Exploring fairness and privacy in machine learning

Carlos Pinzón Henao

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX126

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Exploring Fairness and Privacy in Machine Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Mathématiques et informatique

Thèse présentée et soutenue à Palaiseau, le 6 décembre 2023, par

Carlos Antonio Pinzón Henao

Composition du Jury :

Jean-François Couchot Professeur, Université de Franche-Comté	Président
Jean-Michel Loubes Professeur, Institut de Mathématiques de Toulouse	Rapporteur et examinateur
Josée Desharnais Professor, Laval University	Rapporteuse non examinatrice
Rachel Cummings Associate Professor, Columbia University	Examinatrice
Sara Bouchenak Professeur, INSA Lyon	Examinatrice
Antoine Boutet Assistant Professor, INSA Lyon. Researcher, Inria (Privatics).	Examinateur
Ferdinando Fioretto Assistant Professor, University of Virginia	Examinateur
Catuscia Palamidessi Directrice de recherche, Inria (Comète)	Directrice de thèse
Frank Valencia Chercheur, CNRS	Co-directeur de thèse

Abstract

This dissertation presents four published articles in the field of data ethics that extend our knowledge of fairness in machine learning and advance the state of the art of privacy in data collection and transmission. This document encompasses: (1) a general study of the trade-off between equal opportunity and accuracy of machine learning classifiers along with the proof that these objectives may oppose each other strongly; (2) the empirical proof that when using causal discovery algorithms for fairness assessment, different algorithms may lead to very different conclusions; (3) the proposal of a protocol for longitudinal data collection with guarantees inspired in local differential privacy; and (4) the derivation of two optimal methods for padding transmitted data to protect privacy against network observers.

Résumé en français

Cette thèse présente quatre articles publiés dans le domaine de l'éthique des données qui élargissent nos connaissances sur l'équité dans l'apprentissage automatique et l'état de l'art de la confidentialité dans la collecte et la transmission des données. Ce document englobe (1) le calcul de la limite de Pareto d'égalité des chances et de précision des classificateurs d'apprentissage automatique, la preuve que ces objectifs s'opposent radicalement pour certaines distributions ; (2) la preuve empirique que lors de l'utilisation d'algorithmes de découverte causale pour l'évaluation de l'équité, différents algorithmes peuvent conduire à des conclusions très différentes ; (3) la proposition d'un protocole de collecte de données longitudinales avec des garanties inspirées de la confidentialité différentielle locale ; et (4) la dérivation de deux méthodes optimales de remplissage des données transmises pour protéger la confidentialité contre les observateurs du réseau.

Acknowledgements

I want to thank my academic supervisors, Catuscia Palamidessi, Frank Valencia and Pablo Piantanida, for their constant guidance and help during my academic journey. They played a crucial role in my success by providing valuable support and mentorship. I was fortunate to have Catuscia as a supervisor. She is undoubtedly one of the most brilliant and dedicated individuals I have ever met. Catuscia possesses a rare combination of intelligence, kindness, and attentiveness that sets her apart. I am very grateful that she has always believed in my abilities and my autonomy. I am also very grateful to Frank, for he has given me the opportunity to be where I am today. His steady support and consistent presence have provided a sense of belonging and comfort that makes me feel at home. Beyond being a mentor, Frank has evolved into a dear friend and our shared humor and common research interests have further deepened our connection. His belief in my potential has significantly contributed to my growth, and I want to express my sincere gratitude for the trust he's shown in me. As for Pablo, I thank him that even though we tried several times without success to define a research project on which to work together, he always had his doors open to try new ideas again. I also admire his sincerity and his disciplined approach to research, two qualities that are crucial for productive work.

I extend my appreciation to my colleagues at team Comète and the neighboring team OptimiX for the collaborations and experiences we shared, especially to those who worked directly or shared memorable moments with me, and to my closest friends Hugues, Federica and Héber. I am also grateful for having had the honor of teaching and guiding Sebastian and Cezara, two magnificent students whose skills and potential I admire greatly. Working in the Comète team, with such grateful and motivated individuals, was a

privilege. I am thankful for the opportunities we had to work together and learn from each other, as this greatly contributed to my development as a researcher.

To my dear wife, family and friends, your consistent support, love, and encouragement were my pillars throughout my academic journey. Adriana, I am deeply grateful for the many sacrifices you have made to see me succeed and to succeed at my side. Your presence in my life is my most precious blessing. You are my sunshine. For their constant support, I want to thank my parents, grandparents, aunts, uncles and cousins, including especially my grandfather Antonio Henao who has always been attentive to my academic life, including this PhD without exception. To all of you, your belief in me motivates me to keep going.

Lastly, I am also grateful with Inria, the École Doctorale EDIPP and Catuscia's ERC grant for supplying the resources to carry out my studies.

In summary, I want to express my sincere thanks to my academic supervisors, colleagues, family, friends and financial sponsors for their valuable contributions to my academic and personal growth. I am glad for everything you've done for me.

Financial support

This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation program. Grant agreement n. 835294.

Contents

Chapter 1

Introduction

This dissertation explores four important challenges in the field of data ethics, namely, (1) the trade-off between equal opportunity and accuracy of machine learning classifiers, (2) the use of causal discovery algorithms for fairness assessment, (3) the regular collection of private data using local differential privacy frameworks, and (4) the use of padding schemes for data privacy during transmission. The purpose of this dissertation is to contribute with the development of theoretical insights and practical tools that can enrich the state of the art of data privacy and fairness, and, as a consequence, help society to better formalize and impose guarantees of fairness in machine learning classifiers and privacy in data collection and transmission.

1.1 Domain: data ethics

Data ethics is a multidisciplinary field concerned with the correct management of data, especially, the transparent and fair use of data as well as the security and privacy guarantees of the systems that collect it. It is an interdisciplinary domain that combines technology with human sciences, and it is gaining massive attention and complexity in the last decades due to the exponential increase in data collection and the digital transformation of many human activities.

Data ethics is a domain of crucial importance nowadays because society is facing a fast-paced digital transformation in which digital technologies deeply

infiltrate and permeate numerous spheres of human life. This transformation is caused and accelerated by the growth in data acquisition and the increased range of human activities that we digitalize. As estimated by the International Data Corporation, the volume of digital data created each year is growing exponentially, doubling every 3 years approximately [IDC20, Red21]. This trend is fueled by the continuous production of the semiconductor industry and the steadily growing hardware capacity for data collection, storage, transmission and processing, which doubles roughly every two years, a fact known as Moore's law. In parallel, digital platforms continue to influence communication and social interactions, access to information, entertainment, commerce and professions. With the raise of connected devices and personalized content, digital technologies extract more and more utility from personal data, making it a valuable resource for businesses and organizations, but they also become more and more pervasive and intimate. As a consequence, it is not reasonable nowadays to stop collecting or using data, but it is vital for the benefit of society that data is used with ethical considerations and relying on scientific research.

The research topis in data ethics range from philosophical aspects of big data [Leo19] to technical developments, both theoretical and practical, for ensuring the privacy, fairness and explainability of systems that use data, especially machine learning models, giving birth to the subdomain of Ethical Machine Learning, also known as Ethical Artificial Intelligence (Ethical AI).

Machine Learning (ML) is the main paradigm for processing large collections of data. The success of machine learning comes, on one side, from the simplification of data into a commodity that can be fed to a pipeline, and on the other, from the flexibility of the models it produces. Indeed, ML models can be used for applications requiring prediction, classification, data analysis, artificial data generation, or a combination of them. But also, the unprecedented performance and flexibility of ML models comes often at the cost of complexity and unexplainability of large models, to the point that they are treated as input-output black boxes and this leaves the door open to two unexpected important issues, privacy violations and systematic discrimination, which are detailed below.

Privacy violations in this context are not limited to security, i.e., the correct

access management and encryption of the data by trusted parties, but also to privacy on its own, meaning the correct alteration (also called anonymization or sanitization) of the data before sharing it with untrusted parties by means of non-injective functions, random noise and other information-destructive methods, so that its new content is still useful for certain analyses but does not reveal private information, making it safe even if disclosed.

Several privacy definitions have been proposed historically for the purpose of publishing micro data, i.e., a table concerning one individual per row. In this context, the data holder wants to share their micro-data database with either data analyzers, the open public or clients from the industry in exchange for money or services. The most prominent definitions had been k-anonymity, t-closeness [LLV06] and l-diversity [MKGV07], but all of them were shown to be weak against reidentification or attribute inference, making them mitigations, rather than general solutions [Ngu14]. The key problem of these definitions is that the methods that satisfy them are based mostly on deterministic transformations, e.g. removing a column or collapsing rows into equivalence classes, and this can reveal identities unexpectedly when crossing apparently anonymized information from several sources. Indeed, some researchers de-anonymized a very large fraction of the anonymized Netflix prize dataset in 2008 by using information from another movie platform [NS08]. These fundamental issues led to the development of a probabilistic framework for privacy called differential privacy.

Differential privacy [Dwo06a] (DP) was a breakthrough because of its elegant mathematical properties, which allow to quantify privacy loss (ϵ) in such a way that the total privacy loss when crossing information from several sources is always bounded by the sum of privacy losses of each individual source. This key property, called compositionality, is not satisfied by any of the previous definitions, which can be shown using quantitative information flow [G+23], a framework that aims to unify and quantify information leakage for arbitrary notions of secrets and privacy. DP introduced also the need for queries to observe the data because it assumes that a trusted central server holds private microdata and wants to provide a query service (an API) for computing sanitized aggregations of the data instead of sharing the data itself. With these queries, other new concepts appeared, e.g. query sensitivity and ϵ as an

inspection budget. DP dominates the literature of privacy, and researchers advocate for its use [CD18]. DP is not enforced in the General Data Protection Regulation (GDPR [Par16], an European Union regulation on information privacy in the European Union and the European Economic Area, made in 2016), and some authors [Hol19] sustain that there are semantic conflicts between what DP guarantees and what GDPR mandates, though the main argument is that DP protects against true identification but not against false identification, meaning that, ideally, no attacker should be able to make any inference about your personal data, even incorrect inferences. This argument is debatable because it is logically impossible to prevent incorrect inferences in general, e.g., anyone can (wrongfully) infer random data about anyone, so it is absurd to require this, while it is reasonable to quantify protection as protection against correct inferences.

It should be noted that DP was initially designed assuming that the central server that holds the data is trusted, a perspective that is changing lately for the large companies that collect data. The scandal of Cambridge Analytica [new18] rose concerns, beyond security, about the trust we can put on organizations that collect data. This scandal showed the mechanisms through which private information and digital platforms can be exploited to transform users opinions and behaviors. At the individual level, the so-called *echo chamber* or *filter bubble* [KR19], provides each user with a personalized feed of content that makes them addicted to new content and makes their opinions more extreme. At the global level, polarization becomes extreme and democracy is undermined because the opinions of the population are shaped by whoever is paying most for the information they receive.

These concerns led to increased interest in new definitions based on DP, but which treat the user as the data holder and the server as the client, so users sanitize their data before sending it to the untrusted central server in exchange for personalized services or a statistical analysis at the population level. These new definitions are local differential privacy (LDP) [KLN+11] and distance-based or metric-based differential privacy (d-privacy) [ABCP13], which generalizes in some sense both DP and LDP. The latter is very appropriate, for example, for geolocation data [PBMB18, ABCP13, BP22, SSL+23, SGCR22], in which it is relatively easy to generate false loca-

tions by adding noise. From the perspective of the organizations that hold data, with the introduction of LDP and d-privacy, most of the theory of DP for sanitizing data before sharing has been adapted for the problem of sanitizing before collecting it, and several LDP protocols have been proposed [EPK14, DKY17, EFM+20, ACBX21, ACBX22].

DP has also been studied for machine learning in applications that include using differentially private stochastic gradient descent for training neural networks [SCS13, FFMD22], possibly in a federated learning environment [GBJ+22], using neural network bottleneck representations for creating anonymized representations of the inputs [FPBD18], training ML models with fairness and privacy objectives simultaneously [TFVH21] and training predictive ML models using locally sanitized private data [ACC+20].

Regarding discrimination by machine learning models, it can be the result of biased data or inadequate models or training algorithms. Here, the term *bias* denotes a systematic and unfair preference for or against a group of people, denoted via a categorical feature called the sensitive attribute, which may or not be present in the data, and data is biased if it is not representative of the real-world population or if it reflects historical biases. If the training data is biased, the model will most likely learn and reproduce these biases, causing discrimination. But discrimination may occur from different sources. For instance, a linear regression model might give more weight to certain features, inadvertently amplifying biases present in the data. Also, a very simple model might not be able to capture the complexity of a particular group of people and an overly complex model might overfit the groups disparately, leading to unfair predictions. Lastly, if the data is missing important features, or contains features that are not relevant for the task, or contains undesired proxies of the sensitive attributes, or is treated as cross-sectional when it is longitudinal, disparity may arise. As in privacy, it has been shown that the naïve solution of removing the sensitive attribute is insufficient to prevent discrimination, e.g., a malicious advertiser can create highly discriminatory ads without having access to sensitive attributes [SAV+18].

Examples of applications that incurred in forms of discrimination include hiring [Reu18], facial recognition [BG18], general object recognition [new15] and criminal scoring [LMKA16], some of which use black box models. Generally

speaking, if a facial recognition system is trained primarily on images of lighter-skinned individuals, it will be less accurate for people with darker skin tones. Similarly, if a hiring algorithm is trained on historical hiring data that favors certain demographic groups, it may perpetuate those biases by recommending similar candidates. These unfortunate examples have made society more and more aware of the importance of auditing automated digital tools and the research on fairness in ML has been shaped (via examples and datasets) and fueled (via grants and motivated researchers) by them, a fact that has been stigmatized by some scholars [Och19]. It is important to note, nevertheless, that discrimination by machines is often unintentional, that is, biases emerge from the data or the models without the awareness nor the intention of the developers.

Fairness makes also sense for other applications such as recommender systems [DCDAU22] and natural language processing [CPO19], especially regarding gender bias. Even if we restrict to the simplest task of binary classification, the diversity of scenarios in which discrimination can occur has produced a large set of (sometimes inconsistent) definitions of fairness [MZP21b], of which the most prominent are arguably, statistical parity, conditional statistical parity, calibration and equal opportunity [HPS16].

Statistical parity is the most basic notion of fairness and it requires that the proportion of positive predictions is the same for all groups. It is related to the concept of quotas, i.e., policies requiring that a certain proportion of the population is represented in a certain activity, and it can be shown, from different approaches [GDBFL19, HPS16], that the best classifier that achieves statistical parity corresponds to having a (well-calibrated) score and different thresholds per group, chosen on the basis of having the same quantile across groups. However, this classifier can result in a very low accuracy. For this reason, the notions of equal opportunity (parity of true positive rate, TPR) and equal odds (parity of TPR and TNR) were introduced [HPS16], as they can be satisfied with higher accuracy levels and are still arguably fair in many contexts.

The two fields, privacy and fairness, have been studied together recently because they are both concerned with the correct management of data, and they appear together in many real-world scenarios. For instance, in the afore-

mentioned scandal of Cambridge Analytica [new18], not only the privacy of the users was violated, but also the fairness of the elections, and the use of profiles for targeted advertising can easily raise concerns about discrimination. On top of this, it was shown that the explanations for "why am I seeing this ad?" lack transparency [AVG+18].

Some authors have investigated the trade-offs and incompatibilities between fairness and differential privacy [PMK+20, CGKM19, Aga20], showing that sometimes it is simply impossible to achieve some notions of privacy, fairness and model accuracy simultaneously. It has been also shown under certain conditions that differential privacy has disparate (i.e., different on each sensitive group) but bounded impact on model accuracy [BPS19, MPBT23, TFVHY21]. Pruning also has a disparate effect [TFKN22], however, local differential privacy does not [AMP23]. Moreover, even in the absence of decision variables, fairness and privacy are tied together conceptually, because it is also important to measure the extent to which some subgroups are more susceptible to membership inference attacks than others [KYC+19, YKCT19]. Lastly, there are also attempts to connect causality with privacy and fairness [TSD20, MZP20a, DRBB+23], especially using the causality framework of Pearl [Pea09], and in general, there is great interest on studying the intersection of ethical concepts in machine learning.

In summary, the research in the domain of data ethics, including especially fairness in ML and privacy guarantees in data collection and transmission, has an enormous impact for society, and it has challenges related to the interconnectedness and subjectiveness of different perceptions of fairness and privacy.

1.2 Challenges

This dissertation explores the following challenges that arise in data ethics. The common denominator of these challenges is that they are all motivated by questioning some assumptions that are made in the literature and might not always align with real-world scenarios. It is important to recognize the value of these assumptions when they were first introduced. Thanks to some of the simplifications introduced by these assumptions, the theory has been

able to keep up with the fast-paced development of practical ML and the growing complexity of ML models and datasets. While these assumptions enable robust mathematical analysis and exact solutions and explanation, it is important to acknowledge their limitations and implications, and search for a balance between the rigor and the applicability of the results.

Challenge 1. In the field of fairness in ML, it can be shown that for data distributions with a binary sensitive attribute in the input and a binary output, when the input-output relationship is deterministic, the Bayes classifier achieves equal opportunity and maximal accuracy at the same time. This means that for deterministic data, the Pareto boundary of accuracy and opportunity difference degenerates to a single point, and ML models can be trained with a double objective of fairness and accuracy without compromising any of the two. For probabilistic (non-deterministic) data, however, neither this claim nor the Pareto boundary have been studied, and they should be, because in real life, the relationship between the input and the output is better modeled as probabilistic because input data is just a partial and incomplete representation of reality that does not capture all exogenous factors that can affect the output.

Challenge 2. Also in the field of fairness in ML, most fairness notions that are based on causality assume that the causal graph that explains the causal relationships in the data is known. In practice, the causal graph is estimated by means of causal discovery algorithms and there is a vast amount of these algorithms from which to choose from. A detailed analysis of whether the choice of the causal discovery algorithm affects the fairness conclusions is missing.

Challenge 3. In the field of private data collection, particularly of longitudinal data under local differential privacy, the existing algorithms provide formal guarantees only under the assumption that the data is constant, and not much is said of what they guarantee in absence of this assumption. The community is missing definitions to analyze, capture and quantify the privacy guarantees without this assumption. Moreover, not all the existing tools available for transversal data collection have been explored for longitudinal data collection. Particularly, no longitudinal method based on local hashing

exists and this lack of diversity may narrow-mind the development of future algorithms.

Challenge 4. In the field of privacy in data transmission, padding files and messages is a way of obfuscating the information an attacker may obtain about the data based on its size. Recently, some authors [RR21] derived a strategy for padding data while minimizing Shannon leakage (Shannon mutual information) between the observations and the secrets. The assumption that minimizing Shannon leakage is ideal for privacy applications has been challenged in the last two decades by several scholars [PR18, Smi09] who advocate for the use of Rényi min-leakage over Shannon leakage because the attacker associated to the former is realistic, whereas to the latter, it is not. Therefore, an open problem is that of deriving a strategy for computing padding schemes that minimize Rényi min-leakage.

1.3 Objectives, results and plan

The objective of this dissertation is to contribute to the community of researchers in data ethics by providing theoretical as well as practical results in unexplored areas of the domain, notably those described by Challenges 1, 2, 3 and 4 in Section 1.2. More precisely, this dissertation has four specific objectives, each of which corresponds to a challenge and a chapter of this manuscript. The diversity of topics covered in these 4 chapters is rooted at the rich and complex variety of topics comprehended by data ethics, as well as the wide scope of the main objective.

The 4 specific objectives are presented below, with a brief description of the main results of the corresponding chapter. These chapters resemble their papers of reference as much as possible, only with style changes to fit into this manuscript, to avoid forcing their readers to reread different variants of the same papers here. Chapter 3 is an exception to this, as it presents novel content.

The first objective of this dissertation is to:

Objective 1. (For Challenge 1) Determine whether for all data distributions it is possible to have a classifier that satisfies equal opportunity and non-

trivial accuracy simultaneously, and if not, explain why by studying the boundary of Pareto frontier between accuracy and opportunity difference.

This dilemma is heavily influenced by two contrasting results from the literature. On the one hand, for differentially private training algorithms, it is impossible to guarantee that the output model satisfies these two conditions [CGKM19]. On the other hand, the perfect model, whose prediction matches the correct decision always (something that can only occur with a deterministic input-output relationship), achieves both equal opportunity and maximal accuracy [HPS16], hence, also non-trivial accuracy.

This objective is achieved in Chapter 2. In Chapter 2, we prove that for certain distributions it is impossible to provide equal opportunity and non-trivial accuracy guarantees simultaneously. This theorem is then refined with necessary and sufficient conditions that characterize the data distributions for which this extreme trade-off occurs. Although these conditions are very severe to be found often in arbitrary datasets, they may still hold in practical scenarios, and we illustrate this with an example. In this chapter, we also provide an algorithm for visualizing the Pareto frontier between error and opportunity difference and prove connections with existing ways of visualizing and understanding the trade-off [HPS16].

Next, in Chapter 3, we turn our attention to causality for fairness, for which there are arguments in favor [MZP20b] and against [HPS16] in the literature. The objective of this chapter is to:

Objective 2. (For Challenge 2) Analyze the effect of using different causal discovery algorithms (CDAs) for estimating causal graphs, especially when these are then used for causal based fairness assessment.

This objective is achieved in our paper [BMP+23], and is reinforced with the experiments of Chapter 3. In Chapter 3, we create several examples of very simple distributions for which different CDAs disagree very probably and drastically on their output graphs, a fact that has a dramatic impact on causal based fairness assessment because the fairness conclusions are very sensitive to the graph structure. For this reason, we conclude that causal based fairness assessment should not be fully automated unless the causal graph is known.

In the next two chapters we turn our attention to privacy, which is mentioned in Chapter 2 because of its impact on fairness [CGKM19] but is not directly discussed in the previous chapters. We first take a look in Chapter 4 to Challenge 3 about ensuring privacy for repeated collection of data from the same population (also known as longitudinal data). The objective of this chapter is to:

Objective 3. (For Challenge 3) Provide an alternative to existing protocols [EPK14, DKY17] for longitudinal data collection with LDP guarantees that has similar utility and privacy guarantees, but based on local hashing and with a reduced number of memoized sanitizations of raw data.

In Chapter 4, we propose a protocol named LOLOHA for longitudinal categorical data collection with local differential privacy guarantees. LOLOHA is based on local hashing instead of unary encoding, and for similar levels of error and LDP guarantees for the first report, it provides increased privacy on the users' values, a notion of privacy that we defined for this specific setting, because, as we prove, it is impossible to satisfy pure local differential privacy for arbitrarily large windows of data collection.

Concluding the main body of the dissertation, we discuss the problem of padding files to reduce the information an intruder may infer based on their sizes during transmission. This is performed in Chapter 5, which has for objective to:

Objective 4. (For Challenge 4) Propose an algorithm for designing padding schemes that minimize Rényi min-entropy while also minimizing the average bandwidth overhead and respecting given size constraints.

In Chapter 5, we do exactly this. We derive an algorithm for designing padding schemes that enhance privacy of transmitted data at the expense of some bandwidth overhead. In this context, padding is used to obfuscate the information an attacker may obtain about the data based on its size, assuming that each user is downloading a file out of a pool of publicly known files. The main contribution is that our padding schemes minimize Rényi min-entropy instead of Shannon entropy, which makes the attack model closer to a real life attacker.

Finally, Chapter 6 (discussion) presents the main results of each chapter along with a critical analysis that justifies the relevance, limitations and future work for each of them, and Chapter 7 (conclusion) summarizes more concisely the whole manuscript.

1.4 Summary of contributions

The contributions of this dissertation are summarized below.

- We prove that for certain probabilistic distributions, no model can achieve equal opportunity and non-trivial accuracy simultaneously. In other words, for certain problematic distributions, all fair models have accuracy lower or equal than that of a constant classification model.
- We characterize with a sufficient and necessary condition the scenarios in which non-trivial accuracy and equal opportunity are compatible.
- We illustrate how incompatibility may arise in practice using the Adult dataset and describe phenomena that should be addressed when interpreting the theoretical results of this paper in practical scenarios.
- We show and depict several algebraic and geometric properties about the feasible region in the plane of opportunity-difference versus error.
- We provide an algorithm for computing all the vertices of the feasibility region, including those that form the Pareto-optimal boundary of the accuracy-fairness trade-off.

- We generate and provide minimal examples of distributions for which different causal discovery algorithms disagree strongly on the output graph. This complements our paper [BMP+23], in which its demonstrated that slight differences between causal graphs may have significant impact on fairness/discrimination conclusions.
- We add robustness to the conclusions based on these minimal examples by running all the experiments repeatedly and considering the effect of the sample size.

- We propose the LOLOHA protocol for longitudinal frequency monitoring of categorical data under LDP guarantees.
- We prove the longitudinal privacy and accuracy guarantees of LOLOHA through theoretical analysis and compare it to existing protocols.
- We show the performance of LOLOHA numerically and experimentally, using both real-world and synthetic datasets.

- We derive the algorithms that find the optimal padding schemes for the Pareto frontier between bandwidth overhead and information leakage in two different scenarios called PRP and POP (defined in Section 5.2).
- The code for these algorithms is publicly available at [PPS22]. It includes not only the algorithms we propose, but also the reimplementation of the algorithms of [RR21] to support flexible padding constraints, multiple files having the same size and sparse matrix representations.
- We prove the correctness of the algorithms and test the implementations against brute-force solutions using small synthetic datasets.
- We compare our algorithms with an existing solution [RR21] that uses a different attack model, and discuss how the two approaches are related in terms of the private information leakage type that each attacker represents.

1.5 List of publications

During my doctoral studies, I authored a total of 9 papers, 6 of which were submitted and accepted for publication, 2 are under development and 1 is just a short paper that implements an algorithm in Python. Although all of these papers are related in one way or another with privacy, fairness or ethics in computer science, only four of them are discussed in this dissertation to maintain brevity and consistency. I was the only PhD student authoring these four papers unless indicated otherwise. The exhaustive list of papers is presented below.

1. **On the impossibility of non-trivial accuracy in presence of**

fairness constraints [PPPV22]
This paper was presented by me at the conference AAAI 2022 (virtual) and it was published in the proceedings. It covers some of the topics discussed in Chapter 2. The abstract of this paper is omitted on purpose as this paper is contained in the next one.

2. **On the incompatibility of accuracy and equal opportunity** [PPPV23]
This journal article is an extension of the paper published in AAAI 2022, and it was published in the journal *Machine Learning 2023* by Springer. Chapter 2 presents the contents of this article.

Abstract. One of the main concerns about fairness in machine learning (ML) is that, in order to achieve it, one may have to trade off some accuracy. To overcome this issue, Hardt et al. [HPS16] proposed the notion of equality of opportunity (EO), which is compatible with maximal accuracy when the target label is deterministic with respect to the input features.

In the probabilistic case, however, the issue is more complicated: It has been shown that under differential privacy constraints, there are data sources for which EO can only be achieved at the total detriment of accuracy, in the sense that a classifier that satisfies EO cannot be more accurate than a trivial (i.e., constant) classifier [CGKM19]. In this paper, we strengthen this result by removing the privacy constraint. Namely, we show that for certain data sources, the most accurate classifier that satisfies EO is a trivial classifier. Furthermore, we study the admissible trade-offs between accuracy and EO loss (opportunity difference) and characterize the conditions on the data source under which EO and non-trivial accuracy are compatible.

3. **Causal discovery for fairness*** [BMP+23]
This paper was presented by me as invited talk at the NIPS workshop on algorithmic fairness through the lens of causality and privacy, and it was published in the proceedings of the conference. *For this paper, I worked with two other PhD students (Ruta Binkytė and Karima Makhlouf) who contributed to essential parts of the project. Chapter 3 contains an overview of this paper, but it is dedicated to additional

observations that complement it.

Abstract. Fairness guarantees that the ML decisions do not result in discrimination against individuals or minority groups. Identifying and measuring reliably fairness/discrimination is better achieved using causality which considers the causal relation, beyond mere association, between the sensitive attribute (e.g., gender, race, religion, etc.) and the decision (e.g., job hiring, loan granting, etc.). The big impediment to the use of causality to address fairness, however, is the unavailability of the causal model (typically represented as a causal graph). Existing causal approaches to fairness in the literature do not address this problem and assume that the causal model is available. In this paper, we do not make such an assumption, and we review the major algorithms to discover causal relations from observable data. This study focuses on causal discovery and its impact on fairness. In particular, we show how different causal discovery approaches may result in different causal models and, most importantly, how even slight differences between causal models can have significant impact on fairness/discrimination conclusions.

4. **Frequency estimation of evolving data under local differential privacy** [APPG23]
   I presented this paper at the conference EDBT 2023, and it was published in the proceedings. It is detailed in Chapter 4.

   Abstract. Collecting and analyzing evolving longitudinal data has become a common practice. One possible approach to protect the users' privacy in this context is to use local differential privacy (LDP) protocols, which ensure the privacy protection of all users even in the case of a breach or data misuse. Existing LDP data collection protocols such as Google's RAPPOR [EPK14] and Microsoft's dBitFlipPM [DKY17] can have longitudinal privacy linear to the domain size k, which is excessive for large domains, such as Internet domains. To solve this issue, in this paper we introduce a new LDP data collection protocol for longitudinal frequency monitoring named LOngitudinal LOcal HAshing (LOLOHA) with formal privacy guarantees. In addition, the privacy-utility trade-off of our protocol is only linear with respect to a reduced domain

size $2 \leq g \ll k$. LOLOHA combines a domain reduction approach via local hashing with double randomization to minimize the privacy leakage incurred by data updates. As demonstrated by our theoretical analysis as well as our experimental evaluation, LOLOHA achieves a utility competitive to current state-of-the-art protocols, while substantially minimizing the longitudinal privacy budget consumption by up to k/g orders of magnitude.

5. **Obfuscation padding schemes that minimize Rényi min-entropy for privacy** [SPPP23]
   This paper was presented by Sebastian Simon and Cezara Petrui at the conference ISPEC 2023, and it will be published in the conference proceedings. Sebastian and Cezara are two bachelor students that I met while being professor assistant at a computer programming course at École Polytechnique. They were under my supervision during an internship focused on this research project. The paper is presented in Chapter 5 and proposes obfuscation padding schemes that keep a balance between bandwidth increase and privacy against an attacker measuring network traffic.

   Abstract. Consider a set of users, each of which is choosing and downloading one file out of a central pool of public files, and an attacker that observes the download size for each user to identify the choice of each user. This paper studies the problem of padding the files to obfuscate the exact file sizes and minimize the expected accuracy of the attacker, without exceeding some given padding constraints. We derive the algorithm that finds the optimal padding scheme, prove its correctness, and compare it with an existing solution that uses a similar but different attack model. We also discuss how the two solutions are related in terms of private information leakage.

6. **Computing distributed knowledge as the greatest lower bound of knowledge** [PQRV21]
   This paper was presented by Sergio Ramirez and published in the proceedings of the conference RAMiCS 2021. It characterizes the standard notion of distributed knowledge of a group in lattice theory as the greatest lower bound of the join-endomorphisms representing the knowledge

of each member of the group. This framework is related with polarization and data ethics because it can be used to reason about opinions and model the belief dynamics of a population.

Abstract. Let L be a distributive lattice and $\mathcal{E}(L)$ be the set of join endomorphisms of L. We consider the problem of finding $f \sqcap_{\mathcal{E}(L)} g$ given L and $f, g \in \mathcal{E}(L)$ as inputs. (1) We show that it can be solved in time $O(n)$ where $n = |L|$. The previous upper bound was $O(n^2)$. (2) We characterize the standard notion of distributed knowledge of a group as the greatest lower bound of the join-endomorphisms representing the knowledge of each member of the group. (3) We show that deciding whether an agent has the distributed knowledge of two other agents can be computed in time $O(n^2)$ where n is the size of the underlying set of states. (4) For the special case of $S5$ knowledge, we show that it can be decided in time $O(n\alpha_n)$ where α_n is the inverse of the Ackermann function.

7. **Counting and computing join-endomorphisms in lattices (revisited)** [PQR+22]
This is a journal article under development that extends a paper published in the proceedings of RAMiCS 2020 with the same title. The article explores some properties of lattices and join-endomorphisms. The relationship with data ethics is the same as the previous work in this list, and Santiago Quintero participated in this paper as well.

Abstract. Structures involving a lattice and join-endomorphisms on it are ubiquitous in computer science. We study the cardinality of the set $\mathcal{E}(L)$ of all join-endomorphisms of a given finite lattice L. In particular, we show for $\mathcal{M}_n$, the discrete order of n elements extended with top and bottom, $|\mathcal{E}(\mathcal{M}_n)| = n!\mathcal{L}_n(-1) + (n+1)^2$ where $\mathcal{L}_n(x)$ is the Laguerre polynomial of degree n. We also study the following problem: Given a lattice L of size n and a set $S \subseteq \mathcal{E}(L)$ of size m, find the greatest lower bound $\sqcap_{\mathcal{E}(L)} S$. The join-endomorphism $\sqcap_{\mathcal{E}(L)} S$ has meaningful interpretations in epistemic logic, distributed systems, and Aumann structures. We show that this problem can be solved with worst-case time complexity in $O(mn)$ for distributive lattices and $O(mn + n^3)$ for arbitrary lattices. In the particular case of modular lattices, we

present an adaptation of the latter algorithm that reduces its average time complexity. We provide theoretical and experimental results to support this enhancement. The complexity is expressed in terms of the basic binary lattice operations performed by the algorithm.

8. **Fast Python sampler of the von Mises Fisher distribution** [PJ23] This short article is available at INRIA's HAL service. It implements a method for sampling from the d-dimensional Von Mises Fisher distribution using NumPy, focusing on speed and readability. The implementation differs with those found online and in packages like TensorFlow in that it skips the inversion of a matrix, thus saving some milliseconds, especially when generating thousands of samples each from a VMF distribution with different specified mean. The Von Mises Fisher distribution is related with data ethics because it can be used during the training of neural networks for adding noise to the iterative updates of stochastic gradient descent to achieve differential privacy with respect to the training data.

   Abstract. This paper implements a method for sampling from the d-dimensional Von Mises Fisher distribution using NumPy, focusing on speed and readability. The complexity of the algorithm is O(n d) for n samples, which is theoretically optimal taking into account that n d is the output size.

9. **On the impact of local privacy on the learnability of causal structures**
   This is a paper in development that studies a problem in the intersection between causality and privacy.

   Abstract. Differential privacy is one of the most popular frameworks to protect the sensitive information of the original data providers of a data set. It is based on the application of controlled noise at the interface between the server that stores and processes the data, and the data consumers. Local differential privacy is a variant that allows data providers to apply the privatization mechanism themselves on their data individually. Therefore, it provides protection also in contexts in which the server, or even the data collector, cannot be trusted. The addition of noise, however, affects the utility of the data. In particular,

it can distort the correlation between the individual components of the data, thus hurting tasks such as causal-structure learning. In this paper, we consider various well-known locally differentially private mechanisms and compare the trade-off between the privacy they provide, and the accuracy of the causal structure produced by algorithms for causal learning when applied to data obfuscated by these mechanisms. Our analysis provides valuable insights into selecting appropriate local differentially private protocols for causal discovery tasks. We hope that our work will benefit researchers and practitioners by enabling them to conduct causal discovery while preserving the privacy of users' data.

Chapter 2

On the incompatibility of accuracy and equal opportunity

During the last decade, the intersection between machine learning and social discrimination has gained considerable attention from academia, industry, and the public in general. A similar trend occurred before between machine learning and privacy, and even the three fields have been studied together recently [PMK+20, CGKM19, KR19, Aga20].

Fairness has proven to be harder to conceptualize than privacy, for which differential privacy [Dwo06a] has become the de-facto definition. Fairness is subjective and laws vary between countries. Even in academia, depending on the application, the words fairness and bias have different meanings [Cra17]. The current consensus is that fairness cannot be summarized into a unique universal definition, which has led to a wide range of fairness definitions [MZP21b], and for the most popular definitions, several trade-offs, implementation difficulties, and impossibility theorems have been found [KMR17, Cho17]. One such definition of fairness is equal opportunity [HPS16], which is one of the most common group notions of fairness along with disparate impact, demographic parity, and equalized odds [PS22]. Equal opportunity is restricted to binary classification tasks with binary sensitive attributes.

To contrast equal opportunity (EO) with accuracy, we borrow the notion of

trivial accuracy from [CGKM19]. A *non-trivial* classifier is one that has higher accuracy than any constant classifier. Since constant classifiers are independent of the input, trivial accuracy determines a very low-performance level that any correctly trained classifier should overcome. Yet, as shown in related works [CGKM19, Aga20], under the simultaneous constraints of differential privacy and equal opportunity, it is impossible to have non-trivially accurate classifiers.

In this chapter, we strengthen the result of [CGKM19, Aga20] by showing that, even without the assumption of differential privacy, there are distributions for which equal opportunity implies trivial accuracy. In particular, this is possible when the data source is probabilistic, i.e., the correct label for a given input is not necessarily unique.

Probability plays two different roles in this chapter. On the one hand, we allow classifiers to be probabilistic, i.e., we allow the classification to be influenced by controlled randomness for some inputs. This is needed because satisfying equal opportunity typically requires a probabilistic predictor [HPS16], but also because it has a practical justification. Namely, in some cases, randomness is the only fair way to distribute an indivisible limited resource. For instance, a parent with one candy and two children might throw a coin to decide whom to give it to. This principle is even applied in decisions that have a significant social impact such as the Diversity Visa Program to qualify for a Green Card in the United States [Sta21], and the Beijing lottery for getting a car license plate [Glo18].

On the other hand, we consider probabilistic data sources. This provides a more general framework for studying the trade-off between fairness and accuracy, as there are situations in which reality is more accurately represented by a probabilistic model. For instance, the information carried by the input may be insufficient to conclude definitely the yes-no decision, or there may be constraints that force the decision to be different for identical inputs.

The analysis of this chapter is mostly theoretical and is limited to the notion of equal opportunity, hence to distributions with binary targets and binary sensitive attributes. On the other hand, the results are very general. For instance, whenever we state a property about all predictors, it includes all probabilistic classifiers without exception. Hence, our results hold for clas-

sifiers that do not use the sensitive attribute for prediction, as well as for those that use it to compensate existing biases, or take into account proxy variables, or use multiple threshold mechanisms, or are based on causality, or do not use machine learning at all.

The contributions of this chapter can be summarized as follows.

1. We prove that for certain probabilistic distributions, no predictor can achieve EO and non-trivial accuracy simultaneously.
2. We explain how to modify existing results that assume deterministic data sources to the probabilistic case:
   (a) We prove that for certain distributions, the Bayes classifier does not satisfy EO. As a consequence, in these cases, EO can only be achieved by trading-off some accuracy.
   (b) We give necessary and sufficient conditions for non-trivially accurate predictors to exist.
3. We prove and depict several algebraic and geometric properties about the feasibility region, i.e., the region containing all predictors in the plane of opportunity difference versus error.
4. We determine necessary and sufficient conditions under which non-trivial accuracy and EO are compatible.
5. We develop an algorithm that computes the Pareto-optimal boundary of the accuracy-fairness trade-off, and more generally, the feasibility region.
6. We illustrate how the incompatibility between EO and non-trivial accuracy may arise in practice.
7. We discuss the distortion effect that arises when we use the above algorithm on empirical distributions from sampled data.

For reproducibility, we published a repository [Pin22] with Python code for generating the figures and algorithms mentioned in this chapter, including Algorithms 1 and 2.

The rest of the chapter is organized as follows. Section 2.1 discusses related

work. Section 2.2 recalls the preliminary notions that are used in the rest of the document. Section 2.3 introduces the plane of error versus opportunity difference and shows several geometric properties taking place in this plane. Section 2.4 presents the impossibility result: for certain probabilistic distributions, no predictor can achieve EO and non-trivial accuracy simultaneously. Section 2.5 compares deterministic sources against probabilistic ones, and shows how to modify existing results that hold in the former case to guarantee them in the latter. Section 2.6 presents algorithms for computing the Pareto-optimal frontier and all the vertices of the feasibility region in the plane of error versus opportunity difference. Section 2.7 states the necessary and sufficient conditions under which there exist predictors achieving EO and non-trivial accuracy simultaneously. Section 2.8 shows an example of the impossibility result arising in (a variant of) a real-life dataset. Section 2.9 discusses the distortion on the Pareto-optimal frontier when we compute and evaluate it using empirical distribution from sampled data. Section 2.10 draws the conclusion.

The contents of this chapter were published in the Machine Learning journal (Springer)[PPPV23]. A preliminary and partial version appeared in the proceedings of AAAI 2022 [PPPV22]. The differences with respect to the AAAI-2022 version are that here, we study the Pareto-optimal boundary (Section 2.6), the necessary and sufficient conditions (Theorem 19) that characterize the impossibility between equal opportunity and non-trivial accuracy, and we present a practical example based on the Adult dataset (Figure 2.9).

2.1 Related Work

This chapter contributes to the technical literature about equal opportunity (EO) [HPS16], one of the most common group fairness notions [PS22]. For an overview of when EO is appropriate and how EO relates to other fairness notions, the reader is referred to the survey papers [MZP21b, PS22, CH20, MMS+21] and the moral framework in [HLGK19].

This chapter is strongly related to the following two papers that consider a randomized learning algorithm guaranteeing (exact) EO and also satisfying differential privacy: [CGKM19] shows that, for certain distributions, these

constraints imply trivial accuracy. [Aga20] proves the same claim for any arbitrary distribution and for non-exact EO, i.e., bounded opportunity difference. It also highlights that, although there appears to be an error in the proof of [CGKM19], the statement is still correct. In contrast, in this chapter, we prove the existence of particular distributions in which trivial accuracy is implied directly from the (exact) EO constraint, without any differential privacy assumption.

There are also several works that focus on the incompatibility of fairness constraints. In [KMR17], it is shown that several fairness notions cannot hold simultaneously, except for exceptional cases. Similarly, in [LCM18], it is shown that the two main legal notions of discrimination are in conflict for some scenarios. In particular, when impact parity and treatment parity are imposed, the learned model seems to decide based on irrelevant attributes. These works reveal contradictions when different notions of fairness are imposed together.

In contrast, [CDG18] show issues inherent to anti-classification, classification parity, and calibration, separately, without inducing them simultaneously with another fairness notion. Regarding equal opportunity in the COMPAS case, they show that forcing equal and low false positive rates obliges the system to decide almost randomly (trivially) for black defendants. Our work presents theoretical scenarios in which this problem is even more extreme and the system becomes trivial for both classes. As shown in our sufficiency and necessary conditions, the extreme scenarios are characterized based on six population statistics. In this sense, this chapter is also related to [SYT20], which computes bounds on fairness and accuracy based on population statistics.

Lastly, in comparison to the seminal paper on equal opportunity [HPS16], this chapter uses a different geometric approach. Graphically, their analysis is carried out using ROC curves of fixed predictors. In contrast, we plot directly the error and the difference in opportunity of the two sensitive groups. In Section 2.4, Figure 2.9, we depict side by side the two perspectives. In this sense, we provide a complementary geometric perspective for analyzing equal opportunity and accuracy together.

