competitive as a normative question (for discussion of the relationship between proportionality and competitive districts, see [18]); we simply present the result that it is tractable to achieve anywhere from 0 to d competitive districts on some real-world graphs.

In particular, while having *every* district in a state be competitive makes the makeup of the state’s Congressional delegation highly responsive (that is, small changes in vote share can result in large changes in topline number-of-seats-won), taking this to an extreme (for example, by enacting the plan in Fig. 5(c)) can result in a sharp decline in the proportionality of the results. For example, under the plan in Fig. 5(c), if one party won 55% of the votes statewide, they would sweep *all* of the congressional districts - certainly sending an unrepresentative delegation to represent the state.

We do not expect all states to admit fully-competitive districtings; indeed, states with remarkably homogeneously distributed electorates (such as Massachusetts) have been observed to be impossible to draw competitive districts on [16]. On the other extreme, some states may have voters that are so geographically polarized that drawing competitive districts may require splitting urban areas into unacceptably many districts. For example, Democratic voters in Pennsylvania and Illinois are so heavily concentrated in cities like Philadelphia and Chicago that maximizing competitive districts likely involves splitting them into over a dozen districts - likely an unacceptable result. We leave detailed investigation of these cases for future work.

8 Conclusion

We observe a very large gap between the theoretical intractability of drawing competitive districts (even on fairly natural instances) and the high performance of empirical heuristics on real instances. This is consistent with the literature on optimizing various metrics in redistricting, as well as with the fact that population-balanced districting itself is clearly achievable in reality while being complexity-wise infeasible in the worst case. We attribute the tractability on real instances to the fact that “close to optimal” is an acceptable substitute for “truly optimal” in the context of elections, where a large amount of variation and uncertainty is to be expected.

References

- 1 GerryChain — GerryChain documentation. URL: <https://gerrychain.readthedocs.io/en/latest/>.
- 2 MGGG States. URL: <https://github.com/mggg-states>.

- 3 Redistricting Data Hub. URL: <https://redistrictingdatahub.org/>.
- 4 Dubious Democracy, 2003.
- 5 Konstantin Andreev and Harald Räcke. Balanced graph partitioning. In *Proceedings of the Sixteenth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '04, page 120–124, New York, NY, USA, 2004. Association for Computing Machinery. doi:10.1145/1007912.1007931.
- 6 Eric A. Autry, Daniel Carter, Gregory Herschlag, Zach Hunter, and Jonathan C. Mattingly. Multi-Scale Merge-Split Markov Chain Monte Carlo for Redistricting, August 2020. arXiv:2008.08054.
- 7 Eric A. Autry, Daniel Carter, Gregory J. Herschlag, Zach Hunter, and Jonathan C. Mattingly. Metropolized Multiscale Forest Recombination for Redistricting. *Multiscale Modeling & Simulation*, 19(4):1885–1914, January 2021. doi:10.1137/21M1406854.
- 8 Sachet Bangia, Christy Vaughn Graves, Gregory Herschlag, Han Sung Kang, Justin Luo, Jonathan C. Mattingly, and Robert Ravier. Redistricting: Drawing the Line, May 2017. arXiv:1704.03360.
- 9 Marion Campisi, Andrea Padilla, Thomas Ratliff, and Ellen Veomett. Declination as a Metric to Detect Partisan Gerrymandering. *Election Law Journal: Rules, Politics, and Policy*, 18(4):371–387, December 2019. doi:10.1089/e1j.2019.0562.
- 10 Marion Campisi, Thomas Ratliff, Stephanie Somersille, and Ellen Veomett. Geography and Election Outcome Metric: An Introduction. *Election Law Journal: Rules, Politics, and Policy*, 21(3):200–219, September 2022. doi:10.1089/e1j.2021.0054.
- 11 Tanima Chatterjee, Bhaskar DasGupta, Laura Palmieri, Zainab Al-Qurashi, and Anastasios Sidiropoulos. Alleviating partisan gerrymandering: Can math and computers help to eliminate wasted votes?, April 2018. doi:10.48550/arXiv.1804.10577.
- 12 Maria Chikina, Alan Frieze, and Jonathan C. Mattingly. Separating Effect From Significance in Markov Chain Tests. URL: <https://www.tandfonline.com/doi/full/10.1080/2330443X.2020.1806763>.
- 13 Maria Chikina, Alan Frieze, and Wesley Pegden. Assessing significance in a Markov chain without mixing | PNAS. *PNAS*, 2017.
- 14 Jeanne Clelland, Haley Colgate, Daryl DeFord, Beth Malmskog, and Flavia Sancier-Barbosa. Colorado in context: Congressional redistricting and competing fairness criteria in Colorado. *Journal of Computational Social Science*, 5(1):189–226, May 2022. doi:10.1007/s42001-021-00119-7.
- 15 Daryl DeFord, Moon Duchin, and Justin Solomon. A Computational Approach to Measuring Vote Elasticity and Competitiveness, September 2020. URL: <https://www.tandfonline.com/doi/full/10.1080/2330443X.2020.1777915>.
- 16 Moon Duchin, Taissa Gladkova, Eugene Henninger-Voss, Ben Klingensmith, Heather Newman, and Hannah Wheelen. Locating the Representational Baseline: Republicans in Massachusetts. *Election Law Journal: Rules, Politics, and Policy*, 2018.
- 17 Moon Duchin, Tom Needham, and Thomas Weighill. The (homological) persistence of gerrymandering, July 2020. doi:10.48550/arXiv.2007.02390.
- 18 Moon Duchin and Gabe Schoenbach. Redistricting for Proportionality. *The Forum*, 20(3-4):371–393, December 2022. doi:10.1515/for-2022-2064.
- 19 Thomas G. Hansford and Brad T. Gomez. Estimating the Electoral Effects of Voter Turnout. *American Political Science Review*, 104(2):268–288, May 2010. doi:10.1017/S0003055410000109.
- 20 John A. Henderson, Brian T. Hamel, and Aaron M. Goldzimer. Gerrymandering Incumbency: Does Nonpartisan Redistricting Increase Electoral Competition? *The Journal of Politics*, 80(3):1011–1016, July 2018. doi:10.1086/697120.
- 21 Takehiro Ito, Naoyuki Kamiyama, Yusuke Kobayashi, and Yoshio Okamoto. Algorithms for gerrymandering over graphs, May 2021. URL: <https://www.sciencedirect.com/science/article/pii/S0304397521001894>.

- 22 Jon Kleinberg and Eva Tardos. *Algorithm Design*. Pearson, 2006.
- 23 André Klima, Paul W. Thurner, Christoph Molnar, Thomas Schlesinger, and Helmut Küchenhoff. Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches. *AStA Advances in Statistical Analysis*, 100(2):133–159, April 2016. doi:10.1007/s10182-015-0254-8.
- 24 Shao-Heng Ko, Erin Taylor, Pankaj Agarwal, and Kamesh Munagala. All Politics is Local: Redistricting via Local Fairness. *Advances in Neural Information Processing Systems*, 35:17443–17455, December 2022.
- 25 Richard Kueng, Dustin G. Mixon, and Soledad Villar. Fair redistricting is hard. *Theoretical Computer Science*, 791:28–35, October 2019. doi:10.1016/j.tcs.2019.04.004.
- 26 John F. Nagle. What Criteria Should Be Used for Redistricting Reform? *Election Law Journal: Rules, Politics, and Policy*, 18(1):63–77, March 2019. doi:10.1089/e1j.2018.0514.
- 27 Jose M. Pavia and Rafael Romero. Improving Estimates Accuracy of Voter Transitions. Two New Algorithms for Ecological Inference Based on Linear Programming. *Sociological Methods & Research*, page 00491241221092725, May 2022. doi:10.1177/00491241221092725.
- 28 J. Douglas Smith. *On Democracy's Doorstep: The Inside Story of How the Supreme Court Brought "One Person, One Vote" to the United States*. Farrar, Straus and Giroux, June 2014.
- 29 Ana-Andreea Stoica, Abhijnan Chakraborty, Palash Dey, and Krishna P. Gummadi. Minimizing Margin of Victory for Fair Political and Educational Districting, September 2019. doi:10.48550/arXiv.1909.05583.
- 30 Samuel S.-H. Wang. Three Tests for Practical Evaluation of Partisan Gerrymandering. *Stanford Law Review*, 68(6):1263–1322, 2016.
- 31 Gregory S. Warrington. A Comparison of Partisan-Gerrymandering Measures. *Election Law Journal: Rules, Politics, and Policy*, 18(3):262–281, September 2019. doi:10.1089/e1j.2018.0508.
- 32 David Wasserman. Realignment, More Than Redistricting, Has Decimated Swing House Seats, April 2023. URL: <https://www.cookpolitical.com/cook-pvi/realignment-more-redistricting-has-decimated-swing-house-seats>.

A Further Proofs

A.1 Hardness of Swing Competitiveness

In this section, we prove Theorem 2, which we restate here for convenience:

► **Theorem 2 (Swing Hardness).** *For all positive $\varepsilon < \frac{1}{6}$, the Swing ε -valid District Maximization Problem (Swing-Max(ε)) is NP-hard, even on instances where G is a grid, $d = 2$, and ε -valid districtings (that are not Swing-maximal) are poly-time computable.⁸*

The proof is nearly identical to the proof of Theorem 1.

Proof. Fix ε and let $\delta = 0$. Let $S = [(a_1, b_1) \cdots (a_n, b_n)]$ be an instance of FPSP(0), and let $Z = \sum_{i=1}^n (a_i + b_i)$. Our redistricting instance (“state”) for Swing-Max(ε) will be a $3 \times (n + 2)$ grid of cells $C_{i,j}$, where:

- All cells in the first and third row have zero population: $\text{pop}(C_{1,\cdot}) = \text{pop}(C_{3,\cdot}) = 0$.
- For second-row cells $C_{2,i}$ for $i = 1, 2, \dots, n$,

$$\text{pop}(C_{2,i}) = (a_i + b_i) \quad \text{Party}_A(C_{2,i}) = a_i \quad \text{Swing}(C_{2,i}) = b_i$$

⁸ This constraint of $\varepsilon < \frac{1}{6}$ is extremely lenient: in practice, most districtings have population balance under 1%, and $\varepsilon = \frac{1}{6}$ would allow one district to have *double* the population of another - which is certainly illegal.

7:18 Drawing Competitive Districts in Redistricting

- Finally, for second-row cells $C_{2,j}$ for $j = n + 1, n + 2$,

$$\text{pop}(C_{2,j}) = \frac{1}{2}P, \text{ where } P = \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} (1 + Z)$$

$$\text{Party}_A(C_{2,j}) = \frac{1}{4}P \quad \text{Swing}(C_{2,j}) = \frac{1}{4}P$$

The total population of the state is

$$\text{Pop}_{\text{total}} = 0 + Z + 2 \left(\frac{1}{2}P \right) = Z + P = \frac{\frac{1}{2} - \varepsilon + \frac{1}{2} + \varepsilon}{\left(\frac{1}{2} - \varepsilon\right)} Z + \frac{\frac{1}{2} + \varepsilon}{\left(\frac{1}{2} - \varepsilon\right)} = \frac{1}{\frac{1}{2} - \varepsilon} Z + \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon}$$

Let $\{D_1, D_2\}$ be a ε -valid Swing districting on this state.

The following observations still hold, from the proof of Theorem 1:

Observation 0. One can district this instance in polynomial time.

Observation 1. Cells $C_{2,n+1}$ and $C_{2,n+2}$ must be assigned to different districts. Without loss of generality, let $C_{2,n+1} \in D_1$.

Observation 2. Each district must contain at least one of $C_{2,i}$ for $i \in [n]$.

Claim: $S_1 = \{i : C_{2,i} \in D_1\}, S_2 = \{i : C_{2,i} \in D_2\}$ is a valid solution to the FPSP(ε) instance.

For S_1, S_2 to be a solution to FPSP(ε), we must show (a) both are nonempty (which follows from Observation 2 above), and (b) $\sum_{i \in S_1} a_i = \left(\frac{1}{2} + \delta\right) \sum_{i \in S_1} (a_i + b_i)$.

$\frac{1}{2}$ of the overall population is Party A voters:

$$\begin{aligned} \frac{\text{total Party}_A}{\text{total population}} &= \frac{\sum_{i=1}^n a_i + \left(\frac{1}{2}\right) \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} (1 + Z)}{\sum_{i=1}^n (a_i + b_i) + \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} (1 + Z)} \\ &= \frac{\left(\frac{1}{2}\right) \left(\sum_{i=1}^n (a_i + b_i) + \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} (1 + Z)\right)}{\sum_{i=1}^n (a_i + b_i) + \frac{\frac{1}{2} + \varepsilon}{\frac{1}{2} - \varepsilon} (1 + Z)} = \frac{1}{2} \end{aligned}$$

Since both districts' margins are exactly $\frac{1}{2}$, each one must have *exactly* $\frac{1}{2}$ Party A voters. So,

$$\begin{aligned} \frac{1}{2} &= \frac{\text{Party}_A(D_1)}{\text{Party}_A(D_1) + \text{Party}_B(D_1)} \\ &= \frac{\left(\frac{1}{2}\right) \frac{1}{2}P + \sum_{i \in S_1} a_i}{\frac{1}{2}P + \sum_{i \in S_1} (a_i + b_i)} \\ \left(\frac{1}{2}\right) \left(\frac{1}{2}P + \sum_{i \in S_1} (a_i + b_i)\right) &= \left(\frac{1}{2}\right) \frac{1}{2}P + \sum_{i \in S_1} a_i \\ \left(\frac{1}{2}\right) \frac{1}{2}P + \left(\frac{1}{2}\right) \sum_{i \in S_1} (a_i + b_i) &= \left(\frac{1}{2}\right) \frac{1}{2}P + \sum_{i \in S_1} a_i \\ \sum_{i \in S_1} a_i &= \left(\frac{1}{2}\right) \sum_{i \in S_1} (a_i + b_i) \end{aligned}$$

as required. Thus, S_1, S_2 are a valid solution to the FPSP instance. ◀

A.2 Hardness for an arbitrary number of districts

In this section, we prove Corollary 4:

► **Corollary 4** (Hardness for more than two districts). *For any k, d with $2 \leq k \leq d$, it is NP-hard to generate an ε -valid d -districting plan where at least k districts are competitive (for either δ -VBC or SWING).*

Proof. We reduce from an instance with $d = 2$. Fix d', k with $2 \leq k \leq d'$. We simply add $d' - 2$ cells to the instance. Each cell will have the maximal allowable district population. $k - 2$ of the districts will have half Party A voters and half Party B voters (which makes that singleton district competitive under either Swing or δ -VBC definitions), and the remainder will be entirely Party A voters (i.e., uncompetitive).

Any valid districting with k competitive districts must have the added cells as singleton districts, of which $k - 2$ will be competitive. The other two districts must be a competitive districting of the original instance. ◀

B Polynomial Algorithms and Approximations

In this section, we consider special cases of the Swing District Maximization problem in which either the graph G has a special structure (line, bounded degree tree, etc), the population constraint is relaxed, or the districts satisfy an additional structure.

B.1 Line

We start by considering the case where G is the line graph on n vertices c_1, \dots, c_n , such that there is an edge between c_i and c_{i+1} for $1 \leq i \leq n - 1$. Satisfying the connectivity constraint in this case is easy; simply make sure that every district has only consecutive cells c_j, c_{j+1}, \dots, c_i . A district is feasible if it's connected and satisfies the population constraints. We can find the optimal d -districting by solving a dynamic program. For fixed d , $i \in \{1, \dots, n\}$ and $k \leq d$, let $M(i, k)$ denote the maximum number of swing districts among all valid d -districtings of the subgraph induced by c_1, \dots, c_i . We show that

$$M(i, k) = \max_{\substack{1 \leq j < i \\ (c_{j+1}, \dots, c_i) \text{ feasible}}} \left\{ M(j, k - 1) + \mathbb{1}\{(c_{j+1}, \dots, c_i) \text{ is swing}\} \right\}, \quad (1)$$

and

$$\begin{aligned} M(i, 0) &= -\infty && \forall i \in \{1, \dots, n\} \\ M(i, 1) &= \mathbb{1}\{(c_1, \dots, c_i) \text{ is swing}\} && \text{if } (c_1, \dots, c_i) \text{ feasible, } \forall i \in \{1, \dots, n\} \\ M(i, 1) &= -\infty && \text{if } (c_1, \dots, c_i) \text{ is not feasible, } \forall i \in \{1, \dots, n\} \end{aligned}$$

Equation (1) holds since to get the maximum number of swing districts in a k -districting of c_1, \dots, c_i , we need to decide on the cell c_{j+1} that limits the last district from the left. Once the last district $\{(c_{j+1}, \dots, c_i)\}$ is fixed, we need to pick the remaining $k - 1$ districts from c_1, \dots, c_j . Checking if a subset of cells satisfies population constraints and induces a swing district can be done in $O(n)$ time. Therefore, we can fill the $n \times d$ entries of the matrix $M(i, k)$ top down from left to right. The optimal solution is stored in $M(n, d)$. We have the following lemma.

► **Lemma 6.** *If G is a line on n cells, we can compute the optimal d -districting in $O(n^2d)$ time.*

B.2 Bounded-Degree Trees with Districts of Bounded Depth

In this subsection, we consider bounded-degree trees with the additional assumptions that districts need to have bounded depth, that is, the distance between every two cells in the same district must be less than a parameter $d > 0$. Consider the graph G to be a tree with n vertices and a maximum degree Δ . Let ε be the population tolerance. We require that in a valid districting, all districts must run for a depth of at most d (i.e., the diameter of every district is less than d).

Let $\mathcal{D}(v, \varepsilon, d)$ be the set of ε -valid districts with depth at most d that are rooted at v .

▷ **Claim 7.**

$$|\mathcal{D}(v, \varepsilon, d)| \leq 2^{\Delta^d}$$

Proof. The number of vertices at a distance less or equal than d from the root v is less than Δ^d . For every one of these vertices and every $D \in \mathcal{D}(v, \varepsilon, d)$, D can either contain the vertex or not. ◁

When $d = O(1)$ and $\Delta = O(1)$, we propose a polynomial time dynamic program to solve the swing district maximization problem. For a fixed d and $k \leq d$, let $M(v, k)$ denote the maximum number of swing districts in a ε -valid k -districting for the subtree of G that is rooted at v . To get the optimal districting, we need to first fix the district that v will belong to in $\mathcal{D}(v, \varepsilon, d)$. Because of the depth constraint on districts, the number of possible choices is bounded by 2^{Δ^d} . After we fix the district of v , we need to choose the roots of the remaining $k - 1$ districts.

Let $G(v)$ be the subtree of G that is rooted at v . Let $D \in \mathcal{D}(v, \varepsilon, d)$, and let $\mathcal{R}(D)$ be the roots of the subtrees of $G(v) \setminus D$. If $|\mathcal{R}(D)| > k - 1$, then clearly we cannot assign all the remaining cells to districts. Therefore we need $0 < |\mathcal{R}(D)| \leq k - 1$. Furthermore, to compute the remaining $k - 1$ districts, we need to start from the roots in $\mathcal{R}(D)$ such that, every vertex in $\mathcal{R}(D)$ will give rise to at least one district. To decide how many district every tree rooted at a vertex in $\mathcal{R}(D)$ needs to have, we assign a number $\ell(u) \in \{1, \dots, k - 1\}$ for every $u \in \mathcal{R}(D)$, such that $\sum_{u \in \mathcal{R}(D)} \ell(u) = k - 1$, and the subtree rooted at u contains $\ell(u)$ out of the remaining $k - 1$ districts. This gives rise to the following dynamic program

$$M(v, k) = \max_{\substack{D \in \mathcal{D}(v, \varepsilon, d) \\ |\mathcal{R}(D)| \leq k-1 \\ \sum_{u \in \mathcal{R}(D)} \ell(u) = k-1}} \left\{ \mathbb{1}\{D \text{ is swing}\} + \sum_{u \in \mathcal{R}(D)} M(u, \ell(u)) \right\}$$

If r is the root of G , the $M(r, d)$ will contain the optimal number of swing districts. In order to get $M(r, d)$, we need to top-down fill $O(nk)$ entries of M . To fill an entry $M(v, k)$, we have to choose a district $D \in \mathcal{D}(v, \varepsilon, d)$ and an assignment $u \mapsto \ell(u)$ for $u \in \mathcal{R}(D)$ such that $\sum_{u \in \mathcal{R}(D)} \ell(u) = k - 1$.

▷ **Claim 8.** Once a district $D \in \mathcal{D}(v, \varepsilon, d)$ is fixed, the number of possible assignments $u \mapsto \ell(u)$ for $u \in \mathcal{R}(D)$ such that $\sum_{u \in \mathcal{R}(D)} \ell(u) = k - 1$ is less than $k^{2^{\Delta^d}}$.

Proof. Similarly to the proof of Claim 7, we can show that $|\mathcal{R}(D)| \leq 2^{\Delta^d}$. The number of positive assignments $u \mapsto \ell(u)$ such that $1 \leq \ell(u) \leq k - 1$ is less than $k^{2^{\Delta^d}}$. ◁

The combination of Claim 7 and 8 show that every entry $M(v, k)$ can be computed in $O(2^{\Delta^d} k^{2^{\Delta^d}})$ time given that we already know the previous entries. We therefore have the following lemma.

► **Lemma 9.** *If G is a tree with a maximum degree Δ , and districts can have at most a depth of d , we can compute the optimal d -districting in $O(n2^{\Delta^d} k^{1+2^{\Delta^d}})$ time.*

B.3 Convex Districtings of Grid Graphs

In this subsection, we assume that the graph G is an $m \times n$ grid. If we require all the districts to be x -convex, that is, if two cells c_1 and c_2 of the same row are assigned to the same district, then all the cells between c_1 and c_2 of that same row are also assigned to that district. This case encompasses a compactness constraint since convexity has been used in gerrymandering studies as a measure of compactness to examine how redistricting reshapes the geography of congressional districts [11].

► **Theorem 5.** *Let G be an $m \times n$ grid, and let $P = \text{Pop}(G)$ be the total population on P . There exists an algorithm for computing an x -convex valid d -districting of G with maximum swing districts, with running time $(Pm)^{O(d)}$. In particular, the running time is polynomial when the total population is polynomial and the total number of partitions is a constant.*

Proof. Let D_1, \dots, D_k be an x -convex d -districting of G . For any $i \in \{1, \dots, n\}$, let C_i be the i -th column of G . We observe that for all $i \in \{1, \dots, n\}$, and for all $j \in \{1, \dots, d\}$, we have that $D_j \cap C_i$ is either empty, or consists of a single rectangle of width 1. Let \mathcal{C}_i be the set of all *contiguous* partitions of C_i into exactly d (possibly empty) segments, each labeled with a unique integer in $\{1, \dots, d\}$. We further define

$$\begin{aligned}\alpha_{i,j} &= \text{PartyA}(D_j^* \cap (C_1 \cup \dots \cup C_i)) \\ \beta_{i,j} &= \text{PartyB}(D_j^* \cap (C_1 \cup \dots \cup C_i)) \\ \gamma_{i,j} &= \text{Swing}(D_j^* \cap (C_1 \cup \dots \cup C_i)),\end{aligned}$$

For each column $i \in [n]$, and each district $j \in [d]$, $\alpha_{i,j}$ (resp. $\beta_{i,j}$, $\gamma_{i,j}$) denote the number of Party A (resp. Party B, swing) voters in district D_j between column 1 and column i .

We can enumerate all the possible solutions starting from the first column and moving to the right as follows. For each $i \in \{1, \dots, n\}$, let $I_i = \mathbb{N}^{3d} \times \mathcal{C}_i \times [m]^d$. Let

$$X_i = (\alpha_{i,1}, \beta_{i,1}, \gamma_{i,1}, \dots, \alpha_{i,d}, \beta_{i,d}, \gamma_{i,d}, \mathcal{Z}_{i,i}) \in I_i,$$

where \mathcal{Z}_i is a d -partition of the column C_i and $i = \{F_{i,1}, \dots, F_{i,d}\}$ is a collection of the forbidden indices for every district in the next column C_{i+1} , to ensure the connectivity of the districts as well as the x -convexity constraints.

If $i = 1$, we say that X_i is feasible if $\mathcal{Z}_1 = \{\emptyset, \dots, \emptyset\}$ and, for all $j \in \{1, \dots, d\}$, the set $\mathcal{Z}_1 = \{Z_1, \dots, Z_d\}$ satisfies

$$\alpha_{i,j} = \text{PartyA}(Z_{i,j}), \quad \beta_{i,j} = \text{PartyB}(Z_{i,j}), \quad \text{and} \quad \gamma_{i,j} = \text{Swing}(Z_{i,j})$$

If $i > 1$, we say that X_i with $\mathcal{Z}_i = \{Z_{i,1}, \dots, Z_{i,d}\}$ is feasible if the following holds, there exists some $X_{i-1} = (\alpha_{i-1,1}, \beta_{i-1,1}, \gamma_{i-1,1}, \dots, \alpha_{i-1,d}, \beta_{i-1,d}, \gamma_{i-1,d}, Z_{i-1,i-1}) \in I_{i-1}$ such that

7:22 Drawing Competitive Districts in Redistricting

$$\alpha_{i,j} = \alpha_{i-1,j} + \text{PartyA}(Z_{i,j}), \quad \beta_{i,j} = \beta_{i-1,j} + \text{PartyB}(Z_{i,j}), \quad \text{and} \quad \gamma_{i,j} = \gamma_{i-1,j} + \text{Swing}(Z_{i,j}), \quad (2)$$

$$\mathcal{Z}_{i,j} \cap_{i-1,j} = \emptyset, \quad (3)$$

$$\text{If } Z_{i,j} \not\subset Z_{i-1,j} \text{ then } F_{i,j} = (Z_{i-1,j} \setminus Z_{i,j}) \cup_{i-1,j}, \text{ else } F_{i,j} =_{i-1,j} \quad (4)$$

The constraint (2) simply states that the voter populations in columns $1, \dots, i$ are equal to the the voter populations in columns $1, \dots, i-1$ plus the voters from partition \mathcal{Z}_i . Constraint (3) ensure that the partition of the i -th column has to respect x -convexity and not include any of the forbidden cells from $_{i-1}$. The last constraint (4) shows how to update the forbidden cells for the next column $i+1$. If for a district D_j , the indices of the rows added from column i to D_j is included in the set of rows of D_j from column $i-1$, then there is no need to add any other forbidden row for the next column. If however, $\mathcal{Z}_{i,j} \not\subset \mathcal{Z}_{i-1,j}$, that means that the column i does not “transfer” the rows of D_j from column $i-1$ to column i , then the “non-transferred” rows have to be forbidden in the next column.

For each $i \in \{1, \dots, n\}$ we inductively compute the set of all feasible $X_i \in I_i$. This can be done in time $(Pm)^{O(d)}$ where P is the total population of the map. ◀

Score Design for Multi-Criteria Incentivization

Anmol Kabra¹ ✉

Toyota Technological Institute at Chicago, IL, USA

Mina Karzand ✉

University of California at Davis, CA, USA

Tosca Lechner ✉

University of Waterloo, Canada

Nati Srebro ✉

Toyota Technological Institute at Chicago, IL, USA

Serena Wang ✉

University of California at Berkeley, CA, USA

Google, Palo Alto, CA, USA

Abstract

We present a framework for designing scores to summarize performance metrics. Our design has two multi-criteria objectives: (1) improving on scores should improve all performance metrics, and (2) achieving pareto-optimal scores should achieve pareto-optimal metrics. We formulate our design to minimize the dimensionality of scores while satisfying the objectives. We give algorithms to design scores, which are provably minimal under mild assumptions on the structure of performance metrics. This framework draws motivation from real-world practices in hospital rating systems, where misaligned scores and performance metrics lead to unintended consequences.