2.2 Preliminaries

The notation described in this section is summarized in Table 2.1.

We consider the problem of binary classification with a binary protected feature. *Protected features*, also called sensitive attributes or sensitive features, are input features that represent race, gender, religion, nationality, age, or any other variable that could potentially be used to discriminate against a group of people. A feature may be considered a protected feature in some contexts and not in others, depending on whether the classification task should ideally consider that feature or not. For our purposes, we assume the simple and fundamental case in which there is a single protected attribute that can only take two values, e.g., man or woman, or, religious or non-religious.

Data Source

We consider an observable underlying statistical model consisting of three random variables over a probability space $(\Omega, \mathcal{E}, \mathbb{P})$: the *protected feature* $A : \Omega \to \{0,1\}$, the *non-protected feature vector* $X : \Omega \to \mathbb{R}^d$ for some positive integer d, and the *target label* $Y : \Omega \to \{0,1\}$. We refer to this statistical model as the *data source*.

The distribution of (X, A) is denoted by the measure π that computes for each ((X,A)-measurable) event $E \subseteq \mathbb{R}^d \times \{0,1\}$, the probability $\pi(E) \stackrel{\text{def}}{=} \mathbb{P}[\,(X,A) \in E\,]$. To reduce the verbosity of the discrete case, we denote the probability mass function as $\pi(x,a) \stackrel{\text{def}}{=} \pi(\{(x,a)\})$, i.e., $\pi(x,a) = \mathbb{P}[\,X = x, A = a\,]$.

The expectation of Y conditioned on (X, A) is denoted both as the function $q(x,a) \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X = x, A = a\,]$ and the random variable $Q \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X, A\,] = q(X, A)$. Importantly, the notation $\mathbb{E}[\,Y \mid X = x, A = a\,]$ for defining $q(x,a)$ is not an expectation conditioned on the possibly null event $(X = x, A = a)$. Instead, it is syntactic sugar for the conditional expectation function. Formally speaking, the function q is not necessarily unique in the way it is defined. It is defined almost everywhere uniquely, so that for any alternative conditional expectation function q', we have $q(X,A) = q'(X,A)$ almost surely. Throughout this chapter, we prioritize studying the discrete case to avoid this extreme level of formalism without losing rigor.

The random variable Q plays the role of a soft target label because, since $q(x,a) = \mathbb{P}[\,Y{=}1 \,|\, X{=}x, A{=}a\,]$, then Y can be modeled as a Bernoulli random variable with success probability Q.

The distribution of (X, A, Y) is completely characterized by the pair (π, q), hence we refer to this pair as the distribution of the data source. And we distinguish two cases: the data source is *probabilistic* in general, but if $Q \in \{0,1\}$ (with probability 1), then it is said to be *deterministic*. This distinction is crucial, because several statements hold exclusively in one of the two cases.

(X, A, Y)	Data source
X	Non-protected feature vector in $\mathbb{R}^d$
A	Protected feature in $\{0,1\}$
Y	Target label in $\{0,1\}$
Q, q	Soft target label $Q \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X, A\,]$
π	Distribution of (X, A)
(π, q)	Distribution of (X, A, Y)
$\hat{Q}, \hat{q}$	Predictor $\hat{Q} = \hat{q}(X, A) = \mathbb{E}[\,\hat{Y} \mid X, A\,]$
$\hat{Y}$	Predicted label in $\{0,1\}$
$\mathcal{Q}$	Set of all predictors
$\text{acc}(\hat{Q})$	Accuracy of $\hat{Q}$: $\mathbb{P}[\,\hat{Y}{=}Y\,]$
$\text{oppDiff}(\hat{Q})$	Opportunity difference of $\hat{Q}$: $\mathbb{E}[\,\hat{Q} \mid Y = 1, A = 1\,] - \mathbb{E}[\,\hat{Q} \mid Y = 1, A = 0\,]$

Table 2.1: The notation used in this chapter.

Classifiers and Predictors

Analogously to the data source, we model the estimation $\hat{Y}$ as a Bernoulli random variable with success probability $\hat{Q} = \hat{q}(X, A)$ for some ((X, A)-measurable) function $\hat{q}$. We refer to $\hat{Y}$ as a (hard) *classifier*, and to $\hat{Q}$ or $\hat{q}$ as a (soft) *predictor*. Notice that $\hat{Y}$ is deterministic when $\hat{Q} \in \{0,1\}$ (with probability 1), in which case, $\hat{Y} = \hat{Q}$ (with prob. 1). Hence, all deterministic classifiers are also predictors.

The set of all soft predictors is denoted as $\mathcal{Q}$. We highlight the following predictors in $\mathcal{Q}$:

1. the two constant classifiers, $\hat{0}$ and $\hat{1}$, given by $\hat{0}(x,a) \stackrel{\text{def}}{=} 0$ and $\hat{1}(x,a) \stackrel{\text{def}}{=} 1$,
2. for each $\hat{Q} \in \mathcal{Q}$, the $1/2$-threshold classifier given by $\hat{Q}_{1/2} \stackrel{\text{def}}{=} \mathbf{1}\{\hat{Q} > 1/2\}$,
3. the data source soft target Q, and
4. the Bayes classifier $Q_{1/2} = \mathbf{1}\{Q > 1/2\}$.

It is well known[1] that the Bayes classifier $Q_{1/2}$ has minimal error among all predictors in $\mathcal{Q}$, regardless of whether the data source is deterministic or not.

Evaluation Metrics

To refer to *equal opportunity* [HPS16], we introduce a continuous metric called the *opportunity difference*. The opportunity difference of a predictor $\hat{Q} \in \mathcal{Q}$ is defined as

$$\text{oppDiff}(\hat{Q}) \stackrel{\text{def}}{=} (\,\mathbb{P}[\,\hat{Y}{=}1 \,|\, A{=}1, Y{=}1\,] - \mathbb{P}[\,\hat{Y}{=}1 \,|\, A{=}0, Y{=}1\,]\,),$$

and a predictor $\hat{Q} \in \mathcal{Q}$ is said to satisfy equal opportunity if and only if $\text{oppDiff}(\hat{Q}) = 0$. Alternatively, the opportunity difference can be computed using the formula in Lemma 1.

Lemma 1. *For any predictor $\hat{Q}$, and assuming $\mathbb{P}[\,Y{=}1, A{=}a\,] > 0$ for each $a \in \{0,1\}$, we have*

$$\mathbb{P}[\,\hat{Y}{=}1 \,|\, Y{=}1, A{=}a\,] = \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]},$$

hence also

$$\mathit{oppDiff}(\hat{Q}) = \frac{\mathbb{E}[\,\hat{Q}Q \mid A = 1\,]}{\mathbb{E}[\,Q \mid A = 1\,]} - \frac{\mathbb{E}[\,\hat{Q}Q \mid A = 0\,]}{\mathbb{E}[\,Q \mid A = 0\,]}.$$

As an additional consequence, by considering the symmetric predictor $1 - \hat{Q}$, it is also true that

$$\mathbb{P}[\,\hat{Y}{=}0 \,|\, Y{=}1, A{=}a\,] = \frac{\mathbb{E}[\,(1-\hat{Q})Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]}.$$

[1] See for instance Chapter 3 of [Fuk13].

Proof. Indeed, by applying repetitively the Bayes rule, we get

$$\begin{aligned}\mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}1, A{=}a\,] &= \frac{\mathbb{P}[\,\hat{Y}{=}1, Y{=}1, A{=}a\,]}{\mathbb{P}[\,Y{=}1, A{=}a\,]} = \frac{\mathbb{P}[\,A{=}a\,]\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{P}[\,Y{=}1, A{=}a\,]} \\ &= \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{P}[\,Y{=}1\,|\,A{=}a\,]} = \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]}.\end{aligned}$$

The second equality holds because $(Y, \hat{Y}) \perp A \mid (Q, \hat{Q})$ and $Y \perp \hat{Y} \mid (Q, \hat{Q})$, which, destructuring the conditional independence notation, means that for almost any $q, \hat{q} \in [0, 1]$ (i.e., for all $q, \hat{q} \in S$ for some set S with $\mathbb{P}[\,(Q, \hat{Q}){\in}S\,] = 1$), when conditioning on $Q = q$ and $\hat{Q} = \hat{q}$, we have $Y \perp A$ and $\hat{Y} \perp A$ and $Y \perp \hat{Y}$. □

The *error* and the *accuracy* of a predictor $\hat{Q} \in \mathcal{Q}$ are defined as

$$\begin{aligned}\mathrm{err}(\hat{Q}) &\stackrel{\text{def}}{=} \mathbb{P}[\,\hat{Y}{\neq}Y\,], \\ \mathrm{acc}(\hat{Q}) &\stackrel{\text{def}}{=} 1 - \mathrm{err}(\hat{Q}).\end{aligned}$$

As shown in Lemma 2, the error can also be computed as $\mathrm{err}(\hat{Q}) = \mathbb{E}[\,Q{+}\hat{Q}{-}2Q\hat{Q}\,]$.

Lemma 2. *For every $\hat{Q} \in \mathcal{Q}$,*

$$\mathit{err}(\hat{Q}) = \mathbb{E}[\,|\hat{Q}{-}Y|\,].$$

Proof. Notice that $\mathbb{P}[\,\hat{Y}{\neq}Y\,|\,Y{=}1\,] = \mathbb{P}[\,\hat{Y}{=}0\,|\,Y{=}1\,] = \mathbb{E}[\,1 - \hat{Q} \mid Y = 1\,]$ and $\mathbb{P}[\,\hat{Y}{\neq}Y\,|\,Y{=}0\,] = \mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}0\,] = \mathbb{E}[\,\hat{Q} \mid Y = 0\,]$. In both cases, we may write $\mathbb{P}[\,\hat{Y}{\neq}Y\,|\,Y{=}y\,] = \mathbb{E}[\,|Y - \hat{Q}| \mid Y = y\,]$.

Hence, marginalizing over Y we conclude $\mathbb{P}[\,\hat{Y}{\neq}Y\,] = \mathbb{E}[\,|Y{-}\hat{Q}|\,]$.

□

As mentioned in the previous subsection, the maximal level of accuracy is achieved by the Bayes classifier $Q_{1/2}$. Moreover, as shown in Lemma 3 its exact value is $1/2 + \mathbb{E}[\,|Q{-}1/2|\,]$.

Lemma 3. *(Bayes accuracy)*

$$\mathit{acc}(Q_{1/2}) = 1/2 + \mathbb{E}[\,|Q{-}1/2|\,].$$

Proof. Out of Lemma 2, we know $\mathrm{err}(Q_{1/2}) = \mathbb{E}[\,\epsilon\,]$ where $\epsilon \stackrel{\text{def}}{:=} |Q_{1/2} - Y|$. Let us condition on $Q < 1/2$ and $Q \geq 1/2$ separately (whenever these events have possible probabilities).

For $Q < 1/2$, we have $\mathbb{E}[\,\epsilon \mid Q < 1/2\,] = \mathbb{E}[\,Y \mid Q < 1/2\,] = \mathbb{E}[\,Q \mid Q < 1/2\,]$ and $Q = 1/2 - (1/2 - Q)$. For $Q \geq 1/2$, we have $\mathbb{E}[\,\epsilon \mid Q \geq 1/2\,] = \mathbb{E}[\,1 - Y \mid Q \geq 1/2\,] = \mathbb{E}[\,1 - Q \mid Q \geq 1/2\,]$ and $1 - Q = 1/2 - (Q - 1/2)$.

These cases partition Ω and in both cases we have $\mathbb{E}[\,\epsilon\,] = 1/2 - \mathbb{E}[\,|1/2 - Q|\,]$. It follows that $\mathrm{err}(Q_{1/2}) = 1/2 - \mathbb{E}[\,|Q - 1/2|\,]$.

□

This maximal level of accuracy is also achieved by the *alternative Bayes classifier* given by $\mathbf{1}\{q(x,a) \geq 1/2\}$ ($\geq$ instead of $>$). This is shown in Lemma 5, which uses Lemma 4 to express the error of the predictor when Q is exactly $1/2$.

Lemma 4. *(Conditioning on $1/2$) Let $\hat{Q} \in \mathcal{Q}$ and consider the random variable $|\hat{Q} - Y|$ of the error of $\hat{Q}$ according to Lemma 2. If $\mathbb{P}[\,Q = 1/2\,] > 0$, then $\mathbb{E}[\,|\hat{Q} - Y| \mid Q = 1/2\,] = 1/2$.*

Proof. Define $r \stackrel{\text{def}}{:=} \mathbb{E}[\,\hat{Q}\,]$. Let us condition on $Y = 0$ and $Y = 1$ separately. For $Y = 0$, we have $\mathbb{E}[\,|\hat{Q} - Y| \mid Q = 1/2, Y = 0\,] = \mathbb{E}[\,\hat{Q} \mid Q = 1/2\,] = \mathbb{E}[\,\hat{Q}\,] = r$, and for $Y = 1$, we have $\mathbb{E}[\,|\hat{Q} - Y| \mid Q = 1/2, Y = 1\,] = 1 - \mathbb{E}[\,\hat{Q} \mid Q = 1/2\,] = 1 - \mathbb{E}[\,\hat{Q}\,] = 1 - r$.

Since it holds that $\mathbb{P}[\,Y = y \mid Q = 1/2\,] = 1/2$, we can compute the marginal as $\mathbb{E}[\,|\hat{Q} - Y| \mid Q = 1/2\,] = (1/2)(r + 1 - r) = 1/2$. □

Lemma 5. *(Alternative Bayes) The* alternative Bayes classifier $Q_{1/2}$ *given by $\mathbf{1}\{q(x,a) \geq 1/2\}$ ($\geq$ instead of $>$) has also maximal accuracy.*

Proof. We will prove that $\mathrm{err}(Q_{1/2}) = \mathrm{err}(Q^*_{1/2})$. Following Lemma 2, let $\epsilon \stackrel{\text{def}}{:=} |Q_{1/2} - Y|$ and $\epsilon^* \stackrel{\text{def}}{:=} |Q^*_{1/2} - Y|$.

Conditioned to $Q \neq 1/2$ we have $Q_{1/2} = Q^*_{1/2}$ from their definitions, and thus also $\mathbb{E}[\,\epsilon - \epsilon^* \mid Q \neq 1/2\,] = 0$. It suffices to check the complement event $Q = 1/2$. Suppose $\mathbb{P}[\,Q = 1/2\,] > 0$. Conditioned to $Q = 1/2$, Lemma 4 implies

that $\mathbb{E}[\,\epsilon - \epsilon^* \mid Q = 1/2\,] = 1/2 - 1/2 = 0$.

Hence, $\mathbb{E}[\,\epsilon\,] = \mathbb{E}[\,\epsilon^*\,]$, i.e., $\mathrm{err}(q_{1/2}) = \mathrm{err}(q^*_{1/2})$.

□

With regard to the error and the accuracy, we consider a minimal reference level of accuracy that should be outperformed intuitively by any well-trained predictor. The *trivial accuracy* [CGKM19] is defined as $\tau \stackrel{\text{def}}{=} \max\left\{\mathrm{acc}(\hat{Q}) : \hat{Q} \in \mathrm{Triv}\right\}$, where Triv is the set of (trivial) predictors whose output does not depend on X and A at all, and as a consequence is independent of Y as well. In other words, Triv consists of all constant soft predictors $\mathrm{Triv} \stackrel{\text{def}}{=} \{((x,a) \mapsto c) : c \in [0,1]\}$. According to the Neyman-Pearson Lemma, the most accurate trivial predictor is always hard, i.e., must be either $\hat{0}$ or $\hat{1}$. Thus, τ is well-defined and can be computed as

$$\tau = \max\{\mathbb{P}[\,Y{=}0\,],\, \mathbb{P}[\,Y{=}1\,]\} = 1/2 + |\mathbb{E}[\,Y\,] - 1/2|. \qquad \text{(due to Lemma 6)}$$

Lemma 6. *(Trivial error as an expectation)*

$$\tau = 1/2 + |\mathbb{E}[\,Y\,] - 1/2|.$$

Proof. The constant 0 predictor ($\hat{0}$) has error $\mathbb{E}[\,Y\,]$, while the constant 1 predictor ($\hat{1}$) has error $1 - \mathbb{E}[\,Y\,]$. We can rewrite these quantities respectively as $1/2 - (1/2 - \mathbb{E}[\,Y\,])$ and $1/2 + (1/2 - \mathbb{E}[\,Y\,])$, whose maximum is $\tau = 1/2 + |1/2 - \mathbb{E}[\,Y\,]|$.

□

Lastly, a predictor $\hat{Q} \in \mathcal{Q}$ is said to be *trivially accurate* if $\mathrm{acc}(\hat{Q}) \leq \tau$, and *non-trivially accurate*, or *non-trivial* otherwise. Notice that for a degenerated data source in which the decision Y is independent of X and A, all predictors are forcibly trivially accurate.

2.3 The Feasibility Region

In this section, we analyze the region $M \subseteq [0,1] \times [-1,+1]$ given by

$$M \stackrel{\text{def}}{=} \{(\mathrm{err}(\hat{Q}), \mathrm{oppDiff}(\hat{Q})) : \hat{Q} \in \mathcal{Q}\},$$

which represents the feasible combinations of the evaluation metrics (error and opportunity difference) that can be obtained for a given source distribution (π, q). This region determines the tension between error and opportunity difference. Figure 2.1 shows an example of the region M.

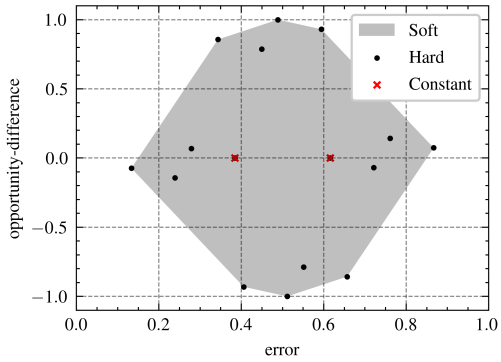


Figure 2.1: Region M for an arbitrary source distribution.

The results presented in this section assume that the data source is discrete, and its range is finite. We will use the following vectorial notation to represent both the distribution (π, q) and any arbitrary predictor $\hat{Q} \in \mathcal{Q}$.

Definition 1. Suppose (X, A) can only take a finite number of outcomes $\{(x_i, a_i)\}_{i=1}^n$ (each with positive probability) for some integer $n > 0$. In order to represent π, q and any $\hat{Q} \in \mathcal{Q}$ respectively, let $\vec{P}, \vec{Q}, \vec{F} \in \mathbb{R}^n$ be the vectors given by

$$\begin{aligned}
\vec{P}_i &\stackrel{\text{def}}{\coloneqq} \mathbb{P}[\,X{=}x_i, A{=}a_i\,], && \text{(Discrete } \pi\text{)} \\
\vec{Q}_i &\stackrel{\text{def}}{\coloneqq} \mathbb{P}[\,Y{=}1 \,|\, X{=}x_i, A{=}a_i\,], && \text{(Discrete } q\text{)} \\
\vec{F}_i &\stackrel{\text{def}}{\coloneqq} \mathbb{P}[\,\hat{Y}{=}1, X{=}x_i, A{=}a_i\,]. && \text{(Discrete } \hat{q}\text{)}
\end{aligned}$$

For notation purposes, let also $\vec{Q}^{(0)}, \vec{Q}^{(1)} \in \mathbb{R}^n$ be given by $\vec{Q}_i^{(a)} \stackrel{\text{def}}{\coloneqq} \vec{Q}_i \cdot \mathbf{1}\{a_i = a\}$, and in order to match (as we will show in Lemma 7) the definition of $\text{err}(\hat{Q})$ and $\text{oppDiff}(\hat{Q})$, let

$$\begin{aligned}
\text{err}(\vec{F}) &\stackrel{\text{def}}{\coloneqq} \langle \vec{P}, \vec{Q} \rangle + \langle \vec{F}, 1{-}2\vec{Q} \rangle, \\
\text{oppDiff}(\vec{F}) &\stackrel{\text{def}}{\coloneqq} \frac{\langle \vec{F}, \vec{Q}^{(1)} \rangle}{\langle \vec{P}, \vec{Q}^{(1)} \rangle} - \frac{\langle \vec{F}, \vec{Q}^{(0)} \rangle}{\langle \vec{P}, \vec{Q}^{(0)} \rangle},
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product $\langle u, v \rangle \stackrel{\text{def}}{\coloneqq} u_1 v_1 + \cdots + u_n v_n$. *(End)*

Regarding Definition 1, we highlight four important remarks:

1. $\vec{Q} \in [0,1]^n$, $\vec{P} \in (0,1]^n$, $\|\vec{P}\|_1 = 1$ and $\vec{F}$ lies in the rectangular n-dimensional box given by

$$0 \preceq \vec{F} \preceq \vec{P},$$

where $\preceq$ denotes the component-wise order in $\mathbb{R}^n$, i.e., $0 \leq \vec{F}_i \leq \vec{P}_i$ for each $i \in \{1, ..., n\}$. Moreover, from the definition of $\vec{P}$ and $\vec{F}$, the vertices of this rectangular box correspond precisely with the deterministic predictors.

2. The vectorial definitions of error and opportunity difference correspond to those of the non-vectorial case. Moreover, their gradients are constant.

3. There is a one-to-one correspondence between the predictors $\hat{q} \in \mathcal{Q}$ and the vectors $\vec{F}$ that satisfy $0 \preceq \vec{F} \preceq \vec{P}$. Indeed, each predictor is uniquely given by its pointwise values $\hat{q}(x_i, a_i) = \frac{\vec{F}_i}{\vec{P}_i}$ and each vector by its pointwise coordinates $\vec{F}_i = \vec{P}_i \hat{q}(x_i, a_i)$. This transformation preserves all the evaluation metrics, a fact proved in Lemma 7, therefore

$$M = \{(\text{err}(\vec{F}), \text{oppDiff}(\vec{F})) : 0 \preceq \vec{F} \preceq \vec{P}\}.$$

4. There is an inherent symmetry in the space of vectors $\vec{F}$ that satisfy $0 \preceq \vec{F} \preceq \vec{P}$. This symmetry is detailed in Lemma 8.

Lemma 7. *(Vectorial metrics) Using the notation of Definition 1, we have*

$$\begin{aligned} \textit{err}(\hat{Q}) &= \textit{err}(\vec{F}), \\ \textit{oppDiff}(\hat{Q}) &= \textit{oppDiff}(\vec{F}). \end{aligned}$$

Proof. For the error, we marginalize over (X, A). Notice

$$\begin{aligned} \mathbb{P}[\, Y \neq \hat{Y} \,|\, X = x_i, A = a_i \,] &= (1 - q(x_i, a_i))\hat{q}(x_i, a_i) + q(x_i, a_i)(1 - \hat{q}(x_i, a_i)) \\ &= (1 - \vec{Q}_i)\frac{\vec{F}_i}{\vec{P}_i} + \vec{Q}_i \frac{\vec{P}_i - \vec{F}_i}{\vec{P}_i} = \frac{\vec{Q}_i \vec{P}_i + \vec{F}_i(1 - 2\vec{Q}_i)}{\vec{P}_i}. \end{aligned}$$

Thus, $\mathbb{P}[\,Y\neq\hat{Y}, X=x_i, A=a_i\,] = \vec{Q}_i\vec{P}_i + \vec{F}_i(1-2\vec{Q}_i)$, and

$$\begin{aligned}\mathrm{err}(\hat{Q}) = \mathbb{P}[\,\hat{Y}\neq Y\,] &= \sum_{i=1}^{n}\mathbb{P}[\,\hat{Y}\neq Y, X=x_i, A=a_i\,]\\ &= \langle\vec{P},\vec{Q}\rangle + \langle\vec{F},1-2Q\rangle = \mathrm{err}(\vec{F}).\end{aligned}$$

For opportunity difference, we also marginalize over (X, A). Notice that

$$\begin{aligned}\mathbb{P}[\,\hat{Y}=1, Y=1, X=x_i, A=a_i\,] &= \vec{P}_i\,\mathbb{E}[\,\hat{Q}Q \mid X=x_i, A=a_i\,]\\ &= \vec{P}_i\frac{\vec{F}_i}{\vec{P}_i}\vec{Q}_i = \vec{F}_i\vec{Q}_i,\end{aligned}$$

hence $\mathbb{P}[\,\hat{Y}=1, Y=1, X=x_i, A=a\,] = \vec{F}_i\vec{Q}_i^{(a)}$. In addition, we have that $\mathbb{P}[\,Y=1, X=x_i, A=a\,] = \vec{P}_i\vec{Q}_i^{(a)}$ and

$$\mathbb{P}[\,\hat{Y}=1\,|\,Y=1, A=a\,] = \frac{\sum_{i=1}^{n}\mathbb{P}[\,\hat{Y}=1, Y=1, X=x_i, A=a\,]}{\sum_{i=1}^{n}\mathbb{P}[\,Y=1, X=x_i, A=a\,]} = \frac{\langle\vec{F},\vec{Q}^{(a)}\rangle}{\langle\vec{P},\vec{Q}^{(a)}\rangle}.$$

Therefore,

$$\begin{aligned}\mathrm{oppDiff}(\hat{Q}) &= \mathbb{P}[\,\hat{Y}=1\,|\,Y=1, A=1\,] - \mathbb{P}[\,\hat{Y}=1\,|\,Y=1, A=0\,]\\ &= \frac{\langle\vec{F},\vec{Q}^{(1)}\rangle}{\langle\vec{P},\vec{Q}^{(1)}\rangle} - \frac{\langle\vec{F},\vec{Q}^{(0)}\rangle}{\langle\vec{P},\vec{Q}^{(0)}\rangle} = \mathrm{oppDiff}(\vec{F}).\end{aligned}$$

□

Lemma 8. *(Metrics symmetry) Using the notation of Definition 1, we have*

$$\begin{aligned}\mathit{err}(\vec{P}-\vec{F}) &= 1 - \mathit{err}(\vec{F}),\\ \mathit{oppDiff}(\vec{P}-\vec{F}) &= -\mathit{oppDiff}(\vec{F}).\end{aligned}$$

Proof. According to Lemma 7, opportunity difference is a linear transformation. Since linear transformations preserve scalar multiplication and vector addition, it follows that $\mathrm{oppDiff}(\vec{P}-\vec{F}) = \mathrm{oppDiff}(\vec{P}) - \mathrm{oppDiff}(\vec{F})$. Moreover, since $\mathrm{oppDiff}(\vec{P}) = 1-1 = 0$, then $\mathrm{oppDiff}(\vec{P}-\vec{F}) = -\mathrm{oppDiff}(\vec{F})$.

According to the same lemma, the error is an affine transformation with offset $\langle\vec{P},\vec{Q}\rangle$. Hence,

$$\begin{aligned}\mathrm{err}(\vec{P}-\vec{F}) &= \mathrm{err}(\vec{P}) - \mathrm{err}(\vec{F}) + \langle\vec{P},\vec{Q}\rangle\\ &= 2\langle\vec{P},\vec{Q}\rangle - \langle\vec{P},1-2\vec{Q}\rangle - \mathrm{err}(\vec{F})\\ &= \langle\vec{P},1\rangle - \mathrm{err}(\vec{F}) = 1 - \mathrm{err}(\vec{F}),\end{aligned}$$

because $\sum_{i=1}^{n} \vec{P}_i = 1$. □

We now make use of results from a different research area in mathematics (geometry) to conclude the main properties of the region M.

Theorem 9. *Assuming a discrete data source with finitely many possible outcomes, the region M of feasible combinations of error versus opportunity difference satisfies the following claims.*

1. *M is a convex polygon.*
2. *The vertices of the polygon M correspond to some deterministic predictors.*
3. *M is symmetric with respect to the point $(1/2, 0)$.*

Proof. The proof is based on the fact that affine transformations map polytopes into polytopes (See Chapter 3 of [Grü13]).

Assume the notation of Definition 1.

Part 1. In geometrical terms, M is the result of applying an affine transformation, i.e., a linear transformation and a translation, to the n-dimensional *polytope* given by $0 \preceq \vec{F} \preceq \vec{P}$.

Affine transformations are known to map polytopes into polytopes (See Chapter 3 of [Grü13]), therefore M must be a 2-dimensional polytope, i.e., the region M is a convex polygon. In theory, this region may also be a 1-dimensional segment, but this can only occur in the extreme case that $Q = 1/2$ (with probability 1).

Part 2. The vertices of a polytope, also called extremal points, are the points in the polytope that are not in the segment between any two other points in the polytope. It is known from geometry theory that affine mappings preserve collinearity, i.e., they map segments into segments, thus they map non-vertices into non-vertices. As a consequence, the vertices of the polygon M correspond to some vertices of the polytope $0 \preceq \vec{F} \preceq \vec{P}$, that is, to some deterministic classifiers.

Part 3. Notice (Lemma 8) that

$$\begin{aligned}\text{err}(\vec{P}-\vec{F}) &= 1-\text{err}(\vec{F}),\\ \text{oppDiff}(\vec{P}-\vec{F}) &= -\text{oppDiff}(\vec{F}).\end{aligned}$$

This implies that for each point $(\text{err}(\vec{F}), \text{oppDiff}(\vec{F})) \in M$, there is another one, namely $(\text{err}(\vec{P}-\vec{F}), \text{oppDiff}(\vec{P}-\vec{F})) \in M$ that is symmetrical to the former w.r.t the point $(1/2, 0)$. Geometrically, this means that the polygon M is symmetric with respect to the point $(1/2, 0)$. □

The reader is invited to visualize the properties of M mentioned in Theorem 9 in Figure 2.1, which depicts the region M for a particular instance [2] of $\vec{P}$ and $\vec{Q}$.

2.4 Strong Impossibility Result

Contrasting with Figure 2.1 in the previous section, Figure 2.2 shows a data source for which the constant classifiers are vertices of the polygon. Figure 2.2 was generated using the theory developed in this section, and it illustrates the strong incompatibility that may occur in certain distributions. Namely, among the predictors satisfying equal opportunity (those in the X-axis), the minimal error is achieved by a constant classifier.

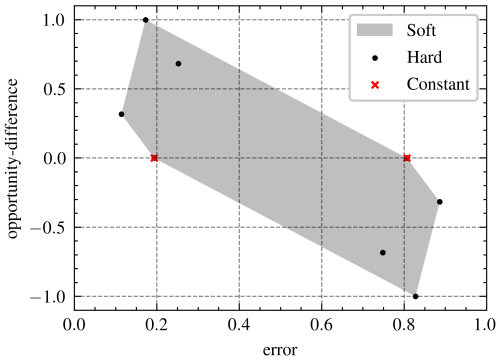


Figure 2.2: In this distribution, the constant classifiers are vertices of the polygon, thus the constraints of equal opportunity and non-trivial accuracy can not be satisfied simultaneously.

In other words, there are data sources for which no predictor can achieve equal opportunity and non-trivial accuracy simultaneously. This is Theorem 11.

[2] Namely, `P=[0.267 0.344 0.141 0.248]`, `Q=[0.893 0.896 0.126 0.207]` and `A=[0 1 0 1]`.

Since Theorem 11 is our strongest result, we also show how to generalize it to non-finite domains. For this purpose, and focusing on formality, we state in Definition 10 very precisely, for which kind of domains it applies.

Definition 10. The *essential range* of a random variable $S : \Omega \to \mathbb{R}^k$ is the set

$$\{\vec{s} \in \mathbb{R}^k : (\forall \epsilon > 0)\ \mathbb{P}[\,\|S - \vec{s}\| < \epsilon\,] > 0\}.$$

We call a set $\mathcal{D} \subseteq \mathbb{R}^k$ an *essential domain* if it is the essential range of any random variable.

Definition 10 excludes pathological domains such as non-measurable sets, the Cantor set, or the irrationals. But it allows for isolated points, convex and closed sets, finite unions of them, and countable unions of them as long as the resulting set is closed. This includes typical domains, such as products of closed intervals $\prod_{i=1}^{n}[l_i, r_i]$, or the whole space $\mathbb{R}^n$.

Theorem 11. *For any essential domain $\mathcal{X} \subseteq \mathbb{R}^d$ with $|\mathcal{X}| \geq 2$, there exists a data source (X, A, Y) whose essential range is $\mathcal{X} \times \{0,1\}^2$ and such that the accuracy $acc(\hat{Q})$ of any predictor $\hat{Q} \in \mathcal{Q}$ that satisfies equal opportunity is at most the trivial accuracy $\tau \in [0, 1)$.*

Proof. The proof is divided into four parts. We will (i) reduce the problem into an algebraic one; (ii) find the linear constraints that solve the algebraic problem when satisfied; (iii) provide an algorithm that generates vectors that satisfy the linear constraints; and finally, (iv) convert the vectorial solution back into a distribution (π, q) for the given domain.

Part 1. Reduction to an algebraic problem.

Partition the non-protected input space $\mathcal{X}$ into two non-empty sets $\mathcal{X}_1, \mathcal{X}_2$, and the input space $\mathcal{X} \times \{0,1\}$ into three regions R_j:

$$R_1 = \mathcal{X}_1 \times \{0\}, \qquad R_2 = \mathcal{X}_2 \times \{0\}, \quad \text{and} \quad R_3 = \mathcal{X} \times \{1\}.$$

For any distribution (π, q) for which these 3 regions have positive probabilities, denote $\vec{P}_j \stackrel{\text{def}}{=} \mathbb{P}[\,(X,A) \in R_j\,] > 0$ and $\vec{Q}_j \stackrel{\text{def}}{=} \mathbb{P}[\,Y = 1 \,|\, (X,A) \in R_j\,]$ for $j \in \{1,2,3\}$. We search for constraints over $\vec{P}$ and $\vec{Q}$ that are feasible and cause $\text{acc}(\hat{Q}) \leq \tau$ for any fair predictor $\hat{Q} \in \mathcal{Q}$ satisfying EO. The first such constraint is

C1. $\vec{P}, \vec{Q} \in (0,1)^3$.

That is, we require $\vec{P}_j$ to be positive, and Y to have at least some degree of randomness in each region.

Given a reference predictor $\hat{Q}$, let $\vec{F} \in [0,1]^3$ be the vector given by $\vec{F}_j \overset{\text{def}}{\coloneqq} \mathbb{P}[\hat{Y}{=}1, (X,A){\in}R_j\,]$. Lemma 7 shows that the accuracy and the opportunity difference of any predictor $\hat{Q}$ can be computed from $\vec{P}$, $\vec{Q}$ and $\vec{F}$ as

$$\begin{aligned} \text{acc}(\hat{Q}) &= \langle \vec{F}, 2\vec{Q}{-}1 \rangle + C_{\vec{Q}}, \\ \text{oppDiff}(\hat{Q}) &= \frac{\vec{F}_3}{\vec{P}_3} - \frac{\vec{F}_1\vec{Q}_1 + \vec{F}_2\vec{Q}_2}{\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2}, \end{aligned}$$

where $C_{\vec{Q}} \overset{\text{def}}{\coloneqq} 1 - \langle \vec{P}, \vec{Q} \rangle$ is a constant and the operator $\langle \cdot, \cdot \rangle$ denotes the inner product explained in Definition 1. Since we are interested in relative accuracies with respect to the trivial predictors, the constant $C_{\vec{Q}}$ is mostly irrelevant. For this reason, we let $L(\vec{F}) \in [-1,1]$ denote the non-constant component of the accuracy $L(\vec{F}) \overset{\text{def}}{\coloneqq} \langle \vec{F}, 2\vec{Q}{-}1 \rangle$.

Both accuracy and opportunity difference are completely determined for any predictor by the vectors $\vec{P}$, $\vec{Q}$ and $\vec{F}$ as shown above. Moreover, both quantities are linear with respect to $\vec{F}$.

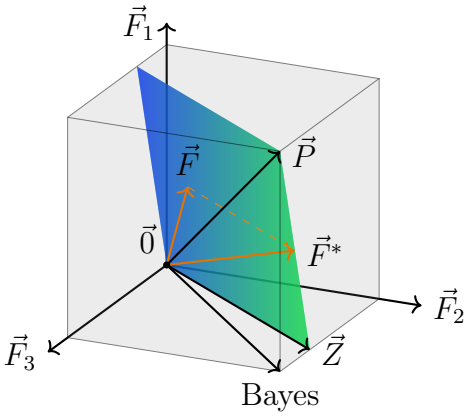


Figure 2.3: In vectorial form, the predictors that satisfy equal opportunity form a plane inside the rectangular box of all predictors.

Regarding equal opportunity, the constraint $\text{oppDiff}(\hat{Q}) = 0$ forms a plane in $\mathbb{R}^3$, depicted in Figure 2.3. This plane passes through the origin, is deter-

mined by $\vec{P}$ and $\vec{Q}$, and contains all vectors $\vec{F}$ (restricted to $0 \leq \vec{F}_j \leq \vec{P}_j$) that satisfy

$$\vec{F}_3(\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2) - \vec{P}_3(\vec{F}_1\vec{Q}_1 + \vec{F}_2\vec{Q}_2) = 0,$$

or equivalently, all vectors F that are normal to the vector $(-\vec{P}_3\vec{Q}_1, -\vec{P}_3\vec{Q}_2, \vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2)$.

Regarding accuracy, the two constant predictors correspond to $\vec{F} = \vec{0}$ and $\vec{F} = \vec{P}$, thus $\tau = C_Q + \max\{L(\vec{0}), L(\vec{P})\}$. Importantly, both of them lie on the equal opportunity plane.

The problem is now reduced to finding vectors $\vec{P}$ and $\vec{Q}$ such that all vectors $\vec{F}$ in the equal opportunity plane satisfy $L(\vec{F}) \leq \max\{L(\vec{0}), L(\vec{P})\}$.

Part 2. Constraints for the algebraic solution.

To fix an orientation, let us impose these constraints:

C2. Among the constant predictors, the accuracy of $\vec{F} = \vec{P}$ is higher than that of $\vec{F} = \vec{0}$. This is $L(\vec{P}) > 0 = L(\vec{0})$.

C3. The Bayes classifier is located at $(0, \vec{P}_2, \vec{P}_3)$ as in Figure 2.3. Algebraically this means $\vec{Q}_1 < 1/2$ and $\vec{Q}_2, \vec{Q}_3 > 1/2$.

In order to derive the constraints that make the scalar field L maximal at $\vec{P}$ over the plane, consider the vector $\vec{Z}$ that lies on the plane and has minimal $\vec{Z}_1$ and maximal $\vec{Z}_2$, i.e.

$$\vec{Z} \stackrel{\text{def}}{=} (0, \vec{P}_2, \vec{P}_3 \frac{\vec{P}_2\vec{Q}_2}{\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2}).$$

Since the gradient of L is given by $2\vec{Q} - 1$ and has signs $(-, +, +)$, then for any vector $\vec{F}$ in the plane, there is $\vec{F}^*$ in the segment between $\vec{P}$ and $\vec{Z}$ such that $\vec{F}_1 = \vec{F}_1^*$ and $L(\vec{F}^*) \geq L(\vec{F})$ (refer to Figure 2.3). This implies that the L attains its maximal value on the segment between $\vec{P}$ and $\vec{Z}$. Hence, for L to be maximal at $\vec{P}$, it would suffice to have $L(\vec{P}) > L(\vec{Z})$. This can be achieved by imposing, in addition,

C4. $\vec{Q}_3 + \vec{Q}_1 \geq 1$, and

C5. $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 < \vec{P}_3\vec{Q}_1$.

because of the following equivalences and implications.

$$
\begin{aligned}
&(\vec{Q}_3+\vec{Q}_1\geq 1)\wedge(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2<\vec{P}_3\vec{Q}_1)\\
\equiv&(1-2\vec{Q}_1\leq 2\vec{Q}_3-1)\wedge(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2<\vec{P}_3\vec{Q}_1)\\
\Rightarrow&(1-2\vec{Q}_1)(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2)<(2\vec{Q}_3-1)\vec{P}_3\vec{Q}_1\\
\equiv&(2\vec{Q}_1-1)(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2)+(2\vec{Q}_3-1)\vec{P}_3\vec{Q}_1>0\\
\equiv&(2\vec{Q}_1-1)\vec{P}_1+(2\vec{Q}_3-1)\vec{P}_3\frac{\vec{P}_1\vec{Q}_1}{\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2}>0\\
\equiv&\langle 2\vec{Q}-1,\vec{P}-\vec{Z}\rangle>0\\
\equiv&\langle 2\vec{Q}-1,\vec{P}\rangle-\langle 2\vec{Q}-1,\vec{Z}\rangle>0
\end{aligned}
$$

Part 3. Solution to the constraints.

Algorithm 1 is a randomized algorithm that generates random vectors. We will prove that the output vectors $\vec{P}$ and $\vec{Q}$ satisfy the constraints of the previous parts of this proof, regardless of the seed and the random sampling function, e.g., uniform. For corroboration and illustration, the distribution in Figure 2.2 presented early was generated using this algorithm[3].

Algorithm 1 Random generator for Theorem 11.

procedure VECTORGENERATOR(seed)
  Initialize random sampler with the seed
  $\vec{Q}_1 \leftarrow$ random in $(0, 1/2)$
  $\vec{Q}_2 \leftarrow$ random in $(1/2, 1)$
  $\vec{Q}_3 \leftarrow$ random in $(1-\vec{Q}_1, 1)$
  $\vec{P}_3 \leftarrow$ random in $(1/2, 1)$
  $a \leftarrow \max\{(1-\vec{P}_3)\vec{Q}_1, 1/2-\vec{P}_3\vec{Q}_3\}$
  $b \leftarrow \min\{(1-\vec{P}_3)\vec{Q}_2, \vec{P}_3\vec{Q}_1\}$
  $c \leftarrow$ random in (a, b)
  $\vec{P}_2 \leftarrow {}^{(c-\vec{Q}_1(1-\vec{P}_3))}/_{\vec{Q}_2-\vec{Q}_1}$
  $\vec{P}_1 \leftarrow 1-\vec{P}_3-\vec{P}_2$
  return $\vec{P}, \vec{Q}$

Two immediate observations about Algorithm 1 are that the construction of $\vec{Q}$ implies that constraints C3 and C4 are satisfied, and the construction of

[3]The algorithm's output was `P=[0.131 0.096 0.772]` and `Q=[0.274 0.858 0.891]`. Also, `A=[0 0 1]` from the partition $\{R_1, R_2, R_3\}$.

$\vec{P}$ implies $\vec{P}_1 + \vec{P}_2 + \vec{P}_3 = 1$. To prove the correctness of the algorithm, it remains to prove that (i) $a < b$ (otherwise the algorithm would not be well-defined), that (ii) $\vec{P}_2 \in (0,1)$ for constraint C1, and also that (iii) constraints C2 and C5 are satisfied. For better readability, the algebraic proof of these claims is moved to Lemma 12.

Part 4. Construction of the distribution.

Generate a pair of vectors $\vec{P}$ and $\vec{Q}$ using the algorithm of the previous part (Part 3). The first goal is to partition $\mathcal{X}$ into $\mathcal{X}_1$ and $\mathcal{X}_2$ to generate the regions R_1, R_2 and R_3. The second goal is to define π in such a way that $\mathbb{P}[\,(X,A) \in R_j\,] = \vec{P}_j$ for each $j \in \{1,2,3\}$. The third and last goal is to define q so that $\mathbb{E}[\,Q \mid (X,A) \in R_j\,] = \vec{Q}_j$ for each j. This can be done immediately by letting $q(x,a) \stackrel{\text{def}}{\coloneqq} \vec{Q}_j$ for all $(x,a) \in R_j$. Thus only the first two goals remain.