2012 ACM Subject Classification Theory of computation → Algorithmic game theory and mechanism design; Theory of computation → Computational geometry

Keywords and phrases Multi-criteria incentives, Score-based incentives, Incentivizing improvement, Computational geometry

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.8

Funding *Anmol Kabra*: Supported in part through the NSF-TRIPODS Institute on Data, Economics, Algorithms and Learning (IDEAL).

Tosca Lechner: Supported by a Vector Research Grant and a Apple Waterloo PhD fellowship for machine learning and data science.

Acknowledgements Anmol Kabra thanks Naren Sarayu Manoj and Max Ovsiankin for pointers on convex analysis and geometry.

1 Introduction

The use of numerical metrics to evaluate performance and guide decision-making is common practice in healthcare, education, business, and public policy. It is common for agencies to design *surrogate scores* that summarize performance metrics, in a way that aligns incentives with performance metrics. Often the scored entities strategically optimize surrogates and end up degrading on metrics, a phenomenon commonly known as *unintended consequences* and pithily conveyed by Goodhart’s law [25, 46]:

“When a measure becomes a target, it ceases to be a good measure.”

¹ Corresponding author. Alphabetical ordering. Submission to Archival track.



Agencies thus aim to ensure that optimizing scores leads to improved metrics. As the number of performance metrics can be large in practice [51, 40], agencies must design *succinct* multi-dimensional surrogate scores. We present a framework to study this *minimal design problem*, and propose score designs that prevent unintended consequences.

Our work is directly motivated by real-world examples in safety-critical domains such as healthcare and education, where manifestations of Goodhart’s law exemplify the serious ramifications of unintended consequences. When Pacificare, a healthcare provider, incentivized hospitals in 2003 to perform certain medical procedures to improve quality of care, several unrepresented metrics deteriorated [35]. Similar misalignment between performance metrics and score-based hospital ratings, used by the Medicare agency (CMS), has been widely critiqued [47, 11, 33, 1, 44, 3]. Even so, CMS uses these score-based ratings to incentivize hospital policies [13, 18]. Hence, it aims to design scores so that improving on scores also improves all performance metrics. This goal motivates the *improvement objective* in our framework. In a similar vein, rating agencies such as USNews aim to incentivize efficient use of hospital resources through published scores [49]. On multi-dimensional metrics, the efficiency goal [41] naturally translates into the notion of pareto-efficiency, which motivates the *optimality objective* in our framework.

We present a framework for designing scores to summarize performance metrics. We give three natural design restrictions that align with real-world interpretability desiderata [15, 49], and propose score designs that satisfy the multi-criteria objectives under these restrictions. Striving for succinct scores, we formulate our design to minimize the dimensionality of scores. We give polynomial-time algorithms to design these succinct scores, which are provably minimal under mild assumptions on the structure of performance metrics. While existing work on score design for incentivization studies scalar scores [34, 28, 43, 52], we design scores of smallest dimensionality to satisfy the multi-criteria objectives. These objectives are unsatisfiable with scalar scores in general.

1.1 Designing surrogate scores from performance metrics

In our model, the agency aims to design a surrogate score function $S : \mathcal{F} \rightarrow \mathcal{S}$ given a set of performance metrics \mathcal{F} of hospitals.

Hospitals report to agencies like CMS and USNews on hundreds of performance metrics such as condition-specific death rates, readmission rates, and percentages of patients receiving satisfactory care [15, 14, 49]. We can denote the values of d metrics of a hospital with a real-valued vector $\mathbf{f} \in \mathcal{F} \subseteq \mathbb{R}^d$. Since d is large and metrics can be related through confounding variables [5, 37], the agency wants to summarize the d metrics as k scores with values $\mathcal{S} \subseteq \mathbb{R}^k$, where k is small as possible. For instance, Example 3 suggests that, to summarize COVID and pneumonia death rate metrics, the agency can choose either of the two metrics as the score, so that $k = 1$. Whereas for pneumonia death rate and excess antibiotic use metrics, Example 4 argues that selecting both metrics as scores is necessary, and so $k = 2$.

Surrogate design objectives

Anticipating that the hospital would target the incentives by optimizing the score function S , the agency wants to design S in such a way that optimizing them ensures that the hospital does well on the performance metrics. We formalize this goal with two design objectives, which utilize an ordering on the sets \mathcal{F} and \mathcal{S} , denoted by $\succ_{\mathcal{F}}$ and $\succ_{\mathcal{S}}$. The two objectives are motivated from CMS and USNews hospital rating agencies [15, 49].

1. **Improvement objective.** Improving on surrogate scores should result in improving on performance metrics. In particular,

$$\text{for } \mathbf{f}, \mathbf{f}' \in \mathcal{F}, \quad \text{if } S(\mathbf{f}') \succeq_{\mathcal{S}} S(\mathbf{f}) \text{ then } \mathbf{f}' \succeq_{\mathcal{F}} \mathbf{f}. \quad (1)$$

2. **Optimality objective.** Pareto-optimal points of surrogate scores should be pareto-optimal points of performance metrics. In particular,

$$\text{ParetoOpt}(S) \subseteq \text{ParetoOpt}(\mathcal{F}). \quad (2)$$

Throughout the paper, we analyze the setting $\mathcal{F} \subseteq \mathbb{R}^d$ and $\mathcal{S} \subseteq \mathbb{R}^k$ and use elementwise order of vectors for $\succeq_{\mathcal{F}}$ and $\succeq_{\mathcal{S}}$.

Surrogate design restrictions

Due to interpretability and public reporting obligations, rating agencies like CMS and USNews design scores by selecting subsets of the list of performance metrics or by taking weighted averages [14, 15, 16, 17, 49]. Moreover, monotonicity of scores in performance metrics is a desirable property for CMS, as it ensures that a hospital striving to improve all performance metrics sees improved score values [14, 17].

We formulate these requirements as three different restrictions on S . These restrictions impose a linear form on $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ with $\mathbf{A} \in \mathbb{R}^{k \times d}$ satisfying certain structural constraints.

1. **Coordinate Selection (Res-CS).** Each of the k coordinates of scores are chosen from d coordinates of performance metrics. That is, for all $i \in [k]$ there exists $j \in [d]$ such that $S(\mathbf{f})_i = \mathbf{f}_j$ for all $\mathbf{f} \in \mathcal{F}$. Equivalently, $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ where rows of \mathbf{A} are 1-hot vectors.
2. **Linear and Monotone (Res-LM).** The k coordinates of scores are linear combinations of d coordinates of performance metrics, and improving on performance metrics should result in improving on surrogate scores. That is, $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ where for $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$, if $\mathbf{f}' \geq \mathbf{f}$ then $\mathbf{A}\mathbf{f}' \geq \mathbf{A}\mathbf{f}$.
3. **Linear (Res-L).** The coordinates of surrogate scores are linear combinations of coordinates of performance metrics. That is, $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ without any further constraints on \mathbf{A} .

Minimal design problem

Since the number of performance metrics d can be large [14, 15, 49], a natural goal is to *succinctly* summarize metrics with scores that are accessible to patients and policymakers. This goal of succinctness translates into designing a multi-dimensional function $S : \mathbb{R}^d \rightarrow \mathbb{R}^k$ with *the smallest output dimension* k . For a combination of design objective and design restriction, the *minimal design problem* is determining the smallest dimensionality k and providing an algorithm outputs a surrogate score function S with this k .

1.2 Our contributions

In this paper, we study the minimal design problem. Our key contributions are:

1. We formalize surrogate score design for incentivizing multiple criteria, motivated from real-world practices of two hospital rating systems, CMS and USNews.
2. We fully determine the minimal design problems of all combinations of objectives and restrictions introduced in Section 1.1, and propose efficient score design algorithms (Algorithms 1 and 2). We summarize our results in Table 1.

- a. We show that the smallest dimensionalities k are dictated by structural properties of the affine hull of performance metrics \mathcal{F} .
- b. Identifying a relationship between improvement and optimality objectives (Theorem 13), we determine the minimal design problem for simultaneously satisfying both objectives.

■ **Table 1** We list smallest dimensionalities k for the minimal design problem of all combinations of objectives and restrictions. Here columns of \mathbf{Z} are an orthonormal basis of the linear subspace associated with r -dimensional affine hull of \mathcal{F} . We define the three matrix ranks `ConeSubsetRank`, `ConeGeneratingRank`, `ConeRank` in Theorem 2. For the improvement objective, the listed dimensionalities are also necessary, when \mathcal{F} has non-empty relative interior (Theorem 7).

Restriction	Improvement (§2)	Optimality (§3)	Both (§4)
Res-CS	<code>ConeSubsetRank</code> (\mathbf{Z})	r	<code>ConeSubsetRank</code> (\mathbf{Z})
Res-LM	<code>ConeGeneratingRank</code> (\mathbf{Z})	1	<code>ConeGeneratingRank</code> (\mathbf{Z})
Res-L	<code>ConeRank</code> (\mathbf{Z})	1	<code>ConeGeneratingRank</code> (\mathbf{Z})

1.3 Related work

Recent work has highlighted the plight of score-based incentivization when scores that do not align with performance metrics. In healthcare, design objectives of hospital rating agencies often vary across agencies. Two popular examples are the Medicare agency (CMS), which incentivizes healthcare investment across care metrics through a five-star score [15, 18], and the USNews agency, which promotes highly-specialized medical departments [49]. When hospitals target these score-based ratings, they often degrade on a few performance metrics [35]. For example, CMS’s score-based ratings have been found to encourage hospitals to selectively treat patients for minimizing readmission rates [3, 20, 12], and have exacerbated unequal access to healthcare [33, 1, 44]. Such unintended consequences are prevalent in fields that use scores as an incentive mechanism [6], for instance, in standardized testing [35] and financial credit ratings [31, 54, 7, 26].

Our framework extends recent work on score design in principal-agent theory [34, 28, 43, 52, 27, 38, 30, 29, 4, 2] by designing scores for multi-criteria objectives. Kleinberg and Raghavan [34] compare linear with monotone scalar score design for incentivizing *effort* from agents. On a similar front, Haghtalab et al. [28] study scalar score design with a linear threshold restriction. Score design has also been studied through a causality lens to optimize the average treated outcome [52, 27, 38]. Finally, Rolf et al. [43] use noisy score observations to approximate the pareto-frontier of performance metrics. Our framework’s optimality objective and design restrictions capture this line of work on scalar scores. However, our improvement objective is a novel contribution, and this objective turns to be unsatisfiable with scalar scores (Theorem 7). Hence, our score design problems are inherently multi-criteria.

Technically, our design algorithms utilize novel techniques to decompose and enclose polyhedral cones, building on work in computational geometry on finding frames of polyhedral cones [21, 39, 53] and enclosing convex hulls [22, 36, 42, 48]. Our definition of `ConeRank` (Theorem 2) is similar to `NonNegativeRank`, which is extensively studied in the context of non-negative matrix factorization [23, 24, 19, 50, 36].

1.4 Notation

We represent scalars as $\lambda, c \in \mathbb{R}$, and vectors and matrices as $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{m \times n}$. We denote the nonnegative orthant with \mathbb{R}_+^n . We generally write matrices as a stack of rows, $\mathbf{W} = [\mathbf{w}_1; \dots; \mathbf{w}_m]$, often denoting the set of rows with W . We say that matrix \mathbf{W} (or set

W) generates cone \mathcal{K}_W if $\mathcal{K}_W = \text{Cone}(W) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \boldsymbol{\lambda}W, \boldsymbol{\lambda} \in \mathbb{R}_+^m\}$. We denote a vector of zeros (or ones) as $\mathbf{0}_n \in \mathbb{R}^n$ (or $\mathbf{1}_n$), and the n -by- n identity matrix as \mathbf{I}_n , dropping subscripts when unambiguous.

2 Minimal design problem for improvement objective

We propose a surrogate score design for satisfying the improvement objective under the three design restrictions. Then we illustrate our design strategy on simple examples of performance metrics \mathcal{F} , highlighting relationships between the geometry of \mathcal{F} and the succinctness of scores. Finally, we show that our proposed design is minimal under a mild assumption on \mathcal{F} , implying that score design for improvement objective is inherently multi-criteria.

We first simplify the improvement objective in Equation (1) to identify geometric objects that represent *movement* and *improvement directions*. Score function $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ on domain \mathcal{F} satisfies improvement when for all $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$, if $\mathbf{A}(\mathbf{f}' - \mathbf{f}) \geq \mathbf{0}$ then $(\mathbf{f}' - \mathbf{f}) \geq \mathbf{0}$. Denoting the *movement directions at center \mathbf{f}* with $\mathcal{F}_{\mathbf{f}} = \{\mathbf{g} = \mathbf{f}' - \mathbf{f} \in \mathbb{R}^d \mid \text{for all } \mathbf{f}' \in \mathcal{F}\}$, we can rearrange terms to get

$$\text{for all centers } \mathbf{f} \in \mathcal{F}, \text{ movement directions } \mathbf{g} \in \mathcal{F}_{\mathbf{f}}, \quad \text{if } \mathbf{A}\mathbf{g} \geq \mathbf{0} \text{ then } \mathbf{I}\mathbf{g} \geq \mathbf{0} \quad (3)$$

Here the *set of score improvement directions* is exactly $\mathcal{K}_A^* = \{\mathbf{g} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{g} \geq \mathbf{0}\}$, which is the dual of polyhedral cone \mathcal{K}_A generated from rows of \mathbf{A} . Similarly, the *set of metric improvement directions* is $\mathcal{K}_I^* = \{\mathbf{g} \in \mathbb{R}^d \mid \mathbf{I}\mathbf{g} \geq \mathbf{0}\} = \mathbb{R}_+^d$, which is the dual of polyhedral cone $\mathcal{K}_I = \mathbb{R}_+^d$ generated from rows of \mathbf{I} . So intuitively, score function $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ satisfies improvement if and only if every movement direction (in $\mathcal{F}_{\mathbf{f}}$) that is a score improvement direction (in \mathcal{K}_A^*) is also a metric improvement direction (in \mathcal{K}_I^*):

$$S \text{ satisfies improvement} \iff \text{for all } \mathbf{f} \in \mathcal{F}, \quad \mathcal{F}_{\mathbf{f}} \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^*. \quad (4)$$

2.1 Design proposal for improvement objective

When performance metrics $\mathcal{F} \subseteq \mathbb{R}^d$ is a full-dimensional set, score design is trivial where the most succinct score design is $S(\mathbf{f}) = \mathbf{f}$. Note that while performance is measured in many dimensions [51, 40], the number of confounding variables of performance metrics is often smaller due to correlated metrics [5, 37]. This typically induces a low-dimensional structure on \mathcal{F} , observed in practice and assumed in theory [6, 8, 5, 37]. We do not assume such low-dimensional structure of \mathcal{F} , but the smallest dimensionality k of score function S is impacted by the intrinsic dimension of \mathcal{F} . The affine hull of \mathcal{F} is a natural geometric choice to capture its intrinsic dimension.

► **Definition 1.** Define the *affine hull* of \mathcal{F} , $\text{aff}(\mathcal{F})$, as the intersection of all affine subspaces in \mathbb{R}^d containing \mathcal{F} . Let \mathcal{L} be the linear subspace associated with $\text{aff}(\mathcal{F})$, i.e. \mathcal{L} is the translation of $\text{aff}(\mathcal{F})$ so that for all centers $\mathbf{f} \in \mathcal{F}$, movement directions $\mathcal{F}_{\mathbf{f}} \subseteq \mathcal{L}$.

By utilizing this subspace \mathcal{L} containing all possible movement directions $\mathcal{F}_{\mathbf{f}}$, we propose a score design in Algorithm 1 with dimensionalities given in Theorem 2. We introduce three *matrix ranks* – ConeSubsetRank (CSR), ConeGeneratingRank (CGR), and ConeRank (CR) – to characterize the score design dimensionalities for the three respective design restrictions – Coordinate Selection (Res-CS), Linear and Monotone (Res-LM), Linear (Res-L). These three matrix ranks capture the geometric properties of performance metrics \mathcal{F} that dictate the dimensionality of optimal score design for the three restrictions.

► **Theorem 2.** *Let columns of \mathbf{Z} be an orthonormal basis of linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. For each design restriction, there exists $S : \mathcal{F} \rightarrow \mathbb{R}^k$, designed using Algorithm 1, that satisfies the improvement objective with the following dimensionalities.*

	Dimensionality $k \geq$
Res-CS	$\text{ConeSubsetRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_Z = \mathcal{K}_V \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r} \text{ s.t. } \mathbf{V} \subseteq \mathbf{Z}\}$
Res-LM	$\text{ConeGeneratingRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_Z = \mathcal{K}_V \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r}\}$
Res-L	$\text{ConeRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_Z \subseteq \mathcal{K}_V \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r}\}$

■ **Algorithm 1** Design strategy for improvement objective.

- 1: Given: performance metrics \mathcal{F} and a design restriction.
- 2: Find \mathbf{Z} whose columns are an orthonormal basis of subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$.
- 3: Find \mathbf{V} that attains² the matrix rank corresponding to the design restriction.
- 4: Find \mathbf{A} that satisfies $\mathbf{V} = \mathbf{AZ}$ and design $S : \mathbf{f} \mapsto \mathbf{Af}$.

Theorem 2 follows from the following key insight of Equation (4): “for $S : \mathbf{f} \rightarrow \mathbf{Af}$ to satisfy the improvement objective, score improvement directions need to be metric improvement directions **only** for movement directions \mathcal{F}_f , which are contained in subspace \mathcal{L} .” In fact, satisfying the improvement objective boils down to ensuring that score improvement directions are a subset of metric improvement directions *in the coefficient space* w.r.t. subspace \mathcal{L} . The respective improvement directions \mathcal{K}_A^* and \mathcal{K}_I^* are generated by rows of \mathbf{A} and \mathbf{I} , which have coefficients that are rows of $\mathbf{V} = \mathbf{AZ}$ and \mathbf{Z} , where columns of \mathbf{Z} are an orthonormal basis of subspace \mathcal{L} . It turns out that improvement directions in the coefficient space are precisely the duals \mathcal{K}_V^* and \mathcal{K}_Z^* of polyhedral cones generated from rows of \mathbf{V} and \mathbf{Z} . So to satisfy the improvement objective, we need to ensure $\mathcal{K}_V^* \subseteq \mathcal{K}_Z^*$, or $\mathcal{K}_Z \subseteq \mathcal{K}_V$.

With the three matrix ranks, we capture the additional structure on \mathbf{A} imposed by the three design restrictions (Section 1.1). Res-L restriction does not further impose structure on \mathbf{A} , and so we only need to *enclose* cone \mathcal{K}_Z with \mathcal{K}_V . Res-LM restriction further requires function S to be monotone in \mathcal{F} , which intuitively means that every metric improvement direction needs to be a score improvement direction, i.e., $\mathcal{K}_Z^* \subseteq \mathcal{K}_V^*$. So to satisfy Res-LM, we must *generate* cone \mathcal{K}_Z with \mathcal{K}_V . Finally, Res-CS restriction requires selecting the k score function coordinates from d metrics. In the coefficient space, this requirement means that rows of \mathbf{V} are chosen from rows of \mathbf{Z} and \mathcal{K}_V generates \mathcal{K}_Z . Hence, the three matrix ranks precisely capture structure on \mathbf{A} imposed by the improvement objective and the design restrictions. We include the proof of Theorem 2 in Theorem A.1.

2.2 Geometry of metrics dictates succinctness of scores

We now illustrate Algorithm 1 with several examples of metrics \mathcal{F} . We instantiate performance metrics in our examples with familiar notions of hospital metrics, to intuitively bridge our analysis and algorithm with practical score design. In doing so, we discuss how the geometry of \mathcal{F} dictates the shape of polyhedral cone \mathcal{K}_Z , influencing the dimensionality of minimal score design for the three design restrictions. Finally, we provide high-level descriptions of techniques to implement Algorithm 1 efficiently.

² For a matrix rank, e.g. CSR, we say that \mathbf{V} “attains” it if $\mathbf{V} \subseteq \mathbf{Z}$ (rows of \mathbf{V} are chosen from rows of \mathbf{Z}), $\mathcal{K}_Z = \mathcal{K}_V$, and the number of rows of \mathbf{V} equals $\text{CSR}(\mathbf{Z})$.



(a) When the two metrics are correlated (Ex. 3), we can choose either metric in $S : \mathcal{F} \rightarrow \mathbb{R}^1$. (b) When the two metrics are anti-correlated (Ex. 4), we must choose both metrics in $S : \mathcal{F} \rightarrow \mathbb{R}^2$.

■ **Figure 1** To design scores for two metrics ($\mathcal{F} \subseteq \mathbb{R}^2$), we can inspect the correlation between metrics – the correlation dictates the succinctness of $S : \mathcal{F} \rightarrow \mathbb{R}^k$ for satisfying improvement.

► **Example 3** (Two correlated metrics \implies choose either for score design). CMS evaluates hospitals on numerous performance metrics like condition-specific death rates, readmission rates, and safety standards [15]. Often comorbidities of medical conditions can lead to positive correlations between metrics. In the case of two *perfectly* positively correlated metrics, Algorithm 1 suggests to choose either of the two metrics to design $S : \mathcal{F} \rightarrow \mathbb{R}^1$.

Consider two metrics – (i) pneumonia death rate and (ii) COVID-19 death rate – that have a positive correlation due to comorbidities. Assume that for a hospital, these two death rates take values $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^2 \mid -f_1 + 2f_2 = 1, -1 \leq f_1 \leq 1\}$, lying in a 1-dimensional affine subspace of \mathbb{R}^2 (Figure 1a, red). As the affine hull $\text{aff}(\mathcal{F}) = \{\mathbf{f} \mid -f_1 + 2f_2 = 1\}$ is 1-dimensional, the associated linear subspace $\mathcal{L} = \{\mathbf{f} \mid -f_1 + 2f_2 = 0\}$ (Figure 1a, blue) containing all movement directions $\mathcal{F}_{\mathbf{f}}$ is 1-dimensional. Per Line 2 of Algorithm 1, we arrange an orthonormal basis for \mathcal{L} as columns of $\mathbf{Z} \propto \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, whose rows generate the polyhedral cone $\mathcal{K}_{\mathbf{Z}} = \{2\lambda_1 + \lambda_2 \mid \lambda_1, \lambda_2 \geq 0\} = \mathbb{R}_+$. Note that the metric improvement directions in the coefficient space are the dual cone $\mathcal{K}_{\mathbf{Z}}^* = \mathbb{R}_+$.

To satisfy improvement objective under a design restriction, we need to find matrix \mathbf{V} that attains the corresponding matrix rank. For all three matrix ranks, the cone $\mathcal{K}_{\mathbf{V}}$ generated by rows of \mathbf{V} needs to *enclose* cone $\mathcal{K}_{\mathbf{Z}}$. Equivalently, in the coefficient space, score improvement directions $\mathcal{K}_{\mathbf{V}}^*$ need to be a subset of metric improvement directions $\mathcal{K}_{\mathbf{Z}}^*$. The choice of $\mathbf{V} = [2] \in \mathbb{R}^{1 \times 1}$ yields the desired property $\mathcal{K}_{\mathbf{Z}} \subseteq \mathcal{K}_{\mathbf{V}}$. In fact, we get $\mathcal{K}_{\mathbf{Z}} = \mathcal{K}_{\mathbf{V}}$ and $\mathbf{V} \subseteq \mathbf{Z}$, and so all three matrix ranks have value 1.

Finally, we can recover $\mathbf{A} = [1, 0]$ such that $\mathbf{V} = \mathbf{AZ}$, and design $S(\mathbf{f}) = [1, 0] \cdot \mathbf{f} = f_1$. It is easy to verify that this S satisfies the improvement objective (we could also have chosen $\mathbf{V} = [1]$ previously to design $S(\mathbf{f}) = [0, 1] \cdot \mathbf{f} = f_2$). Hence, when the two metrics are perfectly positively correlated, choosing one for score design suffices.

► **Example 4** (Two anti-correlated metrics \implies must choose both for score design). Performance metrics used by CMS can also be negatively correlated when a hospital must balance its effort to simultaneously improve all metrics. In the case of two *perfectly* negative correlated metrics, Algorithm 1 suggests to use both metrics to design $S : \mathcal{F} \rightarrow \mathbb{R}^2$, as no 1-dimensional score function can satisfy improvement objective.

Consider two metrics – (i) pneumonia death rate and (ii) excessive antibiotic use – that have a negative correlation as improving on one degrades the other. Assume that these two metrics take values $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^2 \mid -f_1 - 2f_2 = 1, -1 \leq f_1 \leq 1\}$, lying in a 1-dimensional affine subspace of \mathbb{R}^2 (Figure 1b, red). Similar to Example 3, the subspace $\mathcal{L} = \{\mathbf{f} \mid -f_1 + 2f_2 = 0\}$ (Figure 1b, blue) associated to $\text{aff}(\mathcal{F})$ is 1-dimensional. But the

rows of orthonormal basis $\mathbf{Z} \propto \begin{bmatrix} 2 \\ -1 \end{bmatrix}$ generate cone $\mathcal{K}_Z = \{2\lambda_1 - \lambda_2 \mid \lambda_1, \lambda_2 \geq 0\} = \mathbb{R}$, which contains a linear subspace within. This means that the metric improvement directions in the coefficient space are the dual cone $\mathcal{K}_Z^* = \{\mathbf{0}\}$, i.e., there are no non-trivial directions to simultaneously improve both metrics.

To satisfy improvement objective, score improvement directions in the coefficient space \mathcal{K}_V^* need to be a subset of metric improvement directions $\mathcal{K}_Z^* = \{\mathbf{0}\}$, or equivalently $\mathcal{K}_Z \subseteq \mathcal{K}_V$. Hence, we choose $\mathbf{V} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \propto \mathbf{Z}$ with 2 rows. Note that \mathbf{V} with just 1 row would generate either cone \mathbb{R}_+ or cone $-\mathbb{R}_+$, and fail to enclose cone $\mathcal{K}_Z = \mathbb{R}$. Hence, all three matrix ranks have value 2 even though all movement directions \mathcal{F}_f lie in a 1-dimensional subspace \mathcal{L} .

Finally, we can recover $\mathbf{A} = \mathbf{I}_2$ such that $\mathbf{V} = \mathbf{AZ}$ and design the trivial $S(\mathbf{f}) = \mathbf{f}$. Due to the perfect negative correlation in metrics, we must choose both in the score design.

► **Example 5** (Restriction with monotonicity \implies higher dimensionality). When the number of metrics is large, understanding correlations among them can be unintuitive. Hence, we rely on structure of polyhedral cones for score design, specifically improvement directions of scores \mathcal{K}_V^* and metrics \mathcal{K}_Z^* (in the coefficient space). We find that score function dimensionality k under Res-CS and Res-LM restrictions can be much larger than under Res-L, as $\text{CSR}, \text{CGR} \gg \text{CR}$.

Consider the case of four metrics where two of them balance the other two, i.e., a toy example where performance metrics take values $\mathcal{F} = \text{aff}(\mathcal{F}) = \{\mathbf{f} \in \mathbb{R}^4 \mid [1, -1, 1, -1] \cdot \mathbf{f} = 0\}$. Here the four metrics lie in a 3-dimensional linear subspace of \mathbb{R}^4 and $\mathcal{F} = \text{aff}(\mathcal{F}) = \mathcal{L}$. Hence, three orthonormal vectors in \mathbb{R}^4 form a basis of \mathcal{L} such that the rows of \mathbf{Z} generate the “square” cone \mathcal{K}_Z in \mathbb{R}^3 (Figure 2a, red):

$$\mathbf{Z} = \frac{1}{2} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{4 \times 3}.$$

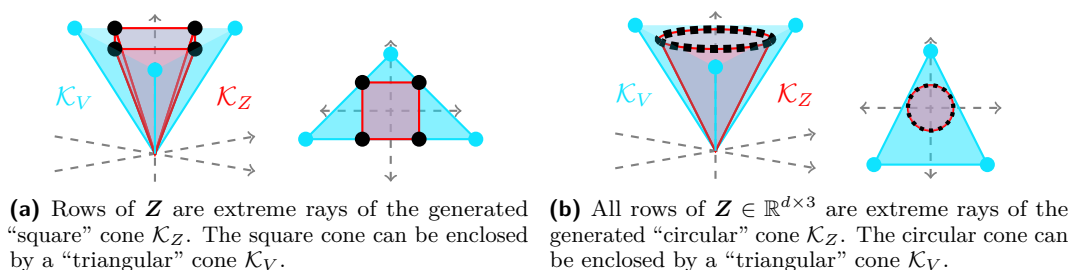
For Res-CS and Res-LM restrictions, we need to find matrix \mathbf{V} such that $\mathcal{K}_V = \mathcal{K}_Z$. As all rows of \mathbf{Z} are *extreme rays* of \mathcal{K}_Z , matrix \mathbf{V} must have four rows $\mathbf{V} = \mathbf{I}_4 \mathbf{Z}$ (any \mathbf{V} with fewer rows would not *generate* the square cone). Hence, $\text{CSR}(\mathbf{Z}) = \text{CGR}(\mathbf{Z}) = 4$. But for Res-L restriction that does not require monotonicity, rows of \mathbf{V} need only ensure $\mathcal{K}_Z \subseteq \mathcal{K}_V$. The following matrix \mathbf{V} with three rows that generates a “triangular” cone \mathcal{K}_V (Figure 2a, blue) *enclosing* the square cone \mathcal{K}_Z :

$$\mathbf{V} = \frac{1}{2} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 1 & 3 & -1 \\ 1 & -3 & -1 \end{bmatrix} \quad \text{and so } \mathbf{V} = \mathbf{AZ} \text{ with } \mathbf{A} = \frac{1}{4} \cdot \begin{bmatrix} 3 & 3 & -1 & -1 \\ 3 & -3 & -1 & 5 \\ -3 & 3 & 5 & -1 \end{bmatrix}.$$

Generally, CSR and CGR can be much larger than CR (Figure 2b). Since these three matrix ranks describe the dimensionality under the three restrictions (Theorem 2), restrictions that require monotonicity (Res-CS, Res-LM) lead to higher dimensionality in score design compared to Res-L. In other words, allowing negative values in matrix \mathbf{A} can significantly reduce dimensionality of score design.

► **Remark 6** (Competing metric improvement directions \implies higher dimensionality under Res-CS). When rows of \mathbf{Z} generate cone \mathcal{K}_Z that is *pointed*³, we get $\text{CSR}(\mathbf{Z}) = \text{CGR}(\mathbf{Z})$. But when

³ A cone \mathcal{K} is pointed if for all nonzero $\mathbf{x} \in \mathcal{K}$, we have $-\mathbf{x} \notin \mathcal{K}$. It is called *non-pointed* otherwise.



■ **Figure 2** Side and top views of cones \mathcal{K}_Z (red) generated by rows of \mathbf{Z} , whose columns are orthonormal basis of 3-dimensional subspace \mathcal{L} . As CSR and CGR require *generating* \mathcal{K}_Z with \mathcal{K}_V , the matrix ranks depend on the number of extreme rays of \mathcal{K}_Z , which can be much higher than $\dim \text{aff}(\mathcal{F}) = 3$. On the other hand, CR only requires *enclosing* \mathcal{K}_Z with \mathcal{K}_V ; and so is independent of the number of extreme rays.

cone \mathcal{K}_Z that is *non-pointed*, we get $\text{CSR}(\mathbf{Z}) > \text{CGR}(\mathbf{Z})$. \mathcal{K}_Z can be non-pointed when improving one metric degrades another, i.e., when metric improvement directions compete among themselves. In this setting, dimensionality under Res-CS is higher than that under Res-LM (see Example A.2).

Efficiently implementing Algorithm 1

Our proposed design strategy in Algorithm 1 can be efficiently implemented with algorithms that utilize the geometry of metrics \mathcal{F} . Elementary linear algebra operations can implement Lines 2 and 4 of Algorithm 1, i.e., finding orthonormal basis \mathbf{Z} and recovering \mathbf{A} from $\mathbf{V} = \mathbf{AZ}$. It is also possible to efficiently implement Line 3, to find matrix \mathbf{V} that attain the matrix ranks – ConeSubsetRank, ConeGeneratingRank, and ConeRank [32]. We briefly discuss algorithms for Line 3, thus ensuring that the full Algorithm 1 can be efficiently implemented. These algorithms leverage a key property of polyhedral cones, *pointedness*.

When the cone \mathcal{K}_Z generated from rows of \mathbf{Z} is pointed, we can easily find \mathbf{V} that attains the matrix ranks. For ConeSubsetRank, we can keep the rows of \mathbf{Z} that are extreme rays of the polyhedral cone \mathcal{K}_Z , as extreme rays minimally generate a pointed cone [9, Prop. 26.5.4]. ConeGeneratingRank turns out to be the same as ConeSubsetRank, as every extreme ray of \mathcal{K}_Z is a row of matrix \mathbf{Z} [9, Prop. 26.5.4]. For ConeRank, the matrix \mathbf{V} attaining it must generate \mathcal{K}_V that encloses \mathcal{K}_Z . An intuitive procedure can find this \mathbf{V} : can scale rows of \mathbf{Z} to lie on a hyperplane, and find a simplex that encloses the convex hull of scaled rows [22].

When the cone \mathcal{K}_Z is non-pointed, the cone contains a linear subspace within. Here we can utilize the unique Minkowski decomposition of polyhedral cones into two orthogonal components: the maximal linear subspace within, and a pointed remnant [45, Sec. 8.2]. Then, for all three matrix ranks, we can generate/enclose non-pointed cone \mathcal{K}_Z , by generating/enclosing the two orthogonal components separately.

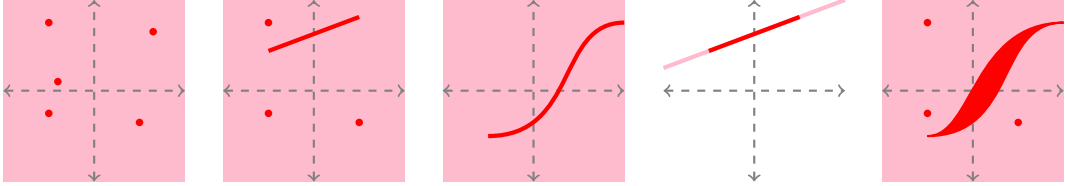
2.3 Proposed design is minimal

Theorem 2 states that dimensionalities determined by the three matrix ranks – ConeSubsetRank, ConeGeneratingRank, and ConeRank – are sufficient for score design. It turns out that these dimensionalities are also necessary under a mild assumption on \mathcal{F} (Theorem 7). Hence, Theorems 2 and 7 together imply that the *three matrix ranks exactly determine the minimal design problem for improvement objective*.

► **Theorem 7.** *Assume metrics $\mathcal{F} \subseteq \mathbb{R}^d$ have non-empty relative interior with respect to $\text{aff}(\mathcal{F})$. Then the listed dimensionalities k in Theorem 2 are necessary.*

We briefly discuss the implication of metrics \mathcal{F} having non-empty relative interior on satisfying the improvement objective. Such a set \mathcal{F} contains a center $\mathbf{f}^* \in \mathcal{F}$ where every direction in subspace \mathcal{L} is a positively-scaled movement direction from $\mathcal{F}_{\mathbf{f}^*}$. Intuitively, all score improvement directions are movement directions in the coefficient space. As a result, we get an equivalence between satisfying improvement in the ambient space and the coefficient space, i.e., satisfying improvement in Equation (4) is equivalent to satisfying $\mathcal{K}_Z \subseteq \mathcal{K}_V$. See Theorem A.3 for the proof.

► **Remark 8.** In Figure 3 we illustrate examples of \mathcal{F} and their relative interior. \mathcal{F} having non-empty relative interior is a reasonable condition in practice, as performance metrics used by rating agencies are often correlated and not isolated points [6, 15, 49, 8, 37, 5]. For instance, CMS uses percentage-rate-based metrics, such as condition-specific death rates, readmission rates, and screening rates [15, 14]. This leads to real-valued metrics $\mathcal{F} = [0, 1]^d$, which has non-empty relative interior. We note that, when the relative interior is *empty*, dimensionality k significantly less than listed values in Theorem 2 can suffice (Proposition A.5).



■ **Figure 3** Examples of $\mathcal{F} \subseteq \mathbb{R}^2$. The left three have empty relative interior, whereas the right two have non-empty relative interior with respect to $\text{aff}(\mathcal{F})$, which is lightly shaded.

► **Remark 9 (Choice of affine subspace and orthonormal basis).** Our design strategy in Algorithm 1 can use *any* orthonormal basis \mathbf{Z} of the linear subspace $\mathcal{L}_{\mathcal{H}}$ associated with *any* affine subspace \mathcal{H} containing metrics \mathcal{F} . To design the *minimal* $S : \mathcal{F} \rightarrow \mathbb{R}^k$, we pick *any* orthonormal basis of subspace \mathcal{L} associated with affine hull $\mathcal{H} = \text{aff}(\mathcal{F})$. This follows from Lemma A.4, which states that three matrix ranks are (1) invariant to the choice of orthonormal basis for a fixed subspace $\mathcal{L}_{\mathcal{H}}$, and (2) minimized with the choice of $\mathcal{H} = \text{aff}(\mathcal{F})$.

3 Minimal design problem for optimality objective

We propose a surrogate score design for satisfying the optimality objective and discuss the minimality of our proposed design. We use the standard definition of pareto-optimality.

► **Definition 10.** *Point $\mathbf{f} \in \mathcal{F}$ is pareto-optimal for maximizing S if no other point in \mathcal{F} both improves $S(\mathbf{f})$ in all coordinates and strictly improves $S(\mathbf{f})$ in at least one coordinate.*

$$\text{ParetoOpt}(S) := \{\mathbf{f} \in \mathcal{F} \mid \text{for all } \mathbf{f}' \in \mathcal{F}, \text{ either } S(\mathbf{f}') \not\geq S(\mathbf{f}) \text{ or } S(\mathbf{f}') = S(\mathbf{f})\}.$$

We write $\text{ParetoOpt}(\mathcal{F})$ to denote the pareto-optimal points in \mathcal{F} w.r.t. the identity map.

We simplify the optimality objective in Equation (2) – $\text{ParetoOpt}(S) \subseteq \text{ParetoOpt}(\mathcal{F})$ – using movement directions $\mathcal{F}_{\mathbf{f}}$ at center \mathbf{f} , score improvement directions $\mathcal{K}_{\mathbf{A}}^*$, and metric improvement directions $\mathcal{K}_{\mathbf{f}}^*$. Intuitively, score function $S : \mathbf{f} \mapsto \mathbf{A}\mathbf{f}$ satisfies optimality if and

only if movement directions \mathcal{F}_f that are *non-strict score improvement directions* are also *non-strict metric improvement directions*:

$$\text{Optimality} \iff \{f \in \mathcal{F} \mid \mathcal{F}_f \subseteq (\mathcal{K}_A^*)^c \cup \ker \mathbf{A}\} \subseteq \{f \in \mathcal{F} \mid \mathcal{F}_f \subseteq (\mathcal{K}_I^*)^c \cup \ker \mathbf{I}\}. \quad (5)$$

3.1 Design proposal for optimality objective

We propose a score design in Algorithm 2 with dimensionalities given in Theorem 11. We note that dimensionalities for score design are much smaller for the optimality objective than for the improvement objective (Theorem 2). Specifically, for Res-LM and Res-L restrictions, a 1-dimensional score function $S : \mathcal{F} \rightarrow \mathbb{R}$ suffices to satisfy optimality whereas multi-dimensional function S is necessary for improvement (Theorem 7). This suggests that the optimality objective is significantly weaker than the improvement objective.

► **Theorem 11.** *For each design restriction, there exists $S : \mathcal{F} \rightarrow \mathbb{R}^k$, designed using Algorithm 2, that satisfies the optimality objective with the following dimensionalities.*

	Dimensionality $k \geq$
Res-CS	$\dim \text{aff}(\mathcal{F})$
Res-LM	1
Res-L	1

■ **Algorithm 2** Design strategy for optimality objective.

-
- 1: Given: \mathcal{F} and a design restriction.
 - 2: **if** Design restriction is Res-LM or Res-L **then**
 - 3: Design $S(f) = \mathbf{a} \cdot f$ with any positive vector \mathbf{a} .
 - 4: **else if** Design restriction is Res-CS **then**
 - 5: Find \mathbf{Z} whose columns are an orthonormal basis of subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$.
 - 6: Let \mathbf{V} be linearly independent rows of \mathbf{Z} .
 - 7: Find \mathbf{A} that satisfies $\mathbf{V} = \mathbf{AZ}$ and design $S : f \mapsto \mathbf{A}f$.
-

For Res-LM and Res-L restrictions, the minimal design is straightforward: design $S : f \mapsto \mathbf{a} \cdot f$ using any vector $\mathbf{a} > \mathbf{0}$ [55]. For Res-CS restriction, we utilize an isomorphism between movement directions \mathcal{F}_f and their coefficients $\mathcal{C}_f \subseteq \mathbb{R}^r$ w.r.t. orthonormal basis $\mathbf{Z} \in \mathbb{R}^{d \times r}$ of subspace \mathcal{L} associated with r -dimensional $\text{aff}(\mathcal{F})$. The columns of \mathbf{Z} span subspace \mathcal{L} and its rows correspond to coordinates of movement directions \mathcal{F}_f . Using this isomorphism, choosing r linearly independent rows of \mathbf{Z} as rows of \mathbf{V} suffices to satisfy the optimality objective. As $\mathbf{V} \subseteq \mathbf{Z}$, we can find $\mathbf{A} \in \mathbb{R}^{r \times d}$ with 1-hot rows such that $\mathbf{V} = \mathbf{AZ}$, and design $S : f \mapsto \mathbf{A}f$ that satisfies the Res-CS restriction. We include the proof in Theorem A.6.

3.2 Discussion of minimality of proposed design

While our proposed design for improvement objective is minimal when \mathcal{F} has non-empty relative interior (Theorem 7), our design for the optimality objective is *not necessarily* minimal under the same condition on \mathcal{F} . The challenge is that $\text{ParetoOpt}(\mathcal{F})$, the optimal trade-off surface [10], depends on the boundary of \mathcal{F} . To demonstrate this, we give three examples of d -dimensional \mathcal{F} with non-empty relative interior – for one of the examples dimensionality $k = \dim \text{aff}(\mathcal{F})$ is necessary for satisfying optimality under Res-CS, whereas for the other two examples, a 1-dimensional S suffices. See Proposition A.7 for the proof.

► **Proposition 12.** Consider designing $S : \mathcal{F} \rightarrow \mathbb{R}^k$ to satisfy optimality objective.

1. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_1 \leq 1\}$, $k \geq 1$ is necessary and sufficient for all design restrictions.
2. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_2 \leq 1\}$, $k \geq 1$ is necessary and sufficient for all design restrictions.
3. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_\infty \leq 1\}$, $k \geq d$ is necessary and sufficient for Res-CS. Moreover, $k \geq 1$ is necessary and sufficient for the Res-LM and Res-L restrictions.

4 Minimal design problem for both objectives simultaneously

So far we have separately analyzed the minimal design problems for improvement and optimality objectives. We now give results for simultaneously satisfying both objectives.

First, we establish a relationship between the improvement and optimality objectives. This result holds even for score functions S that are not linear in \mathcal{F} .

► **Theorem 13.** Let $S : \mathcal{F} \rightarrow \mathbb{R}^k$ be monotone in \mathcal{F} . If S satisfies improvement, then S satisfies optimality.

Proof. Let score function $S : \mathcal{F} \rightarrow \mathbb{R}^k$ be monotone in \mathcal{F} and satisfy improvement. Hence, for all $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$ we have $S(\mathbf{f}') \geq S(\mathbf{f}) \iff \mathbf{f}' \geq \mathbf{f}$, i.e., the function S preserves the ordering on set \mathcal{F} . We prove by contradiction that such an S satisfies optimality. Assume that $\mathbf{f}^* \in \text{ParetoOpt}(S)$ but $\mathbf{f}^* \notin \text{ParetoOpt}(\mathcal{F})$. That is, there exists $\mathbf{f} \in \mathcal{F}$ such that $\mathbf{f} \geq \mathbf{f}^*$ and $\mathbf{f} \neq \mathbf{f}^*$. Because S preserves the ordering, it must be that $S(\mathbf{f}) \geq S(\mathbf{f}^*)$ and $S(\mathbf{f}) \neq S(\mathbf{f}^*)$, which means that $\mathbf{f}^* \notin \text{ParetoOpt}(S)$ and contradicts our assumption. ◀

We utilize Theorem 13 to design S that simultaneously satisfies both objectives. As S is monotone in \mathcal{F} under Res-CS and Res-LM restrictions, it suffices to design S that satisfies the improvement objective. We include the proof in Corollary A.8.

► **Corollary 14.** Let columns of \mathbf{Z} be an orthonormal basis of linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. For each design restriction, there exists score function $S : \mathcal{F} \rightarrow \mathbb{R}^k$ that simultaneously satisfies improvement and optimality objectives with following dimensionalities.

	Dimensionality $k \geq$
Res-CS	ConeSubsetRank(\mathbf{Z})
Res-LM	ConeGeneratingRank(\mathbf{Z})
Res-L	ConeGeneratingRank(\mathbf{Z})

Moreover, for Res-CS and Res-LM restrictions, the score design is minimal when \mathcal{F} has non-empty relative interior.

► **Remark 15.** For simultaneously satisfying both objectives under Res-L restriction, dimensionality $k = \text{CR}(\mathbf{Z})$ is necessary, when \mathcal{F} has non-empty relative interior (Theorem 7). Corollary 14 states that $k = \text{CGR}(\mathbf{Z})$ is sufficient, and $\text{CGR} \gg \text{CR}$ in general (Example 5). We leave to future work to close this gap between necessary and sufficient dimensionality.

5 Conclusion

We propose a framework to design succinct scores to summarize performance metrics \mathcal{F} , and give polynomial-time algorithms that design scores that are provably minimal under mild assumptions on \mathcal{F} . Two future directions are to design scores: (1) when metrics takes discrete high-dimensional values, (2) using incomplete, noisy high data from historical samples of metric values, and (3) when metrics have a non-linear structure. On a technical note, it

remains to identify structural properties of \mathcal{F} and corresponding minimal designs for the optimality objective. Designing minimal scores for simultaneously satisfying both objectives under linear restriction is also an open direction.

References

- 1 Rahul Aggarwal, J Gmerice Hammond, Karen E Joynt Maddox, Robert W Yeh, and Rishi K Wadhera. Association between the proportion of black patients cared for at hospitals and financial penalties under value-based payment programs. *Journal of American Medical Association*, 325(12):1219–1221, 2021.
- 2 Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. Setting fair incentives to maximize improvement. *arXiv preprint arXiv:2203.00134*, 2022.
- 3 Diane Alexander. How do doctors respond to incentives? unintended consequences of paying doctors to reduce costs. *Journal of Political Economy*, 128(11):4046–4096, 2020.
- 4 Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (02), pages 1774–1781, 2020.
- 5 Arlene S Ash, Stephen F Fienberg, Thomas A Louis, Sharon-Lise T Normand, Therese A Stukel, and Jessica Utts. Statistical issues in assessing hospital performance. <https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/downloads/statistical-issues-in-assessing-hospital-performance.pdf>, 2012. Accessed: 2024-01-09.
- 6 Deborah L Bandalos. *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.
- 7 Heski Bar-Isaac and Joel Shapiro. Credit ratings accuracy and analyst incentives. *American Economic Review*, 101(3):120–124, 2011.
- 8 Matthew E Barclay, Mary Dixon-Woods, and Georgios Lyratzopoulos. Concordance of hospital ranks and category ratings using the current technical specification of us hospital star ratings and reasonable alternative specifications. In *JAMA Health Forum*, volume 3(5), pages e221006–e221006. American Medical Association, 2022.
- 9 Kim C. Border. Caltech ec 181, lecture notes: Convex analysis and economic theory. <https://healy.econ.ohio-state.edu/kcb/Ec181/>, 2020.
- 10 Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- 11 Chicago Booth Review. Hospital ratings are deeply flawed. Can they be fixed? <https://www.chicagobooth.edu/review/hospital-ratings-are-deeply-flawed-can-they-be-fixed>, 2020. Accessed: 2024-01-09.
- 12 Jeffrey Clemens and Joshua D Gottlieb. Do physicians’ financial incentives affect medical treatment and patient health? *American Economic Review*, 104(4):1320–1349, 2014.
- 13 CMS.gov. Hospital Value Based Purchasing (VBP) Program. <https://qualitynet.cms.gov/inpatient/hvbp>. Accessed: 2024-01-15.
- 14 CMS.gov. Medicare 2024 Part C & D Star Ratings Technical Notes. <https://www.cms.gov/files/document/2024-star-ratings-technical-notes.pdf>. Accessed: 2024-04-10.
- 15 CMS.gov. Overall Hospital Quality Star Ratings. <https://qualitynet.cms.gov/inpatient/public-reporting/overall-ratings>. Accessed: 2024-01-15.
- 16 CMS.gov. Quality Payment Program: Merit-based Incentive Payment System (MIPS). <https://qpp.cms.gov/mips/reporting-options-overview>. Accessed: 2024-04-10.
- 17 CMS.gov. Report to Congress: Risk Adjustment in Medicare Advantage. <https://www.cms.gov/files/document/report-congress-risk-adjustment-medicare-advantage-december-2021.pdf>. Accessed: 2024-04-10.

- 18 Douglas A Conrad. The theory of value-based payment incentives and their application to health care. *Health Services Research*, 50:2057–2089, 2015.
- 19 David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.
- 20 David Dranove and Paul Wehner. Physician-induced demand for childbirths. *Journal of health economics*, 13(1):61–73, 1994.
- 21 José H Dulá, Richard V Helgason, and N Venugopal. An algorithm for identifying the frame of a pointed finite conical hull. *INFORMS Journal on Computing*, 10(3):323–330, 1998.
- 22 David Gale. On inscribing n-dimensional sets in a regular n-simplex. *Proceedings of the American Mathematical Society*, 4(2):222–225, 1953.
- 23 Nicolas Gillis. *Nonnegative matrix factorization*. SIAM, 2020.
- 24 Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithms for separable non-negative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2013.
- 25 Charles AE Goodhart and CAE Goodhart. *Problems of monetary management: the UK experience*. Springer, 1984.
- 26 Andrew S Grove. *High output management*. Vintage, 2015.
- 27 Luke Guerdan, Amanda Coston, Kenneth Holstein, and Zhiwei Steven Wu. Counterfactual prediction under outcome measurement error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1598, 2023.
- 28 Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 160–166, 2021.
- 29 Jason D Hartline, Yingkai Li, Liren Shan, and Yifan Wu. Optimization of scoring rules. *arXiv preprint arXiv:2007.02905*, 2020.
- 30 Jason D Hartline, Liren Shan, Yingkai Li, and Yifan Wu. Optimal scoring rules for multi-dimensional effort. *arXiv preprint arXiv:2211.03302*, 2022.
- 31 Bengt Holmstrom and Paul Milgrom. The firm as an incentive system. *The American Economic Review*, 84(4):972–991, 1994. URL: <http://www.jstor.org/stable/2118041>.
- 32 Anmol Kabra, Mina Karzand, Tosca Lechner, Nati Srebro, and Serena Wang. Score design for multi-criteria incentivization. To appear on arXiv.
- 33 Hyunmin Kim, Asos Mahmood, Noah E Hammarlund, and Cyril F Chang. Hospital value-based payment programs and disparity in the united states: A review of current evidence and future perspectives. *Frontiers in Public Health*, 10:882715, 2022.
- 34 Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.
- 35 Daniel Koretz. *The Testing Charade: Pretending to Make Schools Better*. The University of Chicago Press, 2017.
- 36 Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *International Conference on Machine Learning*, pages 231–239. PMLR, 2013.
- 37 Nisha Kurian, Jyotsna Maid, Sharoni Mitra, Lance Rhyne, Michael Korvink, and Laura H Gunn. Predicting hospital overall quality star ratings in the usa. In *Healthcare*, volume 9(4), page 486. MDPI, 2021.
- 38 Lydia T Liu, Solon Barocas, Jon Kleinberg, and Karen Levy. On the actionability of outcome prediction. *arXiv preprint arXiv:2309.04470*, 2023.
- 39 Francisco J López. An algorithm to find the lineality space of the positive hull of a set of vectors. *Journal of Mathematical Modelling and Algorithms*, 10(1):1–30, 2011.
- 40 Jerry Muller. *The tyranny of metrics*. Princeton University Press, 2018.
- 41 Committee on Quality of Health Care in America. *Crossing the quality chasm: a new health system for the 21st century*. National Academies Press, 2001.

- 42 Joseph O’Rourke, Alok Aggarwal, Sanjeev Maddila, and Michael Baldwin. An optimal algorithm for finding minimal enclosing triangles. *Journal of Algorithms*, 7(2):258–269, 1986.
- 43 Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Bjorkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, pages 8158–8168. PMLR, 2020.
- 44 Kirsten Schardt, Lorraine Hutzler, Joseph Bosco, Casey Humbyrd, and Matt DeCamp. Increase in healthcare disparities: The unintended consequences of value-based medicine, lessons from the total joint bundled payments for care improvement. *Bulletin of the NYU Hospital for Joint Diseases*, 78(2):93–97, 2020.
- 45 Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.
- 46 Marilyn Strathern. ‘improving ratings’: audit in the british university system. *European review*, 5(3):305–321, 1997.
- 47 The New York Times. The Hype over Hospital Rankings. <https://www.nytimes.com/2013/07/28/sunday-review/the-hype-over-hospital-rankings.html>, 2013. Accessed: 2024-01-09.
- 48 Godfried T Toussaint. Solving geometric problems with the rotating calipers. In *Proc. IEEE Melecon*, volume 83, page A10, 1983.
- 49 U.S. News and World Report. FAQ: How and Why We Rank and Rate Hospitals. <https://health.usnews.com/health-care/best-hospitals/articles/faq-how-and-why-we-rank-and-rate-hospitals>, 2023. Accessed: 2024-01-15.
- 50 Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.
- 51 Rishi K Wadhwa, Jose F Figueroa, Karen E Joynt Maddox, Lisa S Rosenbaum, Dhruv S Kazi, and Robert W Yeh. Quality measure development and associated spending by the centers for medicare & medicaid services. *JAMA*, 323(16):1614–1616, 2020.
- 52 Serena Wang, Stephen Bates, PM Aronow, and Michael I Jordan. Operationalizing counterfactual metrics: Incentives, ranking, and information asymmetry. *arXiv preprint arXiv:2305.14595*, 2023.
- 53 Roger J-B Wets and Christoph Witzgall. Algorithms for frames and lineality spaces of cones. *Journal of Research of the National Bureau of Standards*, 71:1–7, 1967.
- 54 Lawrence J White. Credit rating agencies: An overview. *Annu. Rev. Financ. Econ.*, 5(1):93–122, 2013.
- 55 Lofti Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE transactions on Automatic Control*, 8(1):59–60, 1963.