For the first goal, since $|\mathcal{X}| \geq 2$, we may create a simple Voronoi clustering diagram by choosing two different arbitrary points $s_1, s_2 \in \mathcal{X}$, and letting $\mathcal{X}_1 \stackrel{\text{def}}{\coloneqq} \{s \in \mathcal{X} : \|s - s_1\| \leq \|s - s_2\|\}$ and $\mathcal{X}_2 \stackrel{\text{def}}{\coloneqq} \mathcal{X} \setminus \mathcal{X}_1$.

For the second goal, since $\mathcal{X}$ is an essential domain, there exists a random variable S whose essential range is $\mathcal{X}$. Notice that for each $j \in \{1,2\}$, it holds that $\mathbb{P}[\,S \in \mathcal{X}_j\,] \geq \mathbb{P}[\,\|S - s_j\| < \|s_1 - s_2\|/2\,] > 0$. For each ($(X,A)$-measurable) event E, let $E_a \stackrel{\text{def}}{\coloneqq} \{x : (x,a) \in E\}$, and define $\pi(E)$ as

$$\begin{aligned}
\mathbb{P}[\,(X,A) \in E\,] &\stackrel{\text{def}}{\coloneqq} \sum_{a=0,1} \mathbb{P}[\,X \in E_a, A = a\,], \\
\mathbb{P}[\,X \in E_0, A = 0\,] &\stackrel{\text{def}}{\coloneqq} \sum_{j=1,2} \mathbb{P}[\,S \in E_0 \,|\, S \in \mathcal{X}_j\,]\vec{P}_j, \\
\mathbb{P}[\,X \in E_1, A = 1\,] &\stackrel{\text{def}}{\coloneqq} \mathbb{P}[\,S \in E_1\,]\vec{P}_3.
\end{aligned}$$

This forces $\mathbb{P}[\,(X,A) \in R_j\,] = \vec{P}_j$ for each $j \in \{1,2,3\}$ as desired. □

Lemma 12. *Algorithm 1 is correct.*

Proof. We will prove $a < b$, $\vec{P}_2 \in (0,1)$ and the fulfillment of constraints C2 and C5.

Part 1. Proof that $a < b$.

Recall $a = \max\{(1-\vec{P}_3)\vec{Q}_1, 1/2 - \vec{P}_3\vec{Q}_3\}$ and $b = \min\{(1-\vec{P}_3)\vec{Q}_2, \vec{P}_3\vec{Q}_1\}$.

1. Since $\vec{Q}_1 < 1/2 < \vec{Q}_2$ and $\vec{P}_3 \in (0,1)$, then $(1-\vec{P}_3)\vec{Q}_1 < (1-\vec{P}_3)\vec{Q}_2$.
2. Since $\vec{P}_3 \in (1/2, 1)$, then $(1-\vec{P}_3)\vec{Q}_1 < \vec{P}_3\vec{Q}_1$.
3. Since $\vec{P}_3 \in (0,1)$ and $\vec{Q}_3 \in (1/2, 1)$, then $\vec{P}_3(\vec{Q}_2 - \vec{Q}_3) < 1 \cdot (\vec{Q}_2 - 1/2)$, or equivalently, $1/2 - \vec{P}_3\vec{Q}_3 < (1-\vec{P}_3)\vec{Q}_2$.
4. Since $\vec{P}_3 > \frac{1}{2(\vec{Q}_1+\vec{Q}_3)}$ then $1/2 - \vec{P}_3\vec{Q}_3 < \vec{P}_3\vec{Q}_1$.

Since the inequalities hold for all available choices for a and b, then, in general, $a < b$ holds.

Part 2. Proof that $\vec{P}_2 \in (0,1)$.

We know $c > \vec{Q}_1(1-\vec{P}_3)$ and $c < \vec{Q}_2(1-\vec{P}_3)$. These inequalities imply that $c - \vec{Q}_1(1-\vec{P}_3) \in (0, \vec{Q}_2 - \vec{Q}_1)$, hence also that $\vec{P}_2 \in (0,1)$.

Part 3. Constraint C2 is satisfied.

Since $\vec{P}_1 + \vec{P}_2 = 1 - \vec{P}_3$ and $\vec{Q}_2 > \vec{Q}_1$, then the term $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2$ is minimal when $\vec{P}_1 = 1-\vec{P}_3$ and $\vec{P}_2 = 0$. Thus,

$$\begin{aligned}
\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 + \vec{P}_3\vec{Q}_3 &\geq (1-\vec{P}_3)\vec{Q}_1 + \vec{P}_3\vec{Q}_3 \\
&= \vec{Q}_1 + \vec{P}_3(\vec{Q}_3 - \vec{Q}_1) \\
&> \vec{Q}_1 + \frac{\vec{Q}_3 - \vec{Q}_1}{2} \\
&= \frac{\vec{Q}_3 + \vec{Q}_1}{2} \geq 1/2.
\end{aligned}$$

Part 4. Constraint C5 is satisfied.

Since $b \leq \vec{P}_3\vec{Q}_1$, then $\vec{P}_2(\vec{Q}_2 - \vec{Q}_1) < \vec{P}_3\vec{Q}_1 - \vec{Q}_1(1-\vec{P}_3)$. From this inequality, we may derive constraint C5 as follows.

$$\begin{aligned}
\vec{P}_2(\vec{Q}_2 - \vec{Q}_1) &< \vec{P}_3\vec{Q}_1 - \vec{Q}_1(1-\vec{P}_3) \\
\vec{P}_2\vec{Q}_2 &< (2\vec{P}_3 - 1 + \vec{P}_2)\vec{Q}_1 \\
\vec{P}_2\vec{Q}_2 &< \vec{P}_3\vec{Q}_1 - \vec{P}_1\vec{Q}_1 \\
\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 &< \vec{P}_3\vec{Q}_1.
\end{aligned}$$

□

Finally, to conclude this section we present Example 1, which shows that there are many other scenarios, not necessarily those of Theorem 11, in which EO and non-trivial accuracy are incompatible.

Example 1. Consider a data source (X, A, Y) over $\{0,1\}^3$ whose distribution is given by

x	a	$\pi(x,a)$	$q(x,a)$
0	0	3/8	9/20
0	1	2/8	15/20
1	0	1/8	15/20
1	1	2/8	16/20

Then, (i) there are predictors satisfying equal opportunity, (ii) there are predictors with non-trivial accuracy, but (iii) there are no predictors satisfying both. *(End)*

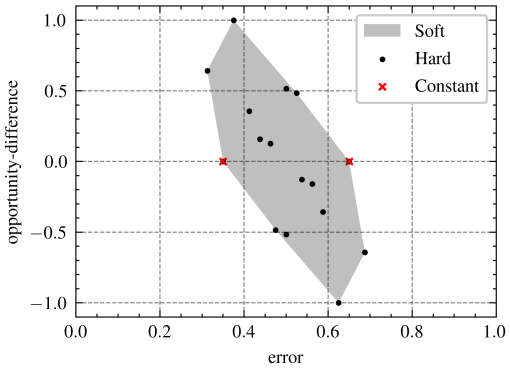


Figure 2.4: Example 1. One of the constant classifiers is Pareto-optimal.

Indeed, Figure 2.4 depicts the region M for Example 1. On the one hand, the set of non-trivially accurate predictors corresponds to the area with an error strictly smaller than the left constant classifier. On the other hand, the set of equal opportunity predictors is (for this particular example) the closed segment between the two constant classifiers. As claimed in Example 1 (and depicted in Figure 2.4), these two sets are non-empty and do not intersect each other.

2.5 Probabilistic versus Deterministic Sources

In this section, we compare the tension between error and opportunity difference when the data source is deterministic and probabilistic. The motivation for studying the probabilistic case is presented in the introduction. Particularly, we show that some known properties that apply for the discrete case may fail to hold for the probabilistic one, and under what conditions this happens.

2.5.1 Deterministic Sources

Under the assumption that the data source is deterministic, there are some important existing results showing the compatibility between equal opportunity and high accuracy:

Fact 13. Assuming a deterministic data source, the Neyman Pearson lemma [Fuk93] implies that if $\tau < 1$, then there is always a non-trivial predictor, for instance, the Bayes classifier $Q_{1/2}$. Otherwise (degenerated case with $\tau = 1$) all predictors are trivially accurate.

Fact 14. Assuming a deterministic data source, the Bayes classifier $Q_{1/2}$ satisfies equal opportunity necessarily [HPS16].

As a consequence, EO and maximal accuracy (thus also non-trivial accuracy) are always compatible provided $\tau < 1$, because the Bayes classifier satisfies both. This is a celebrated fact and it was part of the motivations of [HPS16] for defining equal opportunity, because other notions of fairness, including statistical parity, are incompatible with accuracy.

2.5.2 Probabilistic Sources

If we allow the data source to be probabilistic, the results of the deterministic case change. In particular, Fact 13 is generalized by Proposition 15 and Fact 14 is affected by Proposition 16 and Example 1.

Analogous to τ for deterministic sources, we define a second reference value $\tau^* \in [0, 1]$. We let

$$\tau^* \stackrel{\text{def}}{:=} \max\left\{\mathbb{P}[\,Q \geq 1/2\,],\, \mathbb{P}[\,Q \leq 1/2\,]\right\},$$

highlighting that (i) $Q = q(X, A)$ is a random variable varying in $[0, 1]$, (ii) τ and τ^* are equal when the data source is deterministic, and (iii) the condition $\tau = 1$ implies $\tau^* = 1$, but not necessarily the opposite.

As shown in Proposition 15, the equation $\tau^* = 1$ characterizes the necessary and sufficient conditions on the data source for non-trivially accurate predictors to exist.

Particularly, in the deterministic case, we have $\tau^* = \tau$, and Proposition 15 resembles Fact 13.

Proposition 15. (Characterization of the impossibility of non-trivial accuracy)

For any arbitrary source distribution (π, q), non-trivial predictors exist if and only if $\tau^* < 1$.

Proof. The proof intuition is that if $\mathbb{P}[\,Q \geq 1/2\,] = 1$, then predicting 1 for any input is optimal, and vice versa.

We will prove that all predictors are trivially accurate if and only if $\tau^* = 1$.

($\Leftarrow$) Suppose $\tau^* = 1$, i.e., $\mathbb{P}[\,Q \leq 1/2\,] = 1$ or $\mathbb{P}[\,Q \geq 1/2\,] = 1$.

In the former case, the Bayes classifier $Q_{1/2}$ is the constant predictor $(x, a) \mapsto 0$, thus $\mathrm{acc}(Q_{1/2}) \leq \tau$ necessarily. In the latter case, the alternative Bayes classifier $Q^*_{1/2}$ (defined in Lemma 5) is the constant predictor $(x, a) \mapsto 1$, thus $\mathrm{acc}(Q^*_{1/2}) \leq \tau$. According to Lemma 5, $\mathrm{acc}(Q_{1/2}) = \mathrm{acc}(Q^*_{1/2})$, thus we may conclude $\mathrm{acc}(Q_{1/2}) \leq \tau$ as well.

It follows that $\mathrm{acc}(\hat{Q}) \leq \mathrm{acc}(Q_{1/2}) \leq \tau$ for all $\hat{Q} \in \mathcal{Q}$ because $Q_{1/2}$ has maximal accuracy in $\mathcal{Q}$.

($\Rightarrow$) Suppose that all classifiers, including the Bayes classifier $Q_{1/2}$, are trivially accurate, i.e., $\mathrm{acc}(Q_{1/2}) = \tau$.

According to Lemmas 3 and 6 we may rewrite the known equality $\mathrm{acc}(Q_{1/2}) - \tau = 0$ as $\mathbb{E}[\,|Q - 1/2| - |\mathbb{E}[\,Y\,] - 1/2|\,] = 0$. Using the reverse triangle inequality, we conclude $\mathbb{E}[\,|Q - \mathbb{E}[\,Y\,]|\,] = 0$, thus $Q = \mathbb{E}[\,Y\,]$ is constant.

If $\mathbb{E}[\,Y\,] \leq 1/2$, then $\mathbb{P}[\,Q \leq 1/2\,] = 1$. If $\mathbb{E}[\,Y\,] \geq 1/2$, then $\mathbb{P}[\,Q \geq 1/2\,] = 1$. In any case, we have $\tau^* = 1$.

□

Finally, in Proposition 16 and its proof, we show a simple family of probabilistic examples for which equal opportunity and optimal accuracy (obtained by the Bayes classifier) are not compatible. This issue does not merely arise from the fact that the Bayes classifier is hard while the data distribution is soft. Adding randomness to the classifier does not solve the issue. To justify this, and also for completeness, we considered the soft predictor Q and showed that it also fails to satisfy equal opportunity.

Proposition 16. There are data sources for which neither the Bayes classifier $Q_{1/2}$ nor the predictor Q satisfies equal opportunity.

Proof. Fix any data source with $\mathbb{P}[\,A{=}a, Y{=}1\,] > 0$ for each $a \in \{0,1\}$, pick an arbitrary ((X,A)-measurable) function $c : \mathbb{R}^d \to (0, 1/2)$ and let

$$q\,(x,a) \stackrel{\text{def}}{\coloneqq} \begin{cases} 1/2 - c(x) & \text{if } a = 0 \\ 1/2 + c(x) & \text{if } a = 1 \end{cases}$$

for each $(x,a) \in \mathbb{R}^d \times \{0,1\}$.

Since we know that $Q_{1/2}(x,a) = a$, then the term $\mathbb{E}[\,Q_{1/2}(X,A) \mid A = a, Y = 1\,]$ can be reduced more simply into $\mathbb{E}[\,A \mid A = a, Y = 1\,] = a$. Therefore, the Bayes classifier satisfies $\text{oppDiff}(Q_{1/2}) = 1 - 0 > 0$.

Regarding Q, we have $\mathbb{E}[\,Q \mid A = 1, Y = 1\,] = 1/2 + \mathbb{E}[\,c(X) \mid A = 1, Y = 1\,]$ and $\mathbb{E}[\,Q \mid A = 0, Y = 1\,] = 1/2 - \mathbb{E}[\,c(X) \mid A = 0, Y = 1\,]$. Notice from the range of c, that $\mathbb{E}[\,Q \mid A = 1, Y = 1\,] \in (1/2, 1)$ and $\mathbb{E}[\,Q \mid A = 0, Y = 1\,] \in (0, 1/2)$. Hence $\text{oppDiff}(Q) > 0$.

Therefore, neither $Q_{1/2}$ nor Q satisfy equal opportunity.

□

As a remark, notice that the data sources proposed in the proof of Proposition 16, contrast the extreme case $Y = A$ because they allow some mutual information between X and Y after A is known, as one would expect in a real-life distribution. Nevertheless, there is an evident inherent demographic

disparity in these distributions, and this can be the reason why equal opportunity hinders optimal accuracy for these examples.

2.6 Algorithms for the Pareto Frontier

In this section, we provide an algorithm for computing and depicting the Pareto frontier that optimizes the trade-off between error and the absolute value of opportunity difference (0 being EO). We consider (and aim at minimizing) the absolute value because we regard the difference in opportunity as bias, independently of the sign.

Three algorithms are explained and compared: the brute force, the one we propose, and the double-threshold method based on [HPS16]. The methods are restricted to finite alphabets for the non-protected attributes, i.e., $\mathcal{X} = \{x_1, ..., x_n\}$, so the inputs (x, a) can only take a total of $|\mathcal{X} \times \{0, 1\}| = 2n$ values. For the convenience of the reader, we summarized them in Table 2.2.

Methodology	Complexity	Principle for finding the convex hull
Brute-force	$O(n\ 2^{2n})$	All corners correspond to deterministic classifiers.
Proposed	$O(n \log n)$	Algorithm 2. The n partial derivatives of error and opportunity difference are constant.
Double-threshold	$O(n^3 \log n)$	Algorithm 3. All corners correspond to single-threshold classifiers in V.

Table 2.2: Comparison of methods for finding the Pareto frontier and the feasibility region.

2.6.1 Brute-Force Algorithm

We begin by describing the brute-force algorithm for reference. The brute-force algorithm will compute not only the points that determine the Pareto frontier but all the vertices of the feasibility region M.

Recall that the set of all predictors forms a $2n$-dimensional polytope that is

mapped into the region M when error and opportunity difference are measured. We know that each vertex of the region M corresponds to a deterministic classifier, or equivalently, to one of the 2^{2n} vertices of the polytope.

Therefore, it suffices to compute the error and opportunity difference for the 2^{2n} vertices of the polytope (first part), and then compute their convex hull (second part).

Assuming that each classifier is represented with an array of length $2n$, then the runtime complexity for computing the first part is $O(2n\ 2^{2n})$. For the second part, we may use Graham's scan algorithm [Gra72] to find the vertices of the convex hull. Since there are 2^{2n} points and Graham's scan has complexity $O(N \log N)$ where N is the number points, then the complexity is $O(2^{2n} \log 2^{2n}) = O(2n\ 2^{2n})$. Hence the complexity for the whole algorithm (adding up the first and second parts) is $O(4n\ 2^{2n}) = O(n\ 2^{2n})$.

2.6.2 Proposed Method

The proposed method (Algorithm 2) also computes all the vertices of the feasibility region M, but unlike the brute-force algorithm, it exploits greedily a property that appears to be local (depending on a chosen predictor), but in reality, is global (same for all predictors) in M.

For each predictor in the $2n$-dimensional polytope, let us consider its *taxicab neighbors*, i.e., the set of points that differ with it in at most one coordinate. Since the measurement function from the polytope into M is linear, these neighbors form a *star* in M around the given predictor (Figure 2.5).

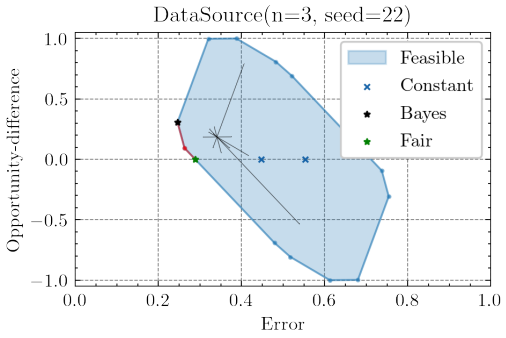


Figure 2.5: Feasibility region for a particular data source showing the Pareto frontier in red and the *taxicab neighbors*' star around an arbitrary predictor.

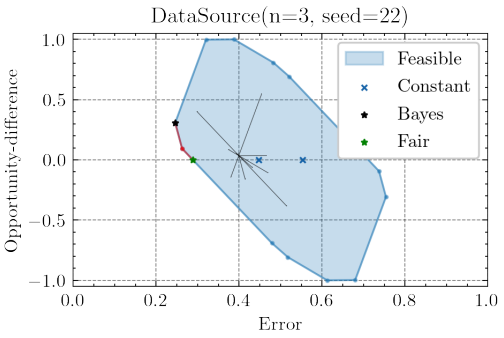


Figure 2.6: Same scenario as in Figure 2.5, but showing a star around a different arbitrary central predictor. The segments of the two stars differ exclusively in offset, not in slope or length. We exploit this fact in Algorithm 2.

Algorithm 2 Fast computation of the feasibility region vertices.

Letting $\alpha_a \overset{\text{def}}{:=} \mathbb{P}[Y=1, A=a] > 0$ and $n \overset{\text{def}}{:=} |\mathcal{X}|$,
procedure CONVEX HULL(α_0, α_1, Q)
    $R \leftarrow [\,]$ ▷ Empty list of rays
    for each (x, a) **do** ▷ $2n$ in total
        sign $\leftarrow -1 + 2 \cdot \mathbf{1}\{a = 1\}$
        $y \leftarrow 1$; $\theta \leftarrow \arctan2(\,1-2q(x,a),\ \text{sign}\cdot q(x,a)/\alpha_a\,)$
        push tuple (θ, x, a, y) into R
        $y \leftarrow 0$; $\theta \leftarrow (\,\theta + \pi \mod (-\pi, \pi]\,)$
        push tuple (θ, x, a, y) into R
    sort(R) ▷ by angle in $(-\pi, \pi]$
    $V \leftarrow []$ (empty list of classifiers)
    $\hat{Q} \leftarrow$ Bayes classifier $Q_{1/2}$
    for each ray (θ, x, a, y) in R **do** ▷ $4n$ in total
        update $\hat{q}(x, a) \leftarrow y$
        push a copy of $\hat{Q}$ into V
    return V ▷ classifiers that are vertices of M

The star consists of at most $4n$ rays ($2n$ segments crossing the middle) that represent the $2n$ degrees of freedom in the polytope. It reveals the possible combinations of error and opportunity difference that we can obtain from

a given predictor by modifying a single component, i.e., the decision for a particular (x, a). In particular, when the central predictor is a vertex of the region M, then two of the rays of the star will land on the two neighboring vertices of the polygon.

The crucial fact exploited by Algorithm 2 is that the inclination and length of the segments of the star are the same regardless of the chosen central predictor. The only variation is the offset (compare Figures 2.5 and 2.6). As a consequence, the $2n$ segments that form the star can be visited in convenient order such that, starting from a vertex of the polygon M, all the visited predictors are vertices (or lie collinearly between two consecutive vertices) of the polygon.

More precisely, Algorithm 2 sorts the rays by angle, starts at the Bayes classifier, and then visits each ray, updating the current classifier according to the ray direction in the polytope. Each angle is computed in Line 6 using the gradients of error and opportunity difference as the x and y arguments respectively (derived from their definitions and Lemma 1). Both gradients were divided by a factor of $\pi(x, a)$ because the arctan2 function is indifferent to linear scales, and this allows the whole Algorithm to become independent of the distribution $\pi(\cdot, \cdot)$, except only for two population values, α_0 and α_1, defined as $\alpha_a \stackrel{\text{def}}{=} \mathbb{P}[\, Y{=}1, A{=}a \,]$.

The runtime complexity of Algorithm 2 is $O(n \log n)$ because of the sort instruction. All other instructions can be computed in linear time. Compared with the complexity of the brute-force algorithm, the proposed method enables the computation and visualization of the feasibility region M or the Pareto boundary for data sources with large (but finite) n. Indeed, Figure 2.7 shows an example with $n = 1000$. Since the method computes all the vertices exactly, the visualization may be zoomed in at any level of detail.

2.6.3 Double-Threshold Method

The following fact was shown by [HPS16]. It allows parametrizing all the Pareto classifiers in a simple manner.

Fact 17. (Six parameters predictors) Any Pareto-optimal predictor $\hat{Q}$ can be written in terms of six parameters $l_0, l_1, r_0, r_1, p_0, p_1 \in [0, 1]$ ($l_a < r_a$, standing

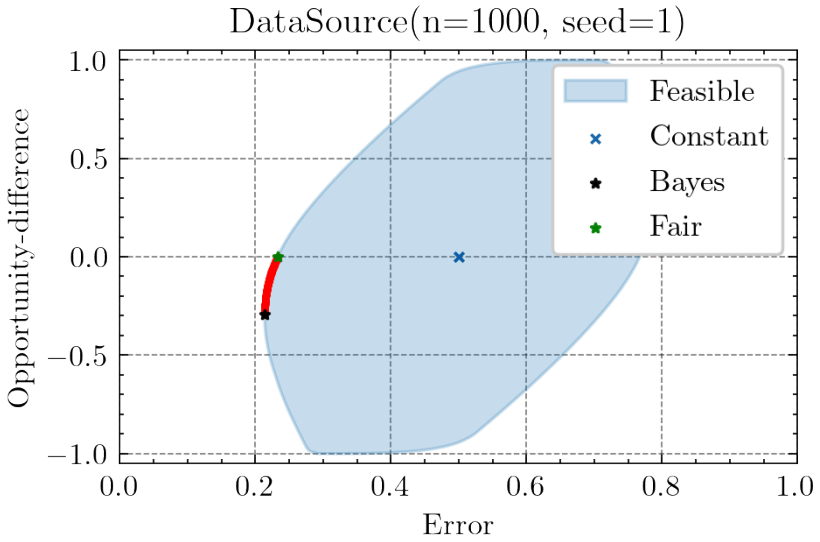


Figure 2.7: Pareto boundary (red) for a more elaborated example with $n = 1000$ in which the $O(n\ 2^{2n})$ brute-force algorithm is inconceivable. The feasibility region is guaranteed to be convex, and although its perimeter looks like a curve, it is a high-resolution piecewise linear path. Also, unlike Figures 2.5 and 2.6, the favored class is $a = 0$ and because of this, the Bayes classifier and the Pareto curve lie in the bottom half.

for left and right thresholds) as

$$\hat{q}(x,a) \overset{\text{def}}{\coloneqq} \begin{cases} 0 & \text{if } q(x,a) \in [0, l_a) \\ p_a & \text{if } q(x,a) \in [l_a, r_a] \\ 1 & \text{if } q(x,a) \in (r_a, 1]. \end{cases}$$

This holds both discrete (as we assume) and non-discrete X.

Following Fact 17, a straightforward algorithm to approximate the Pareto-boundary consists of iterating over a large number of combinations of parameters, e.g., over a six-dimensional grid. This will produce a list of predictors of which we can filter only those that are Pareto optimal (optimal with respect to all other predictors in the list). The filtered predictors will form an approximation of the Pareto boundary.

As shown in Fact 18, if we concentrate on finding only the vertices of the Pareto-boundary and not all the points between them, the search space for

the parameters can be reduced dramatically.

Fact 18. (Double threshold classifiers) For any vertex of the piecewise linear Pareto-boundary, there is a corresponding predictor $\hat{Q}$ (with that error and opportunity difference combination) that can be written in terms of two parameters $t_0, t_1 \in [0,1]$ as either $\hat{q}(x,a) \overset{\text{def}}{\coloneqq} \mathbf{1}\{q(x,a) > t_a\}$, or $\hat{q}(x,a) \overset{\text{def}}{\coloneqq} \mathbf{1}\{q(x,a) \geq t_a\}$, or a combination of the two, e.g.

$$\hat{q}(x,a) \overset{\text{def}}{\coloneqq} \begin{cases} \mathbf{1}\{Q(x,0) > t_0\} & \text{if } a = 0 \\ \mathbf{1}\{Q(x,1) \geq t_1\} & \text{if } a = 1. \end{cases}$$

Proof. Let $l_0, l_1, r_0, r_1, p_0, p_1$ be the six parameters that define $\hat{Q}$ according to Fact 17. Since $\hat{Q}$ is a vertex on the Pareto-boundary it is also a vertex of the region M, and we know from Theorem 9 that the vertices of M correspond to deterministic predictors. Therefore, $\hat{Q}$ can only take values $0, 1$, which implies $p_0, p_1 \in \{0,1\}$. This restriction makes one of the two thresholds l_a or r_a irrelevant for each $a \in \{0,1\}$ and the predictor can be rewritten for each $a \in \{0,1\}$ as either $\hat{q}(x,a) = \mathbf{1}\{q(x,a) \geq l_a\}$ or $\hat{q}(x,a) = \mathbf{1}\{q(x,a) > r_a\}$. □

For our particular case of interest in which the variable for non-protected attributes X is discrete, $q(x,a)$ can only take a finite number of values $r_1, ..., r_m \in [0,1]$ with $r_i < r_{i+1}$. This makes the classifiers $\mathbf{1}\{q(x,a) > r_i\}$ and $\mathbf{1}\{q(x,a) \geq r_{i+1}\}$ equivalent. Therefore, we may unify all the possible cases of Fact 18 without loss of generality using only strict inequalities:

$$\hat{q}(x,a) \overset{\text{def}}{\coloneqq} \mathbf{1}\{q(x,a) > t_a\},$$

for two thresholds $t_0, t_1 \in \{q(x,a) \mid x \in \mathcal{X},\ a \in \{0,1\}\} \cup \{-1\}$. The special value -1 is added to contain the particular case $\mathbf{1}\{q(x,a) \geq 0\}$ for which no strict threshold rule would exist. This is implemented in the 'candidates' procedure in Algorithm 3.

Algorithm 3 Computation of the feasibility region vertices

Letting $\alpha_a \stackrel{\text{def}}{:=} \mathbb{P}[Y=1, A=a] > 0$ and $n \stackrel{\text{def}}{:=} |\mathcal{X}|$,
procedure PARETO VERTICES(α_0, α_1, Q, P)
    $V \leftarrow$ candidates(Q)
    $W \leftarrow [\ (\text{err}(\hat{Q}), \text{oppDiff}(\hat{Q})) \mid \hat{Q} \in V\ ]$ ▷ needs Q, P
    $I \leftarrow$ indices of convex hull of W, sorted clockwise
    $i \leftarrow$ index in I with minimal x-coordinate ▷ V_i is Bayes
    $j \leftarrow$ first (or last) index in I with opposite y-sign to i
        ▷ (first or last depends on the y-sign of W_i)
    $I_{\text{Pareto}} \leftarrow$ indices in I between i and j
    return $[\ V_i \mid i \in I_{\text{Pareto}}\ ]$ ▷ Pareto vertices
procedure CANDIDATES(Q)
    $T_0 \leftarrow \{Q(x, 0) \mid \text{for each } x\} \cup \{-1\}$ ▷ $|T_0| \leq n+1$
    $T_1 \leftarrow \{Q(x, 1) \mid \text{for each } x\} \cup \{-1\}$ ▷ $|T_1| \leq n+1$
    $V \leftarrow [\,]$ ▷ Empty list of threshold classifiers
    for each $t_0 \in T_0$ **do**
        for each $t_1 \in T_1$ **do**
            push $\mathbf{1}\{q(x,a) > t_a\}$ into V
    return V ▷ $|V| \leq (n+1)^2$

Algorithm 3, i.e., the 'Pareto vertices' procedure, computes the error and opportunity difference for each threshold classifiers of interest (each classifier in V) and computes the convex hull to then filter the Pareto boundary. Since $|V| \leq (n+1)^2$, Algorithm 3 is polynomial. The exact complexity depends on the implementation of the computation $(\text{err}(\hat{Q}), \text{oppDiff}(\hat{Q}))$ for a fixed $\hat{Q} \in V$. Normally, this would take $O(n)$ by literally implementing their definition formulas for $|\mathcal{X}| = n$, hence the complexity of Algorithm 3 is $O(n^3 \log n)$.

2.7 Necessary and Sufficient Conditions

In this section, we provide a necessary and sufficient condition (Theorem 19), as well as a simple sufficient (but not necessary) condition (Corollary 20) that guarantees that equal opportunity and non-triviality are compatible. Finally, we discuss when and how a dataset may present this pathological incompatibility.

Theorem 19 (Necessary and sufficient condition for compatibility)**.** *Let*

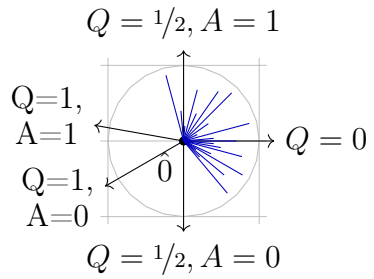
(X, A, Y) be an arbitrary data source. Let $Q_a \overset{\text{def}}{:=} \mathbb{E}[\,Q \mid A = a\,] = \mathbb{E}[\,Y \mid A = a\,]$ *be the output average for each group. Let also*

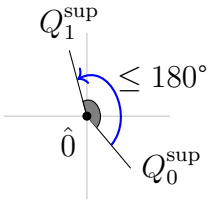
$$Q_a^{\sup} \overset{\text{def}}{:=} \sup\{q \in [0,1] \mid \exists S\ \mathbb{E}[\,Q \mid X \in S \wedge A = a\,] \geq q\}, \text{ and}$$
$$Q_a^{\inf} \overset{\text{def}}{:=} \inf\{q \in [0,1] \mid \exists S\ \mathbb{E}[\,Q \mid X \in S \wedge A = a\,] \leq q\}.$$

Then, equal opportunity and non-triviality are compatible if and only if

$$\begin{aligned}
0 \leq Q_1 Q_0^{\sup}(1 - 2Q_1^{\sup}) &\leq Q_0 Q_1^{\sup}(2Q_0^{\sup} - 1) \quad , \text{ or} \\
0 \leq Q_0 Q_1^{\sup}(1 - 2Q_0^{\sup}) &\leq Q_1 Q_0^{\sup}(2Q_1^{\sup} - 1) \quad , \text{ or} \\
0 \leq Q_1 Q_0^{\inf}(2Q_1^{\inf} - 1) &\leq Q_0 Q_1^{\inf}(1 - 2Q_0^{\inf}) \quad , \text{ or} \\
0 \leq Q_0 Q_1^{\inf}(2Q_0^{\inf} - 1) &\leq Q_1 Q_0^{\inf}(1 - 2Q_1^{\inf}).
\end{aligned}$$

Proof. Recall the star of rays around each classifier explained in Section 2.6.2, and consider the rays around the constant classifier $\hat{0}$ in the plane of error vs. opportunity difference. For each (x, a) in the domain, consider the predictor that maps everything to zero except (x, a) to one. The change in opportunity difference with respect to $\hat{0}$ is $\Delta y = \pi(x,a)\frac{q(x,a)}{Q_a}(2a - 1)$, and the change in error is $\Delta x = \pi(x,a)(1 - 2q(x,a))$. Hence, the angle of this ray is given by $\arctan_2(\Delta y, \Delta x) = \arctan_2(q(x,a)(2a-1),\, Q_a(1 - 2q(x,a)))$. In order to have an impossibility between EO and non-trivial accuracy, the constant classifier $\hat{0}$ must have either minimal error among the classifiers satisfying EO, or maximal error, in which case $\hat{1}$ is minimal. Geometrically, this means that $\hat{0}$ must be part of the convex hull, which holds if and only if all the angles of the rays departing from $\hat{0}$ lie in an interval of length at most $\pi = 180°$.





All the rays for $a = 1$ satisfy $\Delta y \geq 0$ and their angles lie between those of $Q_1^{\inf}$ counter-clockwise to $Q_1^{\sup}$. Similarly, all the rays for $a = 0$ satisfy $\Delta y \leq 0$ and their angles lie between those of $Q_0^{\inf}$ clockwise to $Q_0^{\sup}$. Therefore, checking that all rays lie in an interval of at most π is equivalent to checking that

the counter-clockwise angle from $Q_0^{\sup}$ to $Q_1^{\sup}$ is at most π, or the clockwise angle from $Q_0^{\inf}$ to $Q_1^{\inf}$ is at most π. By replacing the values of Δy and Δx, and considering separately the cases $Q_0^{\sup} \leq 1/2$, $Q_1^{\sup} \leq 1/2$, $Q_0^{\inf} \geq 1/2$, and $Q_1^{\inf} \geq 1/2$, the four inequalities of the theorem statement are obtained. □

From Theorem 19 we can derive a simpler condition for EO and non-trivial accuracy to be compatible. It is only sufficient (i.e., not necessary), but it is easier to check and can be used to verify that a data source (X, A, Y) of a particular application is not pathological for equal opportunity. It is valid for discrete, continuous, and mixed data sources. Therefore, it may be used as a minimal assumption for any research work on equal opportunity dealing with probabilistic data sources.

Figure 2.8 summarizes the sufficiency condition in simple manner. The proof consists of showing that when the 4 events highlighted in Figure 2.8 have positive probabilities, then it is possible to use one of them to improve the performance of the best constant classifier and another one to compensate for equal opportunity.

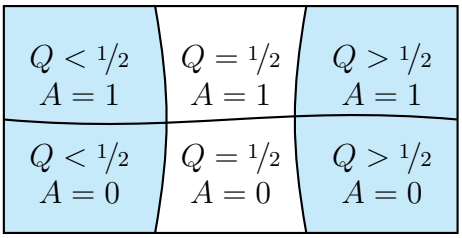


Figure 2.8: Sufficiency condition: If the 4 blue events have positive probability, then equal opportunity and non-triviality are compatible.

Corollary 20 (Sufficient condition)**.** *For any given data source (X, A, Y), not-necessarily discrete, if for each $a \in \{0, 1\}$,*

$$\mathbb{P}[\, Q > 1/2, A = a \,], \mathbb{P}[\, Q < 1/2, A = a \,] > 0,$$

then equal opportunity and non-triviality are compatible. See Figure 2.8

Proof. According to Theorem 19, equal opportunity and non-triviality are compatible if and only if none of its four inequalities hold. If $\mathbb{P}[\, Q > 1/2, A = a \,] > 0$ and $\mathbb{P}[\, Q < 1/2, A = a \,] > 0$ for each $a \in \{0, 1\}$, then

$Q_1^{\sup} > 1/2$, $Q_0^{\sup} > 1/2$, $Q_1^{\inf} < 1/2$ and $Q_1^{\inf} < 1/2$, which respectively violate the four inequalities of Theorem 19. Therefore, compatibility is guaranteed. □

An alternative proof of Corollary 20 that does not use Theorem 19 can be found in the proceedings of AAAI 2022 [PPPV22].

Corollary 20 reveals an important property of the pathological distributions in which EO and non-triviality are incompatible, namely, that they must be already very biased in favor of either $A = 0$ or $A = 1$, and they are highly probabilistic, meaning that the decision Y depends largely on external information, e.g., noise. For instance, if $\mathbb{P}[\,Q{>}1/2, A{=}0\,] = 0$, then for all individuals in the class $A = 0$ the decision that minimizes error is $\hat{Y} = 0$, regardless of their value of X; and the only explanation for individuals with $A = 0$ and $Y = 1$ is external information not contained in X.

2.8 An example based on a real-life dataset

In this section, we show how the incompatibility may occur in practice with a variant of a real-life dataset. A consequence of Corollary 20 is that real world datasets should not incur an incompatibility between EO and non-trivial accuracy if sufficient information about the output is captured in the input features. However, the pathology may still arise when this property is violated. To illustrate this phenomenon, we consider a variant of the Adult dataset [DG17a], where we eliminate some features (thus making it more probabilistic) and artificially reduce the rate of acceptance of the whole population to put the disadvantaged class in a more critical position.

Figure 2.9 shows the Adult dataset after applying the following process: (1) restricting the dataset to the 6 most relevant columns, (2) binarizing the columns using the mean as a threshold, and (3) randomly decreasing the probability of acceptance by 30% for both genders. The purpose of these operations was to illustrate the incompatibility, nevertheless, they are not so arbitrary. Indeed, the first two operations correspond to a simplification of the data, e.g., to perform a simple manual analysis, and the third was applied without direct use of the sensitive attribute (sex), meaning that no additional

gender-specific bias was needed to derive the pathology. In other words, had the acceptance rate been lower for both classes, a simplification of the dataset into 6 binary columns would have sufficed to trigger the incompatibility.

More precisely, Figure 2.9 shows the feasibility region in the plane of error vs opportunity difference (the geometric perspective introduced in this chapter) as well as the associated ROC curve for a classifier (the geometric perspective used in [HPS16]). The left plot shows the constant classifier at the extreme left, on the convex hull of the feasibility region. The plot at the right is the ROC of a standard scikit-learn[PVG+11] random forest classifier of 100 decision trees, using a train-test split of 70%-30%. The parallel lines correspond to constant levels of accuracy and based on the slope and the direction of the gradient, it corroborates that accuracy is maximal at the left-bottom extreme point, which corresponds to $\hat{0}$ with 0 false positives and 0 true positives. The code for processing the dataset and generating the plots is available at [Pin22].

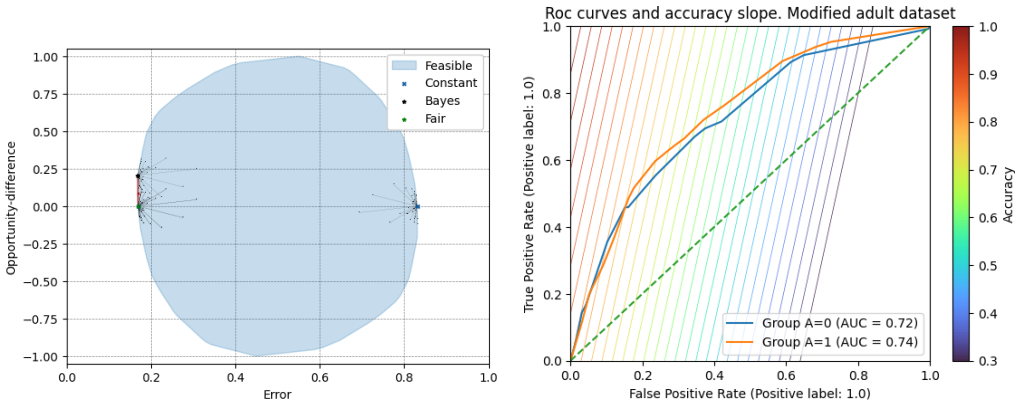


Figure 2.9: Adult dataset after simplification and reduction in acceptance rate. EO and non-trivial accuracy become incompatible.

2.9 Distortion effect of empirical distributions

In many situations, we do not have at our disposal a perfect description of the true data distribution, but only a dataset sampled from the distribution. This is the case, for instance, in machine learning, where the training and the testing are done on the basis of sets of samples. In this section, we discuss how using an empirical distribution from samples may distort the estimation

and the evaluation of opportunity difference and accuracy. This distortion with respect to the true values is a consequence of the fact that an empirical distribution is only an approximation of the true one.

Figure 2.10 shows this mismatch from two points of view on an artificial dataset with $N = 100$ categories for X and $n = 1000$ samples. The dataset was generated by taking samples from a distribution consisting of a fixed categorical distribution for the $2N$ joint categories of X, A, and a binomial distribution for $Y|X, A$ whose parameter depends on the conditioning pair X, A.

Figure 2.10 shows that when the empirical distribution of the dataset is used instead of the true distribution of the data source, the resulting (empirical) Pareto-optimal boundary obtained may mismatch the actual Pareto-optimal boundary, meaning that some classifiers that are empirically deemed as optimal are not optimal, and vice versa.

More precisely, in the left plot of Figure 2.10 the axes represent the measurements of the true error and opportunity difference, and the blue region shows the true convex hull. The orange line represents the empirical Pareto-optimal boundary, computed by applying the algorithms of Section 2.6 on the empirical distribution. As we can see, this boundary does not delimit a convex hull anymore, and it is at some distance from the true Pareto-optimal boundary. In particular, the empirical Fair (max accuracy subject to EO) and empirical Bayes predictors are not at the boundary of the true feasibility region, thus they are suboptimal. Interestingly, the empirical Bayes classifier has less accuracy than the empirical Fair.

Conversely, the right plot of Figure 2.10 depicts the empirical apparent truth that a practitioner would observe in practice. Here, the axes are empirical (apparent) measurements of error and opportunity difference, and the orange area represents the empirical feasible region. The blue line represents the empirical evaluation of the true Pareto-optimal boundary. As we can see, in the empirical view the actual Bayes classifier and the fairest predictor appear to be suboptimal.

Note that the classifiers that form the vertices of the orange convex hull in the bottom plot are exactly the orange points in the top plot and, vice versa,

the classifiers that form the vertices of the blue convex hull in the top plot are exactly the blue points in the top plot.