A Omitted Proofs

A.1 Minimal design problem for improvement objective

► **Theorem A.1** (Theorem 2). *Let columns of \mathbf{Z} be an orthonormal basis of linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. For each design restriction, there exists $S : \mathcal{F} \rightarrow \mathbb{R}^k$, designed using Algorithm 1, that satisfies the improvement objective with the following dimensionalities.*

	Dimensionality $k \geq$
Res-CS	$\text{ConeSubsetRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_{\mathbf{Z}} = \mathcal{K}_{\mathbf{V}} \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r} \text{ s.t. } \mathbf{V} \subseteq \mathbf{Z}\}$
Res-LM	$\text{ConeGeneratingRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_{\mathbf{Z}} = \mathcal{K}_{\mathbf{V}} \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r}\}$
Res-L	$\text{ConeRank}(\mathbf{Z}) := \min_q \{q \mid \mathcal{K}_{\mathbf{Z}} \subseteq \mathcal{K}_{\mathbf{V}} \text{ for some } \mathbf{V} \in \mathbb{R}^{q \times r}\}$

Proof. We give a proof for the Res-CS restriction; proofs for the other two restrictions are similar. We show that, if $k \geq \text{CSR}(\mathbf{Z})$, then there exists $S(\mathbf{f}) = \mathbf{A}\mathbf{f}$ satisfying improvement and Res-CS.

8:16 Score Design for Multi-Criteria Incentivization

Let columns of $\mathbf{Z} \in \mathbb{R}^{d \times r}$ be an orthonormal basis of r -dimensional linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. The definition of CSR states that $k \geq \text{CSR}(\mathbf{Z})$ when there exists $\mathbf{V} \in \mathbb{R}^{k \times r}$ such that (i) $\mathbf{V} \subseteq \mathbf{Z}$ and (ii) $\mathcal{K}_Z = \mathcal{K}_V$. Property (i) means that $\mathbf{V} = \mathbf{AZ}$ for some $\mathbf{A} \in \mathbb{R}^{k \times d}$ with 1-hot rows, and so $S(\mathbf{f}) = \mathbf{A}\mathbf{f}$ satisfies the Res-CS restriction. Property (ii) implies that $\mathcal{K}_Z \subseteq \mathcal{K}_V$, and so S satisfies improvement:

$$\mathcal{K}_Z \subseteq \mathcal{K}_V \xLeftrightarrow{\text{Lem. B.2}} \mathcal{L} \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^* \xrightarrow{\text{Def. 1}} \text{for all } \mathbf{f} \in \mathcal{F}, \mathcal{F}_f \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^* \xLeftrightarrow{\text{Eq. 4}} \text{Improvement.} \quad (6)$$

The proof of Lemma B.2 uses $\mathbf{V} = \mathbf{AZ}$, and the projection of rows of \mathbf{A} and \mathbf{I}_d in subspace \mathcal{L} using orthonormal basis \mathbf{Z} . ◀

► **Example A.2** (Competing metric improvement directions \implies dimensionality for Res-CS $>$ Res-LM). When cone \mathcal{K}_Z generated by rows of \mathbf{Z} is non-pointed, we have $\text{CSR}(\mathbf{Z}) > \text{CGR}(\mathbf{Z})$, implying that the score design dimensionality is higher under Res-CS restriction than under Res-LM. The cone \mathcal{K}_Z can be non-pointed in the presence of competing metric improvement directions, i.e., when improving on one metric degrades another. A non-pointed \mathcal{K}_Z results in a gap between $\text{CSR}(\mathbf{Z})$ and $\text{CGR}(\mathbf{Z})$.

Consider 8 metrics lying in a 5-dimensional subspace, which has the following orthonormal basis (arranged as columns of \mathbf{Z}):

$$\mathbf{Z} = \frac{1}{2} \cdot \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{8 \times 5}.$$

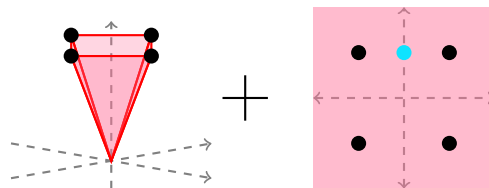
The rows generate a 5-dimensional cone \mathcal{K}_Z with two orthogonal parts: (i) a 2-dimensional linear subspace due to the first 4 metrics, and (ii) a 3-dimensional “square” pointed cone due to the last 4 metrics, as visualized in Figure 4. Since \mathcal{K}_Z contains a 2-dimensional linear subspace within, it is a non-pointed cone.

A matrix \mathbf{V} that attains $\text{CSR}(\mathbf{Z})$ must have rows of \mathbf{V} chosen from rows of \mathbf{Z} and $\mathcal{K}_Z = \mathcal{K}_V$. Excluding any row of \mathbf{Z} shrinks the generated cone – excluding any row of the first 4 generates a halfspace rather than the 2-dimensional subspace, and excluding any row of the last 4 does not generate the “square” pointed cone. So $\text{CSR}(\mathbf{Z}) = 8$. On the other hand, a matrix \mathbf{V} that attains $\text{CGR}(\mathbf{Z})$ need not have rows of \mathbf{V} chosen from rows of \mathbf{Z} ; \mathbf{V} must only satisfy $\mathcal{K}_Z = \mathcal{K}_V$. We need all last 4 rows to generate the “square” cone, but there exists 3 points (the blue and two bottom black points) whose nonnegative combinations generate the 2-dimensional linear subspace. So $\text{CGR}(\mathbf{Z}) = 7$.

► **Theorem A.3** (Theorem 7). *Assume metrics $\mathcal{F} \subseteq \mathbb{R}^d$ have non-empty relative interior with respect to $\text{aff}(\mathcal{F})$. Then the listed dimensionalities k in Theorem 2 are necessary.*

Proof. We give a proof for the Res-CS restriction; proof for the other two restrictions are similar. We show that, when \mathcal{F} has non-empty relative interior, we get:

$$\text{for all } \mathbf{f} \in \mathcal{F}, \quad \mathcal{F}_f \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^* \implies \mathcal{L} \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^*. \quad (7)$$



■ **Figure 4** A 5-dimensional non-pointed cone \mathcal{K}_Z with two orthogonal components: a 2-dimensional linear subspace, and a 3-dimensional “square” pointed cone.

By adding this implication to Equation (6), we prove that, when \mathcal{F} has non-empty relative interior, a score function S satisfies the improvement objective and Res-CS restriction *if and only if* $k \geq \text{CSR}(\mathbf{Z})$.

We now prove the implication in Equation (7). Let $\mathbf{x} \in \mathcal{L} \cap \mathcal{K}_A^*$. Since \mathcal{F} has non-empty relative interior, there exists \mathbf{f}^* in the relative interior. Lemma B.3 states that, as $\mathbf{x} \in \mathcal{L}$, there exists $a > 0$ such that $a\mathbf{x} \in \mathcal{F}_{\mathbf{f}^*}$. Since \mathbf{x} is in cone \mathcal{K}_A^* as well, we have $a\mathbf{x} \in \mathcal{K}_A^*$. Hence, $a\mathbf{x} \in \mathcal{F}_{\mathbf{f}^*} \cap \mathcal{K}_A^*$. According to the premise of Equation (7), we know that $\mathcal{F}_{\mathbf{f}^*} \cap \mathcal{K}_A^* \subseteq \mathcal{K}_I^*$, and so $a\mathbf{x} \in \mathcal{K}_I^*$. As $a > 0$, we get $\mathbf{x} \in \mathcal{K}_I^*$, completing the proof. ◀

► **Lemma A.4.** *Given affine subspace \mathcal{H} containing \mathcal{F} , the matrix ranks are invariant to the choice of orthonormal basis of $\mathcal{L}_{\mathcal{H}}$. Moreover, among all affine subspaces containing \mathcal{F} , the matrix ranks are smallest for $\mathcal{H} = \text{aff}(\mathcal{F})$.*

Proof. We give a proof for CSR, proofs for the other two matrix ranks are similar.

1. We first give a geometric interpretation for invariance to choice of orthonormal basis of $\mathcal{L}_{\mathcal{H}}$. Then we give an algebraic proof.

Geometric interpretation. For any matrix \mathbf{W} , note that $\text{CSR}(\mathbf{W})$ is the minimum cardinality of a subset V of W (set of rows of \mathbf{W}), such that cone \mathcal{K}_V encloses \mathcal{K}_W . By rotating rows of \mathbf{W} without altering the column span of \mathbf{W} , although the row vectors W change, the *relative position of them with respect to each other is the same*. So the cone generated by the rotated vectors is just a rotation of cone \mathcal{K}_W . As a result, the minimum cardinality of a subset of rotated vectors (to enclose the rotated cone) is unchanged, and so $\text{CSR}(\mathbf{W})$ is unchanged.

Algebraic argument. Let columns of \mathbf{Z}_1 and \mathbf{Z}_2 be two sets of orthonormal basis of $r_{\mathcal{H}}$ -dimensional $\mathcal{L}_{\mathcal{H}}$. We will show that $\text{CSR}(\mathbf{Z}_1) = \text{CSR}(\mathbf{Z}_2)$. The two orthonormal bases have the same column span, and are rotations/reflections of each other. So there exists orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{r_{\mathcal{H}} \times r_{\mathcal{H}}}$ such that $\mathbf{Z}_1 = \mathbf{Z}_2\mathbf{Q}$ and $\mathbf{Z}_1\mathbf{Q}^{\top} = \mathbf{Z}_2$.

We prove that $\text{CSR}(\mathbf{Z}_1) \leq \text{CSR}(\mathbf{Z}_2)$. Let $\text{CSR}(\mathbf{Z}_2) = k^*$. Then there exists $\mathbf{V}_2 \in \mathbb{R}^{k^* \times r_{\mathcal{H}}}$ such that $\mathbf{V}_2 \subseteq \mathbf{Z}_2$ and $\mathcal{K}_{\mathbf{Z}_2} \subseteq \mathcal{K}_{\mathbf{V}_2}$. These two properties mean that $\mathbf{V}_2 = \mathbf{A}\mathbf{Z}_2$ for some \mathbf{A} with 1-hot rows, and $\mathbf{Z}_2 = \mathbf{B}\mathbf{V}_2$ for some nonnegative \mathbf{B} . Multiplying with \mathbf{Q} on the right, we get $\mathbf{V}_2\mathbf{Q} = \mathbf{A}\mathbf{Z}_2\mathbf{Q}$ and $\mathbf{Z}_2\mathbf{Q} = \mathbf{B}\mathbf{V}_2\mathbf{Q}$. Therefore, $\mathbf{V}_1 = \mathbf{V}_2\mathbf{Q} \in \mathbb{R}^{k^* \times r_{\mathcal{H}}}$ has the properties $\mathbf{V}_1 \subseteq \mathbf{Z}_1$ and $\mathcal{K}_{\mathbf{Z}_1} \subseteq \mathcal{K}_{\mathbf{V}_1}$. This proves that $\text{CSR}(\mathbf{Z}_1) \leq \text{CSR}(\mathbf{Z}_2)$. With a symmetric argument, we also get $\text{CSR}(\mathbf{Z}_1) \geq \text{CSR}(\mathbf{Z}_2)$.

2. Let \mathcal{H}_1 and \mathcal{H}_2 be two non-empty affine subspaces containing \mathcal{F} such that $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Let \mathcal{L}_1 and \mathcal{L}_2 be linear subspaces corresponding to \mathcal{H}_1 and \mathcal{H}_2 respectively. Since $\mathcal{H}_1 \subseteq \mathcal{H}_2$ and for any $\mathbf{f} \in \mathcal{H}_1$ we can write $\mathcal{L}_1 = \mathcal{H}_1 - \mathbf{f}$ and $\mathcal{L}_2 = \mathcal{H}_2 - \mathbf{f}$, we find that $\mathcal{L}_1 \subseteq \mathcal{L}_2$. According to statement (1), CSR is invariant to the choice of orthonormal basis of linear subspace. Hence, pick columns of \mathbf{Z}_1 and \mathbf{Z}_2 as orthonormal basis of \mathcal{L}_1 and \mathcal{L}_2

respectively, such that columns of \mathbf{Z}_2 are a superset of columns of \mathbf{Z}_1 . In the definition of CSR, adding vectors to \mathbf{Z}_1 only increases the number of constraints to satisfy, and so CSR can only grow. Hence, $\text{CSR}(\mathbf{Z}_1) \leq \text{CSR}(\mathbf{Z}_2)$.

Since $\text{aff}(\mathcal{F})$ is the unique intersection of all affine subspaces containing \mathcal{F} , we have $\text{aff}(\mathcal{F}) \subseteq \mathcal{H}$ for every affine subspace \mathcal{H} containing \mathcal{F} . Thus, $\text{CSR}(\mathbf{Z}) \leq \text{CSR}(\mathbf{Z}_{\mathcal{H}})$, where columns of \mathbf{Z} and $\mathbf{Z}_{\mathcal{H}}$ are orthonormal basis of linear subspaces corresponding to $\text{aff}(\mathcal{F})$ and \mathcal{H} respectively. \blacktriangleleft

► **Proposition A.5.** *For each design restriction, there exists $\mathcal{F} \subseteq \mathbb{R}^d$ with $\dim \text{aff}(\mathcal{F}) = d$ and empty relative interior such that there exists function $S : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies improvement objective.*

Proof. We first give an example of $\mathcal{F} \subseteq \mathbb{R}^2$, and show that there exists $S : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies improvement and the Res-CS restriction. So S will also satisfy the other two design restrictions.

Consider $\mathcal{F} = \{(0, 0), (1, 1), (2, 3)\} \subseteq \mathbb{R}^2$ and let $\mathbf{A} = [1, 0] \in \mathbb{R}^{1 \times 2}$. We now argue that $S(\mathbf{f}) = \mathbf{A}\mathbf{f}$ satisfies the improvement objective. For metric pairs

$$(\mathbf{f}', \mathbf{f}) \in \{((1, 1), (0, 0)), ((2, 3), (1, 1)), ((2, 3), (0, 0))\}$$

we have $\mathbf{A}\mathbf{f}' \geq \mathbf{A}\mathbf{f}$ and $\mathbf{f}' \geq \mathbf{f}$. Hence, improvement objective holds for these pairs. Whereas for metric pairs

$$(\mathbf{f}', \mathbf{f}) \in \{((0, 0), (1, 1)), ((1, 1), (2, 3)), ((0, 0), (2, 3))\}$$

the left-hand side of the implication ($\mathbf{A}\mathbf{f}' \geq \mathbf{A}\mathbf{f}$) is not true. And so improvement objective holds for these pairs *vacuously*. Thus for all $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$ if $\mathbf{A}\mathbf{f}' \geq \mathbf{A}\mathbf{f}$ then $\mathbf{f}' \geq \mathbf{f}$.

We now give a counterexample of $d + 1$ points in $\mathcal{F} \subseteq \mathbb{R}^d$. Let $\mathbf{f}^{(0)} = \mathbf{0}_d$ and $\mathbf{f}^{(1)} = \mathbf{1}_d$. For $i = 2, \dots, d$, construct $\mathbf{f}_j^{(i)} = \left(\mathbf{f}_j^{(i-1)}\right)^2 + j$ for each coordinate $j \in [d]$. For example, the construction in \mathbb{R}^4 is:

$$\mathcal{F} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix} 5 \\ 11 \\ 19 \\ 29 \end{pmatrix}, \begin{pmatrix} 26 \\ 123 \\ 364 \\ 845 \end{pmatrix} \right\}$$

Points $\mathbf{f}^{(1)}, \dots, \mathbf{f}^{(d)}$ are linearly independent, and so $\dim \text{span}(\mathcal{F}) = d$. Let $\mathbf{A} = [1, 0, \dots, 0] \in \mathbb{R}^{1 \times d}$. Following a similar argument as the $d = 2$ case, we find that $S(\mathbf{f}) = \mathbf{A}\mathbf{f}$ satisfies the improvement objective (with dimensionality $k = 1$). \blacktriangleleft

A.2 Minimal design problem for optimality objective

► **Theorem A.6** (Theorem 11). *For each design restriction, there exists $S : \mathcal{F} \rightarrow \mathbb{R}^k$, designed using Algorithm 2, that satisfies the optimality objective with the following dimensionalities.*

	Dimensionality $k \geq$
Res-CS	$\dim \text{aff}(\mathcal{F})$
Res-LM	1
Res-L	1

Proof. For the last two design restrictions, the minimal design is straightforward. Using any vector $\mathbf{a} > \mathbf{0}$ of positive entries, design $S : \mathbf{f} \mapsto \mathbf{a} \cdot \mathbf{f}$ [55]. Clearly, S is linear in \mathbf{f} . To see that S is also monotone, fix $\mathbf{f}, \mathbf{f}' \in \mathcal{F}$ such that $\mathbf{f} \geq \mathbf{f}'$. Taking inner product

with positive vector \mathbf{a} , we get $\mathbf{a} \cdot \mathbf{f} \geq \mathbf{a} \cdot \mathbf{f}'$. To see that optimality objective is satisfied, fix $\mathbf{f}^* \in \text{ParetoOpt}(S)$. Since S is 1-dimensional, by definition of $\text{ParetoOpt}(S)$, we have $\mathbf{a} \cdot \mathbf{f}^* \geq \mathbf{a} \cdot \mathbf{f}$ for all $\mathbf{f} \in \mathcal{F}$. Since \mathbf{a} only has positive elements, for any $\mathbf{f} \in \mathcal{F}$ either $\mathbf{f}^* = \mathbf{f}$ or there exists $j \in [d]$ such that $\mathbf{f}_j^* > \mathbf{f}_j$. Therefore, $\mathbf{f}^* \in \text{ParetoOpt}(\mathcal{F})$.

Res-CS restriction. We now give a design for the Res-CS restriction. We first simplify the optimality objective – $\text{ParetoOpt}(S) \subseteq \text{ParetoOpt}(\mathcal{F})$ using movement directions $\mathcal{F}_{\mathbf{f}} = \{\mathbf{g} = \mathbf{f}' - \mathbf{f} \in \mathbb{R}^d \mid \text{for all } \mathbf{f}' \in \mathcal{F}\}$, definitions of dual cones \mathcal{K}_A^* and \mathcal{K}_I^* , and $\ker \mathbf{A} = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$. We rewrite $\text{ParetoOpt}(S)$ as follows:

$$\begin{aligned} \text{ParetoOpt}(S) &= \{\mathbf{f} \in \mathcal{F} \mid \text{for all } \mathbf{g} \in \mathcal{F}_{\mathbf{f}}, \text{ either } \mathbf{A}\mathbf{g} \not\geq \mathbf{0} \text{ or } \mathbf{A}\mathbf{g} = \mathbf{0}\} \\ &= \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_A^*)^c \cup \ker \mathbf{A}\}. \end{aligned}$$

Similarly, $\text{ParetoOpt}(\mathcal{F}) = \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_I^*)^c \cup \ker \mathbf{I}\}$. Thus we get:

$$\text{Optimality} \iff \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_A^*)^c \cup \ker \mathbf{A}\} \subseteq \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_I^*)^c \cup \ker \mathbf{I}\}. \quad (\text{Eq. 5})$$

We now identify an isomorphism between movement directions $\mathcal{F}_{\mathbf{f}}$ in the ambient space and the coefficient space. Let columns of $\mathbf{Z} \in \mathbb{R}^{d \times r}$ be an orthonormal basis of r -dimensional linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. Fix any $\mathbf{f} \in \mathcal{F}$. Denote with $\mathcal{C}_{\mathbf{f}} \in \mathbb{R}^r$ the set of coefficients of $\mathcal{F}_{\mathbf{f}}$ w.r.t. orthonormal basis \mathbf{Z} , i.e., $\mathcal{C}_{\mathbf{f}} = \mathbf{Z}^\top(\mathcal{F}_{\mathbf{f}})$. This introduces an isomorphism between the sets $\mathcal{F}_{\mathbf{f}}$ and $\mathcal{C}_{\mathbf{f}}$, i.e., for every $\mathbf{g} \in \mathcal{F}_{\mathbf{f}}$ there exists unique $\mathbf{d} \in \mathcal{C}_{\mathbf{f}}$ such that $\mathbf{g} = \mathbf{Z}\mathbf{d}$. With $\mathbf{V} = \mathbf{A}\mathbf{Z}$, we have four equivalences:

$$\begin{aligned} \mathbf{A}\mathbf{g} \geq \mathbf{0} &\iff \mathbf{V}\mathbf{d} \geq \mathbf{0} & \text{and} & \quad \mathbf{A}\mathbf{g} = \mathbf{0} \iff \mathbf{V}\mathbf{d} = \mathbf{0}, \\ \mathbf{g} \geq \mathbf{0} &\iff \mathbf{Z}\mathbf{d} \geq \mathbf{0} & \text{and} & \quad \mathbf{g} = \mathbf{0} \iff \mathbf{Z}\mathbf{d} = \mathbf{0}. \end{aligned}$$

Lemma B.4 uses these equivalences to state that for any $\mathbf{f} \in \mathcal{F}$, we have

$$\mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_A^*)^c \cup \ker \mathbf{A} \iff \mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_V^*)^c \cup \ker \mathbf{V} \quad (8)$$

$$\mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_I^*)^c \cup \ker \mathbf{I} \iff \mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_Z^*)^c \cup \ker \mathbf{Z}. \quad (9)$$

We further simplify the optimality objective (Equation (5)):

$$\text{Optimality} \iff \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_A^*)^c \cup \ker \mathbf{A}\} \subseteq \{\mathbf{f} \in \mathcal{F} \mid \mathcal{F}_{\mathbf{f}} \subseteq (\mathcal{K}_I^*)^c \cup \ker \mathbf{I}\} \quad (10)$$

$$\iff \{\mathbf{f} \in \mathcal{F} \mid \mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_V^*)^c \cup \ker \mathbf{V}\} \subseteq \{\mathbf{f} \in \mathcal{F} \mid \mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_Z^*)^c \cup \ker \mathbf{Z}\} \quad (11)$$

where Equation (11) follows from Lemma B.4.

Now, we choose r linear independent rows of \mathbf{Z} to create $\mathbf{V} \in \mathbb{R}^{r \times r}$. Since \mathbf{Z} has orthonormal columns, we have $\ker \mathbf{V} = \ker \mathbf{Z} = \{\mathbf{0}\}$. Moreover, we have $\mathbf{V} \subseteq \mathbf{Z}$, implying $\mathcal{K}_V \subseteq \mathcal{K}_Z$ and $\mathcal{K}_Z^* \subseteq \mathcal{K}_V^*$ (Lemma B.1). This shows that $\mathcal{K}_Z^* \cup (\ker \mathbf{Z})^c \subseteq \mathcal{K}_V^* \cup (\ker \mathbf{V})^c$. As a result, $(\mathcal{K}_V^*)^c \cup \ker \mathbf{V} \subseteq (\mathcal{K}_Z^*)^c \cup \ker \mathbf{Z}$. Hence, for any $\mathbf{f} \in \mathcal{F}$ for which $\mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_V^*)^c \cup \ker \mathbf{V}$, we also have $\mathcal{C}_{\mathbf{f}} \subseteq (\mathcal{K}_Z^*)^c \cup \ker \mathbf{Z}$. This shows that Equation (11) holds with the proposed choice of \mathbf{V} . As $\mathbf{V} = \mathbf{A}\mathbf{Z}$ for \mathbf{A} with 1-hot rows, this design satisfies optimality and Res-CS restriction. \blacktriangleleft

► Proposition A.7 (Proposition 12). *Consider designing $S : \mathcal{F} \rightarrow \mathbb{R}^k$ to satisfy optimality objective.*

1. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_1 \leq 1\}$, $k \geq 1$ is necessary and sufficient for all design restrictions.
2. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_2 \leq 1\}$, $k \geq 1$ is necessary and sufficient for all design restrictions.
3. For $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_\infty \leq 1\}$, $k \geq d$ is necessary and sufficient for Res-CS. Moreover, $k \geq 1$ is necessary and sufficient for the Res-LM and Res-L restrictions.

Proof. Theorem 11 states $k \geq 1$ is sufficient for Res-LM and Res-L restrictions for any \mathcal{F} ; trivially, $k \geq 1$ is necessary. So, we prove the claims for the Res-CS restriction. For the stated sets \mathcal{F} , we determine $\text{ParetoOpt}(\mathcal{F})$ and discuss choice of S to satisfy $\text{ParetoOpt}(S) \subseteq \text{ParetoOpt}(\mathcal{F})$.

We denote the d coordinates of metric value $\mathbf{f} \in \mathcal{F}$ with $\mathbf{f}_1, \dots, \mathbf{f}_d$. Let \mathbf{e}_j be the j^{th} canonical basis vector of \mathbb{R}^d . We denote the unit ℓ_p -norm ball with $\mathbb{B}_p^d = \{\mathbf{f} \in \mathbb{R}^d \mid \|\mathbf{f}\|_p \leq 1\}$.

1. Let $\mathcal{F} = \mathbb{B}_1^d$, the unit ℓ_1 -norm ball centered at the origin. Note that the j^{th} coordinate of metric value \mathbf{f}_j is maximized when $\mathbf{f} = \mathbf{e}_j$. So vectors $\mathbf{e}_1, \dots, \mathbf{e}_d$ are pareto-optimal w.r.t. \mathcal{F} . In fact, all vectors on the surface of \mathbb{B}_1^d in the nonnegative orthant are pareto-optimal w.r.t. \mathcal{F} . That is, $\text{ParetoOpt}(\mathcal{F}) = \{\mathbf{f} \in \mathbb{R}_+^d \mid \mathbf{1}_d \cdot \mathbf{f} = 1\}$.

We choose any coordinate $j \in [d]$ and design 1-dimensional $S(\mathbf{f}) = \mathbf{f}_j$. Since \mathcal{F} is the unit ℓ_1 -norm ball, $\text{ParetoOpt}(S) = \{\mathbf{e}_j\}$, which a subset of $\text{ParetoOpt}(\mathcal{F})$ as $\mathbf{1}_d \cdot \mathbf{e}_j = 1$. Hence, this design with dimensionality $k = 1$ satisfies the optimality objective under Res-CS restriction.

Trivially, $k \geq 1$ is necessary as well.

2. Let $\mathcal{F} = \mathbb{B}_2^d$, the unit L_2 -ball centered at the origin. Note that the j^{th} coordinate of metric value \mathbf{f}_j is maximized when $\mathbf{f} = \mathbf{e}_j$. So vectors $\mathbf{e}_1, \dots, \mathbf{e}_d$ are pareto-optimal w.r.t. \mathcal{F} . In fact, all vectors on the unit shell in the nonnegative orthant are pareto-optimal w.r.t. \mathcal{F} . That is, $\text{ParetoOpt}(\mathcal{F}) = \mathbb{S}_2^{d-1} \cap \mathbb{R}_+^d = \mathbb{S}_2^{d-1} \cap \mathcal{K}_I$ where \mathbf{I} is the identity matrix. We can similarly determine pareto-optimal points w.r.t. $S(\mathbf{f}) = \mathbf{A}\mathbf{f}$. Let \mathbf{A} have k rows $\mathbf{A} = [\mathbf{a}_1; \dots; \mathbf{a}_k] \in \mathbb{R}^{k \times d}$. The i^{th} coordinate of S is maximized when $\mathbf{f} = \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}$. So vectors $\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \dots, \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|_2}$ are pareto-optimal w.r.t. S . In fact, all vectors on the unit shell and cone \mathcal{K}_A generated by rows of \mathbf{A} are pareto-optimal w.r.t. S . That is, $\text{ParetoOpt}(S) = \mathbb{S}_2^{d-1} \cap \mathcal{K}_A$.

So S satisfies optimality if $\mathbb{S}_2^{d-1} \cap \mathcal{K}_A \subseteq \mathbb{S}_2^{d-1} \cap \mathcal{K}_I$. Any matrix $\mathbf{A} \subseteq \mathbf{I}_d$ implies $\mathcal{K}_A \subseteq \mathcal{K}_I$. Hence, we can choose any coordinate $j \in [d]$ and construct 1-dimensional $S(\mathbf{f}) = \mathbf{f}_j$. This design with dimensionality $k = 1$ satisfies the optimality objective under Res-CS restriction.

Trivially, $k \geq 1$ is necessary as well.

3. Let $\mathcal{F} = \mathbb{B}_\infty^d$, the unit L_∞ -ball centered at the origin. It is easy to see that $\text{ParetoOpt}(\mathcal{F}) = \{\mathbf{1}_d\}$, a singleton set.

Under the Res-CS restriction, $S : \mathcal{F} \rightarrow \mathbb{R}^k$ is such that $S(\mathbf{f}) = [\mathbf{f}_{i_1}; \dots; \mathbf{f}_{i_k}]$ where the every index $i_j \in [d]$. Let I be the set of unique indices. We will now show that if $k < d$, then there does not exist score function S that satisfies optimality. Since $k < d$, we have $|I| < d$. The point $\mathbf{f} \in \mathbb{B}_\infty^d$ is pareto-optimal w.r.t. S if $\mathbf{f}_i = 1$ for every $i \in I$. Precisely, $\text{ParetoOpt}(S) = \{\mathbf{f} \in [-1, 1]^d \mid \mathbf{f}_i = 1 \text{ for all } i \in I\}$. Since there exists $j \in [d]$ that is not in I , $\text{ParetoOpt}(S)$ contains points with $\mathbf{f}_j = -1$. Hence, $\text{ParetoOpt}(S)$ is not a subset of $\text{ParetoOpt}(\mathcal{F})$. Therefore, for $\mathcal{F} = \mathbb{B}_\infty^d$ and $k < d$ it is not possible to design $S : \mathcal{F} \rightarrow \mathbb{R}^d$ that satisfies optimality objective under Res-CS restriction.

Trivially, $k = d$ is sufficient to satisfy the optimality objective under Res-CS restriction: design $S(\mathbf{f}) = \mathbf{f}$. Hence, $k \geq d$ is both necessary and sufficient when $\mathcal{F} = \mathbb{B}_\infty^d$. ◀

A.3 Minimal design problem for both objectives simultaneously

► **Corollary A.8.** *Let columns of \mathbf{Z} be an orthonormal basis of linear subspace \mathcal{L} associated with $\text{aff}(\mathcal{F})$. For each design restriction, there exists score function $S : \mathcal{F} \rightarrow \mathbb{R}^k$ that simultaneously satisfies improvement and optimality objectives with following dimensionalities.*

	Dimensionality $k \geq$
Res-CS	ConeSubsetRank(\mathbf{Z})
Res-LM	ConeGeneratingRank(\mathbf{Z})
Res-L	ConeGeneratingRank(\mathbf{Z})

Moreover, for Res-CS and Res-LM restrictions, the score design is minimal when \mathcal{F} has non-empty relative interior.

Proof. For the first two restrictions (Res-CS and Res-LM), S is monotone in \mathcal{F} . So, Theorems 2 and 13 immediately give the design for simultaneously satisfying both objectives with dimensionality $k = \text{CSR}(\mathbf{Z})$ and $\text{CGR}(\mathbf{Z})$ respectively. Theorem 7 proves the minimality of this design. The design for Res-LM restriction also applies for the Res-L restriction. \blacktriangleleft

B Technical Lemmas

► **Lemma B.1.** For two polyhedral cones \mathcal{K}_1 and \mathcal{K}_2 , we have $\mathcal{K}_1 \subseteq \mathcal{K}_2 \iff \mathcal{K}_2^* \subseteq \mathcal{K}_1^*$.

Proof. Since the two cones are polyhedral, they are closed and convex. For any closed and convex cone \mathcal{K} , the dual of its dual cone is the cone itself: $\mathcal{K}^{**} = \mathcal{K}$. The result then follows from the fact that for any two convex cones $\mathcal{K}_1 \subseteq \mathcal{K}_2 \implies \mathcal{K}_2^* \subseteq \mathcal{K}_1^*$ [10, Sec. 2.6.1]. \blacktriangleleft

► **Lemma B.2.** Let $\mathcal{L} \subseteq \mathbb{R}^d$ be an r -dimensional linear subspace, and let columns of $\mathbf{Z} \in \mathbb{R}^{d \times r}$ be an orthonormal basis of \mathcal{L} . Let \mathcal{K}_{A_1} and \mathcal{K}_{A_2} be cones in \mathbb{R}^d generated by rows of matrices $\mathbf{A}_1 \in \mathbb{R}^{m_1 \times d}$ and $\mathbf{A}_2 \in \mathbb{R}^{m_2 \times d}$ respectively. With $\mathbf{V}_1 = \mathbf{A}_1 \mathbf{Z}$ and $\mathbf{V}_2 = \mathbf{A}_2 \mathbf{Z}$, we have,

$$\mathcal{L} \cap \mathcal{K}_{A_1}^* \subseteq \mathcal{K}_{A_2}^* \iff \mathcal{K}_{V_1}^* \subseteq \mathcal{K}_{V_2}^* \iff \mathcal{K}_{V_2} \subseteq \mathcal{K}_{V_1}.$$

Proof. We can simplify this condition $\mathcal{L} \cap \mathcal{K}_{A_1}^* \subseteq \mathcal{K}_{A_2}^*$ further by expressing vectors in the basis \mathbf{Z} .

First, every $\mathbf{x} \in \mathcal{L}$ has a unique representation in the basis \mathbf{Z} . That is, $\mathbf{x} = \mathbf{Z}\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^r$. Second, every d -dimensional row \mathbf{a} of \mathbf{A}_1 and \mathbf{A}_2 can be written as $\mathbf{a}^{\parallel} + \mathbf{a}^{\perp}$, where $\mathbf{a}^{\parallel} = \mathbf{a}\mathbf{Z}\mathbf{Z}^{\top} \in \mathcal{L}$ and $\mathbf{a}^{\perp} = \mathbf{a}(\mathbf{I} - \mathbf{Z}\mathbf{Z}^{\top}) \in \mathcal{L}^{\perp}$. Therefore, $\mathbf{A}_1 = \mathbf{A}_1^{\parallel} + \mathbf{A}_1^{\perp}$ where $\mathbf{A}_1^{\parallel} = \mathbf{A}_1\mathbf{Z}\mathbf{Z}^{\top} + \mathbf{A}_1(\mathbf{I} - \mathbf{Z}\mathbf{Z}^{\top})$. Note that $\mathbf{A}_1^{\perp}\mathbf{Z} = \mathbf{0}_{m_1 \times r}$. Similarly we can decompose the matrix $\mathbf{A}_2 = \mathbf{A}_2^{\parallel} + \mathbf{A}_2^{\perp}$. Denote the coefficients as $\mathbf{V}_1 = \mathbf{A}_1\mathbf{Z}$ and $\mathbf{V}_2 = \mathbf{A}_2\mathbf{Z}$. Using these simplifications, we get:

$$\mathcal{L} \cap \mathcal{K}_{A_1}^* \subseteq \mathcal{K}_{A_2}^* \iff \text{for all } \mathbf{x} \in \mathcal{L}, \mathbf{A}_1\mathbf{x} \geq \mathbf{0} \implies \mathbf{A}_2\mathbf{x} \geq \mathbf{0} \tag{12}$$

$$\iff \text{for all } \mathbf{c} \in \mathbb{R}^r, \mathbf{A}_1\mathbf{Z}\mathbf{c} \geq \mathbf{0} \implies \mathbf{A}_2\mathbf{Z}\mathbf{c} \geq \mathbf{0} \tag{13}$$

$$\iff \text{for all } \mathbf{c}, (\mathbf{A}_1^{\parallel} + \mathbf{A}_1^{\perp})\mathbf{Z}\mathbf{c} \geq \mathbf{0} \implies (\mathbf{A}_2^{\parallel} + \mathbf{A}_2^{\perp})\mathbf{Z}\mathbf{c} \geq \mathbf{0} \tag{14}$$

$$\iff \text{for all } \mathbf{c}, \mathbf{V}_1\mathbf{Z}^{\top}\mathbf{Z}\mathbf{c} \geq \mathbf{0} \implies \mathbf{V}_2\mathbf{Z}^{\top}\mathbf{Z}\mathbf{c} \geq \mathbf{0} \tag{15}$$

$$\iff \text{for all } \mathbf{c}, \mathbf{V}_1\mathbf{c} \geq \mathbf{0} \implies \mathbf{V}_2\mathbf{c} \geq \mathbf{0} \tag{16}$$

$$\iff \mathcal{K}_{V_1}^* \subseteq \mathcal{K}_{V_2}^* \tag{17}$$

$$\iff \mathcal{K}_{V_2} \subseteq \mathcal{K}_{V_1}. \tag{18}$$

where the last equivalence follows from Lemma B.1. \blacktriangleleft

► **Lemma B.3.** Let \mathcal{L} be the linear subspace corresponding to $\text{aff}(X)$. For any \mathbf{x}^* in the relative interior of X and any $\mathbf{x} \in \mathcal{L}$, there exists a $a > 0$ such that $a\mathbf{x} \in X_{\mathbf{x}^*}$.

8:22 Score Design for Multi-Criteria Incentivization

Proof. We use the definition of relative interior. Since \mathbf{x}^* is in relative interior of X , there exists $R > 0$ such that $(\mathbf{x}^* + R \cdot \mathbb{B}_2^d) \cap \text{aff}(X) \subseteq X$. Centering the sets at \mathbf{x}^* , there exists $R > 0$ such that $R \cdot \mathbb{B}_2^d \cap \text{aff}(X)_{\mathbf{x}^*} \subseteq X_{\mathbf{x}^*}$. We note that $\mathcal{L} = \text{aff}(X)_{\mathbf{x}^*}$.

Let $\mathbf{x} \in \mathcal{L}$. If $\mathbf{x} = \mathbf{0}$ then we are done as $a\mathbf{x} = \mathbf{0} \in X_{\mathbf{x}^*}$ for any $a > 0$. If \mathbf{x} is nonzero, then we can normalize it so that $\tilde{\mathbf{x}} = R \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|} \in R \cdot \mathbb{B}_2^d \cap \mathcal{L}$. From the definition of relative interior, we get that $\tilde{\mathbf{x}} \in X_{\mathbf{x}^*}$. Thus for any nonzero $\mathbf{x} \in \mathcal{L}$ there exists $a = R/\|\mathbf{x}\|$ such that $a\mathbf{x} \in X_{\mathbf{x}^*}$. ◀

► **Lemma B.4.** *Let \mathcal{L} be the linear subspace corresponding to r -dimensional $\text{aff}(X) \subseteq \mathbb{R}^d$, and let columns of $\mathbf{Z} \in \mathbb{R}^{d \times r}$ be an orthonormal basis of \mathcal{L} . For any $\mathbf{x} \in X$, denote with $\mathcal{C}_{\mathbf{x}} \subseteq \mathbb{R}^r$ the preimage of $X_{\mathbf{x}}$ under the orthonormal basis \mathbf{Z} . Let $\mathcal{K}_A \subseteq \mathbb{R}^d$ be generated by rows of $\mathbf{A} \in \mathbb{R}^{m \times d}$, and let $\mathbf{V} = \mathbf{AZ}$. Then for every $\mathbf{f} \in \mathcal{F}$,*

$$X_{\mathbf{x}} \cap \mathcal{K}_A^* \cap (\ker \mathbf{A})^c = \emptyset \iff \mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c = \emptyset.$$

Proof. Note that for every $\mathbf{x} \in X$, the linear subspace spanned by the set $X_{\mathbf{x}}$ is \mathcal{L} , and columns of \mathbf{Z} are an orthonormal basis of \mathcal{L} . That is, for every $\mathbf{y} \in X_{\mathbf{x}}$ there exists unique $\mathbf{d} \in \mathcal{C}_{\mathbf{x}}$ such that $\mathbf{y} = \mathbf{Zd}$. Moreover, we can decompose rows of \mathbf{A} in the linear subspace \mathcal{L} and its orthogonal complement \mathcal{L}^\perp , as in proof of Lemma B.2. We decompose $\mathbf{A} = \mathbf{AZZ}^\top + \mathbf{A}(\mathbf{I}_d - \mathbf{ZZ}^\top)$.

We use these decomposition results to prove the desired result. We first prove the forward direction by contradiction. Let $\mathbf{x} \in X$ and assume that $X_{\mathbf{x}} \cap \mathcal{K}_A^* \cap (\ker \mathbf{A})^c = \emptyset$. Now assume that there exists $\mathbf{d} \in \mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c$. So $\mathbf{Vd} \geq \mathbf{0}$ and $\mathbf{Vd} \neq \mathbf{0}$, implying that $\mathbf{AZd} \geq \mathbf{0}$ and $\mathbf{AZd} \neq \mathbf{0}$. Hence, there exists $\mathbf{y} = \mathbf{Zd} \in X_{\mathbf{x}}$ such that $\mathbf{y} \in \mathcal{K}_A^*$ and $\mathbf{y} \in (\ker \mathbf{A})^c$. This contradicts our assumption that $X_{\mathbf{x}} \cap \mathcal{K}_A^* \cap (\ker \mathbf{A})^c = \emptyset$, and so we must have $\mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c = \emptyset$.

We also prove the backward direction by contradiction. Let $\mathbf{x} \in X$ and assume that $\mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c = \emptyset$. Now assume that there exists $\mathbf{y} \in X_{\mathbf{x}} \cap \mathcal{K}_A^* \cap (\ker \mathbf{A})^c$. So $\mathbf{Ay} \geq \mathbf{0}$ and $\mathbf{Ay} \neq \mathbf{0}$. Using decomposition of rows of \mathbf{A} and \mathbf{y} in the basis \mathbf{Z} , we get that $\mathbf{Ay} = \mathbf{AZd}$ where $\mathbf{y} = \mathbf{Zd}$ for $\mathbf{d} \in \mathcal{C}_{\mathbf{x}}$. So there exists $\mathbf{d} \in \mathcal{C}_{\mathbf{x}}$ such that $\mathbf{AZd} \geq \mathbf{0}$ and $\mathbf{AZd} \neq \mathbf{0}$. Since $\mathbf{V} = \mathbf{AZ}$, we get that there exists $\mathbf{d} \in \mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c$. This contradicts our assumption that $\mathcal{C}_{\mathbf{x}} \cap \mathcal{K}_V^* \cap (\ker \mathbf{V})^c = \emptyset$, and so we must have $X_{\mathbf{x}} \cap \mathcal{K}_A^* \cap (\ker \mathbf{A})^c = \emptyset$. ◀

Privacy Can Arise Endogenously in an Economic System with Learning Agents

Nivasini Ananthkrishnan ✉

University of California, Berkeley, CA, USA

Tiffany Ding ✉

University of California, Berkeley, CA, USA

Mariel Werner ✉

University of California, Berkeley, CA, USA

Sai Praneeth Karimireddy ✉

University of California, Berkeley, CA, USA

Michael I. Jordan ✉

University of California, Berkeley, CA, USA

Abstract

We study price-discrimination games between buyers and a seller where privacy arises endogenously – that is, utility maximization yields equilibrium strategies where privacy occurs naturally. In this game, buyers with a high valuation for a good have an incentive to keep their valuation private, lest the seller charge them a higher price. This yields an equilibrium where some buyers will send a signal that misrepresents their type with some probability; we refer to this as *buyer-induced privacy*. When the seller is able to publicly commit to providing a certain privacy level, we find that their equilibrium response is to commit to ignore buyers’ signals with some positive probability; we refer to this as *seller-induced privacy*. We then turn our attention to a repeated interaction setting where the game parameters are unknown and the seller cannot credibly commit to a level of seller-induced privacy. In this setting, players must learn strategies based on information revealed in past rounds. We find that, even without commitment ability, seller-induced privacy arises as a result of reputation building. We characterize the resulting seller-induced privacy and seller’s utility under no-regret and no-policy-regret learning algorithms and verify these results through simulations.

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

Keywords and phrases Privacy, Game Theory, Online Learning, Price Discrimination

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.9

Related Version *Full Version:* <https://arxiv.org/abs/2404.10767>

Funding TD acknowledges support from the National Science Foundation Graduate Research Fellowship Program under grant no. 2146752, SPK is partially supported by a Mobility Fellowship by the Swiss National Science Foundation. We also acknowledge support from the European Union (ERC-2022-SYG-OCEAN-101071601).

Acknowledgements We thank Alireza Fallah and Stephen Bates for helpful discussions.

1 Introduction

The question of how to define and preserve privacy in the age of machine learning has been a topic of ongoing debate in the computer science and policy communities [11]. The widely accepted theoretical framework of differential privacy [8] formalizes privacy as the ability to



© Nivasini Ananthkrishnan, Tiffany Ding, Mariel Werner, Sai Praneeth Karimireddy, and Michael I. Jordan;

licensed under Creative Commons License CC-BY 4.0

5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 9; pp. 9:1–9:22



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

withstand membership inference attacks. That is, differential privacy ensures that the output of a computation obfuscates whether a particular data point was present in the input.

However, the practical implementations of differential privacy has been fraught with challenges. There has been significant debate around how to interpret the key privacy parameter ϵ and how to choose it [21]. This is especially true when data is continuously collected from users (what does it mean to have a guarantee of $\epsilon = 1$ *per data point* when a user’s data is continuously collected?) This has also led to controversies where companies have claimed their algorithms are private, when in fact the chosen ϵ value confers negligible protection [24]. Further complicating matters, there are multiple variants and extensions of differential privacy – e.g. (ϵ, δ) -DP [8], Reyni-DP [19], Gaussian-DP [7], etc. – each with different parameters and interpretations.

Perhaps more fundamentally, a growing body of work argues that the public’s understanding of privacy is drastically different from differential privacy [23, 18]. While differential privacy focuses on membership inference, privacy is more commonly understood to mean the prevention of the platform using one’s data in ways that are misaligned with the individual’s interests, such as price discrimination or other exploitative practices.

This work seeks to provide a new perspective on privacy that bridges the gap between the theoretical computer science view and the public’s intuitive understanding. We develop a game-theoretic model of privacy that allows us to analyze the effect of privacy choices on all the stakeholders. Additionally, the framework shows how to derive *optimal* privacy mechanisms that balance the gain in privacy with loss of accuracy in order to maximize net utility. In our model, a “principal” (e.g., a platform or seller) can observe signals from “agents” (e.g., users or buyers) and use this information to maximize its own profit, while the agents have an incentive to obfuscate their data to prevent exploitation. We focus on a price-discrimination setting involving interactions between buyers and sellers.

We show that “buyer-induced privacy” behavior, which resembles randomized response, arises endogenously as an equilibrium strategy. Furthermore, we find that the seller is often better off *committing* to not observing the agents’ data at all (“seller-induced privacy”), as the revenue loss from buyer-induced privacy can be substantial. Finally, we extend our analysis to a dynamic setting where the seller is a learning agent who interacts with multiple buyers over time. We demonstrate how a simple external auditing mechanism can implement the sellers’s commitment to privacy and lead to an equilibrium with endogenously arising privacy-preserving behavior.

Our results provide a new framework for understanding privacy that encompasses both the theoretical guarantees of differential privacy and the practical, user-centric notion of privacy. By modeling privacy as an emergent property of an economic system, we hope to offer insights that can inform the design of privacy-preserving platforms and policies.

Motivating example. In the absence of regulation, online retailers may price discriminate based on information they have collected about past purchases of the customers. Some customers may be willing to pay more for a good than others, perhaps due to innate preferences for certain types of good or because they have more disposable income. The retailer wants to identify customers with higher valuations and charge them higher prices in order to maximize their revenue.

Since customers are aware of the potential for price discrimination, they may engage in evasive action to protect their privacy. Customers may avoid choosing goods that signal their true preferences for less consequential purchases, e.g., a high-income customer choosing between an expensive water bottle that is slightly better than a cheaper option may opt to

buy the cheaper bottle in an attempt to obscure their income status. This evasive action imposes a cost on the customer, who misses out on buying their truly preferred product, and also on the retailer, who would have preferred to sell the more expensive product.

What are the behaviors that arise at equilibrium? What if the seller can credibly commit to not price discriminate? How do these behaviors change in more realistic settings where game parameters are not known and strategies must be learned based on past interactions? These are questions we answer in this paper.

1.1 Preview of contributions

We introduce a price-discrimination game in Definition 1 that involves buyers of two types – one with a high valuation and one with a low valuation of an item. A seller may potentially track buyers’ signals that reveal their valuations. We characterize the perfect Bayes Nash equilibrium of this game in Theorem 2 and show that a buyer-induced privacy mechanism emerges in the equilibrium. That is, the buyer with a high valuation, with some probability, chooses an evasive action to appear to have a low valuation.

We then introduce commitment ability for the seller wherein a seller can commit to not track buyers’ signals with some probability. In the price-discrimination game with commitment, the equilibrium response (Corollary 5) results in seller-induced privacy, which obviates the need for buyer-induced privacy. That is, with some probability, the seller chooses to commit to respect privacy and voluntarily does not track signals. Due to this privacy commitment from the seller, it is optimal for buyers to truthfully report their type. We call this seller-induced privacy the “commitment strategy” and denote the resulting utility \mathbb{U}_1^* .

In Section 3, we remove the seller’s commitment ability but give buyers access to the seller’s historical pricing. We model this as a repeated interaction between a seller and buyers with each buyer participating in only one round. The pricing history is used by buyers to construct the seller’s “reputation” (i.e., an estimate of the probability of price discrimination), which buyers then use to inform their signaling strategy. We model the buyers as using a reputation construction procedure that satisfies a consistency condition given in Definition 8, which requires that the reputation is able to differentiate between sellers employing price-discriminating strategies and non-price-discriminating strategies. In Proposition 10, we show the existence of such a reputation mechanism using the available history. We show that consistent reputation can yield seller-induced privacy (i.e., ignoring signals), depending on the model of the seller; we consider no-regret and no-policy-regret sellers. Our findings are:

1. With a no-regret seller, there could be no seller-induced privacy. That is, the seller can use signals and price discriminate in every round and still be no-regret (Proposition 13).
2. Regret minimization achieves strictly less average utility (asymptotically) than \mathbb{U}_1^* (Proposition 14).
3. Employing the commitment strategy in every round is a no-policy-regret algorithm for the seller (Proposition 20).
4. Employing the commitment strategy in every round ensures the seller (asymptotically) an average utility of \mathbb{U}_1^* . This is the highest possible average utility achievable (asymptotically) in the repeated interaction (Proposition 21).

1.2 Related work

Our work sits at the intersection of many areas, ranging from classical economics to online learning.

There is a vast literature on *privacy* in computer science studying mechanisms for notions of privacy such as differential privacy [8]. The mechanisms arising in our setting resemble mechanisms in these works. We observe local privacy (buyer-induced privacy) where users add noise to their data. We also observe central privacy (seller-induced privacy) where the platform ensures similar outcomes for different user data.

Literature in economics studies the economic implications of enacting privacy mechanisms (see [1] for a survey). Within this body of work, there is a literature on privacy and *price discrimination* (e.g., [2, 5, 20, 12]). We build on this work and extend to a setting that relaxes common-prior assumptions for buyers and sellers so that players must now devise strategies based on what they learn from repeated interactions.

In these repeated interactions, we observe the emergence of a *reputation-based privacy mechanism*. This reputation, learned by buyers based on previous interactions, takes the place of the prior that is used in the single-interaction game. There are numerous papers in economics on reputation focusing on sellers' reputations for the quality of the proffered good [15, 22, 9]. We focus on seller's reputation for enacting price discrimination and analyze how this arises in an online learning framework.

We also study the differences in behavior that arise from seller *commitment*, which has been studied in [14], [2], [12] and [16]. We show that even without commitment, similar behavior can arise through repeated interactions where reputation substitutes for the role of commitment.

Finally, we draw upon work on *online learning* and *repeated games*. There are a number of papers [4, 6, 13, 10] on repeated interactions between a principal and an agent where the agent chooses actions based on evolving beliefs about the principal's actions. In our setting, we interpret the evolving beliefs as the reputation of the principal. Our setting differs in two ways. The first is that the principal's actions are not revealed at the end of the round. Instead partial information about the action, depending on the agent's response, is revealed. The second is that our results hold for weaker conditions on the agent's beliefs compared to previous work.

2 A Price-Discrimination Game

We formulate price discrimination as a sequential, incomplete-information game between n buyers and a seller.

► **Definition 1** (PD game). *The price-discrimination game with parameters $n, \alpha, \mu, \bar{\theta}, \underline{\theta}, c_B, c_S$, denoted the $(n, \alpha, \mu, \bar{\theta}, \underline{\theta}, c_B, c_S)$ -PD game, has the following extensive-form representation.*

1. **Nature's move.** *The game begins with Nature assigning types to each participant according to random draws. For $i \in [n]$, the type for buyer i is $\theta_i \in \{\underline{\theta}, \bar{\theta}\}$, representing their valuation of the item being sold, with $\underline{\theta} < \bar{\theta}$. A buyer is type $\bar{\theta}$ with probability μ and type $\underline{\theta}$ with probability $1 - \mu$. The seller's type χ is either signal aware ($\chi = 1$) or signal blind ($\chi = 0$). The seller is signal aware with probability α and signal blind with probability $1 - \alpha$.*
2. **Signaling stage.** *Based on their assigned type θ_i , each buyer signals $s_i \in \{\underline{s}, \bar{s}\}$. Signaling one's true type (\underline{s} for type $\underline{\theta}$ and \bar{s} for type $\bar{\theta}$) incurs no cost, whereas signaling a mismatched type, referred to as "evasion," imposes a cost c_B on the buyer and a cost c_S on the seller.¹*

¹ We can more generally allow for each type of buyer impose a different evasion cost (e.g., if a $\bar{\theta}$ -buyer evades, the costs are $\bar{c}_B, \bar{c}_S \in \mathbb{R}$, and if a $\underline{\theta}$ -buyer evades, the costs are $\underline{c}_B, \underline{c}_S \in \mathbb{R}$. However, as we later

3. **Pricing decision.** The seller chooses a price p_i to set for buyer i . The information the seller can use to set the prices depends on the type of seller. A signal-aware seller can set prices depending on the signals sent by the buyers, that is, they can set one price for all buyers that signaled \underline{s} and a different price for all buyers that signaled \bar{s} . A signal-blind seller must set the same price for all buyers since they have no information to distinguish buyers.
4. **Purchase decisions.** Each buyer, based on the price p_i set for them and their valuation θ_i , makes a choice $b_i \in \{0, 1\}$, to purchase the item ($b_i = 1$) or not ($b_i = 0$).
5. **Utilities.** All players receive their respective utilities. Each buyer's positive utility is zero if they do not buy the item and the difference between their valuation and price otherwise. If they took evasive action in the signaling stage, their negative utility is equal to their cost of evasion c_B . That is, buyer i 's utility is

$$u_B(\theta_i, s_i, p_i, b_i) = (\theta_i - p_i)b_i - c_B e(\theta_i, s_i)$$

where $e(\theta_i, s_i) = \mathbb{1}\{(\theta_i = \underline{\theta} \wedge s_i = \bar{s}) \vee (\theta_i = \bar{\theta} \wedge s_i = \underline{s})\}$ indicates evasion or not. The seller's overall utility is the sum of utilities $u_S(\theta_i, s_i, p_i, b_i)$ from their interactions with each buyer. The positive utility due to buyer i is the revenue p_i if buyer i buys and zero otherwise. If the buyer took evasive action in the signaling stage, the seller incurs negative utility c_S . That is, the seller's utility is

$$u_S((\theta_i, s_i, p_i, b_i)_{i=1}^n) = \sum_{i=1}^n u_S(\theta_i, s_i, p_i, b_i) = \sum_{i=1}^n p_i b_i - c_S e(\theta_i, s_i).$$

Mixed strategies. For simplicity of presentation, our game definition is stated in terms of pure strategies (i.e., players take deterministic actions). However, we can more generally allow players to employ mixed strategies. A *mixed strategy* for a player is a distribution over allowed actions conditioned on the information available when taking the action: buyer i 's mixed signaling strategy induces a conditional distribution over signals $\pi_i^s(\cdot|\theta_i) \in \Delta(\{\underline{s}, \bar{s}\})$; the seller's mixed pricing strategy induces conditional distributions $\pi^P(\cdot|\underline{s}, \chi)$, $\pi^P(\cdot|\bar{s}, \chi)$ over positive reals with the constraint $\pi^P(\cdot|s = \underline{s}, \chi = 0) = \pi^P(\cdot|s = \bar{s}, \chi = 0)$; finally, each buyer i 's mixed buying strategy induces conditional distribution $\pi_i^b(\cdot|\theta_i, p_i) \in \Delta(\{0, 1\})$.