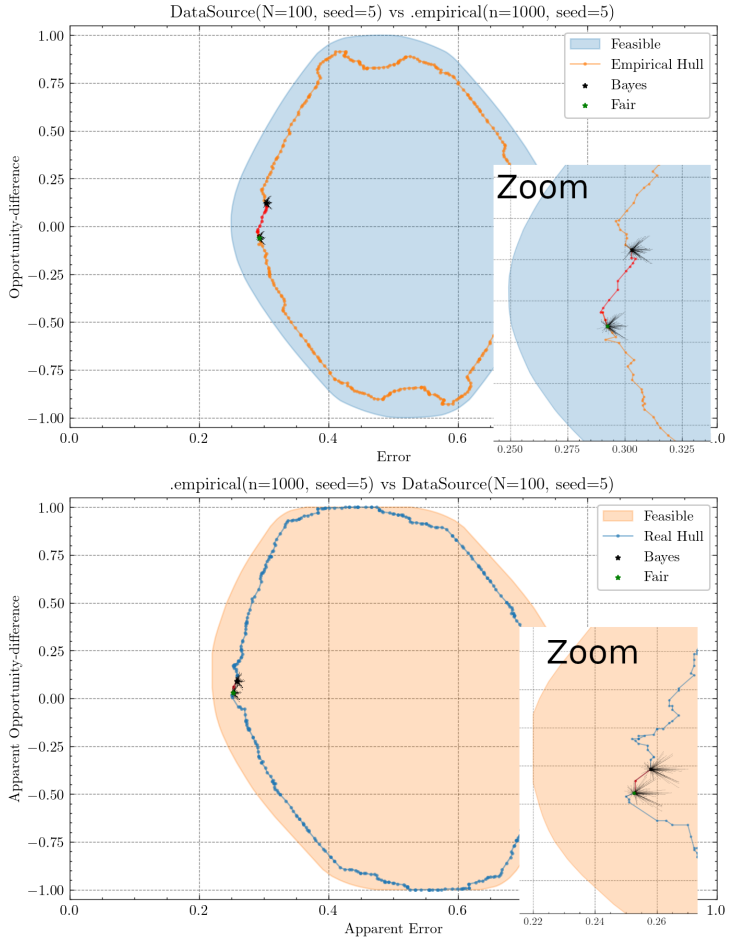


Figure 2.10: True distribution vs empirical dataset. Top: the empirical Pareto boundary mismatches the actual optimal boundary of error and opportunity difference. Bottom: using the dataset for measuring (apparent) error and opportunity difference makes the optimal predictors in the Pareto boundary to appear suboptimal.

The unavailability of the true distribution, which causes a mismatch between the estimated error and opportunity difference and their true values, can have other unexpected consequences. For instance, the left plot in Figure 2.11 shows an example in which the best empirical fair classifier has less error and more opportunity difference than the empirical Bayes classifier. That

is, for that particular data source and sampled dataset, training a model towards maximal accuracy results in more fairness than training taking fairness into account; conversely, training under the fairness constraint results in higher accuracy than training in an unconstrained manner towards maximal accuracy.

This distorting effect has a random nature from the sampling process and is reduced as the number of samples increases, making the empirical measurements closer to their real counterparts. The right plot in Figure 2.11 shows the result of computing the Pareto-optimal boundary on 100 different datasets sampled independently from the same data source distribution of $N = 100$ categories for X and $n = 2500$ samples. The plot shows that, on average, the positions of the empirical Bayes classifier and the empirical Fair classifier match the expected idea of the former having less error and more opportunity difference and vice versa. The empirical Fair classifier has indeed on average an opportunity difference close to zero, suggesting that even though there is no formal guarantee of achieving (true) equal opportunity using the Algorithms in this chapter on (empirical) datasets, one does expect that, with an adequate number of samples, the empirical optimal classifiers will be close to the true optimal ones.

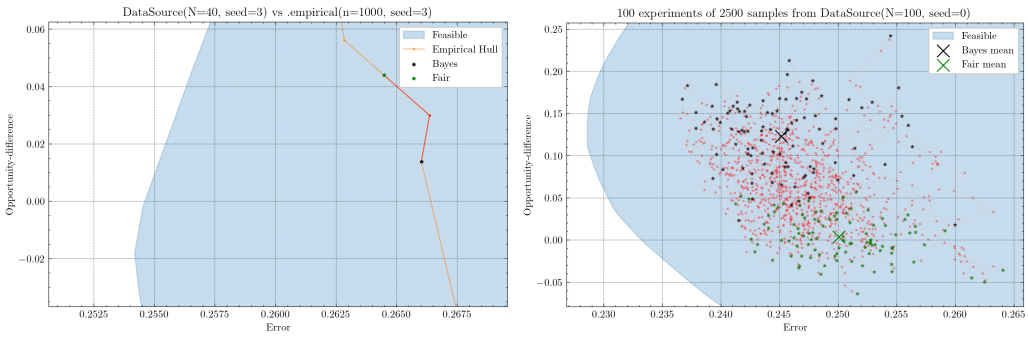


Figure 2.11: Left: a scenario in which the best empirical fair classifier has less error and more opportunity difference than the empirical Bayes classifier. Right: empirical Pareto boundaries for 100 randomly sampled datasets.

2.10 Conclusion

In this chapter, we extended existing results about equal opportunity [HPS16] and accuracy from a deterministic data source to a probabilistic one. The main result, Theorem 11, states that for certain probabilistic data sources, no predictor can achieve equal opportunity and non-trivial accuracy simultaneously. We also characterized in Theorem 19 the conditions on the data source under which EO and non-trivial accuracy are compatible and provided a simple sufficient condition that ensures compatibility (Corollary 20).

The methods used in this chapter rely mostly on geometric properties of the feasibility region in the plane of error vs opportunity difference, thus they are tuned for the fairness notion of equal opportunity, which seeks equal true positive rates TPR. A symmetric analysis can be carried out for equal false positive rates using the same ideas. Since the notion of equal odds seeks both equal true positive rates and equal false positive rates, our methodology and results can be extended to equal odds. In particular, the impossibility theorem holds also for equal odds. However, the geometric methodology that we used was tuned for opportunity difference, they are therefore not directly useful for analyzing statistical parity or individual fairness notions.

The next chapter is related with this one in that it also extends the results and observations in [HPS16], in which apart from establishing the definition of equal opportunity, the authors pose a theoretical argument against the use of graphical/causal models for fairness. These models are nevertheless used, so we investigate this and related issues about the use of causal models for fairness (Problem 2) from a practical point of view in the next chapter.

Chapter 3

Causal discovery for fairness

Causality plays a very important role in the evaluation of fairness criteria because it is not the same to be discriminated while being part of a minority group than to be discriminated *because* of it. For this reason, and despite the abundance of non-causal notions of fairness (statistical parity [Dar71], equal opportunity [HPS16], calibration [Cho17], etc.) [MZP21a], many recent fairness criteria take causality into account [MZP20b].

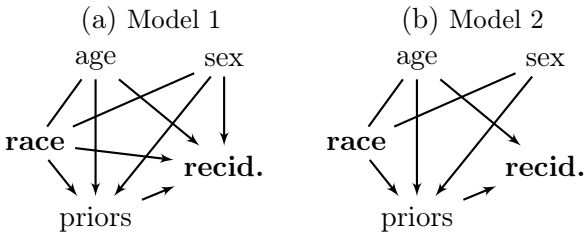


Figure 3.1: Two hypothetical causal graphs for the Compas case.

For illustration, consider the two scenarios depicted in Figure 3.1 for the *Compas* case, which consists of data from Broward County, Florida, initially compiled by ProPublica [ALMK16] that was used to predict the two-year violent recidivism, that is, whether a convicted individual would commit a violent crime in the following two years (1) or not (0). The two models reflect very different scenarios for fairness analysis. In the first causal model, the probability of recidivism is directly affected by the race and the sex of the

offender. If this model was to be the correct one, then it would be accurate to say that recidivism is influenced directly by the individual's race and sex, and it would be mandatory for a prediction to be accurate to take into account the individuals sex and race as a main factor. On the contrary, in the second one, sex and race can still influence recidivism, but only through the total score of prior crimes. In this case, an accurate predictor would not need to use sex nor race as input, and these two attributes would only be needed if an additional group fairness constraint is to be imposed. Overall, these two models serve to highlight the importance of having an accurate causal graph, since two different models have very different meanings for fairness analysis.

The main impediment to causal inference is the unavailability of the true causal graph which indicates the causal relations between variables. Causal graphs can be set manually by experts in the field, but are very often generated using experiments (also called interventions), which might be costly or unfeasible. Alternatively, there are numerous *causal discovery algorithms* (CDAs) in the literature that identify the causal graph based on data, a process known as causal discovery or structure learning.

We showed that using different causal discovery algorithms (CDAs) may lead to dramatic differences on fairness/discrimination conclusions [BMP+23], which means that their outputs should not be trusted blindly. This idea can be split conceptually into two separate factors, namely, that (1) causal graphs produced by different CDAs may differ, and (2) these differences are critical for causal based fairness notions. This chapter presents and complements the idea in [BMP+23] by deepening in the first of these two assertions, including additional results and algorithms.

3.1 Preliminaries

Causal discovery is about finding a graphical model that captures the causal relationships between several random variables based on a sample dataset. Ideally, the graphical model takes the form of a Directed Acyclic Graph (DAG), in which an edge $X \to Y$ indicates that X causes Y, but it can also take the form of a graph mixing directed and undirected edges. The least informative case occurs when the output is an undirected graph, e.g., a

Markov field, as it does not reveal which variables are causing others.

3.1.1 CPDAGs and equivalence classes

Obtaining a DAG from observations is not always possible unfortunately, even for arbitrarily large sample sizes. Consider the procedures $\mathtt{x} = \mathtt{normal}(); \quad \mathtt{y} = (\mathtt{x} + \mathtt{normal}())/\sqrt{2}$ versus $\mathtt{y} = \mathtt{normal}(); \quad \mathtt{x} = (\mathtt{y} + \mathtt{normal}())/\sqrt{2}$, where `normal` is a random number generator of normally distributed samples with mean 0 and standard deviation 1. Semantically, the two procedures are different because in the former, `y` depends on `x`, suggesting a causal relation $\mathtt{x} \to \mathtt{y}$, while the opposite occurs in the latter. But statistically, the two procedures are the same. This is illustrated in Figure 3.2 (left), which depicts the first procedure in blue and the second one in green. It is impossible to distinguish them based on observations, no matter how large the sample is, because the joint distributions they produce coincide.

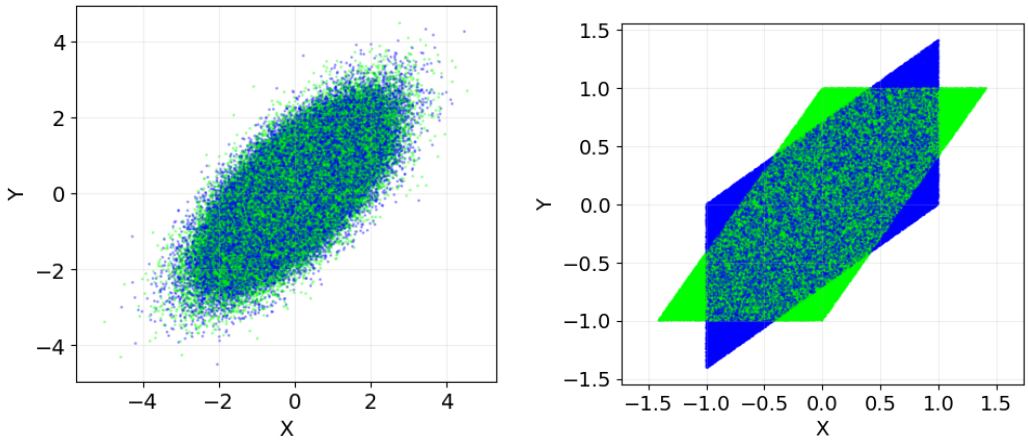


Figure 3.2: Does X cause Y or vice-versa? In the first example (left), $X \to Y$ in blue is indistinguishable from $Y \to X$ in green. In the second (right), they are different, and the causal direction is identifiable under assumptions.

This observation is fundamental in the development of the theory of causality, and it creates a big debate for the use of causality for fairness [HPS16]. But there are cases in which it is indeed possible to distinguish between different causal graphs based on observations. In the aforementioned example,

if the noise was uniform in $[-1, 1]$ instead of normal, Figure 3.2 (right) is produced, and the fact that the variance of Y given $X = x$ depends on x in the green case and does not in the blue one can be used under certain model assumptions to infer the causal directions $X \to Y$ for the blue cloud and $Y \to X$ for the green one. Without model assumptions, however, it is impossible to have an algorithm that determines the causal order of two variables. For three variables, it is still not possible to distinguish the mediators ($\circ \to \circ \to \circ$ and $\circ \leftarrow \circ \leftarrow \circ$) and they can also not be distinguished from a confounder ($\circ \leftarrow \circ \to \circ$). But it is possible to distinguish these three from a so-called v-structure ($\circ \to \circ \leftarrow \circ$). This phenomenon led to the development of equivalence classes of DAGs that can not be distinguished.

These equivalence classes of DAGs are determined by the following equivalence relation: two DAGs are *equivalent* when (i) they have the same set of v-structures and (ii) they can generate the same family of distributions by modifying the model parameters, e.g. if the procedure that generates the data is assumed to be a linear combination of normal random variables, the joint distribution of the output follows necessarily a multivariate Gaussian distribution regardless of the exact parameters that determine the linear combinations. When the variables are all categorical or all continuous with Gaussian joint distribution, these two conditions are equivalent and reduced to just the first one. This assumption is widely used as it simplifies and unifies the theory of equivalence classes of DAGs.

Moreover, from condition (1), each equivalence class can be represented as a partially directed acyclic graph (PDAG), i.e., a graph of directed and undirected edges without directed cycles, where the PDAG $\mathcal{P}$ corresponds to the set of DAGs that agree with the directed edges of $\mathcal{P}$ and assume any direction for the undirected edges without forming a cycle. Not all PDAGs represent an equivalence class. Only those in which directing one of the undirected edges introduces a directed v-structure are, in which case the PDAG is called a complete PDAG (CPDAG). There is therefore a bijection between the set of all CPDAGs and the set of equivalence classes of statistically indistinguishable DAGs. Non-extendable CPDAGs, i.e., CPDAGs with empty equivalence class like $X - Y - Z - X$, whose arrows can not be directed without introducing either a v-structure or a cycle, are the only isolated exception to this

rule, and depending on the author, they may or not be considered CPDAGs.

For instance $\{\circ \to \circ, \quad \circ \leftarrow \circ\}$ is an equivalence class of DAGs, because its elements (DAGs) can not be distinguished, and it is represented with the CPDAG $\circ - \circ$, or with a different notation, $\circ \leftrightarrow \circ$. More examples include $\{\circ\}$ and $\{\circ \to \circ \leftarrow \circ\}$, whose CPDAGs are their respective unique elements, or $\{\circ \leftarrow \circ \to \circ, \quad \circ \to \circ \to \circ, \quad \circ \leftarrow \circ \leftarrow \circ\}$ whose CPDAG is $\circ - \circ - \circ$.

An additional important property of this partition of the set of all DAGs is that there are some model scoring functions such that the score is invariable among DAGs in the same equivalence class, e.g., the BIC score [Hau88].

3.1.2 BIC and CG scores

The BIC score (Definition 21) is a regularized performance score that summarizes in a single real number the extent to which a model is simple and fits some data. In general terms, the BIC score is maximal when the statistical model is both simple and correct, or more technically, when it has few parameters and a high likelihood for the given data. Depending on the reference, the definition may appear multiplied by 2, -2 (minimization in this case) or 1 (compare [Scr16, NC12, Sch78]).

Definition 21. (BIC score) Let $\mathcal{X}$ be an arbitrary domain endowed with a metric and a measure, e.g., euclidean space with the continuous or the discrete measure; let $\hat{p}_\Theta$ be a statistical model of dimension k consisting of a collection $\{\hat{p}_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ of dominated densities $\hat{p}_\theta$; and let $s = (x_i)_{i=1}^N$ be a dataset of N i.i.d. samples $x_i \in \mathcal{X}$, so that $\hat{p}_\theta(s) = \Pi_{i=1}^N \hat{p}_\theta(x_i)$. The *BIC score* is defined as $\mathrm{BIC}(\hat{p}_\Theta, s) \overset{\text{def}}{:=} \ln \hat{p}_{\hat{\theta}}(s) - \frac{k}{2} \ln(N)$, where $\hat{\theta} \overset{\text{def}}{:=} \arg\max_{\theta \in \Theta} \hat{p}_\theta(s)$ is the maximum likelihood estimator of the parameters.

Therefore, the value of the BIC score (Definition 21) consists of the sum between the log-likelihood of the model and a weighted regularization term that penalizes models with large number of tunable parameters.

The technical reasoning behind the formulation of BIC score is driven by the computation of $\int_\Theta \hat{p}_\theta(s) d\theta$ [NC12]. This term is the likelihood of the statistical model $\hat{p}_\Theta$ with unknown parameters and assuming the (possibly improper) uniform prior on θ. Using the first three terms of the second-

order Taylor series expansion of $\ln \hat{p}_\theta(s)$ around $\hat{\theta}$ yields $\ln \int_\Theta \hat{p}_\theta(s) d\theta \approx C + \text{BIC}(s, \hat{p}_\Theta)$, where C is a constant that depends on s and grows with k but is bounded as n grows, so that it can be ignored for large values of n when comparing two models. As a consequence, interpreting the approximation as an equality, choosing a model with higher BIC score than other corresponds to selecting the model $\hat{p}_\Theta$ with higher likelihood. Furthermore, since the BIC score corresponds to a log-likelihood, it is decomposable as the sum of the local BIC scores of each node w.r.t. its parents (directed and undirected parents).

In practice, computing the BIC score for a given predictive model is not straightforward when the target Y is continuous. When Y is categorical, Definition 21 can be used directly because typically, the output of a categorical model is a soft-probability vector, i.e., a categorical density. However, when Y is continuous, the output is a single value $\hat{X} = f(X)$ that estimates $\mathbb{E}[\,[\![\, Y \,]\!] \mid X = x]$ rather than a density. Since Definition 21 cannot be used directly anymore, Fact 22 is used as a proxy for the BIC score.

Fact 22. (BIC score proxy[Gir21, TK86]) Let $\hat{f}$ be a model that estimates $\mathbb{E}[\,[\![\, Y \,]\!] \mid X = x]$ as $\hat{f}(x)$. If the residuals of the true densities $p(\cdot|x)$ are normally distributed for each x, then asymptotically, maximizing the BIC score of a density model that estimates both the mean and the variance of $p(\cdot|x)$ is equivalent to maximizing $\text{BIC}(\hat{f}, S) \stackrel{\text{def}}{=} -N \ln \left(\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{f}(x_i)\right)^2 \right) - \frac{1}{2} \ln(N) \cdot \mathsf{n\_params}(\hat{f})$.

More generally, for mixed scenarios involving both categorical and continuous variables, the TETRAD and gCastle implementations use a score called Conditional Gaussian (CG) score [ARC18] that makes extensive use of Gaussian mixtures and is equivalent to the BIC score under the following assumptions.

A1. The continuous data were generated from a single joint (multivariate) Gaussian mixture where each Gaussian component exists for a particular setting of the discrete variables.

A2. The instances in the data are independent and identically distributed.

A3. All Gaussian mixtures are approximately Gaussian.

It is also crucial in practice to have enough samples when conditioning on

several categorical variables simultaneously, particularly when these variables are parents of the same variable; this and A2 are the most relevant in the all-discrete case. Regarding continuous variables, A1 implies implicitly linearity and gaussianity of the residuals because of the nature of each Gaussian component; this is the most relevant assumption in the all-continuous case.

The CG score implements the likelihood function as follows. For each node, consider the set of that node along with its parents, and filter the continuous and categorical variables separately. Then, consider a single discrete model for all the discrete variables, and for each combination of the discrete variables, consider a separate model for the joint density of the continuous variables. This setting exposes explicitly all the density models needed in the Definition 21 to compute the BIC score.

3.2 Causal discovery algorithms

3.2.1 GES

Greedy Equivalence Search (GES) [Chi02], described in this document as Algorithm 4, consists of searching for a particular state over an abstract space of states and transitions. The states are equivalence classes of DAGs (equivalently CPDAGs) and the search objective is the state that maximizes BIC score, hence as, the output of GES is not a single DAG, but an equivalence class of DAGs represented as a CPDAG.

The transitions of the search space are given by the following rule: a transition from a state to another exists if and only if there are two DAGs, one on each equivalence class, that differ only in the addition or removal of exactly one edge. Hence, there are two types of transitions: forward (adding one edge) and backward (removing one edge). The neighboring states for the state $\mathcal{P}$ are represented with the variable neighbors.

Computing the neighboring states of a given state is carried out by finding edges $X \to Y$ that can be added (or removed) in such a way that the resulting PDAG can be *extended*, i.e., transformed into a DAG by deciding the direction of the undirected edges, and then *completed* to obtain the CPDAG that represents the equivalence class containing it. The completion

Algorithm 4 GES algorithm.

Input: Dataset $\mathcal{D}$ of $|V|$ variables.
Output: CPDAG $\mathcal{P}$ that maximizes BIC score.
$\mathcal{P} \leftarrow$ disconnected CPDAG of $|V|$ nodes
score $\leftarrow 0$
for phase $\in$ [forward, backward] **do**
  while True **do**
    neighbors $\leftarrow \{\mathcal{P}' : \mathcal{P} \rightarrow \mathcal{P}'$ is a phase-transition$\}$
    if |neighbors| $= 0$ **then**
      break
    $\mathcal{P}' \leftarrow \arg\max_{\mathcal{P}' \in \text{neighbors}} \Delta\text{BIC}(\mathcal{P}, \mathcal{P}', \mathcal{D})$
    Δscore $\leftarrow \Delta\text{BIC}(\mathcal{P}, \mathcal{P}', \mathcal{D})$
    if Δscore < 0 **then**
      break
    $\mathcal{P} \leftarrow \mathcal{P}'$
    Add Δscore to score
return $\mathcal{P}$, score

algorithm is simple to implement from the definition of a completed PDAG and is explained in [DT92, Chi02]. In contrast, the extension algorithm is more complex. Briefly, letting N_X denote the undirected neighbors of X, one should find all pairs of variables (X, Y) and sets of variables $S \subseteq N_Y$, such that after adding (removing) the edge $X \rightarrow X$ and setting the directions of S (also of $N_X \cap N_Y$ when removing), the resulting PDAG can be converted into a DAG by smartly deciding the direction of the undirected edges. This algorithm was first introduced with missing details in [DT92], then implemented by [Chi02], although [Gam21] found an error later.

The change in BIC score after following a transition can be computed using a simple rule instead of fitting the whole global model on both states because the BIC score can be decomposed as the sum of the local BIC scores of each of its directed and undirected parents. Since there is a unique variable Y whose parents change during a transition, then the global BIC score difference corresponds exactly with the local BIC score difference for the model of Y and its parents. This optimization corresponds to $\Delta\text{BIC}(\mathcal{P}, \mathcal{P}', \mathcal{D})$ in Algorithm 4.

The greedy strategy of GES consists of repeatedly following the best forward

transition at each state that it encounters until reaching a local maximum, i.e., until the next state reduces the BIC score, (this is the forward phase in Algorithm 4) and then, analogously (backward phase), repeatedly following the best backward transition until a local maximum is reached. No special rule exists to resolve ties for these arg-max operations, although they are extremely unlikely to occur in practice.

The distinctive essential feature of GES is that its greedy technique, which prunes the search space dramatically, is guaranteed to find the optimal state of the whole space, provided that the data distribution matches the assumed statistical model.

In practice, the score implemented in GES is the GC score which is guaranteed to be a proxy for BIC score only under the aforementioned assumptions A1, A2 and A3 (page 66). These assumptions may sound too strict, but the empirical evidence of several articles shows that many implementations of GES performs well even when these assumptions do not hold exactly [ARC18, Chi02, HB12]. This is the case for Tetrad's fges[RZG+18] (written in Java), the pcalg[KMC+12, HB12] library for R (written in C++ underneath), and other Python implementations [Gam21, KG19]. The predictive models of these implementations are therefore pre-configured to linear regression for continuous variables.

3.2.2 PC

The Peter Spirtes and Clark Glymour (PC) algorithm [SGSH00] consists of two main steps. The first, which accounts for most of the computational costs, is to produce a skeleton graph G witch contains only undirected edges, and the second consists of orienting the undirected edges of G to form a CPDAG $\mathcal{P}$.

As shown in Algorithm 5, the PC algorithm starts with the fully connected graph and relies on conditional independence tests in order to either remove or keep edges. For each edge $X \to Y$ and each subset $\mathbf{Z}$ of the neighbors of X and Y, the algorithm checks whether X and Y are independent conditioned on Z using a conditional independence test. The depth d represents the size of the conditioning sets. During the first iteration, all pairs of vertices are

Algorithm 5 PC algorithm.

Input: Dataset $\mathcal{D}$, and significance level α.
Output: CPDAG $\mathcal{P}$.
$G \leftarrow$ totally connected (undirected) skeleton
$d \leftarrow 0$
while $|\text{adj}_G(X) \setminus Y| \geq d$ for every pair of adjacent vertices $X - Y$ in G **do**
  for each adjacent pair $X - Y$ in G **do**
    if $(|\text{adj}_G(X) \setminus Y| \geq d)$ **then**
      for each $Z \subseteq \text{adj}_G(X) \setminus Y$ **do**
        if $|Z| = d$ and $I(X, Y|\mathbf{Z}) \geq \alpha$ **then** Remove edge $X - Y$ in G
          Save $\mathbf{Z}$ as the separating set of $X - Y$
          break
  $d \leftarrow d + 1$
$\mathcal{P} \leftarrow G$ as a partially directed graph
for each triple of vertices (X, C, Y) with $C \in \text{adj}_{\mathcal{P}}(X)$ and $Y \notin \text{adj}_{\mathcal{P}}(X)$ **do**
  if $C \notin \mathbf{Z}$ (separating set of $X - Y$) **then**
    orient $X - C - Y$ as $X \rightarrow C \leftarrow Y$ in $\mathcal{P}$
$\mathcal{P} \leftarrow$ completion of the extension of $\mathcal{P}$ **return** $\mathcal{P}$

tested conditioning on the empty set $\emptyset$, i.e. $d = 0$. Thus, some edges will be removed, and the algorithm will proceed only with the remaining edges in the next iteration ($d = 1$). The size of the conditioning set, d, is incremented after every iteration until d is greater than the size of the adjacent sets of the testing vertices. Note that the graph at hand is updated at each test after edge(s) deletion. Moreover and most importantly, the set of conditioning variables $\mathbf{Z}$ is stored as it is needed later to detect potential presence of v-structures in the causal graph. If the conditional independence tests are reliable, then PC finds the true skeleton graph [LHL+16]. An important observation of this first part of PC is that for high-dimensional sparse graphs, the conditional independence tests are organized in a way that makes the algorithm computationally efficient, since it only needs to test conditional independencies up to order $k - 1$, where k is the maximum size of the adjacency sets of the nodes in the DAG at hand.

For the second part that orients the edges of G (as a PDAG) to form the CPDAG $\mathcal{P}$, PC considers all unshielded triples in G, i.e. triples of nodes (X, C, Y) such that $X - C - Y$ and $X \not- Y$, and orients it as a v-structure

if and only if $C \notin \mathbf{Z}$ (separating set of (X, Y)). After this, $\mathcal{P}$ is a PDAG, possibly not yet a CPDAG, so as in GES, the algorithm finds the completion of any extension of $\mathcal{P}$. This step, can be performed equivalently by orienting as many of the remaining undirected edges as possible by applying repeatedly the rules shown in Algorithm 6 until no more edges can be oriented. The PC-algorithm is proved to be efficient for sparse graphs. The main reason for that is that, once and edge is deleted, the neighbors of a particular node are dynamically updated when saving $\mathbf{Z}$ as the separating set of $X - Y$ in Algorithm 5 [LHL+16].

Algorithm 6 Orientation rules for PC.

Input: PDAG $\mathcal{P}$.
Output: CPDAG (modified in-situ): completion of the extension of $\mathcal{P}$.
while no more edges can be oriented **do**
  for each (X, C, Y) with $X \to C - Y$ and $Y \notin \text{adj}_{\mathcal{P}}(X)$ **do**
    orient $C - Y$ as $C \to Y$ in $\mathcal{P}$ ▷ Rule 1
  for each chain $X \to C \to Y$ **do**
    orient $X - Y$ as $X \to Y$ in $\mathcal{P}$ ▷ Rule 2
  for each pair of chains $X \to C_1 \to Y$ and $X \to C_2 \to Y$ such that $C_2 \notin \text{adj}_{\mathcal{P}}(C_1)$ **do**
    orient $X - Y$ as $X \to Y$ in $\mathcal{P}$ ▷ Rule 3
return $\mathcal{P}$

The conditional independence tests have an α value (input of Algorithm 5) for rejecting the null hypothesis of independence or conditional independence. For continuous variables, PC uses tests of zero correlation or zero partial correlation for independence or conditional independence, respectively. For discrete or categorical variables, it uses either a chi-square or a g-square test of independence or conditional independence. The default value of α is 0.01. However, for discrete searches, using a value of 0.05 is also recommended[1].

Spirtes et al. [SGSH00] provided three different heuristics of selecting the order of conditional tests between variables. Depending on the heuristic, the skeleton phase of the PC algorithm can be either dependent or independent of the order at which the variables are present in the dataset [Tsa19]. The first heuristic tests the variables in lexicographic order. That is changing

[1] https://cmu-phil.github.io/tetrad/manual/

the order of the columns in the dataset results in a different order of how the statistical tests are performed. The second heuristic, at the other hand, relies on dependencies between the variable pairs. In other words, the tests are performed on the pair of variables that are the least dependent. As for the conditioning subsets, they are selected in a lexicographic order. Finally, the third heuristic is based on the idea of performing the tests on the least dependent pairs of variables while conditioning on the subsets that are most dependent on either variable of the pair. Thus, it is clear that the heuristics 1 and 2 are order-dependent while heuristic 3 is not. In other word, the generated causal graph when using heuristics 1 and 2 may change due to a change in the order in which the variables appear in the dataset. However, heuristic 3 is totally order-independent.

According to Spirtes et al. [SGSH00], the v-structure discovery of the PC algorithm should always be applied first (in the orientation step of the algorithm). However, the order of applying the rules shown in lines is not predetermined. That is, these rules can be applied in any order. Thus, in the finite sample size case, statistical errors exist and can result in a misleading skeleton [Tsa19]. Colombo et al. proved by examples that different variable orderings can lead to different orientations and consequently, affecting the output of the step 2 of the PC algorithm even if the skeleton and separating sets are order-independent [CM+14].

3.2.3 Direct LiNGAM

Direct Linear Non-Gaussian Acyclic Model [SIS+11] (LiNGAM) is a CDA that, unlike the previously discussed algorithms, yields a unique directed graph (DAG). The keywords in the name of the algorithm refer to the fact it assumes that the input dataset comes from a joint distribution $(X, Y, ..., Z)$ that can be written as a linear combination $(X, Y, ..., Z)^T = A(X, Y, ..., Z)^T + B$ for some random vector B of independent non-Gaussian exogenous noise B and some matrix A whose non-zero entries form an acyclic adjacency matrix.

Although the theory of CPDAGs tells that it is not possible to distinguish whether $X \to Y$ or $Y \to X$ in general, with these assumptions, it is. Indeed, there are statistical asymmetries in Figure 3.2 (right) between the blue and green clouds of points. For example, in the blue cloud, the variance of $Y|X =$

x is constant when varying x while the variance of $X|Y = y$ is not when varying y. A strictly stronger observation is that Y can be written as $f(X) + \texttt{noise}$, where the noise is independent of X and $f(x) = E[Y|X = x]$, but the converse does not hold, as X can not be written as $f(Y) + \texttt{noise}$ unless the noise amplitude depends on Y. So, being the first model more simple, and the only one following the assumptions of LiNGAM (because f happens to be linear as well), it outputs the causal direction $X \to Y$. The exact opposite occurs in the green cloud, so $Y \to X$ would be chosen.

Algorithm 7 Direct LiNGAM.

Input: Dataset with data columns $\bar{X}, \bar{Y}, ..., \bar{Z}$ representing k variables $X, Y, ..., Z$, and threshold $\alpha > 0$
Output: DAG with weights matrix $W_{X \to Y}$.
$S \leftarrow [\,]$ Empty causal order list
while $|S| < k$ **do**
  $X = \arg\min_{X \notin S} \sum_{Y \notin S \cup \{X\}} I(X; r_{\bar{X} \to \bar{Y}})$
  Push X to the end of S
  for $Y \notin S \cup \{X\}$ **do**
    $\bar{Y} \leftarrow r_{\bar{X} \to \bar{Y}}$ ▷ Remove the effect of X on Y
$W_{X \to Y} \leftarrow 0$ for all $X, Y \in S$
for $Y \in S$ **do**
  $\mathsf{pa} \leftarrow \{X : X \text{ precedes } Y \text{ in } S\}$
  $W_{\mathsf{pa} \to Y} \leftarrow$ linear coefficients $(Y = f(\mathsf{pa}))$
  Set small values in $W_{\mathsf{pa} \to Y}$ to 0 (if abs.$< \alpha$)
return W

In other words, for LiNGAM, the causation $X \to Y$ is strong when the prediction residual $Y - f(X)$ is independent or not sufficiently dependent on X. A dependency score $I(X, Y - f(X))$ is used for this purpose. Mutual information is a good candidate [HS13], although other metrics have been proposed [Shi14]. LiNGAM extends the same analysis to more variables by adding up or averaging scores over all the candidate parents, i.e., by considering $I(X, Y - f(X)) + I(Z, Y - f(Z))$ as the dependency score of the child Y on the parents X and Z. Notice that, unlike PC, LiNGAM does not threshold the dependency score to decide discretely for dependence or

independence. Instead, it compares several scores between them, e.g., for two variables, it will compare the independence score of $X \to Y$ versus $Y \to X$.

LiNGAM (Algorithm 7) learns the causal graph in two steps. First, it finds a causal order of the variables: an ordered list, where the first nodes are more likely to be ancestors and the last, descendants. More precisely, we compute the dependency scores of every variable X on all the other variables, and pick the one with the lowest dependency score (the most independent) as the first node in the list. Then, we subtract the effect of X on all the other variables, and select the next variable using the same procedure.

After the order has been established, all arrows will necessarily go from nodes that appear earlier to nodes that appear later in the list, or put informally, from the most independent to the most dependent ones. Next, from all such arrows, the algorithm proceeds to filter out those whose coefficient in linear regression is smaller (in absolute value) than a threshold α. For this comparison to be fair, the variables should be normalized. The coefficients that were large enough are stored in a matrix W, setting 0 everywhere else. This matrix represents a graph with weights on the arrows, and because of the constraint on the arrows by the ordered list, it is a DAG.

3.3 NoTears and GOLEM

Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NoTears) [ZARX18] is a CDA that avoids the ad-hoc combinatorial problem of exploring the space of DAGs by converting it into an optimization problem. The output of NoTears is a matrix of weights whose non-zero entries form the adjacency matrix of a DAG, just like Direct LiNGAM.

NoTears considers the optimization problem, for a given dataset $D \in \mathbb{R}^{n \times d}$ of n observations and d features, of finding the matrix $W \in \mathbb{R}^{d \times d}$ that represents a DAG and minimizes the linear predictions' mean squared error $\|D - DW\|_2^2$ plus an ℓ_1-regularization term. That is,

$$F(W) \stackrel{\text{def}}{:=} \min_W \quad \frac{1}{2n}\|D - DW\|_2^2 + \lambda\|W\|_1,$$

subject to W representing a DAG, i.e., the matrix $1(W \neq 0)$ must be the

adjoint matrix of a DAG. The novelty introduced with NoTears is the use of an alternative characterization of acyclicity that states that W represents a DAG if and only if

$$h(W) \stackrel{\text{def}}{\coloneqq} \operatorname{tr}(e^{W \circ W}) - d = 0,$$

where $\operatorname{tr}(A)$ is the trace of the squared matrix A, i.e., the sum of the diagonal, $A \circ B$ is the element-wise product of two matrices and e^A denotes matrix exponentiation.

Using the augmented Lagrangian method, the constrained problem can be rewritten as the dual problem $\max_{\alpha \in \mathbb{R}} \min_W L^\rho(W, \alpha)$ with Lagrange multiplier α and step size ρ, and augmented Lagrangian

$$L^\rho(W, \alpha) \stackrel{\text{def}}{\coloneqq} F(W) + \frac{\rho}{2}|h(W)|^2 + \alpha h(W).$$

This double optimization problem can be converted into a sequence of unconstrained problems by means of dual ascent optimization: for a fixed initial α, we find a minimizer W_α of $L^\rho(W, \alpha)$, then ascend by updating $\alpha \leftarrow \alpha + \rho \frac{\partial}{d\alpha}(L^\rho(W, \alpha))|_{W=W_\alpha}$, i.e., $\alpha + \rho h(W_\alpha)$, and repeat until convergence.

It is worth noticing that the constraining set $\{W : h(W) = 0\}$ is a non-convex set, hence NoTears inherits the difficulties of non-convex optimization [ZARX18]. Also, it has been shown that the output of NoTears is susceptible to the scale of the data [KS22] and the authors of [LB14] point out that if a linear model $D = DW + \epsilon$ is assumed, the term $\frac{1}{2n}\|D - DW\|_2^2$ in $F(W)$ should be modified into $\frac{1}{2n}\|(D - DW)\Sigma^{-1/2}\|_2^2$, where $\Sigma = \operatorname{cov}(\epsilon)$ is the variance of the (unobservable) noise terms. This second problem is addressed by the GOLEM algorithm.

Gradient-based Optimization of dag-penalized Likelihood for learning linEar dag Models (GOLEM) [NGZ20] is a method based on NoTears, but with a different choice for the loss term $\|D - DW\|_2^2$ in $F(W)$. The method proposes to replace the loss term with two versions of the (negative) BIC score proxy explained in Fact 22. Recall that this proxy score corresponds with BIC score only under the assumption that model for the data is a linear combination with Gaussian noise. In the first version, assuming equal variances (EV) of the noise, the loss term becomes $\ell_{EV} \stackrel{\text{def}}{\coloneqq} \frac{d}{2}\log(\|(D - DW)\|^2) - \log|\det(I - W)|$. In the second, assuming non-equal variances (NV) of the noise, the

loss term is modified by averaging over the d columns and becomes $\ell_{NV} \stackrel{\text{def}}{:=} \frac{1}{2}\sum_{i=1}^{d} \log(\|(D - DW)_i\|^2) - \log|\det(I - W)|$. By replacing the loss term of NoTears $\|D - DW\|_2^2$ in $F(W)$ with these two losses, two different CDAs are obtained: GOLEM-EV and GOLEM-NV. GOLEM-EV is also affected by rescaling [RSW21], as ℓ_{EV} treats all columns on the same scale, but it has better performance than GOLEM-NV on unscaled data.

3.4 Overview

In this section, we presented 5 popular causal discovery algorithms. There are more CDAs in the literature, including, variants and extensions of PC like FCI [SMR99], variants of Direct LiNGAM like the original ICA-LiNGAM [SHH+06] and alternative CDAs based on alternative definitions of causality like SBCN [BHMR17]. These are not considered in this chapter for brevity, experimental consistency and because the selected ones are already diverse in how they operate. For instance, the outputs of FCI and SBCN are neither DAGs nor CPDAGs, so they are not suitable for a direct comparison with the other algorithms. Table 3.1 summarizes the CDAs included in the experiments.

Algorithm	Output	How it operates
GES	CPDAG	Model likelihood maximization
PC	CPDAG	Independence tests for colliders
Direct LiNGAM	DAG	Residual asymmetries
NoTears	DAG	Constrained optimization
GOLEM	DAG	Constrained optimization

Table 3.1: High level comparison of the CDAs included in the experiments.

It is important to notice from Table 3.1, that the CDAs can be categorized based on the type of its output. The output of GES and PC is a CPDAG, and they have in common that their approach for solving the problem is mostly combinatorial and explicitly takes into account the space of CPDAGs, and their correspondence to equivalence classes of DAGs. On the other hand, the output of LiNGAM, NoTears and GOLEM is a DAG. DAGs have the advantage that they provide a more rich and detailed structure, however, it has been criticized that these structures do not correspond to actual causal

relationships [KS22], and they can be attacked very easily [SZDK22] in the sense that one can make simple modifications to the input dataset that change the structure of the output DAG in targeted manner. These type of experimental attacks are very important as they help to detect conceptual weaknesses in the definitions.

3.5 Experiments

In this section we corroborate experimentally the main thesis of this chapter: the causal graphs produced by different CDAs may differ significantly. To do this, we provide synthetic data distributions that were specially crafted to be very simple (to capture minimal scenarios) and such that when different CDAs are executed on datasets sampled from these distributions, the outputs differ significantly and with high probability w.r.t. sampling noise. Notice that the main paper [BMP+23] contains already several examples of real datasets for which the CDAs have different outputs, but these are based on relatively complex real world datasets, and they were not tested for NoTears nor GOLEM. In particular, the models 1 and 2 shown in Figure 3.1 correspond to the outputs of PC and GES respectively for the Compas dataset, a fact that is very provocative on its own. In this section we find and present far simpler examples in which the same phenomenon occurs.

The objective of this section is not to judge the correctness of any particular CDA, but to find and reveal explicitly simple distributions for which the CDAs' outputs differ. It is expected, from the different operation mechanisms of the CDAs (see Table 3.1), that different CDAs produce similar but not exactly the same causal graphs in all executions. Also, the output of LiNGAM is a DAG, while the output of GES and PC is a CPDAG, so, it is understandable that the outputs differ in general. However, in order to better understand the practical and semantic difference between the CDAs, it is crucial to have at hand simple examples for which the CDAs disagree. This section is dedicated to these examples, as they are not available in the literature.

Since the CDAs have conflicting assumptions, e.g., because the GES requires the data to be (at least approximately) a linear combination of Gaussian ex-

ogenous random variables while LiNGAM requires non-gaussian residuals, we opted for a combination of normal and uniform exogenous distributions that are combined linearly to produce the observed attributes. Concretely, for 3 normally (N_1, N_2, N_3) and 2 uniformly (U_1, U_2) distributed random variables, we searched among all linear combinations (matrices of size 5×3) for those that produce three random variables such that, when sampled, the CDAs disagree very often. The choice of a linear model with just 5 exogenous and 3 observed variables is simplistic on purpose to privilege the conceptual understanding of the distributions that are produced.

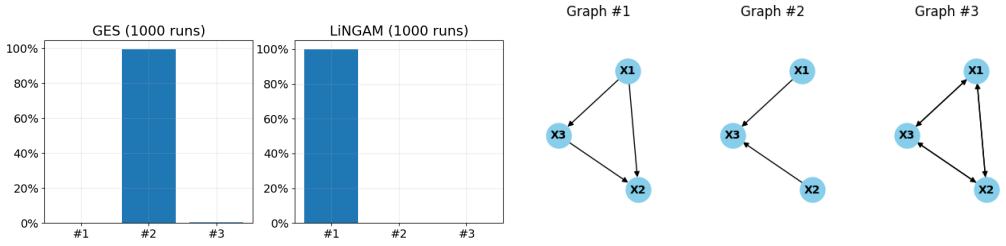


Figure 3.3: Histograms of outputs of LiNGAM and GES for Example 23.

These distributions were obtained by means of a genetic search following the procedure explained in what follows. Recall that a genetic search consists of iterating three basic procedures: scoring candidate solutions, discarding the ones that have low scores and combining the remaining to form new candidates. In our experiments, the searches were executed with a population size of 50 to 300 and a scoring function that seeks to privilege the cases in which GES and LiNGAM report very distant causal graphs very often (for many samples of fixed size). The experiments were run in Python using the CDAs' implementations of the library gCastle, with their default parameters. As we found, for the following simple scheme, the two algorithms report opposite results.

Example 23. Let

$$\begin{aligned} X_1 &= -2N_1 + 0.1U_1 \\ X_2 &= -3.8N_2 - 2.8U_2 \\ X_3 &= -0.4N_3 + 1.8N_1 + 3.8U_2 \end{aligned}$$

where N_1, N_2, N_3, U_1, U_2 are random variables with mean 0 and variance 1, the first three of which are Gaussian and the remaining uniform. Then,

the CDAs LiNGAM and GES have contradictory outputs (Figure 3.3). In particular, for populations of 1000 samples, GES reports almost always that X_2 is a collider while LiNGAM reports also very often (at least 80% of the times) the exact opposite. For LiNGAM, all arrows should be inverted and X_2 is a confounder. For this example, PC agrees with GES very often.

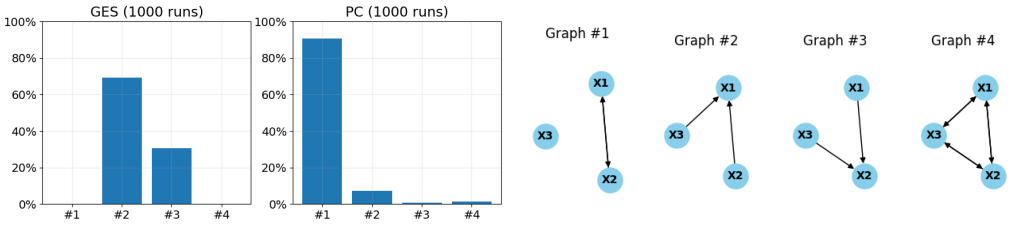


Figure 3.4: Histograms of outputs of GES and PC for Example 24.

Example 24. Defining N_1, N_2, U_1, U_2 as in Example 23, let

$$\begin{aligned} X_1 &= -2.6N_1 + 3.4U_1 \\ X_2 &= -4.7N_1 - 3.7U_1 \\ X_3 &= 2.7N_1 - 3.4U_1 + 1.1N_2 + 0.4U_2 \end{aligned}$$

Then, the outputs of GES and PC are different almost always (Figure 3.4). According to PC, the CPDAG consists of a single undirected edge between X_1 and X_2, whose possible DAGs are either setting X_1 as causing X_2 or vice-versa. Nevertheless, according to GES, there is a collider, either in X_1 or X_3. In this example, PC reports the same output more than 90% of the time while GES reports its two outputs around 60% and 40% of the time respectively, meaning that for this particular distribution, the sampling noise plays a significant role for GES.

The sample size of Examples 23 and 24 was set to 1000, which is large enough to make sample size noise negligible for 3 variables. Indeed, Figure 3.5 shows the effect of sample size on the output stability of several CDAs. For each sample size and CDA and a fixed distribution, we ran 1000 times the CDA on different samples of size 1000 and counted how often the most popular output came out, a measure that reflects robustness of the output. This was repeated for different distributions, including naturally Examples 23 and 24. The main observations are that (1) approximately, the curves do not vary

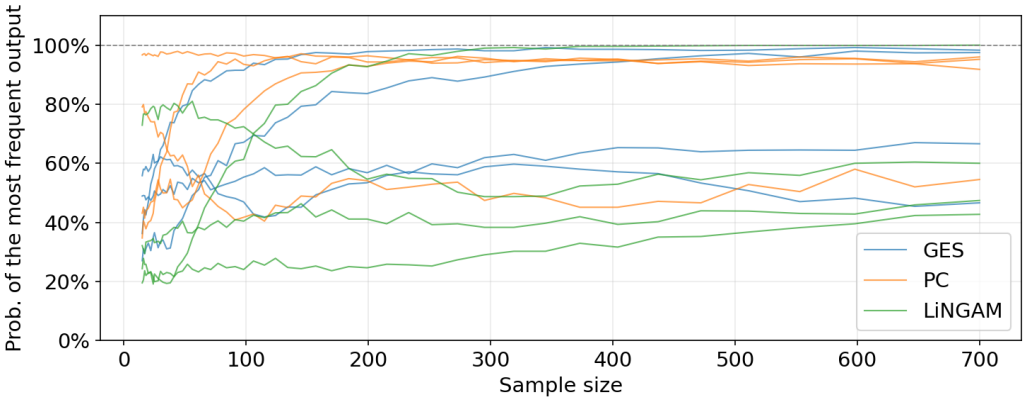


Figure 3.5: Output stability relative to sample size for selected distributions.

much when the sample size exceeds 200, and (2) for some distributions, the tendency is increasing towards values above 90% but for some others, the CDAs are not robust, as the frequencies approach values below or around 60% (as in the example shown in Figure 3.4).

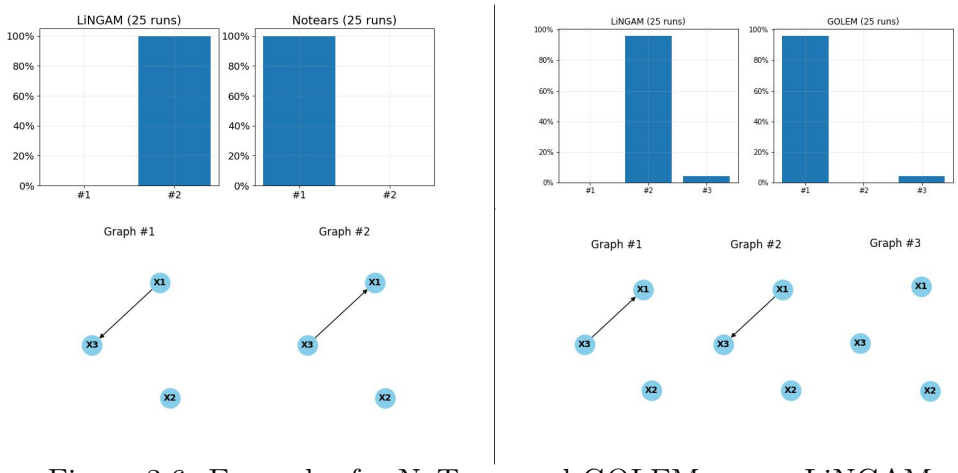


Figure 3.6: Examples for NoTears and GOLEM versus LiNGAM.

Finally, the examples mentioned in this chapter are not limited to PC, GES and LiNGAM. Example 25, along with its corresponding figures 3.6 and 3.7, show the same issues for NoTears and GOLEM versus LiNGAM (and PC for completeness). Overall, these distributions show contradictions between the algorithms, and for some, the algorithms are very confident (with respect to sampling noise) of opposite conclusions. Finding the examples for GOLEM took significantly more time than the others (14 CPU hours) because it was

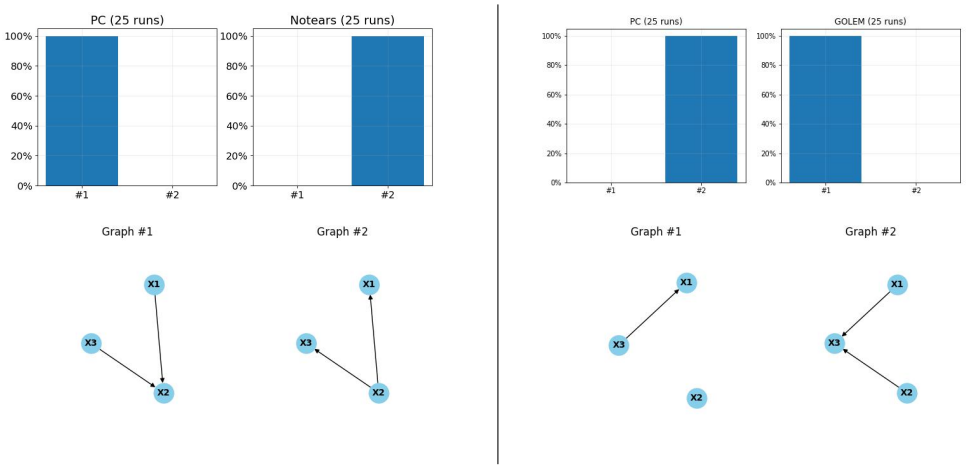


Figure 3.7: Examples for NoTears and GOLEM versus PC.

the slowest of all the algorithms.

Example 25. With the same convention for N_1, N_2, N_3, U_1, U_2 as in Examples 23 and 24, denote the vector of unobservable variables as $Z \stackrel{\text{def}}{:=} [N_1, N_2, N_3, U_1, U_2]^T$, and let $X_L \stackrel{\text{def}}{:=} M_L Z$ and $X_R \stackrel{\text{def}}{:=} M_R Z$ be vectors of three observed real random variables, where

$$M_L \stackrel{\text{def}}{:=} \begin{bmatrix} -2.0 & -2.0 & 0.4 & 0.0 & 3.4 \\ -2.9 & 2.9 & -3.7 & 3.2 & 0.7 \\ -4.2 & -4.2 & 0.0 & 2.0 & -0.9 \end{bmatrix}, \text{and}$$

$$M_R \stackrel{\text{def}}{:=} \begin{bmatrix} -0.7 & -1.7 & -4.8 & -0.2 & 4.7 \\ 0.5 & -2.9 & -3.6 & -4.8 & -2.1 \\ -0.6 & 4.4 & -2.3 & 0.7 & -4.2 \end{bmatrix}.$$

Then, the outputs of LiNGAM and NoTears are different almost always for X_L, and so are those of LiNGAM and GOLEM for X_R. This is shown in Figure 3.6, X_L at the left and X_R at the right. If we let instead

$$M_L = \begin{bmatrix} 3.9 & 4.2 & 0.0 & 0.0 & 0.0 \\ -1.6 & -1.8 & 4.7 & -0.8 & 0.3 \\ 0.0 & 0.0 & 2.7 & 3.2 & -4.0 \end{bmatrix}, \text{and}$$

$$M_R = \begin{bmatrix} -4.3 & -3.6 & 4.1 & 1.1 & 4.0 \\ -2.4 & -1.7 & -4.5 & 0.0 & 0.0 \\ 0.4 & -3.7 & 0.0 & 4.3 & -0.9 \end{bmatrix},$$

then the same phenomenon happens with respect to PC (instead of LiNGAM). Figure 3.7 shows this phenomenon (X_L at the left and X_R at the right). Notice that the differences with PC are not merely of one being a DAG and the other a CPDAG, which is expected and obvious, but true changes in structure.

Perhaps the most interesting fact about these examples is that they are very simple distributions, and the differences between the output graphs are not simply missing edges (whose solution would be to modify a threshold parameter), but also reversed edges and misidentification of v-structures.

3.6 Conclusion

This chapter complements the paper [BMP+23], which shows that different CDAs can generate different causal graphs, and demonstrates how slight differences between causal graphs may have significant impact on fairness/discrimination conclusions. This is done by presenting some particular selected examples of simple distributions for which the disagreement is very high and very frequent between different CDAs, including some that were not reviewed in [BMP+23]. As a consequence, the hypothesis that the choice of the CDA has a critical impact on fairness conclusions is reinforced.

The main takeaway of the paper and this chapter is that causal based fairness assessment should not be fully automated unless the causal graph is known. In other words, CDAs should be used as tools for analysis but not as analysts themselves. It is important to run several of them, varying their parameters to understand which causal graphs are more likely to be correct, and the sample size should be taken into account as a potential source of error. Finally, the distributions presented in this chapter are potentially useful to understand more deeply the differences between the CDAs in practice as they trigger contrasting outputs while being relatively very simple.

With this chapter, we conclude the study of fairness until the discussion section. We now turn our attention into privacy, the second pillar of Ethical AI studied in this dissertation.

Chapter 4

Frequency Estimation of Evolving Data Under Local Differential Privacy

Estimating histograms of evolving categorical data is a fundamental task in data analysis and data mining that requires collecting and processing data in a continuous manner. A typical instance of such a problem is the online monitoring performed on software applications [BEM+17], for example for error reporting [GKG+09], to find commonly typed emojis [App17], as well as to measure the users' system usage statistics [DKY17]. However, the data collected can contain sensitive information such as location, health information, preferred webpage, etc. Thus, the direct collection and storage of users' raw data on a centralized server should be avoided to preserve their privacy. To address this issue, recent works have proposed several mechanisms satisfying Differential Privacy (DP) [Dwo06b, DMNS06, DR+14] in the distributed setting in which an individual can directly randomize her own profile locally, referred to as Local DP (LDP) [KLN+08, DJW13, DWJ13].

One of the strengths of LDP is its simple trust model: since each user perturbs her data locally, user privacy is protected even if the server is malicious. For instance, some big tech companies have chosen to operate some of their applications in the local model, reporting the implementation of LDP protocols to collect statistics on well-known systems such as Google Chrome

browser [EPK14], Apple iOS/macOS [App17], and Windows 10 operating system [DKY17].

Existing LDP protocols for frequency estimation typically focus on one-time computation [FNNT22, ASZ19, WBLJ17, KBR16, KOV16, CMM21, BS15, BNST17]. However, considering both evolving data and the continuous monitoring together, pose a significant challenge under LDP guarantees. For instance, the naïve solution in which an LDP computation is repeated, will quickly increase the privacy loss leading to large values of ϵ due to the sequential composition theorem in DP [DR+14]. To tackle this issue, most state-of-the-art solutions relies on **memoization** [EPK14, DKY17, EFM+20, ACBX21, ACBX22].

Initially proposed by Erlingsson, Pihur, and Korolova [EPK14], the memoization-based RAPPOR protocol allows a user to memorize randomized versions of their true data and consistently reuse it when the same true value occurs. In addition, to improve privacy (*e.g.*, minimize data change detection and/or tracking), the RAPPOR [EPK14] protocol applies a second round of sanitization to the memoized value. However, the longitudinal privacy protection of RAPPOR only works if the underlying true value never or rarely changes (or changes in an uncorrelated fashion), which is unrealistic for evolving data (*e.g.*, the number of seconds an application is used) as the privacy loss is proportional to the number of data changes, *i.e.*, the domain size k in the worst-case.

To address this issue, Ding, Kulkarni, and Yekhanin [DKY17] have proposed a new LDP protocol named dBitFlipPM that improved memoization by mapping several values to the same randomized value. More precisely, dBitFlipPM partitions the original values into $b \leq k$ buckets (*e.g.*, with equal widths), which allows close values to be mapped to the same bucket. Afterwards, each user only samples $d \leq b$ buckets to minimize the number of bits to be randomized. Note that these two steps contributes to the information loss. Another limitation of dBitFlipPM is the possibility of detecting data changes [XYH+22] on the fly since the true value will fall in a different bucket, there will be a higher probability of changing the randomization of the d bits. Even if this only indicates that the user's value has changed, not what it was or is [EPK14, DKY17], there are still some privacy implications with respect to the type of inference an adversary can perform, especially if there are

correlation patterns to be exploited [TKB+17, NV20]. Finally, *d*BitFlipPM's privacy loss can still be proportional to the number of bucket changes, *i.e.*, the new domain size b in the worst case.

A different line of work has taken into account the infrequent data changes on the user side, hereafter referred to as **data change-based** [JRUW18, EFM+19, XYH+22, OWW22]. For instance, Joseph *et al.* [JRUW18] have proposed a new LDP protocol THRESH for monitoring statistics (*e.g.*, frequency) based on two sub-routines: voting and estimation, which requires splitting the privacy budget. The main idea of THRESH is to update through voting the global estimate only when it becomes sufficiently inaccurate. However, privacy budget splitting under LDP guarantees is suboptimal [WBLJ17, ACBX22, WXY+19, ACBX21, NXY+16, EFM+20, ACABX21], which negatively impacts the data utility. Moreover, the authors in [EFM+19, OWW22] proposed the sanitization and report of data changes for frequency monitoring by assuming a limited number of data changes and longitudinal Boolean data, though it can be extended to larger domains. This leads to an accuracy that decays linearly (or sub-linearly) in the number of data changes. Finally, in a recent work, Xue *et al.* [XYH+22] have proposed a new LDP protocol DDRM (Dynamic Difference Report Mechanism) based on difference trees. However, DDRM assumes that the user's private sequence exhibit continuity (*i.e.*, do not fluctuate significantly) and was mainly designed for longitudinal Boolean data. Besides, DDRM requires a privacy budget allocation scheme that depends on the number of data collections as well as to split the privacy budget when extending to a larger domain (*i.e.*, suboptimal).

Main contributions. In this chapter, we address the limitations of memoization-based protocols [EPK14, DKY17, ACBX22, EFM+20] without imposing any restriction on the number of data changes and/or on the number of data collections as in data change-based protocols [JRUW18, EFM+19, XYH+22, OWW22]. More precisely, we propose a novel LDP protocol with formal privacy guarantees for longitudinal frequency estimation of evolving counter (or categorical) data.

Our protocol, hereafter named LOngitudinal LOcal HAshing (LOLOHA), combines a domain reduction approach through local hashing [BS15, WBLJ17] with the memoization solution of RAPPOR using two rounds of sanitiza-

tion [EPK14, ACBX22]. The main strength of LOLOHA is that the longitudinal privacy-utility trade-off is linear only on the new (reduced) domain size g, in which $2 \leq g \ll k$ is a tunable hyperparameter. This way, the worst-case longitudinal privacy loss of LOLOHA has a significant k/g or b/g decrease factor in comparison with RAPPOR and dBitFlipPM, respectively.

Indeed, LOLOHA can be tuned for strong longitudinal privacy by selecting $g = 2$ (BiLOLOHA protocol). To maximize LOLOHA's utility, we also find the optimal g value (OLOLOHA protocol). Experimental evaluations demonstrate the effectiveness of LOLOHA with respect to the quality of frequency estimates, in addition to substantially minimizing the longitudinal privacy loss.

We also show why LDP is generally impossible to achieve when data is longitudinal, which motivates a definition of privacy that better suits the longitudinal scenario. This is in opposition with the common and mathematically equivalent path in the literature of claiming a protocol to be LDP but assuming that the evolving data is uncorrelated or constant in time, which we believe not be realistic in real-life deployments.

In summary, the main contributions of this chapter are three-fold:

- We propose the LOLOHA protocol for longitudinal frequency monitoring under LDP guarantees.
- We prove the longitudinal privacy and accuracy guarantees of LOLOHA through theoretical analysis and compare it to existing protocols.
- We show the performance of LOLOHA numerically and experimentally, using both real-world and synthetic datasets.

Outline. The remainder of this chapter is organized as follows. First, in Section 4.1, we provide the problem definition and review LDP and existing longitudinal LDP protocols. Next, we present and analyze our LOLOHA protocols in Section 4.2. In Section 4.3, we give a theoretical comparison of LOLOHA and state-of-the-art LDP protocols before presenting and interpreting the experimental results in Section 4.4. Finally, in Section 4.5, we review related work before concluding with future perspectives in Section 4.6.

4.1 Preliminaries

In this section, we present the problem considered, and we review the LDP privacy model and relevant protocols.

Notation. For denoting sets, we will use italic uppercase letters V, U, etc., and we write $[1..n] = \{1, \ldots, n\}$. For a vector $\mathbf{x}$ (bold lowercase letters), $\mathbf{x}_i$ represents the value of its i-th coordinate. Finally, we denote randomized protocols as $\mathcal{M}$.

4.1.1 Problem Statement

We consider the situation in which a server collects data from a distributed group of users while requiring the protection of privacy for each user, through LDP. The server collects sanitized data over time from each member of the group with respect to a fixed discrete random variable (*e.g.*, daily usage of a mobile application). Its objective is to estimate the true frequencies, or histograms, of the random variable as well as its evolution over time. We aim to provide the server with an optimized combination of two algorithms: one for the users, who must sanitize locally their data before sending it, and another for the server, which wants to aggregate data and perform the estimation accurately.

Formally, there are n users $U = \{u_1, \ldots, u_n\}$ and a random variable taking values in a set V of size k with true frequencies $\{f(v)\}_{v \in V}$, which may vary over time. Each user $u \in U$, holds a private sequence of values $\mathbf{v}^{(u)} = \left[v_1^{(u)}, v_2^{(u)}, \ldots, v_\tau^{(u)}\right]$, in which $v_t^{(u)}$ represents the discrete value $v \in V$ of user u at time step $t \in [1..\tau]$. At each time step t, upon collecting the sanitized values of all n users, the server will estimate a k-bins histogram $\{\hat{f}(v)\}_{v \in V}$ in a way that minimizes the *Mean Squared Error* (MSE) with respect to $\{f(v)\}_{v \in V}$. For all the algorithms presented hereafter, the estimation $\hat{f}(v)$ is unbiased (*i.e.*, $\mathbb{E}(\hat{f}(v)) = f(v)$). As a consequence, the MSE is equivalent to the variance as:

$$\text{MSE} = \frac{1}{|V|} \sum_{v \in V} \mathbb{E}\left[\left(\hat{f}(v) - f(v)\right)^2\right] = \frac{1}{|V|} \sum_{v \in V} \mathbb{V}[\hat{f}(v)].$$

4.1.2 Local Differential Privacy

Privacy model. In this chapter, we use LDP (Local Differential Privacy) [KLN+08, DJW13, DWJ13] as the privacy model considered, which is formally defined as follows.

Definition 26 (ϵ-Local Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ satisfies ϵ-local-differential-privacy (ϵ-LDP), where $\epsilon > 0$, if for any pair of input values $v_1, v_2 \in Domain(\mathcal{M})$ and any possible output x' of $\mathcal{M}$:

$$\Pr[\mathcal{M}(v_1) = x'] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(v_2) = x'].$$

In essence, LDP guarantees that it is unlikely for the data aggregator to reconstruct the input data. The privacy loss ϵ controls the privacy-utility trade-off for which lower values of ϵ result in tighter privacy protection. Similar to central DP, LDP also has several fundamental properties, such as robustness to post-processing and composition [DR+14].

Proposition 27 (Post-Processing [DR+14])**.** If $\mathcal{M}$ is ϵ-LDP, then $f(\mathcal{M})$ is also ϵ-LDP for any function f.

Proposition 28 (Sequential Composition [DR+14])**.** Let $\mathcal{M}_t$ be ϵ_t-LDP mechanism, for $t \in [\tau]$. Then, the sequence of outputs $[\mathcal{M}_1(v), \ldots, \mathcal{M}_\tau(v)]$ is $\sum_{t=1}^{\tau} \epsilon_t$-LDP. Moreover, if $\mathcal{M}$ is an ϵ-LDP mechanism and $\mathbf{v}$ is a finite sequence of k values, then the sequence of outputs $[\mathcal{M}(v_1), \ldots, \mathcal{M}(v_k)]$ is $k\epsilon$-LDP.

4.1.3 LDP Frequency Estimation Protocols

In this section, we review five state-of-the-art LDP frequency estimation protocols, which are often used as building blocks for more complex tasks (*e.g.*, heavy hitter estimation [BS15, BNST17], machine learning [MABK+20], and private frequency monitoring [DKY17, EPK14, ACBX22]).

Generalized Randomized Response (GRR)

The GRR [KBR16, KOV16] protocol generalizes the Randomized Response (RR) technique proposed by Warner [War65] for $k \geq 2$ while satisfying LDP.

Fix a parameter $\epsilon > 0$ and let $p \coloneqq \frac{e^\epsilon}{e^\epsilon + k - 1} \in (0, 1)$ in which $k = |V|$. For each $v \in V$, let $\eta_{\neq v} \in V$ be a uniform (*i.e.*, exogenous noise) random variable over $V \setminus \{v\}$. We let $\mathcal{M}_{\text{GRR}} : V \to V$ be the random variable given by:

$$\mathcal{M}_{\text{GRR}}(v; \epsilon) \coloneqq \begin{cases} v, & \text{w.p. } p \\ \eta_{\neq v}, & \text{w.p. } 1 - p. \end{cases}$$

This protocol satisfies ϵ-LDP, because $\frac{p}{q} = e^\epsilon$ [KBR16], in which $q \coloneqq {}^{(1-p)}/_{(k-1)}$ determines the probability of the response being any fixed noise value different of v. To estimate the normalized frequency of $v \in V$, one counts how many times v is reported, expressed as $C(v)$, and then computes:

$$\hat{f}(v) = \frac{C(v) - nq}{n(p - q)}, \tag{4.1}$$

in which n is the total number of users. In [WBLJ17], it was proven that Eq. (4.1) is an unbiased estimator (*i.e.*, $\mathbb{E}(\hat{f}(v)) = f(v)$).

Local Hashing (LH)

LH protocols [WBLJ17] can handle a large domain size k by first using hash functions to map an input value to a smaller domain of size $g \geq 2$ (typically $g \ll k$), and then applying GRR to the hashed value.

Fix $\epsilon > 0$ and let $\mathcal{M}_{\text{GRR}} : [1..g] \to [1..g]$ be the GRR mechanism with parameter ϵ and assuming the input-output domain to be $[1..g]$ instead of V, so that the size is g instead of k. In local hashing, each user selects at random a hashing function H from a family of universal hash functions, and reports the pair $\langle \text{H}, \mathcal{M}_{\text{GRR}}(x; \epsilon) \rangle$, in which $x = \text{H}(v)$.

The hash values will remain unchanged with probability $p = \frac{e^\epsilon}{e^\epsilon + g - 1}$ and switch to any different fixed value in $[1..g]$ with probability $q = \frac{1}{e^\epsilon + g - 1}$. This means that for each hash value $x \in [1..g]$, it holds that:

$$\Pr[\mathcal{M}_{\text{GRR}}(\text{H}(v); \epsilon) = x] = \begin{cases} p, & \text{if } x = \text{H}(v) \\ q, & \text{otherwise.} \end{cases}$$

Let $\langle \text{H}^u, x^u \rangle$ be the report from user $u \in U$. The server can obtain the

unbiased estimation of $v \in V$, with Eq. (4.1) by setting $q = \frac{1}{g}$ and $C(v) = |\{u \in U \mid \mathrm{H}^u(v) = x^u\}|$ [WBLJ17].

The authors in [WBLJ17] describe two LH protocols that differ on how g is selected: (1) Binary LH (BLH) that selects $g = 2$ and (2) Optimal LH (OLH) that selects $g = \lfloor e^\epsilon + 1 \rceil$ (rounded to the closest integer).

Unary Encoding (UE)

UE protocols interpret the user's input $v \in V$, as a one-hot k-dimensional vector. More precisely, $\mathbf{x} = \mathtt{UE}(v)$ is a binary vector with only the bit at the position corresponding to v set to 1 and the other bits set to 0. The perturbation function of UE protocols randomizes the bits from $\mathbf{x}$ independently with probabilities:

$$\forall i \in [k]: \quad \Pr[\mathbf{x}'_i = 1] = \begin{cases} p, & \text{if } \mathbf{x}_i = 1, \\ q, & \text{if } \mathbf{x}_i = 0. \end{cases} \tag{4.2}$$

Afterwards, the client sends $\mathbf{x}'$ to the server. The authors in [WBLJ17] describe two UE protocols that depend on the parameters p and q in Eq. (4.2): (1) Symmetric UE (SUE) [EPK14], which selects $p = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $q = \frac{1}{e^{\epsilon/2}+1}$ such that $p + q = 1$, and (2) Optimal UE (OUE), which selects $p = \frac{1}{2}$ and $q = \frac{1}{e^\epsilon+1}$.

The estimation method used in Eq. (4.1) applies equally to both UE protocols, in which $C(v)$ represents the number of times the bit corresponding to v has been reported. Last, both SUE and OUE protocols satisfy ϵ-LDP for $\epsilon = ln\left(\frac{p(1-q)}{(1-p)q}\right)$ [WBLJ17].

4.1.4 Existing Longitudinal LDP Frequency Estimation Protocols

For privately monitoring the frequency of values of a population, the simplest way is that each user adds independent fresh noise to v in each data collection $t \in [1..\tau]$ following one of the LDP protocols described in the previous section. However, this solution is vulnerable to "averaging attacks" in which an adversary can estimate the true value from observing multiple randomized versions

of it. To avoid this averaging attack, the memoization approach [EPK14] was designed to enable longitudinal collections through memorizing a randomized version of the true value v and consistently reusing it [DKY17, ACBX21] or reusing it as the input to a second round of sanitization (*i.e.*, chaining two LDP protocols) [EPK14, ACBX22, EFM+20]. The next four subsections describe state-of-the-art memoization-based protocols.

RAPPOR Protocol

The utility-oriented version of RAPPOR [EPK14] is based on the SUE protocol, which encodes the user's input $v \in V$ as a k-dimensional bit-vector and randomizes each bit independently. More specifically, for each value $v \in V$, the user encodes $\mathbf{x} = \texttt{UE}(v)$ and randomizes $\mathbf{x}$ as follows:

Step 1. Permanent RR (PRR): Memoize $\mathbf{x}'$ such that:

$$\forall i \in [k]: \quad \Pr[\mathbf{x}'_i = 1] = \begin{cases} p_1 = \frac{e^{\epsilon_\infty/2}}{e^{\epsilon_\infty/2}+1}, & \text{if } \mathbf{x}_i = 1, \\ q_1 = \frac{1}{e^{\epsilon_\infty/2}+1}, & \text{if } \mathbf{x}_i = 0, \end{cases}$$

in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP for $\epsilon_\infty = ln\left(\frac{p_1(1-q_1)}{(1-p_1)q_1}\right)$ [EPK14]. This step is carried out only once for each value $v \in V$ that the user has. Thus, the value $\mathbf{x}'$ shall be reused as the basis for all future reports of v.

Step 2. Instantaneous RR (IRR): Generate $\mathbf{x}''$ such that:

$$\forall i \in [k]: \quad \Pr[\mathbf{x}''_i = 1] = \begin{cases} p_2, & \text{if } \mathbf{x}'_i = 1, \\ q_2, & \text{if } \mathbf{x}'_i = 0. \end{cases}$$

This second step is carried out each time $t \in [1..\tau]$ a user report the value v. RAPPOR's deployment selected $p_2 = 0.75$ and $q_2 = 0.25$ [EPK14, WBLJ17] (*i.e.*, also symmetric). The RAPPOR protocol that chains two SUE protocols is referred to as L-SUE in [ACBX22, ACG+22]. We provide the calculation of parameters p_2 and q_2 in the repository [PH22]. Note that ϵ_∞ corresponds to an upper bound for each value v as $t \to \infty$. The privacy guarantees of the IRR step degrade according to the number of reports $t \in [1..\tau]$ [EPK14, EFM+20].

With two rounds of sanitization, each consisting of a transversal LDP protocol parametrized with p and q, the unbiased estimator in Eq. (4.1) is now extended to [ACBX22, EPK14]:

$$\hat{f}_L(v) = \frac{\frac{C(v)-nq_2}{(p_2-q_2)} - nq_1}{n(p_1-q_1)} = \frac{C(v) - nq_1(p_2-q_2) - nq_2}{n(p_1-q_1)(p_2-q_2)}, \tag{4.3}$$

in which p_1 and q_1 are the parameters of the LDP protocol used in the first step while p_2 and q_2 are the parameters of the LDP protocol used in the second step.

In [ACBX22], it was proven that Eq. (4.3) is an unbiased estimator (*i.e.*, $\mathbb{E}(\hat{f}_L(v)) = f(v)$) and that for any value $v \in V$, the variance $\mathbb{V}$ of the estimator $\hat{f}_L(v)$ in Eq. (4.3) is:

$$\begin{gathered} \mathbb{V}[\hat{f}_L(v)] = \frac{\gamma(1-\gamma)}{n(p_1-q_1)^2(p_2-q_2)^2}, \text{ where} \\ \gamma = f(v)\,(2p_1p_2 - 2p_1q_2 + 2q_2 - 1) + p_2q_1 + q_2(1-q_1). \end{gathered} \tag{4.4}$$

In this chapter, we will use the *approximate variance* $\mathbb{V}^*$, in which $f(v) = 0$ in Eq. (4.4), which gives:

$$\mathbb{V}^*\left[\hat{f}_L(v)\right] = \frac{(p_2q_1 - q_2\,(q_1-1))\,(-p_2q_1 + q_2\,(q_1-1) + 1)}{n(p_1-q_1)^2(p_2-q_2)^2}. \tag{4.5}$$

Therefore, one can obtain the RAPPOR approximate variance $\mathbb{V}^*[\hat{f}_{\text{RAPPOR}}(v)]$ by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (4.5).

Optimized Longitudinal UE Protocol

The authors in [ACBX22] analyzed all four combinations between OUE and SUE in both PRR and IRR steps. The optimized protocol named L-OSUE chains the OUE protocol (PRR step) and the SUE protocol (IRR step). Thus, for each value $v \in V$, the user encodes $\mathbf{x} = \texttt{UE}(v)$ and randomizes $\mathbf{x}$ as follows:

Step 1. PRR: Memoize $\mathbf{x}'$ such that:

$$\forall i \in [k]: \quad \Pr[\mathbf{x}'_i = 1] = \begin{cases} p_1 = \frac{1}{2}, & \text{if } \mathbf{x}_i = 1, \\ q_1 = \frac{1}{e^{\epsilon_\infty}+1}, & \text{if } \mathbf{x}_i = 0, \end{cases}$$

in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP as $e^{\epsilon_\infty} = \frac{p_1(1-q_1)}{q_1(q-p_1)}$ [EPK14, ACBX22]. The value $\mathbf{x}'$ shall be reused as the basis for all future reports when the real value is v.

Step 2. IRR: Generate $\mathbf{x}''$ such that:

$$\forall i \in [k]: \quad \Pr[\mathbf{x}''_i = 1] = \begin{cases} p_2, & \text{if } \mathbf{x}'_i = 1, \\ q_2 = 1 - p_2, & \text{if } \mathbf{x}'_i = 0, \end{cases}$$

in which $p_2 = \frac{e^{\epsilon_\infty} e^{\epsilon_1} - 1}{e^{\epsilon_\infty} - e^{\epsilon_1} + e^{\epsilon_\infty + \epsilon_1} - 1}$ and $\mathbf{x}''$ is the report to be sent to the server. Let $p_s = \Pr[\mathbf{x}''_i = 1 | \mathbf{x}_i = 1] = p_1 p_2 + (1 - p_1) q_2$ and $q_s = \Pr[\mathbf{x}''_i = 1 | \mathbf{x}_i = 0] = q_1 p_2 + (1 - q_1) q_2$. For the first report, the L-OSUE protocol satisfies ϵ_1-LDP as $e^{\epsilon_1} = \frac{p_s(1-q_s)}{(1-p_s)q_s}$ [ACBX22, EPK14].

Similar to RAPPOR, the estimated frequency $\hat{f}_{\text{L-OSUE}}(v)$ that a value $v \in V$ occurs, can be computed using Eq. (4.3). One can also obtain the L-OSUE approximate variance $\mathbb{V}^*[\hat{f}_{\text{L-OSUE}}(v)]$ by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (4.5).

Longitudinal GRR (L-GRR)

The L-GRR [ACBX22] protocol chains GRR in both PRR and IRR steps. Therefore, for each value $v \in V$, the user randomizes v as follows:

Step 1. PRR: Memoize x' such that:

$$x' = \begin{cases} v, & \text{w.p. } p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+k-1}, \\ \tilde{v} \in V \setminus \{v\}, & \text{w.p. } q_1 = \frac{1-p_1}{k-1}, \end{cases}$$

in which p_1 and q_1 control the level of longitudinal ϵ_∞-LDP as $e^{\epsilon_\infty} = \frac{p_1}{q_1}$ [KBR16, ACBX22]. The value x' shall be reused as the basis for all future reports on the real value v.

Step 2. IRR: Generate a report x'' such that:

$$x'' = \begin{cases} x', & \text{w.p. } p_2, \\ \tilde{x} \in V \setminus \{x'\}, & \text{w.p. } q_2 = \frac{1-p_2}{k-1}, \end{cases}$$

in which $p_2 = \frac{e^{\epsilon_\infty+\epsilon_1}-1}{-ke^{\epsilon_1}+(k-1)e^{\epsilon_\infty}+e^{\epsilon_1}+e^{\epsilon_1+\epsilon_\infty}-1}$ and x'' is the report to be sent to the server. Let $p_s = \Pr[x'' = v|v] = p_1p_2 + q_1q_2$ and $q_s = \Pr[x'' = v|\tilde{v} \in V \setminus \{v\}] = p_1q_2 + q_1p_2$. For the first report, the L-GRR protocol satisfies ϵ_1-LDP since $e^{\epsilon_1} = \frac{p_s}{q_s}$ [ACBX22].

The estimated frequency $\hat{f}_{\text{L-GRR}}(v)$ that a value v occurs can also be obtained using Eq. (4.3). Besides, one can compute the L-GRR approximate variance $\mathbb{V}^*[\hat{f}_{\text{L-GRR}}(v)]$ by replacing the resulting p_1, q_1, p_2, q_2 parameters into Eq. (4.5).

dBitFlipPM Protocol

The *d*BitFlipPM [DKY17] protocol was proposed to improve the memoization solution of RAPPOR [EPK14] by mapping several true values to the same noisy response at the cost of losing information due to generalization. This is done by first partitioning the original domain V into b buckets (*i.e.*, new domain size $2 \leq b \leq k$) using a function $\texttt{bucket} : V \to [1..b]$, such that close values will fall into the same bucket. Next, each user randomly draws d bucket numbers without replacement from $[1..b]$, denoted by $j_1, j_2, \ldots, j_d$, and fixes them for all future data collections. Then, for each $v \in V$, the user sends a sanitized vector $\mathbf{x}' = [(j_1, x_{j_1}), \ldots, (j_d, x_{j_d})]$ parameterized with the privacy guarantee ϵ_∞ as follows:

$$\forall l \in [1..d] : \Pr[x_{j_l} = 1] = \begin{cases} p = \frac{e^{\epsilon_\infty/2}}{e^{\epsilon_\infty/2}+1}, & \text{if } \texttt{bucket}(v) = j_l \\ q = \frac{1}{e^{\epsilon_\infty/2}+1}, & \text{if } \texttt{bucket}(v) \neq j_l \end{cases}.$$

In other words, users inform the server which bits are sampled as well as their perturbed values, but the server does not receive any information about the remaining $b - d$ bits. The server can estimate the number of times each bucket in $[1..b]$ has been reported with Eq. (4.1) by replacing n with $\frac{nd}{b}$ as each user only sampled d bits among b buckets.

In contrast to RAPPOR, there is no second round of sanitization, which means the user runs dBitFlipPM with ϵ_∞-LDP for all b buckets, with randomization applied to the d fixed bits $j_1, j_2, \ldots, j_d$ and memoizes the response. This approach adds uncertainty to the real value because multiple (close) values will be mapped to the same bucket. The highest protection is given when $d = 1$ [DKY17], which will minimize the chances (to some extent) of detecting high data changes.

4.2 LOLOHA

In this section, we introduce our LOLOHA (Longitudinal Local Hashing) protocol for frequency monitoring throughout time under LDP constraints, and we analyze its utility and privacy.

The privacy analysis of longitudinal protocols requires special treatment because, since they are stateful, they cannot be modeled as mechanisms mapping values into values, but rather sequences into sequences. This makes the LDP constraint too strong in the long term as shown in the following theorem.

Theorem 29. *(LDP cannot be satisfied when $\tau \to \infty$) Consider a randomized longitudinal mechanism $\vec{\mathcal{M}} : [1..n]^\tau \to [1..m]^\tau$ mapping an input sequence $X_1, \ldots, X_\tau$ to an output sequence $Y_1, \ldots, Y_\tau$, for some positive integer τ. For the sake of utility of each reported value Y_t (0-LDP means total detriment of utility), assume some negligible, but positive fixed $\alpha > 0$ such that the mechanism for generating Y_t from X_t and the history $X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}$ is not α-LDP. If $\tau \geq \epsilon/\alpha$ then $\vec{\mathcal{M}}$ is not ϵ-LDP.*

Proof. Let $y_1 = \arg\max_y \frac{\max_x P(X_1=x|Y_1=y)}{\min_x P(X_1=x|Y_1=y)}$, and call x_1^+ and x_1^- to the values that respectively maximize and minimize $P(X_1 = x|Y_1 = y_1)$. By the minimal utility assumption, $\frac{p(x_1^+|y_1)}{p(x_1^-|y_1)} > e^\alpha$.

Let $y_2 = \arg\max_y \frac{\max_x P(X_2=x|Y_2=y,Y_1=y_1,X_1=x_1)}{\min_x P(X_2=x|Y_2=y,Y_1=y_1,X_1=x_1)}$, and call x_2^+ and x_2^- to the values that respectively maximize and minimize $P(X_2 = x|Y_2 = y, Y_1 = y_1, X_1 = x_1)$. Since the output values of the mechanism are reported one by one in temporal order, we have $p(x_1, x_2|y_1, y_2) = p(x_1|y_1)p(x_2|y_2, y_1, x_1)$, hence by the minimal utility assumption and the first step, $\frac{p(x_1^+,x_2^+|y_1,y_2)}{p(x_1^-,x_2^-|y_1,y_2)} =$

$\frac{p(x_1^+|y_1)p(x_2^+|y_2,y_1,x_1^+)}{p(x_1^-|y_1)p(x_2^-|y_2,y_1,x_1^-)} > e^{2\alpha}$.

Repeating this process inductively yields three sequences y_i, x_i^+ and x_i^- of length τ such that $\frac{p(x_1^+,\ldots,x_\tau^+|y_1,\ldots,y_\tau)}{p(x_1^-,\ldots,x_\tau^-|y_1,\ldots,y_\tau)} > e^{\tau\alpha}$.

This makes it impossible for the mechanism to be ϵ-LDP for any $\tau \geq \epsilon/\alpha$. □

For instance, assume that a user has a secret sequence $\mathbf{v} = [1,1,1,3,1,2,1,1,3]$ ($\tau = 9$ time steps), and reports $\vec{\mathcal{M}}(\mathbf{v}) := [\mathcal{M}(v_1),\ldots,\mathcal{M}(v_9)]$, in which $\mathcal{M}$ is the memoization mechanism ($1 \mapsto 2; 2 \mapsto 2; 3 \mapsto 3$) that reuses the sanitized report. The server receives $[2,2,2,3,2,2,2,2,3]$, hence some time-related patterns in the sequence are exposed, but the memoization protects the uncertainty about the user actual values. As the sequence size grows, the vectorized memoization mechanism $\vec{\mathcal{M}}$ that processes temporal data continues to protect the values indefinitely, but fails to satisfy LDP. For this reason, we introduce the following relaxed definition of privacy for longitudinal mechanisms.