Let $\pi = (\pi^s, \pi^P, \pi^b)$ denote a mixed strategy profile. π , along with the probability of player types described in Step 1 of Definition 1 (which we will denote $p(\chi)$ and $p(\theta_i)$) induce a distribution over action profiles with the probability of an action profile $(\chi, (\theta_i, s_i, p_i, b_i)_{i=1}^n)$ given by

$$\mathbb{P}(\chi, (\theta_i, s_i, p_i, b_i)_{i=1}^n) = p(\chi) \prod_{i=1}^n p(\theta_i) \pi_i^s(s_i|\theta_i) \pi^P(p_i|\theta_i, \chi) \pi_i^b(b_i|\theta_i, p_i). \quad (1)$$

Given a mixed strategy profile π , we will denote the expected utility for the seller and buyer i by

$$U_S(\pi) = \mathbb{E}[u_S((\theta_i, s_i, p_i, b_i)_{i=1}^n)] \quad \text{and} \quad U_B^i(\pi) = \mathbb{E}[u_B(\theta_i, s_i, p_i, b_i)],$$

where the expectation is over the joint distribution in (1).

show, the only costs that are relevant are the evasion costs associated with the $\bar{\theta}$ -seller, because the $\underline{\theta}$ seller will never choose to evade, so we can think of $c_B = \bar{c}_B$ and $c_S = \bar{c}_S$.

Solution concept. We study the *perfect Bayes Nash equilibrium (PBNE)*. Mixed strategies of players constitute a PBNE if the following conditions hold: (1) sequential rationality, meaning that each player’s strategy constitutes a best response to their beliefs about the other players’ types and strategies, given the history of the game up to the point of choosing the action and (2) consistency of beliefs, meaning that players’ beliefs about other players’ types are updated following Bayes’ rule.

The following theorem characterizes the PBNE of the price-discrimination game described in Definition 1.

► **Theorem 2.** An $(n, \alpha, \mu, \bar{\theta}, \underline{\theta}, c_B, c_S)$ -PD game has the following unique perfect Bayes Nash equilibrium. Define $\Delta\theta = \bar{\theta} - \underline{\theta}$.

- (a) Buyers with type $\theta_i = \underline{\theta}$ will signal $s_i = \underline{s}$.
- (b) Buyers with type $\theta_i = \bar{\theta}$ will signal

$$s_i = \begin{cases} \underline{s} \text{ w.p. } q^* & \text{if } \alpha > c_B/\Delta\theta \\ \bar{s} & \text{otherwise.} \end{cases} \quad \text{where } q^* = \min \left\{ 1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta} \right\}$$

- (c) The signal-aware seller sets price

$$p_{\text{signalaware}}^*(s) = \begin{cases} \underline{\theta} & \text{if signal } s = \underline{s} \text{ is observed} \\ \bar{\theta} & \text{if signal } s = \bar{s} \text{ is observed.} \end{cases}$$

- (d) The signal-blind seller sets price

$$p_{\text{signalblind}}^* = \begin{cases} \underline{\theta} & \text{if } \underline{\theta} \geq \mu\bar{\theta} \\ \bar{\theta} & \text{otherwise.} \end{cases}$$

- (e) Buyer i buys the good if and only if their price p_i is at most their value, so

$$b_i = \mathbb{1}\{\theta_i \leq p_i\}.$$

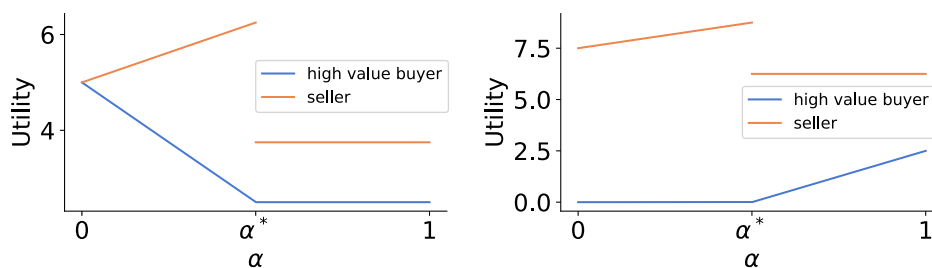
The proof is given in Appendix A.1.

► **Remark 3 (Buyer-induced privacy).** The $\bar{\theta}$ -buyers’ equilibrium response can be interpreted as a privacy-protecting mechanism. This type of buyer is vulnerable to price discrimination, so rather than always signaling their true type, they may choose to randomize their signal. More specifically, if the cost of evasion is very high, the $\bar{\theta}$ -buyer will tell the truth, but if the evasion cost is low enough, the $\bar{\theta}$ -buyer can receive a reduction in price that is higher than their evasion cost. In the latter case, the $\bar{\theta}$ -buyer must then choose the maximum evasion probability q^* such that it is still in the seller’s best interest to take the the buyer’s signal at face value. We call this randomization “buyer-induced privacy.”

Theorem 2 tells us that strategic behavior can only happen if $c_B < \Delta\theta$ (otherwise, we can never have $\alpha > c_B/\Delta\theta$, so buyers will always signal truthfully). For the rest of the paper, we will focus on this setting.

► **Assumption 1.** In all following results, we assume $c_B < \Delta\theta$.

A natural next question is how each player’s utility is affected by the game parameters. In particular, we focus on the effect of α , due to its connection to privacy. In Figure 1, we visualize the utilities of the seller and $\bar{\theta}$ -buyers as α varies from 0 to 1. Observe that the seller’s utility increases for α less than some threshold value α^* , whose exact value we



■ **Figure 1** Plots of the $\bar{\theta}$ -buyer and seller utilities as a function of α in the $\underline{\theta} \geq \mu\bar{\theta}$ setting (left) and the $\underline{\theta} < \mu\bar{\theta}$ setting (right).

give in the corollary below. This corresponds to the set of PD-games where the buyer's equilibrium response is truthful. Beyond α^* , the $\bar{\theta}$ -buyers' equilibrium response changes to being strategic and the seller's utility drops. We formalize the ordering of utilities in the following corollary.

► **Corollary 4.** (*Order of utilities*) Fix $n, \mu, \bar{\theta}, \underline{\theta}, c_B, c_S$ and let $u_S(\alpha), u_B(\alpha)$ denote the seller's and $\bar{\theta}$ -buyers' equilibrium utilities of the $(n, \alpha, \mu, \bar{\theta}, \underline{\theta}, c_B, c_S)$ -PD game. $u_S(\cdot)$ is maximized at $\alpha^* = c_B/\Delta\theta$, and the equilibrium utilities for the settings where the seller is always signal blind ($\alpha = 0$), is always signal aware ($\alpha = 1$), and is signal aware with probability α^* ($\alpha = \alpha^*$) have the following orderings:

(a) When $\underline{\theta} \geq \mu\bar{\theta}$,

$$u_S(\alpha^*) > u_S(0) > u_S(1) \quad \text{and} \quad u_B(0) > u_B(1) = u_B(\alpha^*).$$

(b) When $\underline{\theta} < \mu\bar{\theta}$,

$$u_S(\alpha^*) > u_S(0) > u_S(1) \quad \text{and} \quad u_B(1) > u_B(0) = u_B(\alpha^*).$$

$\underline{\theta}$ -buyers always receive a utility of zero, regardless of the value of α .

2.1 Price discrimination with seller commitment

A key takeaway from Corollary 4 is that the seller's utilities are dependent on the value of α , and if the seller could choose a value of α , they would want to choose $\alpha = \alpha^*$ to maximize their utility. Suppose we are now in a setting where the seller is able to choose and publicly commit to an α . As a motivating example, suppose that the seller must go through a data broker to access signals, and the data broker publishes trusted summaries of what fraction of buyers the seller requests data on. In such a setting, where α is chosen by the seller instead of treated as given, we arrive at the following equilibrium.

► **Corollary 5.** (*Equilibrium of price-discrimination game with commitment*) When the seller has commitment power (i.e., is able to credibly communicate to sellers that they will not price discriminate with some probability), the perfect Bayes Nash equilibrium of the PD-game consists of the following strategies:

(a) The seller commits to not price-discriminating (by playing $p_{\text{signalblind}}^*$ from Theorem 2) with probability $1 - \alpha^*$, where $\alpha^* = c_B/\Delta\theta$.

(b) All buyers always signal truthfully.

The buyers' buying decisions are the same as in Theorem 2.

Proof. (a) follows directly from Corollary 4, which tells us that the seller’s utility is maximized at α^* , and (b) comes from applying Theorem 2 with $\alpha = \alpha^*$. ◀

► **Remark 6.** Commitment ability allows the seller to achieve a higher utility by providing seller-induced privacy. This seller-induced privacy obviates the need for buyers to take evasion action to create buyer-induced privacy, which benefits the seller. We use U_1^* to refer to the seller’s maximum achievable equilibrium utility in the single interaction price discrimination game with commitment. This utility is achieved when the seller plays the strategy given in Corollary 5.

3 Repeated Interactions

In the previous section, we saw the emergence of seller-induced privacy when the seller has commitment ability. If possible, the seller would commit to providing seller-induced privacy (by ignoring signals with probability $1 - \alpha^*$, as in Corollary 5), thereby limiting the extent of price discrimination performed by the seller. However, these results hinge on the buyer believing that the α stated by the seller truly corresponds to the probability of price discrimination. Without this credible commitment from the seller, the story becomes more complicated.

In this section, we study whether seller-induced privacy can still arise in the absence of such commitment ability, through the development of a reputation based on the seller’s historical pricing. We ask the question of how the extent of privacy and resulting utilities differ under reputation-based privacy versus commitment-based privacy. We model the seller as making pricing decisions using an online learning algorithm and show how different models such as *no-regret* and *no-policy-regret* lead to different answers to this question.

In the repeated interaction setting, we also relax the assumptions that the distribution μ over agent types and the probability α that the seller looks at the agent’s signal are publicly known. Rather than playing the single-interaction equilibrium strategies, which require full knowledge of game parameters, the players now have to learn strategies online based on past interactions.

3.1 Setup

We consider repeated interactions between a seller and buyers where a new batch of buyers is drawn at each round. We call this as the *repeated PD protocol*. Each round is similar to the one-shot PD-game from Definition 1 but with the following differences: (1) There is one fixed seller throughout all rounds. (2) When players choose actions, they not only have access to information from the current round (as was the case in the one-shot PD game) but also some information from previous rounds. Specifically, at round t , the seller has access to $((s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1}$, the signals they observed and the prices they set in previous rounds, and each buyer i has access to $((\theta_i^\tau, s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1}$, the buyer types, signals, and prices of all buyers from previous rounds. This modeling of the buyers’ access is appropriate in settings where buyer information is pooled either through crowd-sourcing or by an auditing entity and made available to buyers. (3) The parameter μ (the probability of a type- θ buyer) is not known to the seller. (4) The probability that the seller will price discriminate is not known to buyers, as was assumed in the one-shot PD game; rather, buyers must estimate this probability based on past rounds. We write out the repeated interaction protocol in detail in Appendix B.

3.2 Model of the buyers

Since each buyer participates in only one round of the repeated PD protocol, the equilibrium response is still appropriate to model the buyer's response. However, in the repeated interaction setting, we no longer assume the buyers hold a static, prior belief about the probability of a signal-aware seller. Instead, buyers have evolving beliefs based on the seller's interactions with past buyers.

Some specific buyer strategies we will refer to are π_{truthful}^s , which corresponds to always signaling truthfully, and $\pi_{\text{strategic}}^s$, which corresponds to signaling \underline{s} with probability q^* (as defined in Theorem 2) and signaling \bar{s} with probability $1 - q^*$. We consider the following model of buyer behavior.

► **Definition 7** (Consistent belief based equilibrium responding (CBER) buyers). *Consistent belief based equilibrium responding buyers (or CBER-buyers) form a sequence of beliefs $(\hat{\alpha}_t)_{t=1}^T$ satisfying a consistency property defined below and at round t , choose the corresponding equilibrium strategy (from Theorem 2) of the PD-game with $\alpha = \hat{\alpha}_t$. That is, $\underline{\theta}$ -buyers always signal truthfully, and $\bar{\theta}$ -buyers signal truthfully (play π_{truthful}^s) if $\hat{\alpha}_t \leq \alpha^*$ and signal the opposite type with probability q^* otherwise (play $\pi_{\text{strategic}}^s$).*

We now explain the consistency property. Given a sequence of seller mixed strategies action profiles that induce the sequences of distributions $(\pi_t^p(\cdot|s = \bar{s}))_{t=1}^T$ and $(\pi_t^p(\cdot|s = \underline{s}))_{t=1}^T$ indicating price distributions at each round for signals \underline{s}, \bar{s} respectively, define α_t to be

$$\alpha_t = \mathbb{P}_{\bar{P} \sim \pi_t^p(\cdot|s=\bar{s}), P \sim \pi_t^p(\cdot|s=\underline{s})} [\bar{P} \neq P].$$

That is, α_t denotes the probability of a different price for \bar{s} compared to \underline{s} at round t . The probability here is over the randomness due to the seller's mixed strategy at round t . α_t is a measure of extent of price discrimination by the seller at round t .

► **Definition 8** (Consistent sequence). *Let $\bar{\alpha}_T = (1/T) \sum_{t=1}^T \alpha_t$. We say a sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$ is consistent if $\lim_{T \rightarrow \infty} |\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| = 0$, where the expectation is taken over the randomness of the history $H_T = ((\theta_i^t, s_i^t, p_i^t)_{i=1}^n)_{t=1}^{T-1}$ used to construct $\hat{\alpha}_T$.*

A useful implication of consistency is that $\hat{\alpha}_T$ converges pointwise to $\bar{\alpha}_T$.

► **Lemma 9.** *If $(\hat{\alpha}_t)_{t=1}^T$ is a consistent sequence of beliefs, then for any $\epsilon < 0$ and $\delta > 0$, there exists some positive integer N such that for all $T > N$, we have $\mathbb{P}[|\hat{\alpha}_T - \bar{\alpha}_T| \geq \epsilon] \leq \delta$.*

Proof. Due to consistency and the definition of limits, there exists N such that for all $T > N$, we have $|\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| \leq \delta\epsilon$. Thus, for $T > N$, we can apply Markov's inequality to get $\mathbb{P}(|\hat{\alpha}_T - \bar{\alpha}_T| \geq \epsilon) \leq (|\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T|)/\epsilon \leq \delta\epsilon/\epsilon = \delta$. ◀

The following proposition and associated proof provide an algorithm to construct a consistent sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$.

► **Proposition 10** (Existence of consistent sequence). *Assume that buyers equilibrium-respond to $\hat{\alpha}_t$ at each round t . Then, for any sequence of seller actions, there exists a sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$ that is consistent.*

Proof sketch: Since there are multiple buyers at each round, we can infer whether the seller is price discriminating or not by comparing the prices charged to a buyer who signals \underline{s} and a buyer who signals \bar{s} . However, only some rounds are informative about price discrimination; in rounds where all buyers send the same signal, we are not able

to determine if the seller had a price discriminatory pricing policy in place. The consistent estimator $\hat{\alpha}_t$ we consider is the fraction of past rounds where price discrimination is observed, normalized to account for the probability that a round is likely to be informative about price discrimination. We show that $E[\hat{\alpha}_t] = (1/t) \sum_{\tau=1}^{t-1} \alpha_\tau$, which implies that $\lim_{T \rightarrow \infty} |\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| = \lim_{T \rightarrow \infty} \left| (1/T) \sum_{t=1}^{T-1} \alpha_t - (1/T) \sum_{t=1}^T \alpha_t \right| = \lim_{T \rightarrow \infty} \alpha_T/T = 0$. See Appendix C.1 for the full proof.

3.3 Model of the seller

Since the seller does not a priori know the distribution over buyer types and is engaged in multiple rounds of the repeated interaction, modeling the seller's response by the one-shot equilibrium from Theorem 2 is not reasonable. Instead, we consider the seller as optimizing various common objectives of repeated interactions such as regret minimization and policy-regret minimization.

The seller's mixed strategy at a given round is a pair of probability distributions $\pi_t^p = (\pi_t^p(\cdot|\bar{s}), \pi_t^p(\cdot|\underline{s}))$. Let Π denote the set of possible mixed strategies. For rational sellers, we can focus on distributions supported only on $\{\underline{\theta}, \bar{\theta}\}$ without loss of generality. Prices supported on $\{\underline{\theta}, \bar{\theta}\}$ maximize seller revenue in each round. The seller's effect on future rounds is also not affected by limiting the support. This is because the parameters α_t that the buyers' consistent estimator estimates treats *any* difference in prices as indicating price discrimination, so all price differences are treated the same.

Some specific seller strategies we will refer to are π_{PD}^p and π_{noPD}^p . The former is the "always-price-discriminating strategy," with $\pi_{\text{PD}}^p(\bar{\theta}|\bar{s}) = \pi_{\text{PD}}^p(\underline{\theta}|\underline{s}) = 1$. The latter is the "never-price-discriminating strategy," with $\pi_{\text{noPD}}^p(\underline{\theta}|\underline{s}) = \pi_{\text{noPD}}^p(\underline{\theta}|\bar{s}) = 1$ if $\underline{\theta} \geq \mu\bar{\theta}$ and $\pi_{\text{noPD}}^p(\bar{\theta}|\underline{s}) = \pi_{\text{noPD}}^p(\bar{\theta}|\bar{s}) = 1$ otherwise.

3.3.1 Regret-minimizing seller

The first seller model we consider is a regret-minimizing seller.

► **Definition 11** (Seller's regret). *Given a sequence of mixed strategy profiles $\{\pi_t\} = \{(\pi_t^s, \pi_t^p, \pi_t^b)\}_{t=1}^T$, the seller's average regret is*

$$R_T^S(\{\pi_t\}_{t=1}^T) = \frac{1}{T} \left[\max_{\pi^{p*} \in \Pi} \sum_{t=1}^T U_S(\pi_t^s, \pi_t^{p*}, \pi_t^b) - \sum_{t=1}^T U_S(\pi_t^s, \pi_t^p, \pi_t^b) \right].$$

► **Definition 12** (No-regret algorithm). *Let \mathcal{A}_B be an algorithm employed by the buyer in the repeated PD protocol. A seller algorithm \mathcal{A}_S in the repeated PD protocol is a no-regret algorithm for the seller given \mathcal{A}_B if the sequence of mixed strategies $(\pi_t)_{t=1}^T$ generated by the interaction between \mathcal{A}_B and \mathcal{A}_S has seller's average regret that is sublinear in the number of rounds. That is, $R_T^S((\pi_t)_{t=1}^T) \in o(1)$.*

We will denote by $(\pi_t)_{t=1}^T$ the sequence of random variables denoting the players' mixed strategies in each round. Our results analyze the asymptotic convergence of average seller utility. We say that the average seller utility *asymptotically converges* to some value v if $\lim_{T \rightarrow \infty} \mathbb{E} \left[(1/T) \sum_{t=1}^T U_S(\pi_t) \right] = v$. We write $U_S(\pi^p)$ and $U_S(\pi^s, \pi^p)$ when it is clear what the other arguments are.

If the seller employs a no-regret algorithm, then the seller could end up always price-discriminating i.e., no seller-induced privacy. This is stated below.

► **Proposition 13.** *(Always price-discriminating is regret minimizing) Given CBER-buyers, the seller algorithm that always employs the price-discrimination strategy i.e., $\pi_t^p = \pi_{\text{PD}}^p$ for all timesteps t is a no-regret algorithm for the seller. The seller's average utility asymptotically converges to a value at most $u_S(1)$, where $u_S(1)$ is the seller's equilibrium utility in the single-interaction PD-game with $\alpha = 1$.*

Proof sketch. The strategy of CBER-buyers in each round is either π_{truthful}^s or $\pi_{\text{strategic}}^s$. For both these buyer responses, the seller's optimal strategy is to always price discriminate, as shown in the computation of the seller's equilibrium response in the proof of Theorem 2. In other words, the seller incurs zero regret in each round by always price-discriminating.

Next, we analyze the seller's average utility. Note that when $\pi_t^p = \pi_{\text{PD}}^p$, the probability of seeing different prices for different signals is $\alpha_t = 1$, so $\bar{\alpha}_t = 1$ for all t . By Lemma 9, $\hat{\alpha}_t$ becomes greater than α^* eventually (where α^* is as defined in Corollary 5), which causes $\bar{\theta}$ -buyers to play $\pi_{\text{strategic}}^s$. In other words, eventually the seller and buyers will all be playing their equilibrium strategies for the PD-game with $\alpha = 1$, so their average utilities will converge to the corresponding equilibrium utilities. See Appendix C.2 for the full proof. ◀

The next proposition tells us that regret minimization necessarily causes the seller to achieve a worse expected average utility than the optimal utility they can achieve in the single interaction setting.

► **Proposition 14** (Regret minimization is inherently at odds with achieving \mathbb{U}_1^*). *Given CBER-buyers, for any no-regret seller algorithm, the seller's average utility asymptotically converges to strictly less than \mathbb{U}_1^* .*

Proof sketch. Define $\mathcal{T} = \{t \in [T] : \hat{\alpha}_t \leq \alpha^*\}$ to be the set of rounds where $\bar{\theta}$ -buyers' signaling strategy is π_{truthful}^s . In all other rounds, their signaling strategy is $\pi_{\text{strategic}}^s$. Define $\beta = (1/T) \sum_{t \in \mathcal{T}} \alpha_t$ to be a measure of simultaneous truthfulness from buyers and price-discrimination by the seller. Our proof involves the following parts. We outline the parts and state them as lemmas here and prove them in Appendix C.3

1. Obtaining \mathbb{U}_1^* requires the buyers to be truthful strictly more than α^* fraction of rounds.

► **Lemma 15.** $\lim_{T \rightarrow \infty} |\mathcal{T}|/T \leq \alpha^*$ implies that $\lim_{T \rightarrow \infty} \left(\sum_{t=1}^T U_S(\pi_t) \right) / T < \mathbb{U}_1^*$.

2. The no regret property requires that the seller price discriminates in most rounds where buyers are truthful. So β is close to $|\mathcal{T}|/T$.

► **Lemma 16.** $\lim_{T \rightarrow \infty} |\mathcal{T}|/T \leq \lim_{T \rightarrow \infty} \sum_{t \in \mathcal{T}} \alpha_t / T$.

3. There is a limit on simultaneous price-discrimination and truthful signaling due to the buyers' consistent beliefs. That is, β converges to at most α^* .

► **Lemma 17.** $\lim_{T \rightarrow \infty} \sum_{t \in \mathcal{T}} \alpha_t / T \leq \alpha^*$.

From Lemmas 16, 17, $\lim_{T \rightarrow \infty} |\mathcal{T}|/T \leq \alpha^*$. Lemma 15 shows that this means average seller utility is strictly less than \mathbb{U}_1^* . ◀

3.3.2 Policy-regret-minimizing seller

As we have seen, regret minimization does not guarantee that the seller achieves higher than price-discrimination utility. On the other hand, if we model the seller as minimizing policy regret [3], the seller *necessarily* achieves utility that is higher than the utility achieved by the naive strategy of always price discriminating.

► **Definition 18** (Seller’s policy regret). Consider a buyer algorithm \mathcal{A}_B and a seller algorithm \mathcal{A}_S . Let $(\pi_t(\mathcal{A}_B, \mathcal{A}_S))_{t=1}^T$ be the sequence of mixed strategies generated by the interaction between \mathcal{A}_B and \mathcal{A}_S . Given a sequence of mixed strategies $(\pi_t)_{t=1}^T$, the seller’s average policy regret of $(\pi_t)_{t=1}^T$ relative to a buyer algorithm \mathcal{A}_B and a baseline class \mathbb{A}_S of seller algorithms is

$$PR_T^S((\pi_t)_{t=1}^T; \mathcal{A}_B, \mathbb{A}_S) = \max_{\mathcal{A}_S \in \mathbb{A}_S} \frac{1}{T} \sum_{t=1}^T U_S(\pi_t(\mathcal{A}_B, \mathcal{A}_S)) - \frac{1}{T} \sum_{t=1}^T U_S(\pi_t)$$

► **Definition 19** (No-policy-regret algorithm). Let \mathcal{A}_B be an algorithm employed by the buyer in the repeated PD protocol. An algorithm \mathcal{A}_S is a no-policy-regret algorithm for the seller given \mathcal{A}_B and relative to a class of seller algorithms \mathbb{A}_S if the sequence of mixed strategies $(\pi_t(\mathcal{A}_B, \mathcal{A}_S))_{t=1}^T$ generated by the interaction between \mathcal{A}_B and \mathcal{A}_S satisfies $PR_T^S((\pi_t(\mathcal{A}_B, \mathcal{A}_S); \mathcal{A}_B, \mathbb{A}_S)_{t=1}^T) \in o(1)$.

Consider a baseline class \mathbb{A}_S^{MS} consisting of seller algorithms that employ the same mixed strategy in each round, that is, $\pi_t^p(\cdot|\bar{s})$ is the same distribution for all t and similarly for $\pi_t^p(\cdot|\underline{s})$.

► **Proposition 20** (Policy-regret-minimizing seller achieves \mathbb{U}_1^*). Given CBER-buyers, if the seller achieves sub-linear policy regret relative to \mathbb{A}_S^{MS} , then the seller’s average utility asymptotically converges to at least \mathbb{U}_1^* .

Proof sketch. Under the conditions of this proposition, the seller’s utility must, by definition of policy regret, approach a utility at least as high (or better) than the utility of any strategy in \mathbb{A}_S^{MS} as $T \rightarrow \infty$. Recall that \mathbb{U}_1^* is the seller utility achieved in the PD game when $\alpha = \alpha^*$. Consider the PD game that results in a seller utility of at least $\mathbb{U}_1^* - \epsilon$, which is achieved by the seller price-discriminating with probability $\tilde{\alpha} < \alpha^*$. Then the repeated-interaction strategy of always price-discriminating with probability $\tilde{\alpha}$ has an average expected utility of at least $\mathbb{U}_1^* - \epsilon$ (this must be true due to the consistency of buyer beliefs; see the full proof in Appendix C.4 for details). Taking ϵ to 0 gives the desired result. ◀

Combining the previous result with the following result tells us that a no-policy regret seller’s algorithm will cause the seller’s average utility to asymptotically converge to *exactly* \mathbb{U}_1^* . In fact, this result tells us the stronger result that there does not exist *any* seller algorithm that can achieve utility higher than \mathbb{U}_1^* .

► **Proposition 21.** Given CBER-buyers, for any seller algorithm, the seller’s average utility asymptotically converges to at most \mathbb{U}_1^* .

Proof sketch. This proof is similar to the argument of the proof of Proposition 14 and the full proof is in Appendix C.4. The key ideas again are that for high seller utility, there must be sufficiently many rounds where simultaneously, the seller price discriminates and the buyer reports truthfully. Since the buyers’ belief estimators are consistent, this cannot be the case. The difference between the average seller utility and \mathbb{U}_1^* is a constant times the following quantity: $\frac{1}{T} \sum_{t \in \mathcal{T}} (\pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\underline{\theta}|\underline{s})) - \alpha^*$, where \mathcal{T} is the set of rounds where the buyer signals truthfully. Lemma C.1, 17 (from the proof of Proposition 14) show that the consistency property implies that this difference converges to most zero. ◀

4 Experiments

In this section, we simulate the *repeated PD protocol* with $\mu = 0.5$, $\underline{\theta} = 5$, $\bar{\theta} = 15$, $c_B = c_S = 5$, and $n = 10$ and empirically verify our theoretical claims from Section 3. We report the convergence of buyer and seller utilities, seller actions, and buyer estimators. The seller and buyer algorithms we consider are described below. Code is available at <https://github.com/nivasini/PrivacyDynamics>.

4.1 Algorithms

Seller.

1. **Signal-blind seller.** The seller plays the regret-minimizing Exp3 algorithm (specifically Exp3-IX in Chapter 12 of [17]). At round t the seller sets a price $p_t \in \{\underline{\theta}, \bar{\theta}\}$ according to the algorithm's current sampling distribution, charges p_t to all buyers and updates the sampling distribution based on the resulting average utility from the buyers' purchase decisions.
2. **Signal-aware seller.** The seller plays a contextual version of Exp3, which we call CExp3, in which the algorithm maintains two sampling distributions over prices $\{\underline{\theta}, \bar{\theta}\}$, conditioned on the received signal, \underline{s} or \bar{s} . At each round, the seller samples once from each distribution and charges one price \underline{p}_t to all buyers who signal \underline{s} and \bar{p}_t to all buyers who signal \bar{s} . Depending on the sampling distributions, \underline{p}_t and \bar{p}_t may or may not be equal.
3. **Stackelberg equilibrium seller.** The seller commits to an $\alpha^* = c_B/\Delta\theta$ level of price-discrimination, i.e., they play the $(\alpha = 1)$ -PD equilibrium strategy (Theorem 2) with probability α^* and the $(\alpha = 0)$ -PD equilibrium with probability $1 - \alpha^*$.

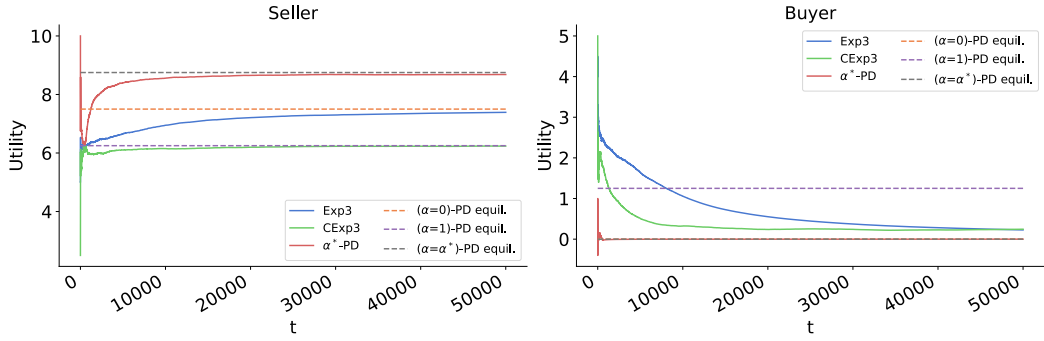
CBER-Buyer. Using a sequence of consistent estimators $\{\hat{\alpha}_\tau\}_{\tau=1}^{t-1}$ (Def. 8) to estimate the seller's probability of price-discrimination at each round, each buyer plays the $(\alpha = \hat{\alpha}_\tau)$ -PD equilibrium strategy. For our simulations, buyers use the estimator described in (3) to estimate the seller's probability of price discrimination at each round. All buyers in a single round use the same estimator.

4.2 Discussion

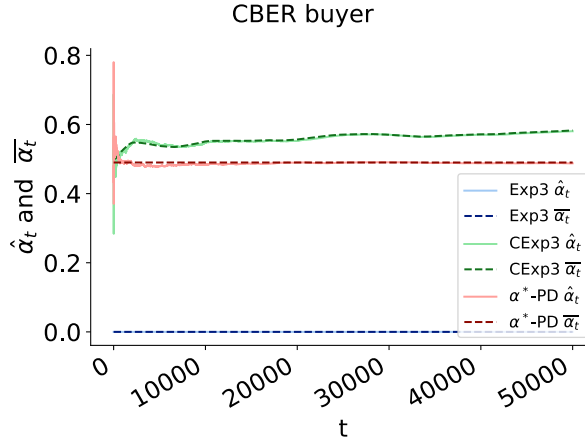
Convergence of Utilities. Figure 2 shows convergence of seller and buyer utilities for each of the seller's algorithms played against a CBER-buyer. As expected, when a seller plays Exp3 (which ignores signals) against a CBER-buyer, the players' utilities converge to the $(\alpha = 0)$ -PD equilibrium utility (Theorem 2). When the seller plays CExp3 (which observes signals) against a CBER-buyer, the seller's utility converges to the $(\alpha = 1)$ -PD equilibrium utility. Given our experiment parameters, multiple different distributions $\pi_t^p(\cdot | s = \underline{s})$ reward the seller equivalently, while some are more favorable for the buyer than others. Therefore, while the seller's utility will always converge to $(\alpha = 1)$ -PD, the buyer's utility may converge to something less than $(\alpha = 1)$ -PD. Finally, when the seller plays the Stackelberg equilibrium against a CBER-buyer, the players' utilities converge to the $(\alpha = \alpha^*)$ -PD equilibrium utility.

Consistency of $\hat{\alpha}$. Figure 3 illustrates the consistency of the buyer's estimator ((3)). Our simulations show that the buyer's estimate $\hat{\alpha}_t$ of the seller's probability of price discrimination converges to 0 against a seller playing Exp3, to 0.5 against a seller playing α^* -PD (where $\alpha^* = c_B/\Delta\theta = 0.5$ given our simulation parameters), and to higher-than-0.5 against a seller playing CExp3. Importantly, $\hat{\alpha}_t$ aligns with the seller's true average probability of price-discrimination, $\bar{\alpha}_t$, giving empirical evidence for Lemma 9.

Convergence of Seller Actions. In Figure 4, we track the cumulative proportion of the seller's price-discriminatory vs. non-price-discriminatory actions. Specifically, we track four seller actions: 1) charging a high price regardless of signal, 2) charging a low price regardless of signal, 3) charging a high price for a low signal and low price for a low signal (PD), and 4) charging a low price for a high signal and a high price for a low signal (reversePD). Given our



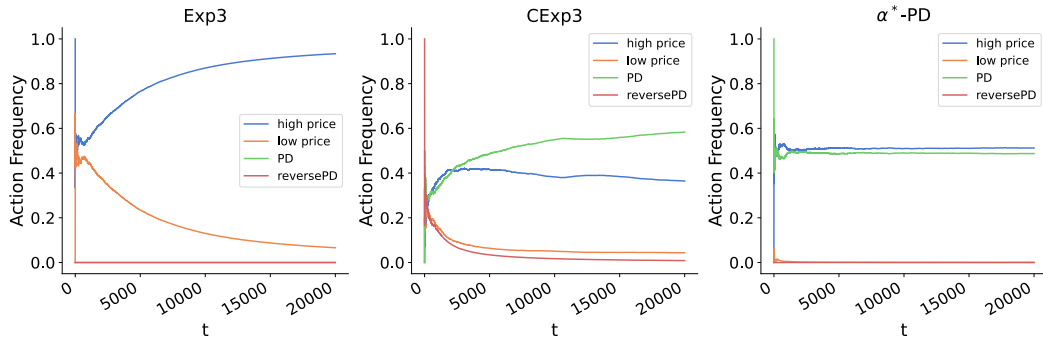
■ **Figure 2** Convergence of seller and buyer utilities for various algorithms. $\underline{\theta} < \mu\bar{\theta}$ with our experiment parameters, so the buyer's $(\alpha = 0)$ -PD and $(\alpha = \alpha^*)$ -PD utilities are the same (see Corollary 4).



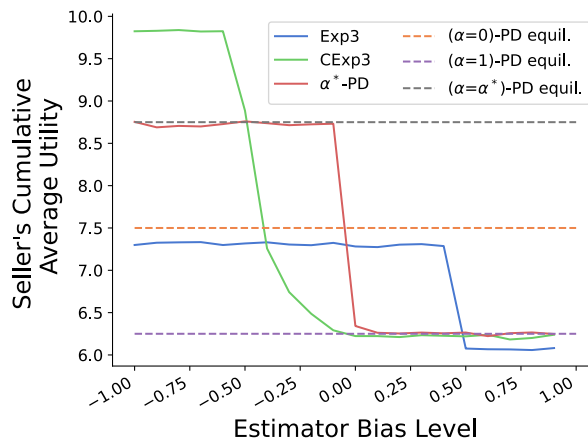
■ **Figure 3** $\hat{\alpha}_t$ and $\bar{\alpha}_t$ over time when seller is playing Exp3, CExp3, or α^* -PD. In all cases, $\hat{\alpha}$ is a consistent estimator of the seller's true probability of price discrimination.

parameter values for these simulations (i.e. $\underline{\theta} < \mu\bar{\theta}$ and $\alpha^* = 0.5$), in equilibrium we would expect that, for each batch of n buyers at a single round: 1) a signal-blind seller sets a high price for all n buyers, 2) a signal-aware seller sets a high price for high-signal buyers and a low price for low-signal buyers, and 3) a α^* -PD seller sets a high price for all high-signal buyers and low price for all low-signal buyers with probability 0.5 and sets a high price for all n buyers with probability 0.5. Figure 4 gives empirical evidence for this intuition.

Biased $\hat{\alpha}$. In realistic settings, the buyer may not have a consistent estimate of price discrimination and instead only have access to a biased $\hat{\alpha}$. Figure 5 examines whether a seller can benefit from non-consistency in the buyer's estimate. The y -axis of the figure tracks the seller's cumulative average utility after 20,000 rounds of interaction with CBER-buyers. We partition the interval $[-1, 1]$ into twenty segments γ_i of width 0.1, and the buyers use estimator $\hat{\alpha}_t + \epsilon_t$, where $\epsilon_t \sim \text{Unif}(\gamma_i)$. The plot then tracks the seller's cumulative average utility after 20,000 rounds of interaction with buyers for each bias interval γ_i . If $\hat{\alpha}_t + \epsilon_t$ is less than 0 or greater than 1, we clip it at those values respectively. In all cases, the seller is hurt by a $\bar{\theta}$ -buyer who overestimates the probability of price discrimination (high values of



■ **Figure 4** Relative frequency of actions for the seller playing Exp3, CExp3 and α^* -PD. The number of PD and reversePD actions for the Exp3 seller are both 0, as is expected.



■ **Figure 5** Cumulative average utility of the seller playing against CBER-buyers using biased $\hat{\alpha}$'s.

ϵ_t) and is thus more likely to evade, costing the seller the evasion cost. Against a buyer who underestimates the probability of price discrimination (low values of ϵ_t), neither the Exp3 nor α^* -PD seller gains utility, since the equilibrium behavior of the buyer with consistent $\hat{\alpha}_t$ aligns with the no-price-discrimination equilibrium (see Figure 2). By contrast, the CExp3 seller benefits from a buyer who underestimates the probability of price discrimination, since the seller benefits from discriminatory pricing without incurring the evasion cost. Against a CBER-buyer with consistent estimates, this advantage is impossible at equilibrium.

5 Conclusion

Since the type and level of privacy desired generally depends on the utilities of stakeholders and forms of interaction among them, we propose a game theoretic framework for privacy in this paper. We analyzed the perfect Bayes Nash equilibrium in a single-interaction setting as well as no-regret and no-policy-regret dynamics emerging over repeated interactions. In both these settings, we show how the different components of the game – utilities, actions and information sets (information available to players when choosing actions) impact the privacy levels that emerge.

Our results shed light on the impacts of different privacy-related interventions – we showed that enabling a seller to credibly commit to privacy (e.g., through privacy legislation like the GDPR) or revealing the seller’s past behavior (e.g., through privacy auditing) can surprisingly improve their utility. Thus, we believe our framework can be used to help analyze and craft privacy policies.

References

- 1 Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.
- 2 Alessandro Acquisti and Hal Varian. Conditioning prices on purchase history. *Marketing Science*, 2004.
- 3 Raman Arora, Michael Dinitz, Teodor V. Marinov, and Mehryar Mohri. Policy regret in repeated games. *Advances in Neural Information Processing Systems*, 2020.
- 4 Modibo K. Camara, Jason D. Hartline, and Aleck Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, 2020.
- 5 Vincent Conitzer, Curtis Taylor, and Liad Wagman. Hide and seek: Costly consumer privacy in a market with repeated purchases. *Marketing Science*, 31(2):277–292, 2012.
- 6 Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. *Advances in Neural Information Processing Systems*, 32, 2019.
- 7 Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.
- 8 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- 9 Jeffrey Ely and Juuso Valimaki. Bad reputation. *The Quarterly Journal of Economics*, 118(3):785–814, 2003.
- 10 Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.
- 11 John Bellamy Foster and Robert McChesney. Surveillance capitalism. *Monthly review*, 66(3):1–31, 2014.
- 12 Drew Fudenberg and J Miguel Villas-Boas. Behavior-based price discrimination and customer recognition. *Handbook on Economics and Information Systems*, 1:377–436, 2006.
- 13 Nika Haghtalab, Chara Podimata, and Kunhe Yang. Calibrated Stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36, 2024.
- 14 Oliver D Hart and Jean Tirole. Contract renegotiation and coasian dynamics. *The Review of Economic Studies*, 55(4):509–540, 1988.
- 15 Johannes Horner. Reputation and competition. *American Economic Review*, 92(3):644–663, 2002.
- 16 Shota Ichihashi. Online privacy and information disclosure by consumers. *American Economic Review*, 110(2):569–595, 2020.
- 17 Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2020.
- 18 Katrina Ligett and Kobbi Nissim. We need to focus on how our data is used, not just how it is shared. *Communications of the ACM*, 66(9):32–34, 2023.
- 19 Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.
- 20 Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso Valletti. The value of personal information in markets with endogenous privacy. *Center for Economic and International Studies*, 13(352), 2015.

- 21 Kobbi Nissim, Aaron Bembenek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R O'Brien, Thomas Steinke, and Salil Vadhan. Bridging the gap between computer science and legal approaches to privacy. *Harvard Journal of Law & Technology*, 31:687, 2017.
- 22 Carl Shapiro. Premiums for high quality products are returns to reputation. *Quarterly Journal of Economics*, 98(4):659–679, 1983.
- 23 Alicia Solow-Niederman. Information privacy and the inference economy. *Northwestern University Law Review*, 117:357, 2022.
- 24 Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in Apple's implementation of differential privacy on macOS 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

A Proofs from Section 2

A.1 Proof of Theorem 2

Proof. Part (a) comes from the fact that $\underline{\theta}$ buyers have no reason to pretend to have a higher valuation for the good than they actually do. Part (e) comes from the fact that buyers are utility maximizing.

Part (c) comes from the following reasoning: since signal blind sellers cannot see the buyers' signals, they must choose one price to set for all buyers. The seller wants to maximize their revenue, so they would ideally want to set the highest price that the buyer is willing to pay ($\bar{\theta}$ for $\bar{\theta}$ -buyers and $\underline{\theta}$ for $\underline{\theta}$ -buyers). However, the seller does not know the type of the buyer; all they know is the probability μ that the buyer is $\bar{\theta}$. The seller has to make a decision between charging $\bar{\theta}$ or $\underline{\theta}$. If the seller charges $\underline{\theta}$, both $\underline{\theta}$ and $\bar{\theta}$ agents would be willing to buy, so the expected revenue is $\underline{\theta}$. If the seller charges the higher price $\bar{\theta}$, only $\bar{\theta}$ agents would be willing to buy, so the expected revenue is $\mu\bar{\theta}$, which corresponds to

$$p_{\text{signalblind}}^* = \begin{cases} \underline{\theta} & \text{if } \underline{\theta} \leq \mu\bar{\theta} \\ \bar{\theta} & \text{if } \underline{\theta} > \mu\bar{\theta}. \end{cases}$$

Part (c) and (d) come from the following best-response arguments. Our goal is to show $p_{\text{signalaware}}^*$ is a best response given q^* and vice versa, where

$$p_{\text{signalaware}}^*(s) = \begin{cases} \underline{\theta} & \text{if } s = \underline{s} \text{ is observed} \\ \bar{\theta} & \text{if signal } s = \bar{s} \text{ is observed.} \end{cases} \quad \text{and} \quad q^* = \min \left\{ 1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta} \right\}$$

What is the signal aware seller's best response after seeing \bar{s} ? From part (a), we know that $\underline{\theta}$ buyers never signal $\bar{\theta}$, so the seller knows that a \bar{s} signal implies that the buyer is type $\bar{\theta}$ and should therefore set a price of $\bar{\theta}$ after seeing \bar{s} , i.e., $p_{\text{signalaware}}^*(\bar{s}) = \bar{\theta}$.

What is the signal aware seller's best response after seeing \underline{s} ? In order for $p_{\text{signalaware}}^*$ to be a best response, it must maximize the seller's expected utility, where the expectation is over the seller's posterior belief over the buyer's type given that they have signaled \underline{s} . Given probability q^* that the $\bar{\theta}$ buyer sends signal \underline{s} , the seller's posterior belief $\hat{\mu}$ that the buyer is type $\bar{\theta}$ is

$$\hat{\mu} = \mathbb{P}(\theta = \bar{\theta} | s = \underline{s}) = \frac{\mathbb{P}(s = \underline{s} | \theta = \bar{\theta})\mathbb{P}(\theta = \bar{\theta})}{\mathbb{P}(s = \underline{s} | \theta = \bar{\theta})\mathbb{P}(\theta = \bar{\theta}) + \mathbb{P}(s = \underline{s} | \theta = \underline{\theta})\mathbb{P}(\theta = \underline{\theta})} = \frac{q^*\mu}{q^*\mu + 1 - \mu}.$$

9:18 Privacy Can Arise Endogenously

Let $f(p)$ denote the seller's expected utility from charging price p after observing signal \underline{s} , so

$$f(p) = \begin{cases} p - \hat{\mu}q^*c_S & \text{if } p < \underline{\theta} \\ \hat{\mu}p - \hat{\mu}q^*c_S & \text{if } p \in [\underline{\theta}, \bar{\theta}]. \end{cases}$$

In order for $p_{\text{signalaware}}^*(\underline{s})$ to be a best response, it must be the value that maximizes f :

$$p_{\text{signalaware}}^*(\underline{s}) = \max_p f(p) = \begin{cases} \underline{\theta} & \text{if } q^* \leq \min \left\{ 1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta} \right\} \\ \bar{\theta} & \text{else.} \end{cases} = \underline{\theta},$$

where the last equality comes from the choice of q^* . This shows that $p_{\text{signalaware}}^*(\underline{s}) = \underline{\theta}$ is a best response for the seller. We now turn our attention to the $\bar{\theta}$ -buyer.

What is the optimal probability q^ of evasion for the $\bar{\theta}$ -buyer?* Let $g(q)$ denote the expected utility for the $\bar{\theta}$ buyer when they evade with probability q , given that the seller is playing $p_{\text{signalblind}}^*$ if they are signal blind and $p_{\text{signalaware}}^*$ if they are signal aware, so

$$g(q) = \mathbb{P}(\text{seller is signal blind})(\bar{\theta}\text{-buyer utility if seller plays } p_{\text{signalblind}}^*) \\ + \mathbb{P}(\text{seller is signal aware})(\bar{\theta}\text{-buyer utility if seller plays } p_{\text{signalaware}}^*). \quad (2)$$

- If $1 \leq (1-\mu)\underline{\theta}/\mu\Delta\theta$, this implies that $\underline{\theta} \geq \mu\bar{\theta}$, so (2) simplifies to

$$u_B = (1 - \alpha)\Delta\theta + (\alpha\Delta\theta - c_B)q.$$

- If $(1-\mu)\underline{\theta}/\mu\Delta\theta \leq 1$, this implies $\theta < \mu\bar{\theta}$, so (2) simplifies to

$$u_B = \begin{cases} (\alpha\Delta\theta - c_B)q & \text{if } q \leq (1-\mu)\underline{\theta}/\mu\Delta\theta \\ -c_Bq & \text{else.} \end{cases}$$

Combining everything, we see that the $\bar{\theta}$ -buyer's optimal probability of evasion is q^* as written in the theorem statement. \blacktriangleleft

B Repeated PD Protocol

The detailed algorithm is described in the arXiv version.

C Proofs from Section 3

C.1 Proof of Proposition 10

Proof. For each round t , let $I_t = \mathbb{1} \{ \exists i \text{ s.t. } s_i^t = \bar{s} \text{ and } \exists j \text{ s.t. } s_j^t = \underline{s} \}$ be an indicator for whether both types of signals are observed at round t , i.e., whether round t is “informative” about if there is price discrimination. For rounds t with $I_t = 1$, we additionally define the following random variables: $\bar{P}_t = p_i^t$ for the smallest $i \in [N]$ such that $s_i^t = \bar{s}$; $\underline{P}_t = p_j^t$ for the smallest $j \in [N]$ such that $s_j^t = \underline{s}$; and $X_t = \mathbb{1} \{ \bar{P}_t \neq \underline{P}_t \}$, an indicator for observed price discrimination. Note that the choice to define \bar{P}_t and \underline{P}_t to correspond to the *smallest* index satisfying the corresponding condition is simply for concreteness; we could equivalently sample uniformly from the set of indices satisfying the condition.

Recall that $H_t = ((\theta_i^\tau, s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1}$ is the history known by buyers at the beginning of round t . Consider the following estimator:

$$\hat{\alpha}_t = \frac{1}{t} \sum_{\tau=1}^{t-1} \frac{X_\tau I_\tau}{\mathbb{E}[I_\tau | H_\tau]} \quad (3)$$

The expectation $\mathbb{E}[I_\tau|H_\tau]$ is over the randomness at round τ . Note that $\hat{\alpha}_t$ is computable based on the history H_t , because $\mathbb{E}[I_\tau|H_\tau]$ is computable for any $\tau < t$. We will now show that $\hat{\alpha}_t$ satisfies Definition 8. We start by computing the expectation of $\hat{\alpha}_t$:

$$\begin{aligned} \mathbb{E}[\hat{\alpha}_t] &= \mathbb{E} \left[\frac{1}{t} \sum_{\tau=1}^{t-1} \frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]} \right] \\ &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E} \left[\frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]} \right] && \text{linearity of expectation} \\ &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E} \left[\mathbb{E} \left[\frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]} \middle| H_\tau \right] \right] && \text{tower rule} \\ &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E} \left[\frac{\mathbb{E}[X_\tau I_\tau|H_\tau]}{\mathbb{E}[I_\tau|H_\tau]} \right] \end{aligned}$$

Observe that X_τ and I_τ are independent given H_τ . To see why, note that the randomness in $X_\tau|H_\tau$ comes only from the randomness in the seller's mixed strategy at round τ , whereas the randomness in $I_\tau|H_\tau$ comes only from the randomness in the buyers mixed strategy at round τ . The mixed strategies are fixed given H_τ , and the additional randomness is independent. Thus,

$$\begin{aligned} &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E} \left[\frac{\mathbb{E}[X_\tau|H_\tau] \mathbb{E}[I_\tau|H_\tau]}{\mathbb{E}[I_\tau|H_\tau]} \right] \\ &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E}[\mathbb{E}[X_\tau|H_\tau]] \\ &= \frac{1}{t} \sum_{\tau=1}^{t-1} \mathbb{E}[\mathbb{E}[\mathbb{1}\{\bar{P}_\tau \neq \underline{P}_\tau\} | H_\tau]] && \text{by definition of } X_\tau \end{aligned}$$

Since $\bar{P}_\tau|H_\tau \sim \pi_t^p(\cdot|s = \bar{s})$ and $\underline{P}_\tau|H_\tau \sim \pi_t^s(\cdot|s = \underline{s})$ by definition of the game, we have

$$= \frac{1}{t} \sum_{\tau=1}^{t-1} \alpha_\tau$$

Finally, plugging in the above expression with $t = T$ into the criterion for consistency, we have

$$\lim_{T \rightarrow \infty} \left| \mathbb{E}[\hat{\alpha}_T] - \frac{1}{T} \sum_{t=1}^T \alpha_t \right| = \lim_{T \rightarrow \infty} \left| \frac{1}{T} \sum_{t=1}^{T-1} \alpha_t - \frac{1}{T} \sum_{t=1}^T \alpha_t \right| = \lim_{T \rightarrow \infty} \frac{\alpha_T}{T} = 0$$

as desired. The last equality comes from the fact that α_T is a probability, so it is bounded between 0 and 1 for all T . \blacktriangleleft

C.2 Proof of Proposition 13

Proof. First, we will show that always price-discriminating ($\pi_t^p = \pi_{\text{PD}}^p$ for all $t \in [T]$) is no-regret against CBER-buyers. For CBER-buyers, their strategy π_t^s at each round t is either π_{truthful}^s or $\pi_{\text{strategic}}^s$. For both these buyer responses, the seller's optimal strategy is to always price discriminate as shown in the computation of the seller's equilibrium response in the proof of Theorem 2. In other words, the seller incurs zero regret in each round and thus zero average regret.

Next, we will analyze the seller's average utility. Note that when $\pi_t^p = \pi_{\text{PD}}^p$, the probability of seeing different prices for different signals is $\alpha_t = 1$, so $(1/t) \sum_{\tau=1}^t \alpha_\tau = 1$ for all t . By the consistency property, $\hat{\alpha}_t$ becomes greater than α^* eventually (where α^* is as defined in Corollary 5) and the buyer plays $\pi_{\text{strategic}}^s$. In other words, eventually the seller and buyers will all be playing their equilibrium strategies for the PD-game with $\alpha = 1$, so their average utilities will converge to the corresponding equilibrium utilities. We make this argument formal below.

Define $\kappa < \infty$ to be the maximum utility that can be achieved by a seller in any round. The finiteness of κ is guaranteed by definition of the seller's utility function. Define $A_T = \{\exists t > \sqrt{T} \text{ s.t. } \hat{\alpha}_t > \alpha^*\}$ and let $A_T^C = \{\hat{\alpha}_t > \alpha^* \text{ for all } t > \sqrt{T}\}$ denote the complement. Let $\gamma_T = \mathbb{P}(A_T)$ and $1 - \gamma_T = \mathbb{P}(A_T^C)$ denote the corresponding probabilities. Then, we can decompose the expected average seller's utility as

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T U_S(\pi_t) \right] = \gamma_T \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T U_S(\pi_t) \middle| A_T \right] + (1 - \gamma_T) \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T U_S(\pi_t) \middle| A_T^C \right]. \quad (4)$$

The first term of (4) is trivially upper bounded by $\gamma_T \kappa$.