Definition 30 (Longitudinal LDP)**.** For a longitudinal memoizing mechanism $\mathcal{M} : A^\tau \to B^\tau$, in which $A = [1..k]$, let $\mathcal{M}^\star$ denote a mechanism that takes as input a permutation x of A and outputs $\mathcal{M}^\star(x) := x''$ by shuffling the k entries of x, yielding x', and letting $x_i'' := \mathcal{M}(x_i')$ for each $i = 1..k$, sequentially. $\mathcal{M}$ is said to be ϵ-LDP on the users' values iff $\mathcal{M}^\star$ is ϵ-LDP.

Definition 30 discards all information contained in time correlation by shuffling the input and aggregates the total privacy loss after all input values have been memoized. Moreover, Definition 30 corresponds to the total privacy budget that will be consumed for sanitizing all the values of the user.

Previous influential works, such as RAPPOR [EPK14] and dBitFlipPM [DKY17], handle the negative consequences of Theorem 29 implicitly by assuming that the data values (or buckets) never change or change in an uncorrelated manner. We consider the former to be unrealistic, and the latter is insufficient to guarantee LDP, though it makes users indistinguishable. In this chapter, we privileged Definition 30 over extreme assumptions on the data to be able to explain at least what is actually being protected by the mechanism when the assumptions do not hold. Hence, we present long term guarantees in terms of LDP on the users' values, but also,

single-report LDP guarantees, as done in the literature, which are equivalent to LDP assuming constant values.

4.2.1 Overview of LOLOHA

LOLOHA is inspired by the strengths of RAPPOR [EPK14] (double sanitization to minimize data change detection) and dBitFlipPM [DKY17] (several values are mapped to the same randomized value) protocols. More precisely, LOLOHA is based on LH for the PRR step to satisfy ϵ_∞-LDP (upper bound), which significantly reduces the domain size. Thus, the user will uniformly choose at random a universal hash function H that maps the original domain $V \to [1..g]$, with $g \geq 2$ typically much smaller than $k = |V|$. Indeed, given a general (universal) family of hash functions $\mathscr{H}$, each input value $v \in V$ is hashed into a value in $[1..g]$ by hash function $\mathrm{H} \in \mathscr{H}$, and the universal property requires:

$$\forall v_1, v_2 \in V, v_1 \neq v_2 : \quad \Pr_{\mathrm{H} \in \mathscr{H}} [\mathrm{H}(v_1) = \mathrm{H}(v_2)] \leq \frac{1}{g}.$$

In other words, approximately k/g values $v \in V$ can be mapped to the same hashed value $\mathrm{H}(v)$ in $[1..g]$ due to collision. After the hashing step, to satisfy ϵ_∞-LDP, the user invokes the GRR protocol to the hashed value $x = \mathrm{H}(v)$ and memoizes the response $x' = \mathcal{M}_{\mathrm{GRR}}(x; \epsilon_\infty)$. Then, the value x' will be reused as the basis for all future reports on the **hashed value** x, which supports all values in set $X_H = \{v \in V \mid \mathrm{H}(v) = x\}$. The intuition is that the user only leaks ϵ_∞ for each hashed value $x \in [1..g]$ as they support all values $v \in V$ that collide to $x = \mathrm{H}(v)$. Notice that instead of memoization, users could also pre-compute the mapping for each input value. These two methods would be equivalent in terms of the functionality provided. An implementation of LOLOHA is available in the `multi-freq-ldpy` Python package [ACG+22].

Moreover, in contrast with the dBitFlipPM protocol in which only close values are mapped to the same bucket, any two values in V can collide with probability at most $1/g$. Therefore, even if the user's value changes periodically, correlated or in an abrupt manner, there will still be uncertainty on the actual value v. However, with only this PRR step, it would be possible

to detect some of the data changes due to the randomization of a different hash value. Therefore, LOLOHA also requires the user to apply a second round of sanitization (*i.e.*, IRR step) to the memoized values x' with the GRR protocol such that the first report satisfies ϵ_1-LDP, for some chosen positive $\epsilon_1 < \epsilon_\infty$.

4.2.2 Client-Side of LOLOHA

Algorithm 8 displays the pseudocode of LOLOHA on the client-side, which receives as input: the true sequence of values $\mathbf{v} = [v_1, v_2, \ldots, v_\tau]$ of the user that is running the code, a universal family $\mathscr{H}$ of hash functions $\mathrm{H} : V \to [1..g]$, and the constants $\epsilon_1, \epsilon_\infty$, with $0 < \epsilon_1 < \epsilon_\infty$, that represent respectively the leakage of the first report and the maximal longitudinal leakage.

Algorithm 8 Client-Side of LOLOHA.

Input: User longitudinal values $[v_1, v_2, \ldots, v_\tau]$, family $\mathscr{H}$ of hash functions and constants $0 < \epsilon_1 < \epsilon_\infty$.

Output: None. Sends data to server during execution.

$\mathrm{H} \leftarrow_R \mathscr{H}$ ▷ Hash function chosen at random
Send H.
$\epsilon_{\mathrm{IRR}} \leftarrow \ln\left(\frac{e^{\epsilon_\infty+\epsilon_1}-1}{e^{\epsilon_\infty}-e^{\epsilon_1}}\right)$
for each time $t \in [1..\tau]$ **do**:
  $x \leftarrow \mathrm{H}(v_t)$. ▷ Hash step
  if x *is not* memoized **then**:
    $x' \leftarrow \mathcal{M}_{\mathrm{GRR}}(x; \epsilon_\infty)$ over $[1..g]$. ▷ PRR step
    Memoize output x' for x.
  else:
    Get memoized output x' for x.
  end if
  $x''_t \leftarrow \mathcal{M}_{\mathrm{GRR}}(x'; \epsilon_{\mathrm{IRR}})$ over $[1..g]$. ▷ IRR step
  Send x''_t. ▷ Sanitized data
end for

Privacy analysis. The privacy guarantees of Algorithm 8 are detailed in Theorems 31, 32 and especially 33.

Theorem 31. *(Single report LDP of memoization)*

Let $\mathcal{M} : V \to \mathscr{H} \times [1..g]$ denote the process of applying the hash and PRR steps of LOLOHA to a single element $v \in V$, producing $\mathcal{M}(v) = (\mathrm{H}, x')$. Then $\mathcal{M}$ is ϵ_∞-LDP.

Proof. The parameters for the PRR step are $p = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+g-1}$ and $q = \frac{1}{e^{\epsilon_\infty}+g-1}$. For any two possible input values $v_1, v_2 \in V$ and any reported output (H, x'), we have

$$\frac{\Pr\left[(\mathrm{H}, x')|v_1\right]}{\Pr\left[(\mathrm{H}, x')|v_2\right]} \leq \frac{p}{q} = \frac{\frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+g-1}}{\frac{1}{e^{\epsilon_\infty}+g-1}} = e^{\epsilon_\infty}.$$

□

Theorem 32. *(Single report LDP of LOLOHA)*
Let $\mathcal{M} : V \to \mathscr{H} \times [1..g]$ denote the process of applying the hash, PRR, and IRR steps of LOLOHA to a single element $v \in V$, producing $\mathcal{M}(v) = (\mathrm{H}, x'')$. Then $\mathcal{M}$ is ϵ_1-LDP.

Proof. Let (p_1, q_1) denote the parameters for the PRR step and (p_2, q_2), the parameters for the IRR step. That is, $p_1 = \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+g-1}$, $q_1 = \frac{1}{e^{\epsilon_\infty}+g-1}$, $p_2 = \frac{e^{\epsilon_{\mathrm{IRR}}}}{e^{\epsilon_{\mathrm{IRR}}}+g-1}$, and $q_2 = \frac{1}{e^{\epsilon_{\mathrm{IRR}}}+g-1}$. If $x'' \neq \mathrm{H}(v)$, it must have changed during either the PRR or the IRR step, and if $x'' = \mathrm{H}(v)$, either it was not changed during either step or it was changed during both. From this analysis, it can be concluded that for each $y \in [1..g]$, we have

$$\Pr[x'' = y] = \begin{cases} p_1 p_2 + q_1 q_2, & \text{if } y = \mathrm{H}(v), \\ p_1 q_2 + q_1 p_2, & \text{if } y \neq \mathrm{H}(v). \end{cases}$$

Therefore, for any two possible input values $v_1, v_2 \in V$ and any output (H, x''), we have,

$$\frac{\Pr\left[(\mathrm{H}, x'')|v_1\right]}{\Pr\left[(\mathrm{H}, x'')|v_2\right]} \leq \frac{p_1 p_2 + q_1 q_2}{p_1 q_2 + q_1 p_2} = \frac{e^{\epsilon_\infty} \cdot e^{\epsilon_{\mathrm{IRR}}} + 1 \cdot 1}{e^{\epsilon_\infty} \cdot 1 + 1 \cdot e^{\epsilon_{\mathrm{IRR}}}}.$$

Moreover, since $e^{\epsilon_{\mathrm{IRR}}} = \frac{e^{\epsilon_\infty+\epsilon_1}-1}{e^{\epsilon_\infty}-e^{\epsilon_1}}$, then $e^{\epsilon_{\mathrm{IRR}}} e^{\epsilon_\infty} + 1 = e^{\epsilon_1}(e^{\epsilon_{\mathrm{IRR}}} + e^{\epsilon_\infty})$. Hence,

$$\frac{\Pr\left[(\mathrm{H}, x'')|v_1\right]}{\Pr\left[(\mathrm{H}, x'')|v_2\right]} \leq e^{\epsilon_1}.$$

□

Theorem 33. *(Privacy protection as $\tau \to \infty$)*
The client-side of LOLOHA is $g\epsilon_\infty$-LDP on the users' values.

Proof. The non-vectorized memoization mechanism (hash and PRR steps) of LOLOHA is a function $\mathcal{M} : V \to [1..g]$ that can memorize at most g reports. For each separate individual report, we know that $\mathcal{M}$ satisfies ϵ_∞-LDP (Theorem 31). Therefore, by sequential composition of at most g results (Proposition 28), $\mathcal{M}$ satisfies $g\epsilon_\infty$-LDP, and LOLOHA satisfies $g\epsilon_\infty$-LDP on the users' values. □

The privacy guarantees of the IRR step (Theorem 32) degrade according to the number of reports $t \in [1..\tau]$ [EPK14, EFM+20]. If we let ϵ_t be the privacy guarantee on the users' values of Algorithm 8 for a fixed user using the data in times $[1..t]$, so that $t = 1$ matches exactly ϵ_1 (Theorem 32), then we have $\epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_\tau \leq g\epsilon_\infty$.

Besides, from Algorithm 8, one can remark that instead of leaking a new ϵ_∞ for each $v \in V$, LOLOHA will only leak ϵ_∞ for each hashed value $x \in [1..g]$. Therefore, unlike RAPPOR that has a worst-case guarantee of $k\epsilon_\infty$-LDP on the users' values, the overall privacy guarantee of our LOLOHA solution will grow proportionally to the new domain size $2 \leq g \ll k$, with worst-case longitudinal privacy of $g\epsilon_\infty$-LDP on the users' values.

4.2.3 Server-Side of LOLOHA

The server-side algorithm of LOLOHA is described in Algorithm 9, which takes the reported values by n users and aggregates them to estimate the frequencies of each $v \in V$ at each point in time.

For large n, the estimations of Algorithm 9 are guaranteed to be close to the true population parameters with high probability as explained in Proposition 34. Moreover, one can also compute the LOLOHA approximate variance $\mathbb{V}^*[\hat{f}_{\text{LOLOHA}}(v)]$ by replacing the server parameters in Algorithm 9 into Eq. (4.5).

Proposition 34. (Asymptotic utility guarantee of LOLOHA)
Fix any arbitrary $t \in [1..\tau]$. For each $v \in V$, let $f(v)$ be the true population probability of producing the value v at time t, and let $\hat{f}(v) \in [0, 1]$ be the

Algorithm 9 Server-Side of LOLOHA.

Input: Constants $0 < \epsilon_1 < \epsilon_\infty$, and for each user $u \in U$, a hash function $H_u : V \to [1..g]$ and a sequence of hash values $[x_1^{\prime\prime(u)}, \ldots, x_\tau^{\prime\prime(u)}]$.

Output: Matrix with estimations $\hat{f}_{\text{LOLOHA}}(v)_t$ for each $v \in V$ at each $t \in [1..\tau]$.

Compute parameters:

$\epsilon_{\text{IRR}} \leftarrow \ln\left(\frac{e^{\epsilon_\infty+\epsilon_1}-1}{e^{\epsilon_\infty}-e^{\epsilon_1}}\right)$; $n \leftarrow |U|$

$p_1 \leftarrow \frac{e^{\epsilon_\infty}}{e^{\epsilon_\infty}+g-1}$; $q_1' \leftarrow \frac{1}{g}$

$p_2 \leftarrow \frac{e^{\epsilon_{\text{IRR}}}}{e^{\epsilon_{\text{IRR}}}+g-1}$; $q_2 \leftarrow \frac{1}{e^{\epsilon_{\text{IRR}}}+g-1}$

for each time $t \in [1..\tau]$ **do**:

  for each $v \in V$ **do**:

    $C(v) \leftarrow |\{u \in U \mid \text{H}_u(v) = x_t^{\prime\prime(u)}\}|$

    $\hat{f}_L(v)_t \leftarrow \frac{C(v)-nq_1'(p_2-q_2)-nq_2}{n(p_1-q_1')(p_2-q_2)}$ ▷ Eq. (4.3) with q_1'.

  end for

end for

return matrix $[\hat{f}_L(v)_t]_{t,v}$

estimation produced by Algorithm 9 for time t. For any $\beta \in (0,1)$, it holds with probability at least $1-\beta$ that:

$$\max_{v \in V} |\hat{f}(v) - f(v)| < \sqrt{\frac{k}{4n\beta(p_1-q_1')(p_2-q_2)}}.$$

Proof. Fix $v \in V$, and let Δ be the random variable given by $\Delta := \hat{f}(v) - f(v) \in [-1,1]$ be a random variable. Since $\hat{f}(v)$ is unbiased, we have $\mathbb{E}[\Delta] = 0$ and $\mathbb{V}[\Delta] = \mathbb{V}[\hat{f}(v)]$. We remark that for any $\delta, \beta \in (0,1)$, among all random variables Δ' defined in $[-1,1]$ such that $\mathbb{E}[\Delta']$ and $\Pr[|\Delta'| \geq \delta] = \beta$, the one with minimal variance is the random variable Δ^* that concentrates a mass of $1-\beta$ at $\Delta' = 0$ and two masses of $\beta/2$ at $-\delta$ and δ. This random variable has variance $\mathbb{V}[\Delta^*] = \beta\delta^2$. Hence, for arbitrary $\delta \in (0,1)$ and letting in particular $\beta := \Pr[|\hat{f}(v) - f(v)| \geq \delta]$, we conclude that $\mathbb{V}[\hat{f}(v)] = \mathbb{V}[|\hat{f}(v) - f(v)|] \geq \mathbb{V}[\Delta^*] = \Pr[|\hat{f}(v) - f(v)| \geq \delta] \cdot \delta^2$. In other words,

$$\Pr[|\hat{f}(v) - f(v)| \geq \delta] \leq \mathbb{V}[\hat{f}(v)]/\delta^2.$$

Now, considering all $v \in V$ simultaneously, we obtain $\Pr[\max_{v\in V} |\hat{f}(v) - f(v)| \geq \delta] \leq \sum_{v\in V} \Pr[|\hat{f}(v) - f(v)| \geq \delta] = (1/\delta^2) \sum_{v\in V} \mathbb{V}[\hat{f}(v)]$. By rewriting this equation in terms of confidence, we conclude that with probability at least $1 - \beta$,

$$\max_{v\in V} |\hat{f}(v) - f(v)| < \sqrt{\textstyle\sum_{v\in V} \mathbb{V}[\hat{f}(v)]/\beta}.$$

Lastly, from Eq. (4.4) it can be concluded that $\mathbb{V}[\hat{f}(v)] \leq 1/4n(p_1-q_1')(p_2-q_2)$ because the product $\gamma(1-\gamma)$ is maximal at $\gamma = 1/2$. As a consequence, $\max_{v\in V} |\hat{f}(v) - f(v)| < \sqrt{k/4n\beta(p_1-q_1')(p_2-q_2)}$. □

4.2.4 Selecting and Optimizing Parameter g

Binary LOLOHA (BiLOLOHA). Following Theorem 33, the strongest longitudinal privacy protection of LOLOHA is when $g = 2$.

Optimal LOLOHA (OLOLOHA). To maximize the utility of LOLOHA, we find the optimal g value by taking the partial derivative of $\mathbb{V}^*[\hat{f}_{\text{LOLOHA}}(v)]$ with respect to g. Let $\epsilon_1 = \alpha\epsilon_\infty$, for $\alpha \in (0, 1)$. This partial derivative is a function in terms of ϵ_∞ and α, or alternatively, in terms of $a = e^{\epsilon_\infty}$ and $b = e^{\alpha\epsilon_\infty}$, and it is minimized when g equals (*cf.* development in repository [PH22]):

$$g = 1+\max\left(1, \left\lfloor \frac{1-a^2+\sqrt{a^4-14a^2+12ab(1-ab)+12a^3b+1}}{6(a-b)} \right\rceil\right), \tag{4.6}$$

in which $\lfloor . \rceil$ means rounding to the closest integer. Fig. 4.1 illustrates the optimal g selection with Eq. (4.6) by varying the longitudinal privacy guarantee $\epsilon_\infty = [0.5, 1, \ldots, 4.5, 5]$ and $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$. From Fig. 4.1, one can remark that in high privacy regimes (*i.e.*, low ϵ values), the optimal g is binary (*i.e.*, our BiLOLOHA protocol with $g = 2$). As ϵ_∞ or/and $\epsilon_1 = \alpha\epsilon_\infty$ get(s) higher (low privacy regimes), the optimal g is non-binary, which can maximize utility with a cost in the overall longitudinal privacy $g\epsilon_\infty$-LDP on the users' values, for $g > 2$.

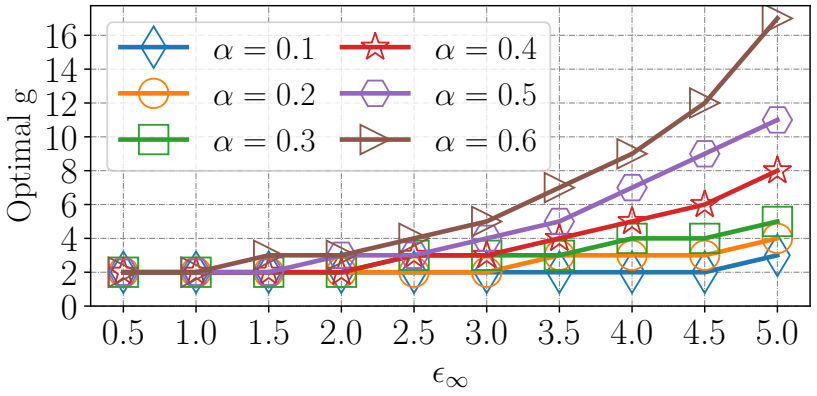


Figure 4.1: Optimal g selection for our OLOLOHA protocol by varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$.

4.3 Theoretical Comparison

In this section, we compare LOLOHA with the state-of-the-art protocols described in the previous Section 4.1.4 from a theoretical point of view. Table 4.1 shows a summary of the main characteristics of these protocols, excluding utility.

For the theoretical utility, numerical analysis is preferred over an analytical one because the formulas of variance and approximate variance are excessively complex. For L-OSUE and dBitFlipPM, the approximate variances are $\frac{4e^{\epsilon_1}}{n\left(e^{2\epsilon_1}-2e^{\epsilon_1}+1\right)}$ and $\frac{b}{2dn\sinh\left(\frac{\epsilon_\infty}{2}\right)}$ respectively, but for the other protocols, the formulas are provided only in the repository [PH22] since they are excessively verbose for this document.

In order to evaluate numerically the approximate variance $\mathbb{V}^*$ of LOLOHA in comparison with state-of-the-art ones [EPK14, ACBX22], for each protocol, we set the longitudinal privacy guarantee ϵ_∞ (upper bound) and the first report privacy guarantee $\epsilon_1 = \alpha\epsilon_\infty$ (lower bound), for $\alpha \in (0, 1)$. This allows to obtain parameters p_1, q_1, p_2, q_2 for each protocol, which are then used to compute their approximate variance with Eq. (4.5).

Fig. 4.2 illustrates the numerical values of the approximate variance for our LOLOHA protocols, RAPPOR [EPK14], and L-OSUE [ACBX22] with

Protocol	**Comm.** bits per user per time step	**Server** run-time complexity	**Privacy loss** budget consumption
LOLOHA	$\lceil \log_2 g \rceil$	$n\,k$	$g\,\epsilon_\infty$
L-GRR [ACBX22]	$\lceil \log_2 k \rceil$	n	$k\,\epsilon_\infty$
RAPPOR [EPK14]	k	$n\,k$	$k\,\epsilon_\infty$
L-OSUE [ACBX22]	k	$n\,k$	$k\,\epsilon_\infty$
dBitFlipPM [DKY17]	d	$n\,b$	$\min(d+1, b)\,\epsilon_\infty$

Table 4.1: Theoretical comparison of the protocols.

$n = 10000$, $\epsilon_\infty = [0.5, 1, \ldots, 4.5, 5]$, and $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$. From Fig. 4.2, one can remark that all protocols have similar variance values when $\alpha \leq 0.3$ with only a small difference when ϵ_∞ is high. However, in low privacy regimes, *i.e.*, when ϵ_∞ and α are high, BiLOLOHA is the least performing protocol in terms of utility, accompanied by RAPPOR. Indeed, our OLOLOHA protocol has a very similar utility as the optimized L-OSUE [ACBX22] protocol, which indicates a clear connection also found between their one-round versions [WBLJ17], *i.e.*, OLH and OUE.

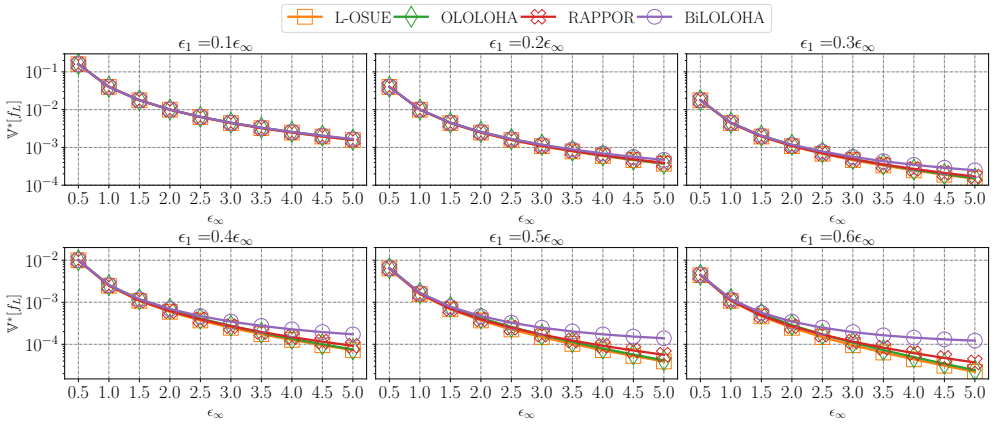


Figure 4.2: Numerical values of the approximate variance $\mathbb{V}^*[\hat{f}_L(v)]$ in Eq. (4.5) of our LOLOHA protocols, RAPPOR [EPK14], and L-OSUE [ACBX22] varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.1, 0.2, \ldots, 0.6\}$.

Though not included in our analysis, the L-GRR protocol from [ACBX22] has

shown to be very sensitive to k (a parameter on which its variance depends on), leading to extremely high values that would obfuscate the curves of the other protocols in Fig. 4.2. However, L-GRR is ideal when k is small, which is the case for instance for binary attributes. Besides, we also did not numerically compare our protocols with dBitFlipPM as it only has a single round of sanitization. A proper comparison with dBitFlipPM would be only considering the PRR step of our LOLOHA protocols. Therefore, by comparing the approximate variances of double randomization protocols, we can conclude that our LOLOHA protocols preserve as much utility as state-of-the-art protocols [EPK14, ACBX22].

Moreover, from Table 4.1, LOLOHA has less communication cost than L-UE and similar server time computation, which is advantageous for large-scale system deployment to monitor frequency longitudinally. In addition, one clear limitation of RAPPOR, L-OSUE, and L-GRR is that they do not support even small data changes of the user's actual data [DKY17], which requires to invoke the whole algorithm again on the new value. Therefore, following Definition 30 and Proposition 28, the overall privacy guarantee of RAPPOR, L-OSUE, and L-GRR, for all user's true value $v \in V$ (assuming the user's value will change periodically) will grow proportionally to the number of *data changes*, with worst-case longitudinal privacy of $k\epsilon_\infty$-LDP on the users' values.

On the other hand, with dBitFlipPM, its overall privacy guarantee for all user's true value $v \in V$ (assuming the user's value will change periodically) will grow proportionally to the number *of bits* d or the number of *bucket changes*, with worst-case longitudinal privacy of $\min(d+1, b)\epsilon_\infty$-LDP on the users' values (*cf.* Definition 30 and Proposition 28). However, there is a loss of information due to both the generalization of the original domain size k to b buckets and due to sampling only d bits. Besides, the dBitFlipPM protocol is vulnerable to detecting high data changes (*i.e.*, change of real bucket) as there is no second round of sanitization (*i.e.*, IRR step) [XYH+22]. This data change detection problem is (to some extent) minimized when d is small.

4.4 Experimental Evaluation

In this section, we present the setup of our experiments and the experimental results of our LOLOHA protocols in comparison with the state-of-the-art.

4.4.1 Setup of Experiments

The main goal of our experiments is to study the effectiveness of our proposed LOLOHA protocols on longitudinal frequency estimates through multiple $\tau > 1$ data collections. In particular, we aim to show that our LOLOHA protocols (i) maintain competitive utility to state-of-the-art memoization-based LDP protocols [EPK14, DKY17, ACBX22] while (ii) minimize longitudinal privacy loss. With these objectives in mind, we run experiments using both synthetic and real-world datasets.

Environment. All algorithms are implemented in Python 3 with Numpy and Numba libraries. The codes we develop for all experiments are available in the repository [PH22]. Since LDP algorithms are randomized, we report average results over 20 runs.

Datasets. We use the following real and synthetic datasets.

- **Syn.** To simulate the deployment of [DKY17] to collect data every 6 hours, we generate a synthetic dataset with $k = 360$ (*i.e.*, the number of minutes in 6 hours), $n = 10000$ users, and $\tau = 120$ data collections (*i.e.*, 4x over 30 days). For each user, the value at the first timestamp follows a Uniform distribution. For each subsequent time, a change can occur with probability $p_{ch} = 0.25$, with value following a Uniform distribution too.
- **Adult.** This is a classical dataset from the UCI machine learning repository [DG17b] with $n = 45222$ samples after cleaning. We only selected the "hours-per-week" attribute with $k = 96$. To simulate multiple data collections, we randomly permuted the data $\tau = 260$ times (*i.e.*, 52 weeks over 5-years). Note that the real frequency remains the same, but each user has a random private sequence.
- **DB_MT.** This dataset is produced by the `folktables` Python package [DHMS21] that provides access to datasets derived from the US

Census. We selected the survey year 2018 and the "Montana" state, which results in $n = 10336$ samples. To simulate $\tau = 80$ counter data collections, we selected all person record-replicate weights attributes[1], *i.e.*, PWGTP1, ..., PWGTP80. The total number of unique values among all columns is $k = 1412$.

- **DB_DE.** Similar to DB_MT, we selected the "Delaware" state, which results in $n = 9123$, $\tau = 80$, and $k = 1234$.

Methods evaluated. We consider for evaluation the following longitudinal LDP protocols:

- *RAPPOR*. The utility-oriented protocol from [EPK14] based on SUE (*cf.* Section 4.1.4).
- *L-OSUE*. The optimized L-UE protocol from [ACBX22] (*cf.* Section 4.1.4).
- *L-GRR*. The optimized longitudinal protocol from [ACBX22] when k is small (*cf.* Section 4.1.4).
- *dBitFlipPM*. The one-round randomization mechanism from [DKY17] with $d \in \{1, b\}$, referred respectively as 1BitFlipPM and bBitFlipPM, in which the former 1BitFlipPM is tuned for privacy and the latter bBitFlipPM for utility (*cf.* Section 4.1.4)
- *LOLOHA*. Our protocols following Algorithm 8, which are BiLOLOHA with $g = 2$ adjusted for privacy and OLOLOHA with g following Eq. (4.6) tuned for utility.

Privacy metrics. We vary the longitudinal privacy parameter in the range $\epsilon_\infty = [0.5, 1, \ldots, 4.5, 5]$ and $\epsilon_1 = \alpha\epsilon_\infty$, for $\alpha \in \{0.4, 0.5, 0.6\}$, to compare our experimental results with numerical ones from Section 4.3 (with higher visibility).

Performance metrics. To evaluate our results, we use the MSE averaged by the number of data collection τ, denoted by MSE_{avg}. Thus, for each time $t \in [1..\tau]$, we compute for each value $v \in V$ the estimated frequency $\hat{f}_L(v)_t$

[1] https://www.census.gov/programs-surveys/acs/microdata/documentation.html.

and the real one $f(v)_t$ and calculate their differences before averaging by τ. More formally,

$$MSE_{avg} = \frac{1}{\tau} \sum_{t \in [1..\tau]} \frac{1}{|V|} \sum_{v \in V} \left(f(v)_t - \hat{f}_L(v)_t \right)^2 . \tag{4.7}$$

We also assess the averaged longitudinal privacy loss for all users, denoted by $\check{\epsilon}_{avg}$. More precisely, after the end of all data collections τ, we compute for each user $u \in U$ their overall longitudinal privacy loss $\check{\epsilon}_\infty^{(u)}$ and average by n. For example, RAPPOR (and L-GRR and L-OSUE) leaks a new ϵ_∞ in each data change with $\check{\epsilon}_\infty \leq k\epsilon_\infty$, while LOLOHA protocols leak a new ϵ_∞ in each hash value change with $\check{\epsilon}_\infty \leq g\epsilon_\infty$. More formally,

$$\check{\epsilon}_{avg} = \frac{1}{n} \sum_{u \in U} \check{\epsilon}_\infty^{(u)}. \tag{4.8}$$

Finally, for the dBitFlipPM protocol, we also evaluate the percentage of users in which an attacker can identify **all** (bucket) data change points (*i.e.*, *worst-case* analysis) due to different PRR reports throughout the τ data collections.

4.4.2 Results

First, we compare the utility performance of our LOLOHA protocols with all four state-of-the-art memoization-based protocols for frequency monitoring under LDP guarantees, namely, RAPPOR [EPK14], L-OSUE [ACBX22], L-GRR [ACBX22], and dBitFlipPM [DKY17], for $d \in \{1, b\}$. Fig. 4.3 illustrates the MSE_{avg} metric in Eq. (4.7) for all methods and all Syn, Adult, DB_MT, and DB_DE datasets, by varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.4, 0.5, 0.6\}$. On the one hand, since $k \leq 360$ for Syn and Adult datasets, when implementing dBitFlipPM, we select $b = k$ to estimate the same k-bins histogram as all other methods in Figs. 4.3a and 4.3b. On the other hand, we select $b = \lfloor k/4 \rfloor$ bins for both DB_MT ($k = 1412$) and DB_DE ($k = 1234$) datasets, but we did not include the error metric of dBitFlipPM in Figs. 4.3c and 4.3d as the error is five orders of magnitude higher due to histograms of different sizes ($b < k$).

Fig. 4.3 shows that the experimental results with all datasets match the numerical results of variance values from Fig. 4.2 for our LOLOHA protocols, RAPPOR, and L-OSUE. More specifically, our OLOLOHA protocol has similar utility to the optimized L-OSUE protocol, a relationship also find between their one-round versions OLH and OUE in [WBLJ17]. In high privacy regimes, all four protocols, *i.e.*, RAPPOR, L-OSUE, BiLOLOHA, and OLOLOHA have very similar utility. In low privacy regimes, L-OSUE and OLOLOHA outperforms both RAPPOR and BiLOLOHA. The least performing longitudinal LDP protocols are L-GRR and 1BitFlipPM, the former due to high domain sizes k, as shown in [ACBX22], and the latter due to sampling only a single $d = 1$ bit out of b ones. The bBitFlipPM protocol outperforms all experimented longitudinal LDP protocols due to having only a single round of sanitization (*i.e.*, the PRR step) and by reporting all $d = b$ bits, which is consistent with [DKY17] (the larger d the greater the utility).

However, increasing the number of bits d the users must report negatively impacts privacy, as each new input value has a high probability of generating a new output value, which will be detected by the server. For instance, for both dBitFlipPM protocols, for $d \in \{1, b\}$, Table 4.2 exhibits the percentage of users in which **all** bucket changes were detected by the server due to different PRR responses throughout τ data collections, for all Syn, Adult, DB_MT, and DB_DE datasets. Remark that when $d = 1$, the protocol is adjusted for privacy, thus being less vulnerable with respect to privacy with only a small percentage ($< 1\%$) of users that the server always detected a different randomized output due to different input values. Besides, one can note that the percentage of attacked users decreases as ϵ_∞ gets higher when $d = 1$. The intuition is that the probability of randomizing the single bit will be smaller with high ϵ_∞, thus generating the same report many times. On the other hand, the bBitFlipPM protocol is tuned for utility, which increased the probability of *always* generating a new randomized output due to new input values and, thus leading to 100% of detection for all four datasets. Though we only perform both extreme cases (lower $d = 1$ and upper $d = b$ bounds), one can picture the privacy-utility trade-off of dBitFlipPM for other d values in between our results of Fig. 4.3 and Table 4.2.

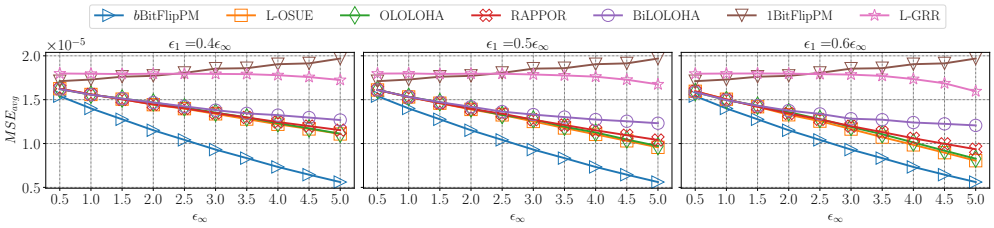
We now analyze the longitudinal privacy guarantees of our LOLOHA proto-

Table 4.2: Percentage of users in which the server detected **all** data change points for *d*BitFlipPM, for $d \in \{1, b\}$, and all Syn, Adult, DB_MT, and DB_DE datasets.

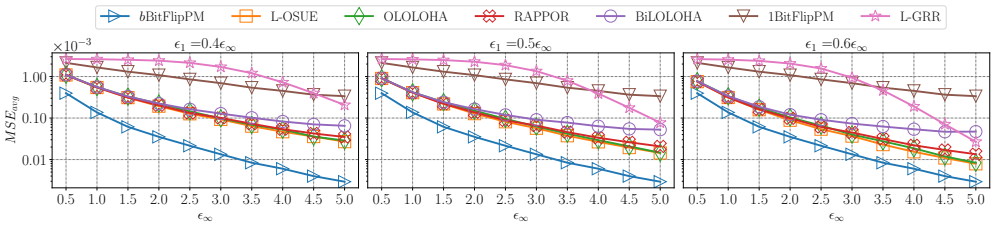
ϵ_∞	$d = 1$				$d = b$			
	Syn	Adult	DB_MT	DB_DE	Syn	Adult	DB_MT	DB_DE
0.5	0%	0%	0.0048%	0%	100%	100%	100%	100%
1.0	0%	0%	0.0044%	0%	100%	100%	100%	100%
1.5	0%	0%	0.0048%	0%	100%	100%	100%	100%
2.0	0%	0%	0.0039%	0%	100%	100%	100%	100%
2.5	0%	0%	0.0024%	0%	100%	100%	100%	100%
3.0	0%	0%	0.0024%	0%	100%	100%	100%	100%
3.5	0%	0%	0.0024%	0%	100%	100%	100%	100%
4.0	0%	0%	0.0019%	0%	100%	100%	100%	100%
4.5	0%	0%	0.0010%	0%	100%	100%	100%	100%
5.0	0%	0%	0.0010%	0%	100%	99.99%	100%	100%

cols in comparison with the state-of-the-art memoization-based LDP protocols. Fig. 4.4 illustrates the $\check{\epsilon}_{avg}$ metric in Eq. (4.8) for all methods and all Syn, Adult, DB_MT, and DB_DE datasets, by varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.4, 0.5, 0.6\}$. Notice that the results of *d*BitFlipPM protocols in Figs. 4.4a and 4.4b are with $b = k$ buckets and in Figs. 4.4c and 4.4d are with $b = \lfloor k/4 \rfloor$ buckets.

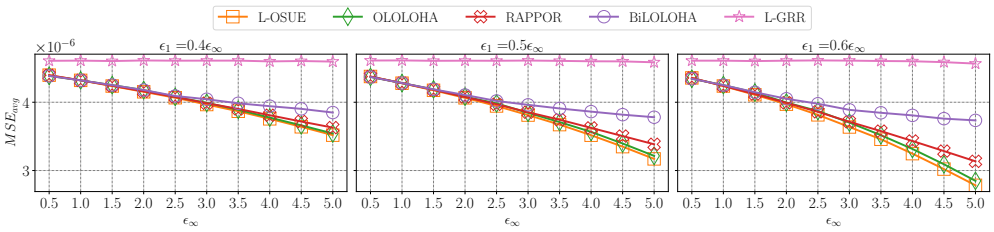
From Fig. 4.4, one can remark that all four LDP protocols, RAPPOR, L-OSUE, L-GRR, and *b*BitFlipPM (when $b = k$ in Figs. 4.4a and 4.4b), have an averaged longitudinal privacy loss linear to the number of data changes the users performed throughout the τ data collections. Fig. 4.4a presents the smallest $\check{\epsilon}_{avg}$ as both $k = 360$ and the change rate $p_{ch} = 0.25$ are small. However, in a worst-case scenario in which the users change their values significantly or $\tau \to \infty$, the overall privacy loss of RAPPOR, L-OSUE, L-GRR, and *b*BitFlipPM can grow to values as large as $k\epsilon_\infty$ for all datasets. Note that in Figs. 4.4a and 4.4b, naturally, setting $b = k$ does not benefit from the *d*BitFlipPM advantage for enhancing longitudinal privacy protection by mapping several close values to the same bin, which leads to higher $\check{\epsilon}_{avg}$. In contrast, in Figs. 4.4c and 4.4d, the longitudinal privacy loss of *b*BitFlipPM protocols is lower than RAPPOR, L-OSUE, and L-GRR because $b = \lfloor k/4 \rfloor$



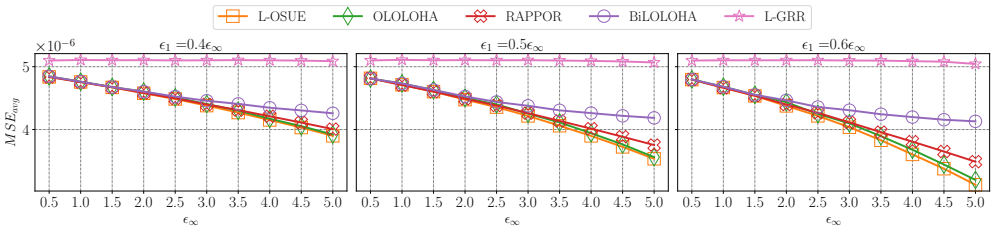
(a) Syn dataset: $k = 360$, $n = 10000$, and $\tau = 120$.



(b) Adult dataset: $k = 96$, $n = 45222$, and $\tau = 260$.



(c) DB_MT dataset: $k = 1412$, $n = 10336$, and $\tau = 80$.

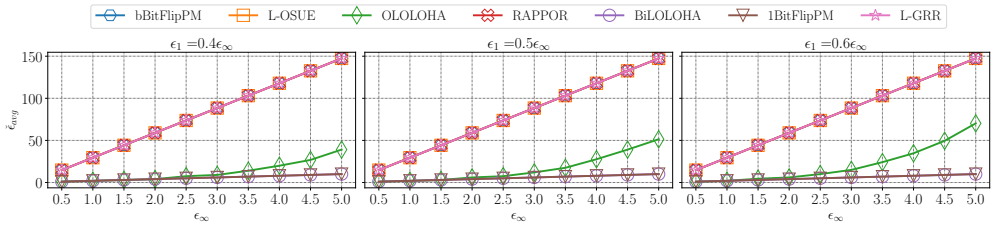


(d) DB_DE dataset: $k = 1234$, $n = 9123$, and $\tau = 80$.

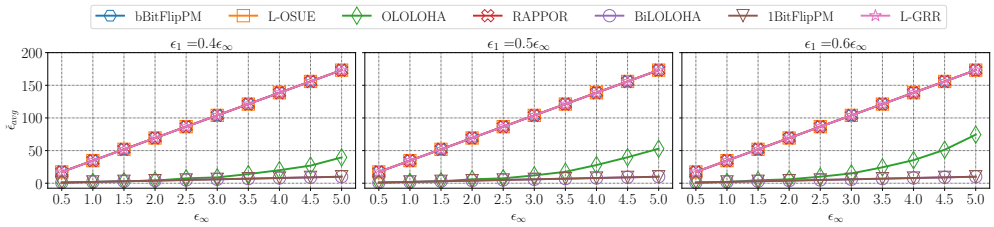
Figure 4.3: Averaged MSE for τ data collections in Eq. (4.7) by varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.4, 0.5, 0.6\}$, on (a) Syn, (b) Adult, (c) DB_MT, and (d) DB_DE datasets. The evaluated methods are: dBitFlipPM [DKY17], L-OSUE [ACBX22], RAPPOR [EPK14], L-GRR [ACBX22], and our LOLOHA protocols.

buckets, but still significantly higher than our LOLOHA protocols.

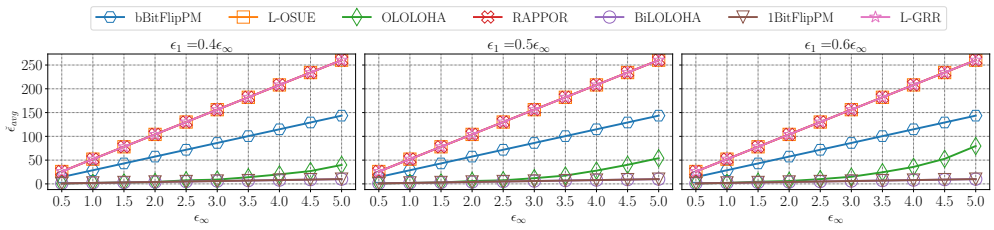
Indeed, the privacy loss of our LOLOHA protocols depends only on the new



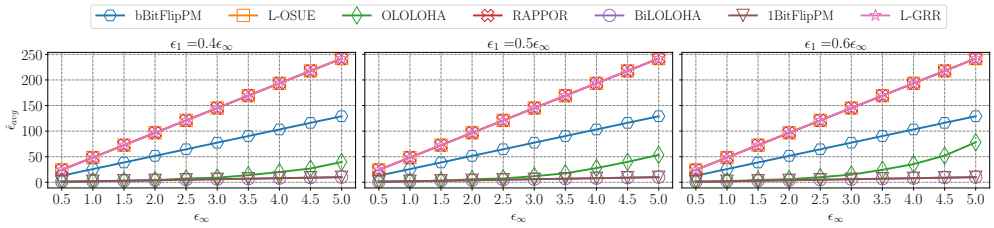
(a) Syn dataset: $k = 360$, $n = 10000$, and $\tau = 120$.



(b) Adult dataset: $k = 96$, $n = 45222$, and $\tau = 260$.



(c) DB_MT dataset: $k = 1412$, $n = 10336$, and $\tau = 80$.



(d) DB_DE dataset: $k = 1234$, $n = 9123$, and $\tau = 80$.

Figure 4.4: Averaged longitudinal privacy loss for τ data collections in Eq. (4.8) by varying the longitudinal ϵ_∞ and first report $\epsilon_1 = \alpha\epsilon_\infty$ privacy guarantees, for $\alpha \in \{0.4, 0.5, 0.6\}$, on (a) Syn, (b) Adult, (c) DB_MT, and (d) DB_DE datasets. The evaluated methods are: dBitFlipPM [DKY17], L-OSUE [ACBX22], RAPPOR [EPK14], L-GRR [ACBX22], and our LOLOHA protocols.

domain size $g \geq 2$, which is agnostic to k. For this reason, our BiLOLOHA protocol with $g = 2$ leaked about 15 to 25 orders of magnitude less than the

state-of-the-art LDP protocols considering the experimented τ values. These are similar results achieved by the 1BitFlipPM protocol, which agrees with the theoretical analysis in Table 4.1, although BiLOLOHA consistently and considerably outperforms 1BitFlipPM in terms of utility loss (see Fig. 4.3). Besides, since our OLOLOHA protocol has privacy loss depending on the optimal g value in Eq. (4.6), which can be $g > 2$ in low privacy regimes, it only resulted in about 2 to 5 order of magnitude less privacy loss than the state-of-the-art LDP protocols, for the experimented τ value. More specifically, when ϵ_∞ is high and $\alpha = 0.6$ (see Fig. 4.4d), OLOLOHA leaked about 2 orders of magnitude less privacy loss than the bBitFlipPM protocol. However, as the number of data collections $\tau \to \infty$, bBitFlipPM privacy loss will go to $b\epsilon_\infty$, which is b/g times higher than the one from OLOLOHA with $g\epsilon_\infty$. Besides, in practice, lower values of ϵ_∞ and $\alpha \ll 1$ should be used to ensure strong longitudinal privacy guarantees since the first $\epsilon_1 = \alpha\epsilon_\infty$-LDP report. As shown in Fig. 4.1, this will mean lower values of g, which will substantially decrease the longitudinal privacy loss of OLOLOHA.

4.4.3 Discussion

In brief, we evaluated in our experiments the performance of our LOLOHA protocols in comparison with four state-of-the-art memoization-based LDP protocols [EPK14, DKY17, ACBX22] for frequency monitoring on different datasets and varying different parameters. We now summarize the main findings that help justify the many claims of this chapter.

More precisely, the conclusions we stated in Section 4.3 are based on the analytical variances of the LDP protocols. To corroborate these conclusions, our empirical experiments in Section 4.4.2, which measured the MSE metric, do indeed correspond to the numerical results of the variances.

Furthermore, the main disadvantage of RAPPOR [EPK14] and the two others optimized protocols from [ACBX22], (*i.e.*, L-GRR and L-OSUE), is the linear relation on k for the overall longitudinal privacy loss, *i.e.*, $k\epsilon_\infty$, as each data change needs to be memoized. Thus, for the monitoring of large-scale systems (*e.g.*, application usage, calories ingestion, preferred webpage, etc.), the overall privacy loss of such protocols will be tremendous, being unrealistic for private frequency monitoring.

Even though the dBitFlipPM [DKY17] generalizes the original domain size k to b buckets, there is still a linear relation on the new domain size $b \leq k$ for the overall longitudinal privacy loss, *i.e.*, $b\epsilon_\infty$, as each bucket change needs to be memoized *when the mechanism is tuned for utility*. What is more, this generalization naturally leads to loss of information and one has to carefully choose the bucket numbers/width for the best privacy-utility trade-off. Besides, the privacy-utility trade-off of dBitFlipPM also depends on the number of bits $d \leq b$ each user samples. However, even when $d = 1$, which offers the strongest protection [DKY17], in our experiments, the server was still able to detect **all** bucket change of a small portion of users (see Table 4.2). Hence, as one adjusts d for utility, *i.e.*, $1 < d \leq b$, the higher the attacker's success rate to detect all user's data changes will be.

The best choice for adequately balancing privacy and utility for frequency monitoring is with our LOLOHA protocols, as the privacy loss is only linear to the new (reduced) domain size $2 \leq g \ll k$. Though we only experiment with $80 \leq \tau \leq 260$ data collections in Section 4.4.2, in the worst case, this represents a significant k/g decrease factor of privacy loss by our LOLOHA protocols. Intuitively, LOLOHA can be tuned to satisfy the strongest longitudinal privacy protection by selecting $g = 2$ (*i.e.*, our BiLOLOHA protocol). In this setting, there is loss of utility in the encoding step through local hashing since the output is just one bit. For instance, even if this bit is transmitted correctly after the two rounds of sanitization, the server can only obtain one bit of information about the input (*i.e.*, to which half of the input domain the value belongs to [WBLJ17]). Nevertheless, from the analytic variance analysis in Fig. 4.2 and empirical experiments in Fig. 4.3, LOLOHA is optimal with $g = 2$ in high privacy regimes, *i.e.*, low ϵ_∞ values, which is desirable for practical deployments.

As a limitation, users fix their randomly selected hash function $\mathrm{H} \in \mathscr{H}$ with our LOLOHA protocols (*cf.* Algorithm 8), which can be regarded as a unique identifier in longitudinal data collection. However, this is a common assumption of the LDP model, which assumes the server already knows the users' identifiers [BEM+17, WXY+19, EFM+19, EFM+20], but not their private data. One way to counter this link between the user's randomized report and their identifier is to assume a trusted intermediate, such as a shuffler,

that does not collude with the server, *e.g.*, the Shuffle DP model [BEM+17, EFM+19, EFM+20], which we let the investigation for future work.

4.5 Related Work

Differential privacy [DMNS06, Dwo06b, DR+14] has been increasingly accepted as the current standard for data privacy. The central DP model assumes a trusted curator, which collects the clients' raw data and releases sanitized aggregated data. The LDP model [KLN+08, DJW13, DWJ13] does not rely on collecting raw data anymore, which has a clear connection with the concept of randomized response [War65]. In recent years, there have been several studies on the local DP setting, *e.g.*, for frequency estimation of a single [WBLJ17, ASZ19, FNNT22, KBR16, KOV16, NV20, CMM21] and multiple [ACABX21, VCS22, LTH+23] attributes; mean estimation [NXY+16, WXY+19], heavy hitter estimation [BS15, BNST17], and machine learning [MABK+20, ZT21].

As for locally differentially private monitoring, Erlingsson, Pihur, and Korolova [EPK14] proposed the RAPPOR algorithm for frequency monitoring that is based on the *memoization* solution described in Section 4.1.4. The recent study of Arcolezi *et al.* [ACBX22] generalizes this framework for optimally chaining two LDP protocols, proposing the L-GRR protocol that is optimized for small domain size k and the L-OSUE protocol for higher k (see Figs. 4.2 and 4.3). Moreover, Erlingsson *et al.* [EFM+20] formalize the privacy guarantees of using two rounds of sanitization under both local and shuffle DP guarantees. Naor and Vexler [NV20] also formalized the privacy guarantees of chaining two LDP protocol as well as introduced a new Everlasting privacy definition.

An alternative approach for memoization named dBitFlipPM has been proposed by Ding, Kulkarni, and Yekhanin [DKY17], discussed in Section 4.1.4. The dBitFlipPM protocol allows frequent but only *small* changes in the original data since a high change (*i.e.*, a different bucket) can be detected by an attacker (*cf.* Table 4.2). Although an attacker that is able to identify a data change can still not infer the user's actual data (controlled by ϵ_∞), the overall LDP guarantees can be highly reduced if these changes are

correlated [EPK14, DKY17, EFM+19]. For instance, the authors in [TKB+17] performed a detailed analysis of Apple's LDP implementation and examined its longitudinal privacy implications. Naor and Vexler [NV20] also investigated the trackability of RAPPOR following their new Everlasting privacy definition.

LOLOHA leverages the best of RAPPOR and dBitFlipPM, which can inherently minimize these inference attacks. More precisely, on the one hand, LOLOHA uses LH [WBLJ17] for domain reduction, which allows many values to collide (universal hashing property) and thus creates uncertainty about the user's actual value. Indeed, LH protocols are the least attackable LDP protocols in the recent studies of Arcolezi *et al.* [AGCP22] and Emre Gursoy *et al.* [GLC+22] considering a Bayesian adversary. Besides that, LOLOHA also has two rounds of sanitization following RAPPOR's framework, which can improve privacy to minimize data change detection. Finally, another line of work for frequency monitoring under LDP is data change-based [JRUW18, EFM+19, XYH+22, OWW22], motivated by the fact that, generally, users' data changes infrequently. A similar idea was proposed much earlier in the work of Chatzikokolakis, Palamidessi, and Stronati [CPS14], which proposed a predictive mechanism for location-based systems to utilize privacy budget only for new "hard" location points (*i.e.*, with bad predictions). However, these approaches normally impose restrictions on the number of data collections τ and on the number of data changes as their accuracy degrades linearly or sub-linearly with the number of changes in the underlying data distributions, which can limit their applicability and scalability to real-world systems.

4.6 Conclusion and Perspectives

In this chapter, we study the fundamental problem of monitoring the frequency of evolving data throughout time under LDP guarantees. We proposed a new locally differentially private protocol named LOLOHA, which is built on top of domain reduction to minimize longitudinal privacy loss up to a k/g factor and double randomization to enhance privacy. Through theoretical analysis, we have proven the longitudinal privacy (Theorems 31, 32, and 33) and accuracy guarantees (Proposition 34) of our LOLOHA protocols.

In addition, through extensive experiments with synthetic and real-world datasets, we have shown that our proposed LOLOHA protocols preserve competitive utility as state-of-the-art LDP protocols [EPK14, DKY17, ACBX22] by considerably minimizing longitudinal privacy loss (from 2 to 25 orders of magnitude with the experimented τ values). As future work, we intend to identify reasonable conditions of the input data (not constant as in [EPK14, ACBX22]) in which one can satisfy the standard ϵ-LDP definition. Besides, we intend to identify attack-based approaches to longitudinal LDP frequency estimation protocols (*e.g.*, data change detection or correlated data) and to extend the analysis of our LOLOHA protocols to the shuffle DP model.

In the next chapter, we discuss a different problem in which the server is not collecting data but serving it, and the privacy is not in the contents of the transmitted data, but on their sizes. Nevertheless, there are strong conceptual connections between the two problems, e.g., there is also a trade-off involving privacy and differential privacy is again excessively restrictive, so a different notion of privacy based on attackers is used.

Chapter 5

Obfuscation padding schemes that minimize Rényi min-entropy for privacy

Consider a set of users, each of which is choosing and downloading one file out of a central pool of public files, and an attacker that observes the download size for each user and is willing to identify the choice of each user. The files are public, but the choices are private. The objective is to pad the files with some small overhead to obfuscate the information gained by the attacker and reduce his chances of discovering the choices of the users. This chapter studies the problem of minimizing the expected accuracy of the attacker by padding the files without exceeding some given padding constraints.

On one extreme, if the files are not padded at all, the attacker might easily map the observed download sizes with the original files; e.g., if there is just one file of size 10.32MB and the attacker observes that the network traffic of some user corresponds to a file of size 10.32MB, he will immediately know what file was chosen. This can be prevented by padding several files to common sizes to obfuscate the information gained by the attacker. On the other extreme, if all files are padded to a common size, this common size should be large enough to cover the largest file in the set, and, as a consequence, many small files will be padded excessively, increasing the bandwidth use. The ideal solution lies between these two extreme cases. For this reason, this

chapter considers the problem of maximizing privacy while respecting some flexible padding constraints, like, for example, that no file can increase its size more than 10%.

The attacker we consider makes just one attempt to re-identify the file, and to maximize his chances, he will of course guess a file that has the maximum posterior probability given the observed (obfuscated) size. This model of attack is known in literature as *one-try attack* [Smi09], and it has been characterized in information-theoretic terms using *Rényi min-entropy*. More specifically, entropy in general represents the (lack of) information content of a discrete probability distribution, and Rényi min-entropy is a form of entropy that emphasizes the highest probability value. The prior and posterior entropies represent the probabilistic knowledge of the attacker before and after he observes the obfuscated size, respectively. In particular, Rényi posterior min-entropy is related to *hypothesis testing* and, as a measure, it closely corresponds to the *Bayes error*. The difference between the prior and posterior entropies represents how much the knowledge of the attacker (and hence his probability of success) increases thanks to the observation, and it is, therefore, a measure of the efficiency of the padding scheme. In literature this difference is known as Rényi min-entropy *leakage*.

The padding problem considered in this chapter might also apply to equivalent scenarios in which an attacker exploits time side-channel information. For illustration, consider an intelligence service that is surveilling people entering and exiting a building. They can use the time each user took inside to infer the type of service he received, e.g., whether he was at the bank, shopping, or at the cinema in the mall. In this case, the users can waste some time inside the building on purpose to confuse the observer. Equivalently, a server can delay its responses in a planned manner to prevent an attacker from inferring the chosen type of request. More generally, an algorithm can sleep on purpose to prevent leaking information about the input, as exploited by timing attacks [Sch00, Son01].

Contributions

- We propose two algorithms that derive the optimal padding schemes, one for the deterministic case, and one for the randomized case (PRP and POP, defined in Section 5.2).

- We prove the correctness of the algorithms and test the implementations against brute-force solutions using small synthetic datasets.
- Likewise, we compare our algorithms with an existing solution [RR21] that uses an attack model based on Shannon entropy, and discuss how the two approaches are related in terms of the type of private information leakage that each attacker represents.
- The code is publicly available at [PPS22]. It includes not only the algorithms we propose, but also the reimplementation of the algorithms of [RR21] to support flexible padding constraints, multiple files having the same size, and sparse matrix representations.

5.1 Related Work

The model of attacker we use has been well investigated in the field of *Quantitative Information Flow* (QIF), which is a branch of security aimed at studying inference attacks, namely attackers that try to infer the value of the secret from related observations. The QIF theory actually formalizes a variety of models, each of them characterized by parameters that represent the capabilities and the goal of the attacker. For a detailed coverage of the topic we refer to [AM+20].

This chapter is strongly related with the work of Reed and Reiter [RR21], in which the authors consider the same problem with a different attack model, based on Shannon entropy, and more specifically, on measuring the leakage in terms of Shannon mutual information. Shannon mutual information is a well known notion that has been shown to be very useful in several scientific fields. In security and privacy, however, it does not seem the right notion for modeling the attacker. Indeed, its operational interpretation corresponds to an attacker that can try to guess the exact secret by making an unbound number of attempts, and his objective is to minimize the expected number of attempts before he identifies it correctly. This seems a less natural model of attacker than those of QIF (and hence than the one we use, based on Rényi min-entropy), and it also sometimes leads to conclusions that are contrary to common sense. For a detailed discussion about this issue, refer to [Smi09].

Reed and Reiter [RR21] propose three padding algorithms, called `PrpSh`, `PopSh` and PwoD (padding without a distribution), for finding padding schemes that minimize Shannon leakage under different bandwidth constraints. These algorithms do not support, however, multiple files having the same size nor flexible padding constraints as defined in this chapter. We re-implemented their algorithms with these additional details before comparing them with our proposed solutions, and we explained in terms of attack models and information leakage the core difference between them.

In [Deg21] they consider the BREACH/CRIME [GHP15] security attack in which the attacker observes sizes and can also control a malicious script that runs in the browser of the victim. By exploiting the greedy mechanism of the Huffman encoder in the compression stage of the cookies, the attacker is able to use repeatedly the size information to discover the cookie secret and impersonate the victim. As they show, random gaussian padding can be used and is better than uniform padding to reduce the attacker's probability of success from 1.0 to 0.0026. Although this chapter is more related with security than privacy, it shows how important padding can be to obfuscate information.

Lastly, one of the main conclusions in [WCM09] is that the optimal way to reduce information obtained by an attacker that monitors traffic is to modify the traffic patterns so that they are confused with other patterns. We draw a similar conclusion formally in our problem (Proposition 36), proving that it is optimal to pad messages to reach the sizes of other existing files.

5.2 Problem formalization

The collection of public files is denoted as $E = \{e_1, e_2, \ldots, e_n\}$, where E is sorted non-decreasingly by the sizes $|e_i| \in \mathbb{N}$. For the sake of generality, we allow different files to have the same size, hence the set of file sizes $S \stackrel{\text{def}}{:=} \{|e| \,|\, e \in E\}$ has $m \leq n$ unique elements, which we enumerate in increasing order as $S = \{s_1, s_2, \ldots, s_m\}$.

A *padding function* or padding scheme is a function $f : E \to \mathbb{N}$ respecting $f(e_i) \geq |e_i|$ that tells to what size each file should be padded. The padding constraints are expressed with the proposition $\forall i,\ f(e_i) \in [|e_i|, b_i]$, where each

$[|e_i|, b_i] = \{|e_i|, |e_i|+1, ..., b_i\}$ is an integer interval.

The sequence of users with their respective choices is modelled as a sequence of i.i.d. samples coming from the marginal distribution of the files. File e_i is chosen with frequency $p_i \in [0,1]$, where $\sum_{i=1}^{n} p_i = 1$. We let X be a random variable satisfying $\mathbb{P}(X{=}e_i) \stackrel{\text{def}}{=} p_i$, thus, a sequence of users with choices can be represented as a sequence of i.i.d. choices following the distribution of X.

The attacker will predict, upon seeing a download of size $z \in \mathrm{Im}(f)$ (where the image $\mathrm{Im}(f) \stackrel{\text{def}}{=} \{z \in \mathbb{N} \mid \mathbb{P}(f(X){=}z) > 0\}$ denotes the set of possible outputs of f), that the secret value of X is the file e_i that maximizes $\mathbb{P}(f(e_i){=}z)$. To do this, he uses the public information he has access to and the information he can infer. The files and their sizes before padding are public, and he can determine the padding scheme by requesting each of the files himself, possibly multiple times in case of a randomized padding scheme. In addition, considering the worst-case scenario, we assume that he knows or has estimated the frequencies p_i with which files are chosen on average. With this information, the attacker can always find a file e_i that maximizes $\mathbb{P}(f(e_i){=}z)$ for the observed z, and his expected probability of success is therefore

$$\sum_{z \in \mathrm{Im}(f)} \max_{i \in [1..n]} \mathbb{P}(X{=}e_i \wedge f(X){=}z) = \sum_{z \in \mathrm{Im}(f)} \max_{i \in [1..n]} p_i \cdot \mathbb{P}(f(e_i){=}z). \tag{5.1}$$

The objective is to find a padding function $f : E \to \mathbb{N}$ that minimizes the accuracy of the attacker while respecting the given padding constraints. In addition, two scenarios are considered separately: *per-object-padding* (POP) refers to the case when f is deterministic, hence the files are padded once and forever; *per-request-padding* (PRP) refers to the case when the padding is done on demand and f is probabilistic.

5.2.1 Presentation in terms of privacy leakage

The objective of minimizing the attacker accuracy can equivalently be presented in terms of minimizing privacy leakage. There are several definitions for leakage $\mathbb{I}(|X|, f(X))$ of a padding function $f : E \to \mathbb{N}$. Particularly, Rényi min-entropy leakage [Smi09], which we call *Rényi leakage* in this chapter, is

defined using Rényi min-entropy $\mathbb{H}_\infty$ as follows:

$$\mathbb{I}_\infty(f) \stackrel{\text{def}}{\coloneqq} \mathbb{I}_\infty(|X|, f(X)) = \mathbb{H}_\infty(|X|) - \mathbb{H}_\infty(|X| \mid f(X)), \tag{5.2}$$

$$\mathbb{H}_\infty(|X|) = -\log_2 \max_{z \in \mathrm{Im}(f)} \mathbb{P}(|X| = z), \tag{5.3}$$

$$\mathbb{H}_\infty(|X| \mid f(X)) = -\log_2 \sum_{z \in \mathrm{Im}(f)} \max_{i \in [1..n]} (p_i \cdot \mathbb{P}(f(e_i) = z)). \tag{5.4}$$

The importance of Rényi leakage in more general contexts can be found in [PR18] and [Smi09]. Basically, Rényi leakage is a special case ($\alpha = \infty$) of a family of leakages $\mathbb{I}_\alpha$ based on α-Rényi entropy $\mathbb{H}_\alpha$. Since Rényi-min entropy $\mathbb{H}_\infty(|X|)$ is constant in regard to the padding-scheme, minimizing Equation (5.2) is equivalent to maximizing Equation (5.4), which is in turn equivalent to minimizing Equation (5.1). Therefore, Rényi leakage is in direct one-to-one correspondence with the probability of success of the attacker.

Another important case ($\alpha = 1$) is Shannon leakage, which is given by: $\mathbb{I}(|X|, f(X)) = \sum_{i,z} p_i \, \mathbb{P}(f(e_i){=}z) \log_2 \frac{\mathbb{P}(f(e_i)=z)}{\mathbb{P}(f(X)=z)}$. With some effort, this leakage can also be interpreted in terms of an attacker that we call Shannon attacker. The Shannon attacker is assumed to have access to an oracle that answers queries of the type "is the file in *this set of files*?" for each user, and his objective is to find the right files using the minimal number of queries, as in a 20Q game. Although the oracle assumption makes the Shannon attacker unrealistic, defenses against him are useful against the Rényi attacker of this chapter because, intuitively, the more queries the Shannon attacker needs, the harder it is to guess the correct file in a single try.

For this particular application, the direct pragmatic connection between Rényi leakage and a simple adversary success makes it more appealing than the Shannon attacker. The same argument is used in [Che17], whose privacy measure is closely related with ours. More generally in the privacy community, leakage functions are better described in terms of their associated attacker rather than their information theoretic properties [ACPS12, Rom20].

5.2.2 Why not differential privacy?

Differential privacy [Dwo06a], is one of the most prevalent formalizations of privacy. For this particular problem, a padding scheme f satisfies ϵ-

differential privacy if and only if for all input files $e_i, e_j \in E$ and all output sizes $z \in \mathrm{Im}(f)$, we have $\mathbb{P}(f(e_1){=}z) \leq \exp(\epsilon)\,\mathbb{P}(f(e_2){=}z)$.

This notion of privacy represents an attacker whose success function is given by how much more likely one input file is *with respect to another one* for a given observation. However, this is excessively strong for the problem under consideration. Indeed, as Theorem 35 shows, differential privacy can only be achieved at the total detriment of bandwidth use.

Theorem 35. *For any $\epsilon > 0$, the padding scheme that satisfies ϵ-differential privacy and minimizes bandwidth is the one that pads all input files to the size of the largest one.*

Proof. Fix $\epsilon > 0$ and let $e_j \stackrel{\text{def}}{=} \arg\max_{e_i \in E} |e_i|$ be the largest file in E. For all sizes $z < |e_j|$, we have $\mathbb{P}(f(e_j){=}z) = 0$ because e_j can not be padded to smaller sizes than $|e_j|$. Moreover, the differential privacy constraint forces every other file $e_i \neq e_j$ to satisfy $\mathbb{P}(f(e_i){=}z) \leq \exp(\epsilon)\mathbb{P}(f(e_j){=}z) = 0$ whenever $z < |e_j|$. In other words, all files must be padded to sizes at least as large as $|e_j|$, i.e., $\mathbb{P}(f(X) \geq |e_j|) = 1$. Among all the mappings f that have this property, the one that minimizes bandwidth is the one that pads all files exactly to the largest file size $|e_j|$, and it satisfies ϵ differential privacy trivially because it is a constant function. □

Theorem 35 is the reason why we exclude differential privacy from the analysis and focus on the privacy notions discussed in the previous section. This theorem is a direct consequence of the inevitable fact that padding can only enlarge files and not reduce their sizes. Apart from putting in evidence the abusive overhead required by differential privacy, this theorem also shows that its parameter ϵ is irrelevant as a measure of privacy for the problem under consideration, making it inappropriate.

5.2.3 Simplification of the output set

We conclude this section by proving that optimal padding functions always map to sizes in S. This is a key-fact for the derivation of the algorithms and their proofs. Intuitively, if a set of files can be padded to a common certain size z, but can also be padded to $z - 1$, we can pad them to $z - 1$ and win

some bandwidth without leaking any additional information. This forces the optimal padding functions to always pad to the sizes z for which it is not possible to pad to $z-1$ without sacrificing privacy, which are precisely the sizes in S. The same holds true for padding schemes that minimize Shannon leakage, as shown in [RR21].

Proposition 36. For any padding-scheme $f : E \to \mathbb{N}$, there exists a padding-scheme $f^* : E \to S$ such that $\mathbb{I}(f^*) \leq \mathbb{I}(f)$. Moreover, $\mathbb{P}(f^*(X) \leq f(X)) = 1$, hence f^* uses less padding (bandwidth) than f.

Proof. Define f^* as the composition $f^* \stackrel{\text{def}}{=} g \circ f$, where $g(z) = \max\{s \in S : s \leq z\}$, that is, $f^*(X) = g(f(X))$. The function g is defined only for $z \geq \min S$ and f^* is well-defined because the padding constraints force $\mathbb{P}(f(X) \geq \min S) \leq \mathbb{P}(f(X) \geq |X|) = 1$. By definition, $g(z) \leq z$, thus $\mathbb{P}(f^*(X) \leq f(X)) = 1$. Let us now show, regarding privacy leakage, that $\mathbb{I}(f^*) \leq \mathbb{I}(f)$. Let I^*_{xs} denote $\mathbb{P}(X{=}x \wedge f^*(X){=}s)$ and I_{xz} denote $\mathbb{P}(X{=}x \wedge f(X){=}z)$. We will show that the accuracy of the attacker (Eq. 5.1) is smaller or equal for f^* than for f. This can be expressed as $\sum_s \max_x I^*_{xs} \leq \sum_s \sum_{z:g(z)=s} \max_x I_{xz}$. On the left and right-hand sides, we have summations on $s \in S$, so it suffices to prove that this inequality holds for each fixed s. At each $s \in S$, since $I^*_{xs} = \sum_{z:g(z)=s} I_{xz}$, the inequality becomes $\max_x \sum_{z:g(z)=s} I_{xz} \leq \sum_{z:g(z)=s} \max_x I_{xz}$, which is necessarily true. Indeed, letting $x^{(s)} \stackrel{\text{def}}{=} \arg\max_x \sum_{z:g(z)=s} I_{xz}$ for the left-hand side, we have for each z with $g(z) = s$ that $I_{x^{(s)}z} \leq \max_x I_{xz}$. □ □

Proposition 36 can be seen as an instance of the Data Processing Inequality, which can be found as Theorem 8 of [ES11], or more generally for privacy contexts in [MCPS12].

Corollary 37. *A padding function that has minimal leakage must pad each file to the size of another file in the initial set.*

Having Corollary 37 in mind, the padding scheme f can be represented as an obfuscation channel matrix P where $p_{ij} = \mathbb{P}(f(e_i){=}s_j)$, in which case, the problem can be specified as shown below, and the attacker accuracy becomes

$$\sum_j \max_{i \in [1..n]} p_i \cdot p_{ij}. \tag{5.5}$$

Problem input: (1) A set E of n files $\{e_i \,|\, i \in [1..n]\}$ with frequencies p_i, sorted sizes $|e_i|$ and set of unique sizes $S = \{s_1, ..., s_m\}$. (2) Padding constraints of the form $\forall i,\ s_{l_i} \leq f(e_i) \leq s_{r_i}$, parametrized with pairs of indices $l_i, r_i \in [1..m]$.

Desired output: A padding function $f : E \to S$ in the form of a channel matrix $p_{ij} = \mathbb{P}(f(e_i){=}s_j)$ that minimizes Rényi leakage $\mathbb{I}_\infty(f)$ or equivalently Eq. (5.5). Depending on the problem variant, f must be deterministic (POP) or randomized (PRP).

5.3 Algorithms

In this section, we derive the algorithms `PopRe` and `PrpRe` that minimize the Rényi leakage (5.2) for the POP and PRP cases respectively. They contrast those for Shannon mutual information minimization found in the paper [RR21], denoted here as `PopSh` and `PrpSh`. The complexities of these algorithms are summarized in Table 5.1.

Algorithm	Minimizes	WC Runtime complexity	Memory
`PopRe`	Rényi leakage	$O(n^2\,\bar{b})$	$n\,\bar{b}$
`PrpRe`, `PrpReBa`	Rényi leakage	$O(n\,\bar{b})$	$n\,\bar{b}$
`PopSh`	Shannon leakage	$O(n\,\bar{b})$	$n\,\bar{b}$
`PrpSh`	Shannon leakage	$O(\text{ITERS} \cdot n\,m)$	$n\,m$

Table 5.1: Complexities, where $\bar{b} \overset{\text{def}}{:=} (1/n) \sum_{i=1}^{n} r_i - l_i + 1$ is the matrix average band size. For practical reference, with reasonable padding constraints, if the files are diverse with a large and spread spectrum of sizes, one expects $\bar{b} \ll m \approx n$.

Algorithm `PrpSh` is an approximation algorithm and has a runtime complexity that depends on the degree of accuracy imposed by the user and the limit number of iterations ITERS allowed. Also, the complexities of the dynamic programming algorithms correspond to the theoretical worst-case and might overestimate the actual implementations. For instance, although `PopRe` has two parameters varying in $[1..n]$, not all combinations need to be calculated in a top-down implementation.

5.3.1 Per-object-padding scenario, PopRe

In this section we develop the algorithm that minimizes Rényi leakage in the POP variation, in which the matrix P is constrained to $p_{ij} \in \{0, 1\}$. Before describing the algorithm, we will prove Remark 38, which will be used as the main update of the entries of the channel-matrix.

Remark 38. Let f be a Rényi optimal padding-scheme and e_i be the file with the highest associated frequency p_i, and assume that $p_{ij} = 1$ for some $j \in [1..m]$. Then there exists a padding-scheme f^* with the same Rényi leakage such that $p_{kj} = 1$ for all $k \in [1..n]$ such that $j \in [l_k..r_k]$.

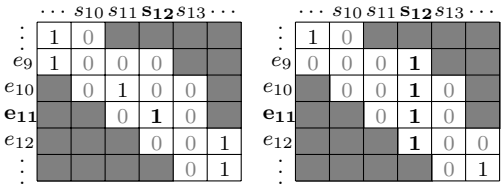


Figure 5.1: Remark 38: if the file with maximal frequency is e_{11} and the left matrix (f) is optimal, the right one (f^*) must be as well.

Proof. We consider the padding-scheme f to be represented as the channel-matrix between the secrets and the observables. When we want to minimize (5.5) we sum over each column of the matrix P. In particular, on the column j we have $\max_{a\in[1..n]}(p_a \cdot p_{aj}) = p_i$ since p_i is the highest frequency among the frequencies of the files and $p_{ij} = 1$. Now, let us consider the padding-scheme f^* whose matrix P^*, consists on moving every 1 that we can to column j:

$$p^*_{ab} = \begin{cases} p_{ab} & \text{if } b \neq j \text{ and } a \in [1..n] \text{ such that } j \notin [l_a..r_a] \\ 1 & \text{if } b = j \text{ and } a \in [1..n] \text{ such that } j \in [l_a..r_a] \\ 0 & \text{otherwise} \end{cases}$$

On the column j of the matrix P^* we will still have $\max_{a\in[1..n]}(p_a \cdot p^*_{aj}) = p_i$ because the padding-scheme f^* preserves the maximum on column j. Moreover, on the rest of the columns, the maximum either decreases or stays the same since we created more entries $p^*_{ab} = 0$, which means that the product $p_a \cdot p^*_{ab} = 0$. However, we chose f to be the Rényi optimal padding-scheme and with the remarks above, f and f^* give the same leakage. □ □

Figure 5.1 depicts an example of a sub-matrix of P as described in Remark 38. In the figure, we have exactly one entry equal to 1 in each line because the

Algorithm 10 Per-object-padding pseudocode. This implementation uses recursion both for computation and reconstruction.

procedure RÉNYI POP ▷ Main function
  MEMO $\leftarrow \{\}$ ▷ Empty map
  $p_{ij} \leftarrow 0$ ▷ A matrix p full of zeros
  renyi $\leftarrow$ RECONSTRUCT$(0, n)$
  return (p, renyi) ▷ Output matrix p and its Rényi leakage
procedure RECONSTRUCT(a, b)
  $(\text{renyi}, k, a^\star, b^\star) \leftarrow f(a, b)$
  for $j = a^\star..b^\star$ **do** $p_{jk} \leftarrow 1$ **end for**
  if $a < a^\star$ **then** RECONSTRUCT$(a, a^\star)$ **end if**
  if $b^\star < b$ **then** RECONSTRUCT$(b^\star, b)$ **end if**
  return renyi
procedure $f(a, b)$
  if $(a, b) \in$ MEMO **then return** MEMO$[(a, b)]$ **end if**
  if $a = b$ **then return** $(0, \infty, a, b)$ **end if**
  best $\leftarrow (\infty, \infty, \infty, \infty)$
  $i_{\max} \leftarrow \arg\max_{i=a..b} p_i$
  for $k = l_{i_{\max}}..r_{i_{\max}}$ **do**
    $j_{\min}, j_{\max} \leftarrow$ range of files $e_{j_{\min}}..e_{j_{\max}}$ that can be padded to size s_k
    $a^\star \leftarrow \max(a, j_{\min})$
    $b^\star \leftarrow \min(b, j_{\max})$
    renyi $\leftarrow f(a, a^\star)[0] + p_{i_{\max}} + f(b^\star, b)[0]$ ▷ Index [0] is the Rényi component
    this $\leftarrow$ (Rényi, $k, a^\star, b^\star$)
    best $\leftarrow \min(\text{best}, \text{this})$ ▷ Lexicographic (compares first by Rényi)
  $(\text{renyi}, k, a^\star, b^\star) \leftarrow$ best ▷ Unpack tuple
  MEMO$[(a, b)] \leftarrow (\text{renyi}, k, a^\star, b^\star)$
  return $(\text{renyi}, k, a^\star, b^\star)$

channel-matrix is stochastic, and we are in the POP case. Additionally, the quantity in (5.5) represents the sum of the maximum over columns where each 1 counts for the frequency of the file. Then, the update does not increase the (5.5) because the 1 with maximal frequency dominates its column, and moving all possible 1's above or below it does not increase Rényi leakage.

Using Remark 38 we can divide the padding problem into sub-problems that minimize (5.5) and leverage dynamic programming: $\forall a \leq b \in [1..n]$, we define

$$D[a][b] = \min_{\text{P channel matrix}} \sum_{j\in[1..m]} \max_{i\in[a+1..b]} (p_i \cdot p_{ij}),$$

i.e., $D[a][b]$ gives the minimal leakage for the sub-problem that pads files from e_{a+1} to e_b, under the general constraints.

By convention, we consider $D[i][i] = 0$, which will be the base case. To write the recurrence formula, we need to take the file $e_{i_{\max}}$ with maximum frequency $p_{i_{\max}}, i_{\max} \in [a+1, b]$. We go through every size index $k \in [1..m]$ such that $e_{i_{\max}}$ can be padded to the size of s_k, and we update the channel-matrix according to Remark 38, i.e., add 1's on k-th column if we can (taking into consideration the padding constraints) and complete the lines that have a fixed 1 with 0's on the remaining entries. Then, we apply the recurrence on the rows which are not updated, i.e., from a to $a^* \stackrel{\text{def}}{=} \max(a, \max_{i\in[1..n]}\{i | r_i < k\})$, and, respectively, from $b^* \stackrel{\text{def}}{=} \min(k, b)$ to b. Hence,

$$D[a][b] = p_{i_{\max}} + \min_{k\in[l_{i_{\max}}..r_{i_{\max}}]} (D[a][a^*] + D[b^*][b])$$

After applying the dynamic algorithm program with the aforementioned recurrence, we get the minimization of (5.5) in $D[0][n]$, from which we can compute the minimal Rényi leakage. If we want to recover the channel-matrix itself, in $D[a][b]$ we pass on the index k for which the maximum happens, as an argument. In case of a tie, we choose the smallest index $k \in \{1, \ldots, n\}$ in order to reduce average padding. Hence, we know in each sub-interval $[a, b]$ what we pad everything to, so the information is enough to recover the channel matrix. A pseudocode summarizing all the logic is shown in Algorithm 10. A concrete optimized implementation can be found in [PPS22].

In Figure 5.2 we depict the channel-matrix of the files with sizes $S = \{1000, 1050, 1100, 1110, 1120, 1140\}$ and associated frequencies $\{22\%, 5\%, 23\%, 12\%, 18\%, 20\%\}$. As shown in the visual representation of the padding-scheme in the right, we observe that, for both of the existing padded sizes, there are multiple files that are padded to the same element, making them indistinguishable for an attacker. Moreover, the blue bars on the graph indicate the frequencies of the files, and the red bars, the maximum frequency among the frequencies of the files padded to each specific size. The red bars are effectively highlighting the terms of the sum (5.5).

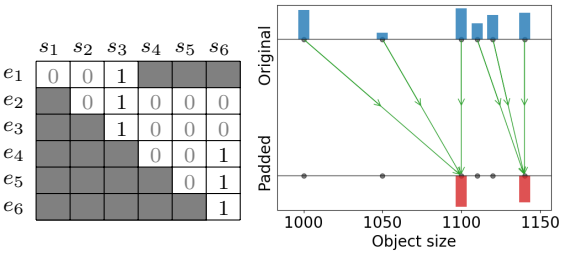


Figure 5.2: `PopRe` on a dataset of 6 files.

5.3.2 Per-request-padding scenario, `PrpRe`

In this section, we treat the case of Per-Request-Padding and provide an algorithm for finding the probabilistic channel-matrix P which minimizes the Rényi leakage. We will look at the joint distribution matrix I with entries $I_{ij} = p_i \cdot p_{ij}, \forall i \leq n, j \leq m$, for which $\sum_{j=1}^{m} I_{ij} = p_1$ for each $i \in [1..n]$.

We proceed by finding iteratively, for each of the m columns, starting from the last one, the Rényi optimal manner of setting the entries of I given the padding constraints. Furthermore, we define the *optimal distribution of* p_i *across the* i*-th row,* $1 \leq i \leq n$ to be the way we fill in the entries $p_{i1}, \ldots, p_{im}$ such as to obtain the minimum sum of the type (5.5) and preserve the relation $p_{i1} + \ldots + p_{im} = p_i$.

The proof of our algorithm requires us to consider sub-problems in which the sequence $(p_i)_{1 \leq i \leq n}$ is updated at each step of the algorithm, thus being different from the initial set of frequencies associated to each file. Hence, we rewrite the problem as a more general one in terms of a *budget* sequence $(b_i)_{1 \leq i \leq n}$ of length n (initialized as $(p_i)_{1 \leq i \leq n}$), which dictates the remaining value to be distributed across each row i, for $i \in [1..n]$. The general problem is "Given a non-negative budget sequence $(b_i)_{i=1}^{k}$ of length $k \in [1..n]$, find a solution matrix $I_{k \times m}$ that minimizes Equation (5.5), under the padding constraints

for rows $i \in [1..k]$, namely the set $\{[l_1, r_1], \ldots, [l_k, r_k]\}$ and $\sum_{j=1}^{m} I_{ij} = b_i$".

We will design the algorithm to solve the general problem recursively by returning the matrix I for the budget sequence $\{p_1, \ldots, p_n\}$ with n terms. The recurrence relationship can be described using the following observation that is used when creating the probabilistic channel-matrix for the padding-scheme f:

Remark 39. The solution $I_{k\times m}$ for a given $(b_i)_{i=1}^{k}$ that minimizes Rényi leakage satisfies the recurrence relationship

$$I_{ij} = \begin{cases} b_i & \text{if } j = m \text{ and } i \in [1..k], |e_i| = s_m \\ b_i - b'_i & \text{if } j = m \text{ and } i \in [1..k-1], |e_i| \neq s_m, \\ & m \in [l_i..r_i] \\ I'_{ij} & \text{otherwise} \end{cases}$$

where $I'_{(k-t)\times(m-1)}$ is the solution to the same minimization problem for the sequence $(b'_i)_{i=1}^{k-t}$ of length $k-t$, t = number of files from E which can be padded to s_m, such that for any $i \in [1..k-t]$, it is defined as:

$$b'_i = \begin{cases} \max(b_i - b_{t_{\max}}, 0) & \text{if } m \in [l_i..r_i] \text{ and} \\ & b_{t_{\max}} = \max\{b_i || e_i| = s_m\} \\ b_i & \text{otherwise} \end{cases}$$

Proof. If there are no files among $\{e_1, \ldots, e_k\}$ which can be padded to s_m, we set $t = 0$ and solve the minimization problem for the same budget sequence and for the set of $m-1$ sizes $\{s_1, \ldots, s_{m-1}\}$.

If there are files that can be padded to s_m, then due to the padding constraints, the element e_i can only be padded to s_m, so the entry I_{im} must necessarily be equal to b_i, for all i such that $|e_i| = s_m$. Let us denote by $T = \{k-t+1, \ldots, k\}$ the set of indices satisfying $|e_i| = s_m, \forall i \in T$ and $b_{t_{\max}} = \max\{b_i | i \in T\}$. Clearly, for every $i \in T$, $I_{ij} = 0, \forall j \in \{1, \ldots, k-1\}$. On the m-th column of the matrix I, we have $\max_{i\in[1..k]} I_{im} \geq b_{t_{\max}}$.

In order to minimize the sum (5.5) and taking into consideration that the maximum entry on column m is at least $b_{t_{\max}}$, we aim to distribute for every i such that e_i can be padded to s_m and $|e_i| \neq s_m$, a quantity equal to $b_{t_{\max}}$

(or, if $b_i < b_{t_{\max}}$, then we distribute the whole b_i) on the entry I_{im}, so that we preserve the maximum on this last column to be $b_{t_{\max}}$. This way, we can assure that, among the other columns, we'll have to distribute a smaller fraction of b_i, which means that the maximum on each column between 1 and $m-1$ will decrease, and so will (5.5).

The problem reduces to find the optimal sub-matrix $I'_{(k-t)\times(m-1)}$ to complete the first $k-t$ rows of I, and with the aforementioned remark, we can actually consider I' to be the solution given the updated sequence $(b'_i)_{1\leq i\leq k-t}$ which is defined, for every i such that file e_i that can be padded to s_m, as either 0, if $b_i \leq b_{t_{\max}}$, or as $b_i - b_{t_{\max}}$, if $b_i \geq b_{t_{\max}}$. When we reconstruct the matrix I, on the m-th column we will have the value $I'_{im} + b_{t_{\max}}$ or $I'_{im} + b_i$ (depending on whether b_i is smaller, respectively larger, than $b_{t_{\max}}$).