To bound the second term of (4), first note that for any round t where $\hat{\alpha}_t > \alpha^*$, the buyer's strategy will be equivalent to their equilibrium strategy with $\alpha = 1$. Thus, the best utility that the seller can achieve for those rounds is $u_S(1)$. It follows that under the condition that $\hat{\alpha}_t > \alpha^*$ for every $t > \sqrt{T}$, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T U_S(\pi_t) &= \frac{1}{T} \sum_{t=\sqrt{T}}^T U_S(\pi_t) + \frac{1}{T} \sum_{t=1}^{\sqrt{T}} U_S(\pi_t) \\ &\leq \frac{1}{T} \sum_{t=\sqrt{T}}^T u_S(1) + \frac{1}{T} \sum_{t=1}^{\sqrt{T}} \kappa \\ &= \frac{T - \sqrt{T}}{T} u_S(1) + \frac{\sqrt{T} \kappa}{T} \\ &\leq u_S(1) + \frac{\kappa - u_S(1)}{\sqrt{T}}. \end{aligned}$$

Plugging back into (4), we get

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T U_S(\pi_t) \right] \leq \gamma_T \kappa + (1 - \gamma_T) \left(u_S(1) + \frac{\kappa - u_S(1)}{\sqrt{T}} \right).$$

By the consistency property (Lemma 9), we know $\lim_{T \rightarrow \infty} \gamma_T = 0$, which yields the stated asymptotic bound on the seller's average utility. \blacktriangleleft

C.3 Missing Proofs of Lemmas in Proof of Proposition 14

Proof of Lemma 15. Based on the utility orderings from Corollary 4, note the following ordering of seller utilities for different combinations of buyer and seller policies:

$$U_S(\pi_{\text{truthful}}^s, \pi_{\text{PD}}^p) > U_S(\pi_{\text{truthful}}^s, \pi_{\text{noPD}}^p) > U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p) > U_S(\pi_{\text{strategic}}^s, \pi^p),$$

where π^p is any other pricing strategy besides π_{PD}^p and π_{noPD}^p . We can then write

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T U_S(\pi_t) &\leq \frac{1}{T} \sum_{t \in \mathcal{T}} U_S(\pi_{\text{truthful}}^s, \pi_{\text{noPD}}^p) + \sum_{t \in [T] \setminus \mathcal{T}} U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p) \\ &= \frac{|\mathcal{T}|}{T} U_S(\pi_{\text{truthful}}^s, \pi_{\text{noPD}}^p) + \left(1 - \frac{|\mathcal{T}|}{T}\right) U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p) \\ \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T U_S(\pi_t) &\leq U_S(\pi_{\text{truthful}}^s, \pi_{\text{noPD}}^p) \lim_{T \rightarrow \infty} \frac{|\mathcal{T}|}{T} + U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p) \left(1 - \lim_{T \rightarrow \infty} \frac{|\mathcal{T}|}{T}\right) \end{aligned}$$

Since $U_S(\pi_{\text{truthful}}^s, \pi_{\text{PD}}^p) > U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p)$, the above upper bound on the limit of the average utility is increasing as $\lim_{T \rightarrow \infty} |\mathcal{T}|/T$ is increasing. When $\lim_{T \rightarrow \infty} |\mathcal{T}|/T \leq \alpha^*$,

$$\begin{aligned} &\leq \alpha^* U_S(\pi_{\text{truthful}}^s, \pi_{\text{PD}}^p) + (1 - \alpha^*) U_S(\pi_{\text{strategic}}^s, \pi_{\text{PD}}^p) \\ &< \alpha^* U_S(\pi_{\text{truthful}}^s, \pi_{\text{PD}}^p) + (1 - \alpha^*) U_S(\pi_{\text{truthful}}^s, \pi_{\text{PD}}^p) = \mathbb{U}_1^* \end{aligned}$$

◀

Proof of Lemma 16. Let R_T^S denote the average seller utility in the T rounds. Since the seller is no regret, $\lim_{T \rightarrow \infty} R_T^S = 0$.

Consider the regret due to the seller deviating to π_{PD}^p in each round. The gain in utility in each round due to this deviation is non-negative since π_{PD}^p is the best-response to both possible buyer strategies $\pi_{\text{truthful}}^s, \pi_{\text{strategic}}^s$. We can then lower bound the regret by considering regret accumulated in rounds where $\hat{\alpha}_t \leq \alpha^*$. In such rounds, all buyers are truthful, so whenever the seller does not charge a buyer the price corresponding to their signal type, they incur regret. The probability that the seller observes \bar{s} but charges $\bar{\theta}$ is $\mu \pi_t^p(\underline{\theta}|\bar{s})$, and this yields a loss of utility of $\Delta\theta$, because the buyer is type $\bar{\theta}$. Similarly, the probability that the seller observes \underline{s} but charges $\bar{\theta}$ is $(1 - \mu) \pi_t^p(\bar{\theta}|\underline{s})$, and this yields a loss of utility of $\underline{\theta}$, since the buyer is type $\underline{\theta}$.

$$\begin{aligned} R_T^S &\geq \frac{1}{T} \sum_{t: \hat{\alpha}_t \leq \alpha^*} \mu \Delta\theta \pi_t^p(\underline{\theta}|\bar{s}) + (1 - \mu) \underline{\theta} \pi_t^p(\bar{\theta}|\underline{s}) \\ &\geq \frac{1}{T} \sum_{t: \hat{\alpha}_t \leq \alpha^*} \kappa (1 - (\pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s}))) \quad \text{where } \kappa := \min\{\mu \Delta\theta, (1 - \mu) \underline{\theta}\} \\ \implies \frac{1}{T} \sum_{t \in \mathcal{T}} (\pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s})) &\geq \frac{|\mathcal{T}|}{T} - \frac{R_T^S}{\kappa} \end{aligned}$$

The above inequality shows that $|\mathcal{T}|/T$ is bounded above by some measure of simultaneous truthfulness and price discrimination. Each quantity $\pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s})$ is a measure of price-discrimination in each round and is related to α_t as described in the following lemma.

► **Lemma C.1.** *When seller pricing strategies are supported on $\{\bar{\theta}, \underline{\theta}\}$, $\alpha_t \geq \pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s})$*

Proof of Lemma C.1. Since seller pricing strategies are supported on $\{\bar{\theta}, \underline{\theta}\}$, α_t which is the probability of seeing different prices for different signals is

$$\begin{aligned} \alpha_t &= \pi_t^p(\bar{\theta}|\bar{s}) \pi_t^p(\underline{\theta}|\underline{s}) + \pi_t^p(\underline{\theta}|\bar{s}) \pi_t^p(\bar{\theta}|\underline{s}) \\ &= \pi_t^p(\bar{\theta}|\bar{s}) (1 - \pi_t^p(\bar{\theta}|\underline{s})) + (1 - \pi_t^p(\bar{\theta}|\bar{s})) \pi_t^p(\bar{\theta}|\underline{s}) \\ &= \pi_t^p(\bar{\theta}|\bar{s}) + \pi_t^p(\bar{\theta}|\underline{s}) - 2\pi_t^p(\bar{\theta}|\bar{s}) \pi_t^p(\bar{\theta}|\underline{s}) \\ &= \pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s}) + 2\pi_t^p(\bar{\theta}|\underline{s}) (1 - \pi_t^p(\bar{\theta}|\bar{s})) \\ &\geq \pi_t^p(\bar{\theta}|\bar{s}) - \pi_t^p(\bar{\theta}|\underline{s}) \end{aligned}$$

◀

9:22 Privacy Can Arise Endogenously

By inequality 1, Lemma C.1, and since $\lim_{T \rightarrow \infty} R_T^S / \kappa = 0$, $\lim_{T \rightarrow \infty} \frac{|\mathcal{T}|}{T} \leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t \in \mathcal{T}} \alpha_t$. ◀

Proof of Lemma 17. Consider the last index $t^* \in \mathcal{T}$. Let us consider two cases. The first case is $\lim_{T \rightarrow \infty} t^*/T < \alpha^*$. Then, $\sum_{t \in \mathcal{T}} \alpha_t / T \leq |\mathcal{T}|/T \leq t^*/T$. This implies $\lim_{T \rightarrow \infty} \sum_{t \in \mathcal{T}} \alpha_t / T \leq \alpha^*$. In the second case, $\lim_{T \rightarrow \infty} t^* = \infty$. Consider $\bar{\alpha}_{t^*} = \frac{1}{t^*} \sum_{t \leq t^*} \alpha_t \geq \sum_{t \in \mathcal{T}} \alpha_t / T$. By the consistency property, $\lim_{T \rightarrow \infty} \bar{\alpha}_{t^*} = \hat{\alpha}_{t^*}$. $\hat{\alpha}_{t^*} \leq \alpha^*$ since $t^* \in \mathcal{T}$. ◀

C.4 Proofs of Propositions 20, 21

Please see arXiv version for proofs.

Online Algorithms with Limited Data Retention

Nicole Immorlica ✉

Microsoft Research, Cambridge, MA, USA

Brendan Lucier ✉

Microsoft Research, Cambridge, MA, USA

Markus Mobius ✉

Microsoft Research, Cambridge, MA, USA

James Siderius ✉

Tuck School of Business at Dartmouth, Hanover, NH, USA

Abstract

We introduce a model of online algorithms subject to strict constraints on data retention. An online learning algorithm encounters a stream of data points, one per round, generated by some stationary process. Crucially, each data point can request that it be removed from memory m rounds after it arrives. To model the impact of removal, we do not allow the algorithm to store any information or calculations between rounds other than a subset of the data points (subject to the retention constraints). At the conclusion of the stream, the algorithm answers a statistical query about the full dataset. We ask: what level of performance can be guaranteed as a function of m ?

We illustrate this framework for multidimensional mean estimation and linear regression problems. We show it is possible to obtain an exponential improvement over a baseline algorithm that retains all data as long as possible. Specifically, we show that $m = \text{POLY}(d, \log(1/\epsilon))$ retention suffices to achieve mean squared error ϵ after observing $O(1/\epsilon)$ d -dimensional data points. This matches the error bound of the optimal, yet infeasible, algorithm that retains all data forever. We also show a nearly matching lower bound on the retention required to guarantee error ϵ . One implication of our results is that data retention laws are insufficient to guarantee the right to be forgotten even in a non-adversarial world in which firms merely strive to (approximately) optimize the performance of their algorithms. Our approach makes use of recent developments in the multidimensional random subset sum problem to simulate the progression of stochastic gradient descent under a model of adversarial noise, which may be of independent interest.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms

Keywords and phrases online algorithms, machine learning, data, privacy, law

Digital Object Identifier 10.4230/LIPIcs.FORC.2024.10

Category Extended Abstract

Related Version *Full Version*: <https://arxiv.org/abs/2404.10997>

Funding *James Siderius*: This research was initiated while the author was a Research Intern at Microsoft Research.

Acknowledgements The authors thank Rad Niazadeh, Stefan Bucher, the Simons Institute for the Theory of Computing, participants at the CS and Law conference, seminar participants at the 2024 SIGecom Winter Meetings and the 2022 C3.ai DTI Workshop on Data, Learning, and Markets.

1 Introduction

Modern algorithms run on data. But as the potential uses for large datasets have expanded, so too have concerns about personal data being collected, retained, and used in perpetuity. Data protection laws are one response to these concerns, providing individuals the right to have their data removed from datasets. The EU’s GDPR is a flagship example, encoding a “right



© Nicole Immorlica, Brendan Lucier, Markus Mobius, and James Siderius;
licensed under Creative Commons License CC-BY 4.0

5th Symposium on Foundations of Responsible Computing (FORC 2024).

Editor: Guy N. Rothblum; Article No. 10; pp. 10:1–10:8

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to be forgotten” and mandating compliance with data deletion requests [14]. Similar policies have taken effect across the United States, such as the California Consumer Privacy Act [5] and Virginia’s Consumer Data Protection Act [6]. These policies specify rules governing deletion requests, but data removal is a complicated process. Data is not just stored: it is used to make decisions; it touches a vast array of metrics; it trains machine learning models. What, then, should it mean to remove data from a system? And how do such requests impact an algorithm’s ability to learn?

A growing body of literature approaches these questions through the “outcome-based” lens of constraining the observed behavior and outcomes of an algorithm. For example, one might require that once a piece of data has been “removed” in response to a request, the algorithm’s behavior should be indistinguishable (in a cryptographic sense) from one that does not have access to the data. Formalizing this idea leads to a myriad of details and modeling choices, and multiple notions of deletion-respecting algorithms have been proposed [4, 13, 7]. An alternative approach is to directly impose restrictions on an algorithm’s internal implementation that regulate and define the data removal process. This “prescriptive” approach is especially appealing from a regulatory perspective, since such restrictions provide clear guidance on what is and is not allowed (and, by extension, what constitutes an enforceable violation). But on the other hand, the actual implications of any given implementation restriction are not necessarily clear *a priori*. Constraints that appear very restrictive at first glance may still allow undesirable behavior through clever algorithm design. This undesirable behavior may be exhibited even by non-adversarial firms that simply wish to optimize their performance. Thus, for any given definition of what is meant by an implementation that respects data deletion, it is crucial to explore the outcomes that are generated by optimal (or near-optimal) algorithm design.

In this paper we explore the latter approach. We consider a stark framework in which an online algorithm can retain *no state* beyond stored data, which is subject to deletion requests. We show that even under such a restriction and even for simple statistical tasks like mean estimation, an algorithm that can preemptively delete data from its dataset can effectively retain information about data that was supposedly removed while still following the letter of the law (i.e., limited data retention). Moreover, we show that one can use this flexibility to substantially improve performance on statistical tasks relative to naive baseline algorithms that follow the spirit of the law (i.e., the right to be forgotten).¹ These results suggest that even in a world where an algorithm can retain no internal state whatsoever beyond its dataset, the curation of the dataset itself can be used to encode substantial information about data that has been supposedly removed, and even non-adversarial designers who seek only to maximize performance may naturally develop algorithms that leak information that was requested to be deleted. These results emphasize the importance of laws that regulate outcomes as well as process.

1.1 A Framework for Limited Data Retention

We propose a framework for algorithm design built upon a literal interpretation of a request to remove data. A sequence of data points is observed by a learning algorithm that actively maintains a subset of the data that has been observed so far. Each data point can come with a request that it be stored for only m rounds, after which it must be removed from the

¹ For example, by retaining all data as long as is allowed by regulation and then using optimal statistical estimators on the retained dataset.

algorithm’s subset.² We think of m as a legally-mandated period of time after which the algorithm is obligated to fulfill the request.³ The algorithm is free to discard data earlier, if desired; the only constraint is that data cannot be retained beyond the m rounds.

Of course, removing data points from the “official” dataset has no bite without additional restrictions on what else the algorithm can store. To clarify the impact of removing data, we impose a crucial modeling assumption: the algorithm *cannot retain any state* between rounds other than the dataset itself. In other words, any statistics or intermediate calculations performed by the algorithm must be recomputed, when needed, using only the data currently in the dataset.⁴

Such an algorithm can be described by two procedures: one that maintains the dataset (i.e., given the current subset and an incoming data point, choose which subset to keep) and one that answers a query about the full data stream given the current subset, possibly employing some non-standard estimator tailored to the data retention strategy.

We initiate an exploration of this framework through the lens of two standard statistical tasks: mean estimation and linear regression. In the case of mean estimation, each data point is drawn from an unknown distribution over \mathbb{R}^d and the algorithm’s goal is to recover the distribution’s mean. In the case of linear regression, each data point is a pair (x, y) where x is a d -dimensional characteristic vector and y is generated through a linear function of x plus random noise, and the goal is to simulate the linear function on challenge queries. In each case, the mean squared error achievable by an estimator that can retain an entire data stream of T data points (without any requirement to remove data) improves linearly with T . We ask: what error is achievable by an algorithm that respects requests to remove incoming data points within m rounds?

One baseline algorithm is to simply retain all data as long as possible. That is, the algorithm retains all of the previous m data points, then returns the maximum likelihood estimator given the sample for the target query. This approach is equivalent to keeping a uniform subsample of m draws from the underlying distribution. For the mean estimation and linear regression tasks, a uniform subsample of size m yields an average squared error no better than $\Theta(1/m)$, even for draws from a Gaussian distribution. In other words, this baseline would need to retain data for $m = O(T)$ rounds to achieve error comparable to what is attainable from the entire data stream.

1.2 An Improved Data Retention Policy

We show that it is possible to achieve an exponential improvement relative to the baseline solution described above. We present an algorithm for mean estimation that achieves a loss guarantee comparable to the optimal estimator over all T data points, but that retains each data point for only $m = \text{POLY}(d, \log(T))$ rounds. In more detail, if m is at least $\Theta(d \log(d/\epsilon))$, then for any query time $T > Cd/\epsilon$ (where C is a constant depending on the input distribution) the expected squared error will be at most ϵ . For linear regression we achieve a similar guarantee, with $m = \Theta(d^2 \log(d) \log(d/\epsilon))$. Our algorithms are polytime: each update step takes time linear in d and $1/\epsilon$.

² All of our results extend directly to model where a removal request can be made in any round after the data arrives (not just at the moment of arrival), and the data must be removed within m rounds of the request.

³ For example, under GDPR Article 12, any request to delete personal data must be honored “Without undue delay and in any event within one month of receipt of the request” [14].

⁴ One can equivalently think of this as a policy describing which statistics can be kept between rounds; namely, those that could be directly recomputed using only the retained data.

We also present a nearly-matching lower bound: if $m = o\left(\frac{d \log(1/\epsilon)}{\log(d) \log \log(1/\epsilon)}\right)$ then the algorithm must have error greater than ϵ with constant probability, regardless of the output function used to map the final subsample to an estimate of the mean.

1.3 Related Work

There is a substantial line of literature that explores definitions of data removal, especially as it relates to data protection and privacy laws. The literature on machine unlearning, initiated by [4], explores the process of updating a trained machine learning model so that it cannot leak information about to-be-deleted data. This has led to a vast body of work exploring different definitions and designs; see [19, 23] for some recent surveys of this literature. Beyond machine learning contexts, a notion of data deletion in terms of not leaking information about the data and maintaining secrecy, termed deletion-as-confidentiality, was proposed by [13]. A more permissive notion that constrains the leakage of information only after a removal request, deletion-as-control, was explored by [7]. Such works employ outcome-based constraints on data leakage, often in combination with internal state restrictions. In contrast, we explore a prescriptive framework that directly restricts an algorithm’s implementation and explore the extent to which these restrictions do (or do not) constrain the algorithm’s achievable performance and observable outcomes. While our algorithm respects certain notions of random differential privacy [15], we show that simple implementation restrictions to delete data points is not sufficient to retain full differential privacy of the deleted data.

Our work is also related to a line of literature on non-uniform subsampling for linear regression. The typical goal is to draw a sample from a large (or infinite) pool of potential data items (x, y) to increase accuracy of resulting models. Early works employed leverage scores to weight the predictor vector x [10, 17]. This approach has been extended to other norms via low-distortion embeddings [18] and improved by including outcomes y via importance weighting [9, 24, 22]. In contrast, our approach is not based on independent sampling but rather adaptive sample maintenance with elements added and removed over time.

Our approach is also closely related to coresets construction [11], in which the goal is to develop a highly compressed summary of a large dataset that retains the ability to answer queries from a given query class. Effective constructions are known for many learning problems, including variations of regression for numerous risk functions [1]. In principle a coresets can retain additional information beyond an (unweighted) subset of the original data, whereas our framework motivates us to focus specifically on unweighted subsampling.

From a technical perspective, our constructions use online implementations of stochastic gradient descent (SGD), which itself makes heavy use of sampling [3]. Our algorithms effectively simulate the progression of SGD using subsamples to approximate estimates. These approximations introduce some poorly-controlled noise to the SGD process, which necessitates an analysis that is robust to adversarial noise; for this we provide a slight variation on an SGD analysis due to [21]. To show that small subsets of data suffice to approximate the evolution of a sequence of improving estimates of regression coefficients, we employ recent advances in the theory of the random subset sum problem (RSS) [2, 16, 8]. The application of the RSS problem in contexts where SGD is used has also been explored in literature related to the Strong Lottery Ticket Hypothesis (SLTH) in learning theory [12, 2, 20]. However, the application of RSS to our setting requires a novel analysis.

2 Techniques

The full version of our paper, which can be found at <https://arxiv.org/abs/2404.10997>, contains our formal model, theorem statements, and proofs. Here, we describe our techniques.

Our approach is to simulate the progression of stochastic gradient descent (SGD). Consider the mean estimation task, where the goal is to learn the distribution mean θ . As new data points arrive, a (non-subsample-based) SGD algorithm would maintain an estimate for θ ; continuously updating the estimate in proportion to the noisy gradient estimate provided by fresh samples. Of course, such an SGD algorithm cannot be directly implemented as a subsampling algorithm because we are not allowed to directly maintain an estimate for θ , only subsets of data samples.

Our algorithm must therefore *simulate* the desired gradient steps: given a current estimator for θ (implied by the current subsample) and a proposed update (suggested by a gradient step), our algorithm will search recently-seen data points for a subset whose average approximates the target update. This approximation via a subset of data introduces noise into the gradient step. This noise is challenging to control, since the set of all averages of a subset of data points are heavily correlated with each other. We therefore treat this noise as adversarial, and note that SGD guarantees are robust to such adversarial noise as long as it is appropriately bounded. It turns out that having squared ℓ_2 distance of approximately ϵ between the target update and the closest subset-average would suffice to achieve our desired error rate.

How much memory is necessary to ensure that there is a subset of data points whose average is within ϵ of a given target point? As it turns out, recent developments in the random subset sum problem provides a surprising answer: for the single-dimensional case we only need that $m = \Omega(\log(1/\epsilon))$ and that the target point is not “too far” from the distributional average θ , where the asymptotic notation hides dependencies on parameters of the input distribution. To put this in perspective, this is asymptotically the same as what would be achievable if each of the 2^m subsets of samples were drawn independently of every other subset. To extend to d dimensions, we can either employ a multidimensional variant of the random subset sum problem, or target an error of ϵ/d on each dimension separately.

Extending this approach to linear regression tasks brings a new challenge. While SGD can still be used in this setting, the optimal estimator for the regression coefficients is not an empirical mean, so one cannot directly apply solutions to the random subset sum problem to encode an update step. For mean estimation, the natural estimator for θ is precisely an empirical mean of collected data points, so it is straightforward to encode an estimate of θ with a subsample that solves a random subset sum problem. For linear regression, the estimator for the coefficient vector is a specific transformation of a set of input (x, y) pairs, so we cannot directly apply the same trick. Instead, we will reduce to the mean estimation problem in a different way. Our proposed algorithm collects data points together into small groups that each generate an independent maximum likelihood estimate for the regression coefficients. As long as these groups are sufficiently large ($\Omega(d \log d)$ data points per group is enough) their corresponding estimates will be smoothly distributed near the true state. We can then think of these per-group maximum likelihood estimates as inputs to the random subset sum problem, and find a subset of them whose average approximates a proposed gradient step. This allows us to encode a gradient step by preserving the corresponding groups in our subsample.

To this point we have described ways to simulate gradient descent using subsampling. The algorithm effectively learns the desired statistics at the same rate as an algorithm with unbounded memory, and by maintaining the subsample judiciously it can encode

what has been learned. One might naturally wonder at this point whether the memory requirements could be substantially improved with more clever encodings. Since we put no restriction on the mapping from subset to algorithm output, in principal an algorithm could use the retained subsample to encode complex features of the full data stream in some Byzantine manner, then decode this information at query time. While we do not rule this out, we show a lower bound: any algorithm that satisfies the recency property *requires* $m = \Omega((d/\log(d))(\log(1/\epsilon)/\log\log(1/\epsilon)))$ in order to achieve squared error ϵ , even for mean estimation. Roughly speaking, this bound follows because even if the algorithm succeeds in perfectly learning θ , it will not achieve error less than ϵ if for *every* subset of m data points, the output function applied to that subset falls outside the ϵ -ball centered at θ . For m smaller than our bound, we can take a union bound over all subsets to show that the probability of this bad event will be large no matter what output function is used.

3 Conclusions and Future Work

In this work we introduced a framework for online algorithms subject to strict data retention limits. The algorithms in our framework retain no state other than a subsample of the data, and each data point must be removed from the subsample after at most m rounds. We provide upper and lower bounds on the value of m needed to achieve error ϵ for mean estimation and linear regression. We find that it is possible to substantially outperform a naive maximal-storage baseline by adaptively and proactively curating the algorithm’s dataset in order to improve its representativeness of the full data stream (including data that was to have been dropped).

Many technical questions are left open for future pursuits. We take a worst-case perspective that all data must be removed after m rounds, but one might consider a model where some data points can be retained for much longer. Does the presence of long-lived data alongside data that must be removed quickly enable different algorithmic approaches? Our subsampling framework can also be extended to other statistical tasks like non-linear regression, estimating higher moments, classification tasks, and so on. In each case, the algorithmic challenge is to dramatically reduce the size of a training set, online, so that a (perhaps specially-tailored) training process executed on the subsample can achieve performance approximately matching what is possible on the full data.

One could also apply our framework to non-stochastic or partially stochastic environments, where data is not necessarily generated according to a stationary process. Such environments can amplify the impact of individual data points (and their removal) on an algorithm’s output and state. Of course, the achievable algorithmic guarantees might vary substantially depending on the assumptions made on the data. But even so, understanding the structure of optimal (or near optimal) algorithms can shed light on the manner in which algorithm designers may be incentivized to build systems in the face of data retention limitations.

Finally, one can explore whether alternative frameworks for data removal lead to different types of behavior in optimal algorithm designs. A step in this direction is to quantify the extent to which optimal algorithms in a given framework are “undesirable” from the perspective of data removal, and use this to directly compare frameworks. Such an endeavor can help to build a toolkit for building up algorithmic restrictions that align well with stated policy goals.

References

- 1 Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coresets constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.
- 2 Luca Becchetti, Arthur Carvalho Walraven da Cunha, Andrea Clementi, Francesco d’Amore, Hicham Lesfari, Emanuele Natale, and Luca Trevisan. On the multidimensional random subset sum problem. *arXiv preprint arXiv:2207.13944*, 2022.
- 3 Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.
- 4 Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.
- 5 California consumer privacy act. https://leginfo.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5, 2018. Cal. Civ. Code §§ 1798.100 et seq.
- 6 Consumer data protection act, 2021 h.b. 2307/2021 s.b. 1392. <https://lis.virginia.gov/cgi-bin/legp604.exe?ses=212&typ=bil&val=Hb2307>, 2021.
- 7 Aloni Cohen, Adam Smith, Marika Swanberg, and Prashant Nalini Vasudevan. Control, confidentiality, and the right to be forgotten. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3358–3372, 2023.
- 8 Arthur da Cunha, Francesco d’Amore, Frédéric Giroire, Hicham Lesfari, Emanuele Natale, and Laurent Viennot. Revisiting the random subset sum problem. *arXiv preprint arXiv:2204.13929*, 2022.
- 9 Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. *Advances in neural information processing systems*, 26, 2013.
- 10 Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.
- 11 Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.
- 12 Damien Ferbach, Christos Tsirigotis, Gauthier Gidel, and Joey Bose. A general framework for proving the equivariant strong lottery ticket hypothesis. In *The Eleventh International Conference on Learning Representations*, 2022.
- 13 Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 373–402. Springer, 2020.
- 14 General data protection regulation, 2016. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.
- 15 Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Random differential privacy. *arXiv preprint arXiv:1112.2680*, 2011.
- 16 George S. Lueker. Exponentially small bounds on the expected optimum of the partition and subset sum problems. *Random Structures & Algorithms*, 12(1):51–62, 1998. doi:10.1002/(SICI)1098-2418(199801)12:1<51::AID-RSA3>3.0.CO;2-S.
- 17 Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning*, pages 91–99. PMLR, 2014.
- 18 Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100, 2013.
- 19 Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

10:8 Online Algorithms with Limited Data Retention

- 20 Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in neural information processing systems*, 33:2599–2610, 2020.
- 21 Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1571–1578, 2012.
- 22 Daniel Ting and Eric Brochu. Optimal subsampling with influence functions. *Advances in neural information processing systems*, 31, 2018.
- 23 Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.
- 24 Rong Zhu. Gradient-based sampling: An adaptive importance sampling for least-squares. *Advances in neural information processing systems*, 29, 2016.