Now, let us show that, for the sub-matrix I', we have 0's on every entry of the m-th column. By definition, I' must be a Rényi optimal solution for the updated sequence of b'_i's. Using Proposition 36, there exists a Rényi optimal padding-scheme f' which maps $e_i, i \in [1..k-t] \to \{s_1, \ldots, s_{k-t}\}$, for any set of files $\{e_1, \ldots, e_{k-t}\}$ with the associated frequencies $\{b'_1, \ldots, b'_{k-t}\}$. Consequently, for every $i \in [1..k-t], \mathbb{P}(f'(e_i) = s_m) = 0 \Rightarrow I'_{im} = 0$. □ □

Algorithm 11 Per-request-padding pseudocode.

```
procedure RÉNYI PRP
    ∀i, b_i ← p_i                                   ▷ budget array
    I ← Joint prob. matrix of zeros
    for j=m, m-1, ..., 1 do
        t_max = arg max_{i | |e_i|=s_j} b_i
        if b_{t_max} > 0 then
            j_min, j_max ← range of files e_{j_min}..e_{j_max} that can be padded to s_j
            for i = j_max, j_max − 1, ..., j_min do
                I[i, j] ← min(b_{t_max}, b_i)
                b_i = b_i − I[i, j]
    P ← channel matrix after dividing each row i of I by p_i
    return P
```

Therefore, we have proved that the matrix I can be recursively expressed using the sub-matrices obtained when we update the budget sequence ac-

cordingly, at each step decreasing by 1 the number of columns and by at least 1 the number of rows of the matrix returned from the algorithm, until we reduce a problem to finding the Rényi optimal scheme for a budget sequence with a single element. Since we want to minimize (5.5) in the case of n files with frequencies $\{p_1, \ldots, p_n\}$ and the associated set of sizes $\{s_1, \ldots, s_m\}$, we proceed the induction on the number of rows and columns as described in Remark 39 and eventually fill in all the entries of the solution $I_{n\times m}$. The channel-matrix P is then computed as $p_{ij} = I_{ij}/p_i$, and this is the output of `PrpRe`.

This algorithm is presented in Algorithm 11 in the form of pseudocode, and it is implemented in [PPS22] with some optimizations.

Bandwidth minimization

Once `PrpRe` has found a channel matrix that minimizes Rényi leakage, it is still possible to use heuristics to search for other channel matrices with the same (minimal) leakage but with less bandwidth use. We call `PrpReBa` to be the algorithm that runs `PrpRe` and the bandwidth reduction heuristics afterwards.

Let the list C of maximums on each column after running `PrpRe`, i.e., $C = \{\max_{i\in[1..n]} I_{ij} | j \in [1..m]\}$, where $C_j = \max_{i\in[1..n]} I_{ij}$ for every $j \in [1..m]$. Define a *move* to be a change in the matrix I performed on two of the entries of the matrix at line i, for some $i \in [1..n]$ such that (I_{ia}, I_{ib}) becomes $(I_{ia} - \alpha, I_{ib} + \alpha)$ while keeping the entries of I positive, i.e., $\alpha \leq I_{ia}$.

Now, we will describe an *update* on the line i, which will consist of a series of *moves* and will return a new matrix I^*. We start with I^* to be the matrix I, but with 0's on the i-th line. Since the sum on row i is equal to p_i, we start with this quantity and go through the columns in order from $j = 1$ to $j = m$. For each column, we set:

$$I_{ij} = \begin{cases} C_j & \text{if } C_j + \sum_{k=1}^{j-1} I_{ik} \leq p_i \\ p_i - \sum_{k=1}^{j-1} I_{ik} & \text{otherwise} \end{cases}$$

5.4 Experiments and Comparison

Several experiments were carried out for three distinct purposes, namely, (1) to test the correctness of the implementations against brute-force algorithms for small sized problems, (2) to corroborate the direct link between Rényi leakage and the success rate of an attacker and (3) to compare the runtime, bandwidth and leakages of all the algorithms on a public dataset. The code of all the experiments is available in [PPS22].

5.4.1 Brute-force tests for correctness

To complement and corroborate the theory developed in this chapter, all the algorithms were tested against brute-force implementations for small datasets (with at most 10 elements). More precisely, for each randomly generated test case of file sizes and frequencies, we explored (exhaustively) all the POP padding schemes satisfying the constraints, and chose among them, the ones that minimized Rényi leakage, Shannon leakage or bandwidth, with the purpose of comparing them with the solutions returned by our algorithms.

We ran ten thousand experiments (code available in [PPS22]), all corroborating that: among all POP schemes, `PopRe` achieves minimal Rényi leakage, `PopSh` achieves minimal Shannon leakage, and `PrpRe` leaks at most the Rényi leakage of `PopRe`.

5.4.2 Attacker test for illustration

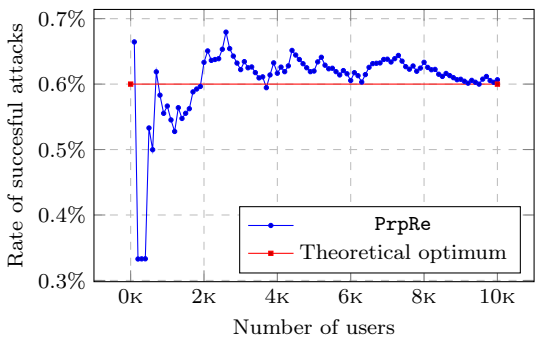


Figure 5.3: Attacker's success convergence.

We simulated the attacker described in this chapter by Equation (5.1), who always guesses the original file with maximum probability given the priors and the padding scheme. Figure 5.3 shows that as the number of user increases, the success rate of the attacker against the padding proposed by `PrpRe` approaches the expected theoretical minimal pos-

sible success rate. This is a direct consequence of the law of large numbers as well as the equivalence between minimizing the expected success of the attacker (5.1) and the Rényi leakage, via Eq. (5.5).

5.4.3 Dataset tests for comparison

We used the dataset of NodeJS, proposed originally in [RR21]. This dataset consists of a list of 423,450 JavaScript packages provided by NPM for browser and nodeJS applications, each with its associated file size and access frequency, as of August 2021. Taking into account the large number of files and the availability of the access frequencies, we used the NodeJS dataset to benchmark the algorithms.

We used two versions of the NodeJS dataset: the *large* NodeJS dataset is the original dataset with 423,450 files, and the *small* consists of only the 1000 most frequently accessed files. The small NodeJS dataset allowed us to benchmark and compare the algorithms with large complexity, which timed-out on the large dataset. In all experiments, we parametrize the padding constraints with a single constant $c > 0$ that represents the constraint $|X| \leq f(X) \leq (1+c) \cdot |X|$.

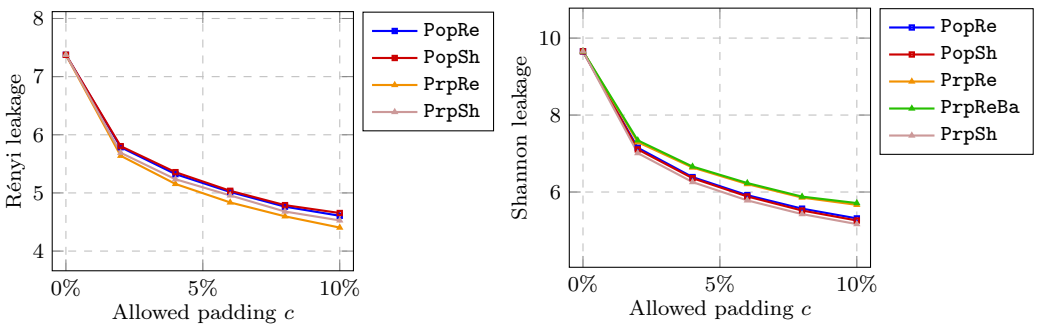


Figure 5.4: Rényi and Shannon leakage on the small dataset.

Figure 5.4 depicts the variation of privacy leakage as a function of c on the small dataset. The trend is approximately equal in the large dataset, except that `PopRe` times out. The Rényi plot does not include `PrpReBa` to reduce redundancy, as it coincides with `PrpRe`. In the figure, we can appreciate the expected trend that larger c allows for more padding and less leakage of privacy, both in Rényi and Shannon definitions. It can also be verified

that the algorithms tuned to minimize Rényi leakage, have a very small (but not minimal) Shannon leakage, and vice-versa. For instance, the differences between `PopRe` and `PopSh` in both leakages are inferior to 2%. This is a consequence of the information theoretical connection between the two types of leakage.

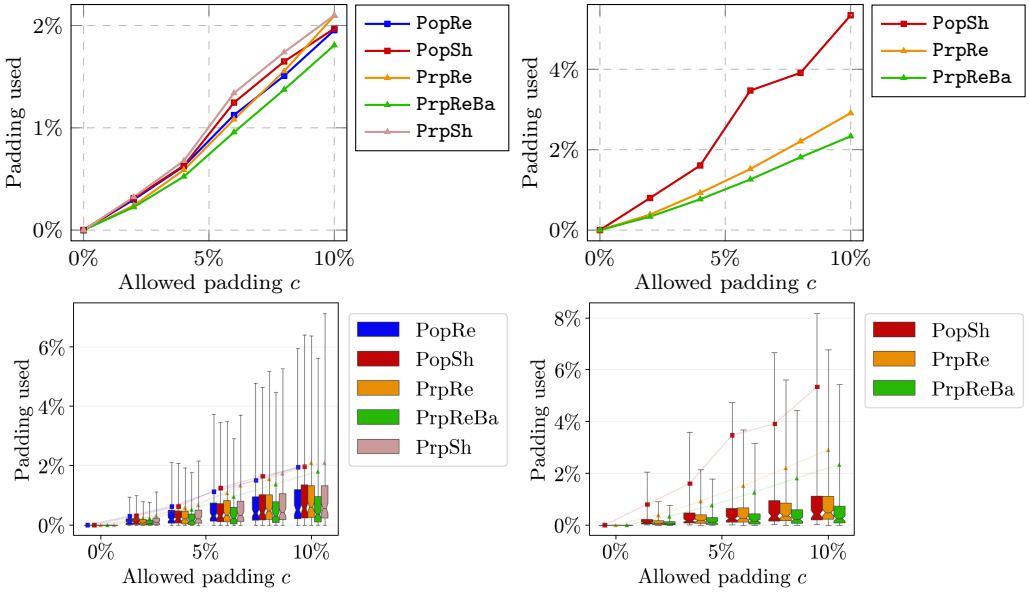


Figure 5.5: Bandwidth increase on the small (left) and large (right) datasets. The top plots show expected values and the bottom plots show, in addition, confidence intervals for a single random request. For each box, the body (Q1 and Q3 quartiles) corresponds to 50% confidence and the whisker (5% and 95% quantiles) to 90%.

The bandwidth increase generated by the padding of the files can be analyzed in Figure 5.5. For reference, the average file size in the dataset, weighted by frequency is 52.5 KB, so 1% increase, means around 5.3 additional kilobytes. Several observations can be made out of Figure 5.5. First, as anticipated, the larger the c, the larger the paddings on average. Second, the algorithms do not pad as much as they are allowed. Instead, when 10% is allowed, the optimal paddings lie at around 2% for the small dataset and 4% for the large dataset. For this particular example, the algorithms used more of the available padding on the large than in the small dataset, but we did not explore in depth in our experiments whether this pattern holds in general. Third, the

improvements of `PrpReBa` over `PrpRe` can be corroborated, and estimated to approximately 20% less bandwidth use with the same Rényi leakage. Lastly, it appears empirically that the solutions that minimize Rényi leakage use less padding on average than those that minimize Shannon leakage.

Furthermore, the box plots in Figure 5.5 show that the padding use (with respect to the average file size) is most often below its average, meaning that there are a few files that contribute significantly more than the others to bandwidth excess. These files must be the largest, as they are the files for which the additional bandwidth can be the largest compare, even possibly exceeding the average file size. Note that the computation of confidence intervals can not be made for privacy leakages (Figure 5.4), as they are global guarantees of privacy that do not make sense for individual files.

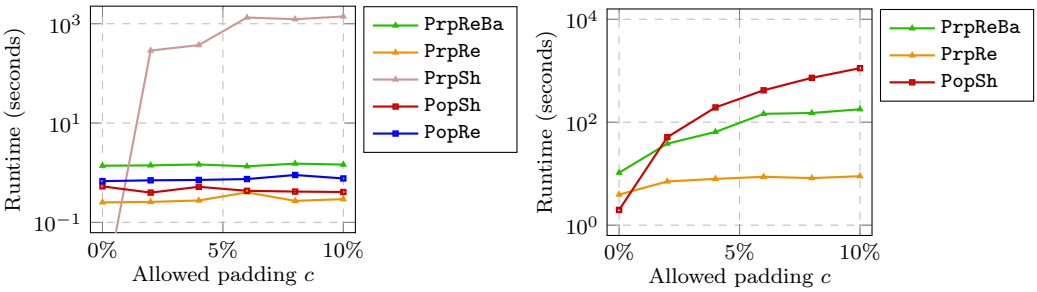


Figure 5.6: Runtime plots on small (left) and large (right) datasets. The plots ignore the 7 additional seconds needed for JIT compilation.

Figure 5.6 depicts the runtime of the algorithms under analysis. We refer the reader to Table 5.1 about the runtime complexities for a richer analysis of the plots. The analysis could have been even richer, if we included confidence intervals as in Figure 5.5, but we did not do it because of time constraints (the execution of `PopSh` in the large dataset is in the order of several hours of CPU time), and because the added value in this case is very little, as we already have a theoretical derivation of the complexities.

The left plot in Figure 5.6 does not have a clear tendency of longer executions for more relaxed padding constraints (higher c, thus also higher $\bar{b}$), meaning that for small datasets, all algorithms are suitable. In this regime, the runtime is not yet affected significantly by the growth of $\bar{b}$, possibly due to large constants that are masked by the complexity class and implementation

details, especially for PrpReBa. Nevertheless, the difference between PopRe versus PrpRe and PopSh is already visible, and indeed, PopRe times out (several hours) for the large dataset. The right plot highlights the scalability of the algorithms. For all values of c plotted in this graph, the runtime for PrpRe is under 7 seconds, which makes it the fastest algorithm. PrpReBa peaks at $c = 10\%$ with around 3 minutes while PopSh needed 15 minutes. In this regime, the effect of increasing $\bar{b}$ via c on the runtime is clear.

5.5 Conclusion

We designed and proved the optimality of several algorithms (PopRe, PrpRe, PrpReBa) that minimize the expected success rate of an attacker. The algorithms were compared with existing solutions (PopSh, PrpSh) that consider a different attack model. The comparison was done both numerically via experiments and theoretically via privacy leakage.

Prioritizing scalability, we recommend using either PrpRe or PrpReBa for the PRP problem, as they are much faster and provide protection against a more reasonable attacker than the existing solutions (PopSh, PrpSh). Nevertheless, for the POP problem, we recommend any of either the existing solution PopSh or our algorithm PopRe that minimizes Rényi leakage, because even though our attack model is more realistic, the complexity of PopSh makes it more practical.

In general terms, the two attack models are correlated in the sense that the optimizing against one of them results in a strong, though not optimal, protection against the other one (with empirical differences of less than 2%). In more detail, however, the Rényi attacker is more realistic than the Shannon attacker, and the padding schemes that minimize Rényi leakage seem to use less bandwidth in practice, making our proposed algorithms even more appealing.

This chapter concludes the main body of the manuscript. In the next chapter we discuss the main results of each chapter along with a critical analysis that justifies the relevance, limitations and future work for each of them.

Chapter 6

Discussion and future work

In this section, we explain and discuss the main takeaways from Chapters 2, 3, 4 and 5 justifying their relevance and the prospects for future work.

In Chapter 2, we proved (Theorem 11) that for certain biased distributions it is impossible to provide equal opportunity (EO) guarantees without degrading accuracy to trivial levels. This theorem improves the theoretical understanding of the notion of equal opportunity and its impact on accuracy.

For other fairness notions like statistical parity (e.g., 50% women, 50% men), the potential incompatibility between fairness and accuracy is less surprising because it forces the promotion of candidates from all the protected groups, which can hinder accuracy if there are no qualified candidates in one of them. For EO, however, this result is counterintuitive because EO was designed to improve its trade-off with accuracy on the first place.

The full potential of our Theorem 11 of incompatibility between accuracy and equal opportunity can be appreciated in statistical learning theory by rewriting it as Theorem 40 below.

Theorem 40. *No machine learning training algorithm can guarantee for all data distributions of* (X, A, Y) *with* $\mathbb{P}[\,A{=}a, Y{=}y\,] > 0$ *for all* $a, y \in \{0, 1\}$*, that the trained model* $\hat{Q}$ *satisfies probably and approximately*[1] *as* $n \to \infty$*,*

[1]Meaning that any level of confidence < 1 and closeness > 0 can be guaranteed with a

that $oppDiff(\hat{Q}) \to 0$ *and* $err(\hat{Q}) \to \alpha$ *for some non-trivial value* $\alpha < \tau = \min\{err(\hat{0}), err(\hat{1})\}$.

Proof. Let's assume from Theorem 11 that the population (X, A, Y) follows a distribution for which the feasibility region M is a polygon not containing any combination of error and opportunity difference of the form $(\beta, 0)$ with $\beta < \tau$, which includes $(\alpha, 0)$. Let $\hat{Q}_1, \hat{Q}_2, \ldots$ be a stochastic sequence of predictors obtained by training with a fixed learning algorithm on datasets of increasing size. By means of contradiction, suppose that the error and opportunity difference of this sequence approaches probably and approximately to the combination $(\alpha, 0)$. Then, there are sequences of points in M that approach $(\alpha, 0)$, making $(\alpha, 0)$ an accumulation point of M. Since the region M is a convex set that contains its convex hull (this follows from Theorem 9), it must be closed, hence it contains all its accumulation points, including $(\alpha, 0)$, which contradicts the initial assumption. □

This statement is similar in spirit to an existing fundamental theorem in statistical learning theory [Vap99] called the no-free-lunch theorem [SSBD14] for ML, which states that for infinite domains, no single training algorithm can guarantee probably approximately optimal classifiers. However, the causes behind this fundamental theorem and Theorem 40 are very different. In our result, the limitation arises from some particular highly biased distributions for which the bias can not be fixed with any classifier. In the no-free-lunch theorem for ML, the limitation comes from the combinatorial explosion in the number of problems that should be solved using a single training algorithm. Notice also that the no-free-lunch theorem is about not being able to achieve *maximal* accuracy after the learning process, while our result is about not being able to achieve a (merely) non-trivial level of accuracy rule when EO is enforced.

We complemented our incompatibility result with sufficient and necessary conditions that characterize the data distributions for which the incompatibility occurs. In addition, we also provided an algorithm for visualizing the Pareto frontier between error and opportunity difference and proved connections with existing ways of visualizing and understanding the trade-off.

sufficiently large (or larger) n

This algorithm is designed for discrete distributions over populations with a relatively small number of individuals, and it can find the most accurate classifier that satisfies EO, as well as enumerating all the deterministic classifiers that lie on the Pareto-optimal boundary. The probabilistic classifiers in this boundary are infinite, but they can all be expressed as linear combinations of two consecutive optimal deterministic classifiers in the enumeration, hence our algorithm does find all the classifiers in the Pareto frontier.

For larger and finite as well as infinite domains, an approximation algorithm based on Algorithm 3 can be used (setting T_0 and T_1 to be 1D grids of $[0, 1]$ and assuming some estimators for error and oppDiff instead of the exact values), and this would be very similar to the existing algorithm proposed by [HPS16] from which Algorithm 3 was inspired. The key differences are that their algorithm is tuned to find the optimal classifier as precisely as possible, while our algorithm is tuned to find all (up to an approximation grid) optimal classifiers with a certain level of precision.

The main limitation of these results is the extent to which the incompatibility applies in real life settings. We detailed on how the theorem can have an effect in practical scenarios by means of an example, but as we show, the bias conditions are very strict, making our theorem a result of relatively low practical impact (it is strong but rarely applies) but very high theoretical interest.

Also, it is worth mentioning that our analysis focuses on *exact* equal opportunity. Nevertheless, the insights that we derived about the geometry of the feasibility region can be used to deduce that, regarding approximate equal opportunity, the incompatibility theorem does not hold anymore. More precisely, there is always compatibility in that case, unless the Bayes classifier is constant, which is a pathological and dull scenario.

Taking into account that this work is comprehensive enough, we consider it unnecessary to extend it further. Instead, future prospects and possibilities could be to carry out the same analysis for different group fairness notions, excluding exact statistical parity, a well studied [EGGL20, GDBFL19, BdBG+22, GLR20] fairness notion, for which the classifier with minimal error is obtained by thresholding the quantile of the individuals in their sensitive group, a condition that reduces accuracy severely when a sensitive group is at great

disadvantage with respect to the other.

In Chapter 3, we illustrated the strong dependency of the causal graph on the choice of the causal discovery algorithm (CDA). This claim reinforces the idea presented in [BMP+23], that fairness metrics that rely on the causal graph depend on the choice of the CDA used for its estimation.

Causal based fairness notions are ideal for fairness assessment as they take into account the actual causal effect of the sensitive attribute on the decision instead of mere correlations. But most of these notions require the causal graph that best describes the data generation process to be known. However, causal graphs are not known in general for arbitrary populations and tasks, and it could be a matter of debate between experts whether a given graph is indeed the correct one. This is apparently one advantage of CDAs, compared to experts, that they are unsupervised learning algorithms and select one causal graph with the least subjectivity possible. But there are several CDAs and parameters to choose, and as we show in Chapter 3, it can happen that using one CDA or another leads to very different outputs.

The surprising fact shown in Chapter 3 is that this can happen with very simple distributions, and the differences between the output graphs are not simply missing edges (whose solution would be to modify a threshold parameter), but also reversed edges. These conflicts between the algorithms is not very common, indeed the search for them required genetic algorithms, but their existence questions the causal principles, and the use of causal vocabulary for arbitrary graphical models, a fact that has been warned and criticized before for models that produce DAGs [KS22].

The impact of the experiments in Chapter 3 is particularly intense if the causal graph is used as input to a fairness assessment. As we showed in [BMP+23] and supported with the experiments in this chapter, there is a potential error propagation in this pipeline that can lead to a fairness misconclusion. For this reason, we conclude that causal based fairness assessment should not be fully automated unless the causal graph is known.

Although we provided a detailed technical overview on how several CDAs work, and this serves as an explanation of why different CDAs may produce different causal graphs, it would be ideal to have a stronger conceptual

framework to explain this phenomenon in detail. Meanwhile, as a practical solution to the low level of trust that can be put on each CDA, we propose to run several of them and decide whether there is consensus. Subsequent studies could focus on how to proceed when there is no consensus between the algorithms.

In Chapter 4, we proposed a protocol named LOLOHA that uses local hashing for longitudinal categorical data collection with local differential privacy guarantees. The protocol consists of a frequency estimation algorithm used by a central server and a client-side sanitization algorithm used by each user. The main property of LOLOHA is that it maps the private input data of each user into a very small domain, which makes the collected information very uninformative (protective) about the exact private values of the users.

LOLOHA has two crucial features. First, LOLOHA has increased privacy on the users values with respect to state-of-the-art algorithms, while having similar levels of error and LDP guarantees for the first report. Second, unlike the existing longitudinal data collection protocols in the literature that rely solely on unary encoding, LOLOHA stands out by using local hashing. As a result, LOLOHA expands the diversity of techniques for longitudinal data collection —a crucial aspect for fostering the development of future ideas in the field.

The main caveat of LOLOHA is that local hashing allows the server to track the users based on the hash salts that they use. To put it differently, the server knows the sequence of reports sent by each user, but these reports are guaranteed to be sanitized, so that the server can not ever be certain about the true values. Even though the protocol guarantees LDP on the users' values, which is a strong guarantee of privacy, tracking might be seen as a privacy disadvantage by the users. As a result, an interesting area for further investigation is how to modify the protocol to prevent tracking by combining ideas from the Shuffle DP model [BEM+17, EFM+19, EFM+20, JGAP23].

Privacy *on the users' values* is a notion of privacy that we defined for this specific setting, because pure local differential privacy is too strong for regular data collection. As we proved in Theorem 29, Chapter 4, it is impossible to satisfy pure LDP in this setting. As a result, the main direction for future work in this area is to find consensus on a single notion of privacy

for longitudinal settings. This can be done by exploring more deeply the definition we proposed, or proposing new ones.

We support our definition because the obfuscation of the data sequence of each user is very strong, and extremely uninformative for small g, but we acknowledge that attacks with sufficient external knowledge can infer some entries about the memoization cache, and break the LDP guarantees. For instance, if the attacker happened to knew your true value from external sources for several days and observed the obfuscated data you reported, he may use this knowledge to predict your true data in the future, although never with 100% confidence, if you report similar obfuscated data again.

The discussion of what privacy notion suits best is not a simple one [Ngu14]. As shown in Chapter 4, even though local differential privacy theoretically rich and very protective for the users, it is too strong to be used with streams of data without losing utility all along. Alternative solutions that we did not consider could be to let the user tune their actual privacy expectations [CKR21] versus the statistical quality incentives they may receive in return [CGL15], or even to monetize them [LLMS14, BJP21, JP19]. It would be interesting to explore a definition of LDP that takes some of these additional dimensions into account, and to find its corresponding protocols.

Lastly, in Chapter 5, we proposed a method for designing padding schemes that enhance privacy of transmitted data at the expense of some bandwidth overhead. We consider a set of users, each of which is choosing and downloading one file out of a central pool of public files, and an attacker that observes the download size for each user to identify the choice of each user. Padding is used to obfuscate the information an attacker may obtain about the data based on its size.

The main contribution of our work is that, unlike previous work that protected against an information theoretical attacker by minimizing Shannon entropy, our padding schemes minimize Rényi min-entropy, which represents a more practical attacker. Concretely, Rényi min-entropy corresponds to an attacker who makes a guess about the private file of the user and wants to maximize the probability of being correct, while Shannon entropy corresponds to an attacker that can ask binary questions about the file to an oracle and wants to minimize the number of questions needed to guess the

file with entire certainty. So, although both entropy measures are mathematically rich and correspond axiomatically to information attackers [ACM+16], for its simplicity, our attacker is central in privacy research [CCPT20], and it has been used as reference to report security/privacy attack tests in the real world [CJT22].

We also proved that our padding schemes achieve the global minimum of Rényi min-entropy, and tested this fact extensively as well with virtual experiments.

Our analysis can be extended in two different ways. On the one hand, our results are limited to the scenario in which each user downloads exactly one file and the attacker measures very precisely the amount of transferred data, which can be restrictive assumptions. In practice, data is split into packets of fixed size and the attacker obtains information about the size of the downloaded file indirectly, based on the number of transferred packets and the time delays between them, so if a user makes two consecutive downloads, the attacker will need the timing information of the packets to be able to distinguish them. As a consequence, a future direction of this work would be to explore the same problem under multiple downloads and exploiting timing information.

On the other hand, our solution to the double minimization problem of leakage and bandwidth consists of setting a bandwidth constraint and minimizing leakage. Conversely, the problem of setting a leakage constraint and minimizing bandwidth could be studied as well, or more generally, the problem of finding all Pareto optimal solutions.

Overall, this manuscript describes a variety of new results in the field of data ethics, some related to fairness in machine learning with binary sensitive features and binary outcomes (Chapters 2 and 3) and the others to the collection and transmission of private data (Chapters 4 and 5). More precisely, this manuscript presents general theoretical theorems, relevant observations and proposes several algorithms with proofs of correctness and optimality.

We highlight the following claims as they are very general theorems or interesting observations for the theory of fairness and privacy: Theorems 9, 19 and 11, Fact 18, Example 1 and the example of section 2.8 from Chap-

ter 2; Theorem 29 (impossibility of LDP for regular data collection) from Chapter 4; Theorem 35 (unsuitability of differential privacy for the padding problem) from Chapter 5; and Examples 23, 24 and 25 from Chapter 3.

Regarding algorithms, this manuscript proposes several algorithms. Algorithms 2 and 3, listed in Table 2.2, compute the Pareto boundary of accuracy and equal opportunity; Algorithms 8 and 9 define the LOLOHA protocol; and Algorithms 10 and 11 compute optimal padding schemes. All of these algorithms are equipped with proofs of correctness and complexity. For LOLOHA, we also prove the privacy-utility bounds, and for Rényi POP and PRP algorithms we prove that bandwidth is minimized.

The chapters have some slight relationships, although they are not extensions or particularizations of the others. For instance, Chapters 4 and 5 are about privacy, they deal with inherent trade-offs between quality of service and privacy, and they provide practical solutions.

All chapters have in common that they highlight the existence of unnoticed properties in the theory and practice of data ethics. Chapters 2 and 3, which are on fairness, are perhaps more provocative in this regard as they contradict the intuition. Also, the first three chapters have an ethical baseline trade-off: be it accuracy versus fairness, statistical precision of collected data versus privacy, or network bandwidth versus privacy; and the last three chapters propose a solution to an existing challenge: be it collecting data while guaranteeing some privacy, padding files for privacy without increasing network bandwidth over some thresholds, or deciding whether a causal graph is trustable (by running several CDAs instead of choosing a single one).

Chapter 7

Summary

This dissertation explores four important challenges in the field of data ethics, namely, the trade-off between equal opportunity and accuracy of machine learning classifiers, the regular collection of private data, the leakage of information during data transmission, and the use of causal discovery algorithms for fairness assessment. Its goal is to contribute in the development of theoretical insights and practical tools that can foster the state of the art of fairness in machine learning and privacy in data collection and transmission.

In Chapter 2, we proved a theorem that improves our theoretical understanding of the notion of equal opportunity and its impact on the accuracy of machine learning models. We proved that for certain highly biased distributions it is impossible to provide equal opportunity guarantees without degrading accuracy. This theorem was then refined, and we found necessary and sufficient conditions that characterize the data distributions for which this extreme trade-off occurs. Although these conditions are very severe to be found in arbitrary datasets, they may hold in practical scenarios, and we illustrated this with an example. We also provided an algorithm for visualizing the Pareto frontier between error and opportunity difference and proved connections with existing ways of visualizing and understanding the trade-off. Perhaps the same analysis could be carried out with other fairness notions.

In Chapter 3, we illustrated the strong dependency of the causal graph on

the choice of the causal discovery algorithm (CDA), reinforcing the idea that causal based fairness metrics depend on the choice of the CDA used for the estimation of the causal graph [BMP+23]. We created several examples of very simple distributions for which different CDAs disagree very probably and drastically on their outputs, a fact that has a dramatic impact on causal based fairness assessment because the fairness conclusions are very sensitive to the structure of the causal graph. For this reason, we conclude that causal based fairness assessment should not be fully automated unless the causal graph is known. Although we provided a detailed technical overview on how several CDAs work, and this serves as an explanation of why different CDAs may produce different causal graphs, it would be ideal to have a stronger conceptual framework to explain this phenomenon in detail, and the challenge of how to proceed when there is no consensus between the algorithms remains open.

In Chapter 4, we proposed a protocol named LOLOHA for longitudinal categorical data collection with local differential privacy guarantees. LOLOHA is based on local hashing unlike the main state-of-the-art protocols, which are based on unary encoding, and it provides increased privacy on the users' values while having similar levels of error and LDP guarantees for the first report. Privacy on the users' values is a notion of privacy that we defined for this specific setting, because, as we prove, it is impossible to satisfy pure local differential privacy for arbitrarily large windows of data collection. The main future work in this area is to find consensus on a single notion of privacy for longitudinal settings, and to find the protocols for it. This can be done by exploring more deeply the definition we proposed, or proposing new ones.

In Chapter 5, we proposed a method for designing padding schemes that enhance privacy of transmitted data at the expense of some bandwidth overhead. In this context, one party is sending a file or a message from a known set to another trusted party, and padding is used to obfuscate the information an attacker may obtain about the data being shared based on its size. The main contribution is that our padding schemes minimize Rényi min-entropy instead of Shannon entropy, which makes the attack model closer to a real life attacker. However, to apply our methods into a wider range of scenarios, it is necessary to explore more complex cases in which several files are sent

and timing information is used by the attacker.

In sum, this dissertation presents four published articles in the field of data ethics that complement the state of the art of data privacy and fairness by providing practical algorithms and theoretical insights.

Résumé général en français

Cette thèse explore quatre défis importants dans le domaine de l'éthique des données, à savoir, (1) le compromis entre l'égalité d'opportunité et l'exactitude des classificateurs d'apprentissage automatique, (2) l'utilisation d'algorithmes de découverte causale pour l'évaluation de l'équité, (3) la collecte régulière de données privées, et (4) la fuite d'informations lors de la transmission de données. Son objectif est de contribuer au développement de aperçus théoriques et d'outils pratiques qui améliorent notre connaissance en matière d'équité dans l'apprentissage automatique et l'état de l'art en matière de la confidentialité dans la collecte et transmission de données.

Pour le premier défi, nous avons prouvé un théorème qui améliore notre compréhension théorique de la notion d'égalité d'opportunité (equal opportunity—EO) et de son impact sur l'exactitude des modèles d'apprentissage automatique. Nous avons prouvé que pour certaines distributions fortement biaisées, il est impossible de garantir l'EO sans dégrader l'exactitude. Ce théorème a ensuite été affiné et nous avons trouvé des conditions nécessaires et suffisantes qui caractérisent les distributions de données pour lesquelles ce compromis extrême se produit. Bien que ces conditions soient très strictes pour être trouvées dans des ensembles de données arbitraires, elles peuvent être valables dans des scénarios pratiques et nous l'avons illustré avec un exemple. Nous avons également fourni un algorithme pour visualiser la frontière de Pareto entre l'erreur et la différence d'opportunité, et nous avons prouvé des liens avec les méthodes existantes de visualisation et de compréhension du compromis.

Pour le deuxième, nous avons illustré la forte dépendance du graphe causal du choix de l'algorithme de découverte causale (CDA), renforçant l'idée que les mesures de l'équité basées sur la causalité dépendent du choix du CDA utilisé pour l'estimation du graphe causal. Nous avons créé plusieurs exem-

ples de distributions très simples pour lesquelles différents CDAs divergent très probablement et radicalement dans leurs résultats, un fait qui a un impact dramatique sur l'évaluation de l'équité basée sur la causalité car les conclusions en matière d'équité sont très sensibles au choix du CDA. Pour cette raison, nous concluons que l'évaluation de l'équité basée sur la causalité ne devrait pas être entièrement automatisée à moins que le graphe causal ne soit connu. Bien que nous ayons fourni un aperçu technique détaillé du fonctionnement de plusieurs CDAs différents, et que cela serve d'explication à la raison pour laquelle différents CDAs peuvent produire différents graphes causaux, il serait idéal d'avoir un cadre conceptuel plus solide pour expliquer ce phénomène en détail, et le défi de savoir comment procéder lorsqu'il n'y a pas de consensus entre les algorithmes reste ouvert.

Pour le troisième défi, nous avons proposé un protocole appelé LOLOHA pour la collecte de données catégorielles longitudinales avec des garanties de confidentialité différentielle locale (local differential privacy—LDP). LOLOHA est basé sur le hachage local contrairement aux protocoles principaux de l'état de l'art, qui sont basés sur le codage unaire, et il offre une confidentialité accrue sur les valeurs des utilisateurs tout en ayant des niveaux d'erreur et des garanties de LDP similaires pour le premier rapport. La confidentialité sur les valeurs des utilisateurs est une notion de confidentialité que nous avons définie pour ce cadre spécifique, car, comme nous le prouvons, il est impossible d'avoir de la LDP pure dans la collecte régulière de données. Le principal travail futur dans ce domaine consiste à trouver un consensus sur une seule notion de confidentialité pour la collecte longitudinal. Cela peut être fait en approfondissant la définition que nous avons proposée ou en en proposant de nouvelles.

En dernier, pour le quatrième défi, nous avons proposé une méthode pour concevoir des schémas de remplissage qui améliorent la confidentialité des données transmises au détriment d'une certaine surcharge de bande passante. Dans ce contexte, une partie envoie un fichier ou un message d'un ensemble connu à une autre partie de confiance, et le remplissage est utilisé pour masquer les informations qu'un attaquant peut obtenir sur les données partagées en fonction de leur taille. La principale contribution est que nos schémas de remplissage minimisent la min-entropie de Rényi au lieu de l'entropie

de Shannon, ce qui rend le modèle d'attaque plus proche d'un attaquant réel. Cependant, pour appliquer nos méthodes à un plus large éventail de scénarios, il est nécessaire d'explorer des cas plus complexes dans lesquels plusieurs fichiers sont envoyés et des informations temporelles sont utilisées par l'attaquant.

Bibliography

[ABCP13] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

[ACABX21] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 47–57, New York, NY, USA, 2021. Association for Computing Machinery.

[ACBX21] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Longitudinal collection and analysis of mobile phone data with local differential privacy. In Michael Friedewald, Stefan Schiffner, and Stephan Krenn, editors, *Privacy and Identity Management*, pages 40–57, Cham, 2021. Springer International Publishing.

[ACBX22] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. *Digital Communications and Networks*, 2022.

[ACC+20] Héber H Arcolezi, Jean-François Couchot, Selene Cerna,

Christophe Guyeux, Guillaume Royer, Béchara Al Bouna, and Xiaokui Xiao. Forecasting the number of firefighter interventions per region with local-differential-privacy-based data. *Computers & Security*, 96:101888, 2020.

[ACG+22] Héber H. Arcolezi, Jean-François Couchot, Sébastien Gambs, Catuscia Palamidessi, and Majid Zolfaghari. Multi-freq-ldpy: Multiple frequency estimation under local differential privacy in python. In Vijayalakshmi Atluri, Roberto Di Pietro, Christian D. Jensen, and Weizhi Meng, editors, *Computer Security – ESORICS 2022*, pages 770–775, Cham, 2022. Springer Nature Switzerland.

[ACM+16] Mário S Alvim, Konstantinos Chatzikokolakis, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith. Axioms for information leakage. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 77–92. IEEE, 2016.

[ACPS12] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *Proceedings of the 25th IEEE Computer Security Foundations Symposium (CSF)*, pages 265–279, 2012.

[Aga20] Sushant Agarwal. Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo, 2020.

[AGCP22] Héber H. Arcolezi, Sébastien Gambs, Jean-François Couchot, and Catuscia Palamidessi. On the risks of collecting multidimensional data under local differential privacy. *arXiv preprint arXiv:2209.01684*, 2022.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

[AM+20] Mário S. Alvim, Konstantinos Chatzikokolakis 0001, Annabelle McIver, Carroll Morgan, Catuscia Palamidessi, and Geoffrey Smith 0001. *The Science of Quantitative Information Flow*. Information Security and Cryptography. Springer, 2020.

[AMP23] Héber H Arcolezi, Karima Makhlouf, and Catuscia Palamidessi. (local) differential privacy has no disparate impact on fairness. *arXiv preprint arXiv:2304.12845*, 2023.

[App17] Apple Differential Privacy Team. Learning with privacy at scale, 2017. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, (accessed January 2023).

[APPG23] Héber Hwang Arcolezi, Carlos Pinzón, Catuscia Palamidessi, and Sébastien Gambs. Frequency estimation of evolving data under local differential privacy. In *EDBT 2023-26th International Conference on Extending Database Technology*, pages 512–525, 2023.

[ARC18] Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Scoring bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1):3–18, 2018.

[ASZ19] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 16–18 Apr 2019.

[AVG+18] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *NDSS 2018-Network and Distributed System Security Symposium*, pages 1–15, 2018.

[BdBG+22] Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-

Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198, 2022.

[BEM+17] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP '17, page 441–459, New York, NY, USA, 2017. Association for Computing Machinery.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[BHMR17] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21, 2017.

[BJP21] Sayan Biswas, Kangsoo Jung, and Catuscia Palamidessi. An incentive mechanism for trading personal data in data markets. In *Theoretical Aspects of Computing–ICTAC 2021: 18th International Colloquium, Virtual Event, Nur-Sultan, Kazakhstan, September 8–10, 2021, Proceedings 18*, pages 197–213. Springer, 2021.

[BMP+23] Rūta Binkytė, Karima Makhlouf, Carlos Pinzón, Sami Zhioua, and Catuscia Palamidessi. Causal discovery for fairness. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, pages 7–22. PMLR, 2023.

[BNST17] Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Thakurta. Practical locally private heavy hitters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2285–2293, Red Hook, NY, USA, 2017. Curran Associates Inc.

[BP22] Sayan Biswas and Catuscia Palamidessi. Privic: A privacy-preserving method for incremental collection of location data, 2022.

[BPS19] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[BS15] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA, 2015. Association for Computing Machinery.

[CCPT20] Konstantinos Chatzikokolakis, Giovanni Cherubin, Catuscia Palamidessi, and Carmela Troncoso. The bayes security measure. *arXiv preprint arXiv:2011.03396*, 2020.

[CD18] Rachel Cummings and Deven Desai. The role of differential privacy in gdpr compliance. In *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 20, 2018.

[CDG18] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *ArXiv*, abs/1808.00023, 2018.

[CGKM19] 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.

[CGL15] Michela Chessa, Jens Grossklags, and Patrick Loiseau. A short paper on the incentives to share private information for population estimates. In *Financial Cryptography and Data Security: 19th International Conference, FC 2015, San Juan, Puerto Rico, January 26-30, 2015, Revised Selected Papers 19*, pages 427–436. Springer, 2015.

[CH20] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[Che17] Giovanni Cherubin. Bayes, not naïve: Security bounds on website fingerprinting defenses. *Proceedings on Privacy Enhancing Technologies*, 2017(4):215–231, oct 2017.

[Chi02] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CJT22] Giovanni Cherubin, Rob Jansen, and Carmela Troncoso. Online website fingerprinting: Evaluating website fingerprinting attacks on tor in the real world. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 753–770, 2022.

[CKR21] Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. " i need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3037–3052, 2021.

[CM+14] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[CMM21] Graham Cormode, Samuel Maddock, and Carsten Maple. Frequency estimation under local differential privacy. *Proceedings of the VLDB Endowment*, 14(11):2046–2058, July 2021.

[CPO19] Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, 2019.

[CPS14] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A predictive differentially-private mechanism for mobility traces. In Emiliano De Cristofaro and Steven J. Murdoch, editors, *Privacy Enhancing Technologies*, pages 21–41, Cham, 2014. Springer International Publishing.

[Cra17] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[Dar71] Richard B Darlington. Another look at "cultural fairness". *Journal of educational measurement*, 8(2):71–82, 1971.

[DCDAU22] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Online certification of preference-based fairness for personalized recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6532–6540, 2022.

[Deg21] Jean Paul Degabriele. Hiding the lengths of encrypted messages via gaussian padding. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1549–1565, 2021.

[DG17a] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DG17b] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml, (accessed January 2023).

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

[DJW13] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages

429–438. IEEE, October 2013.

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3571–3580. Curran Associates, Inc., 2017.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DRBB+23] Natalia Díaz-Rodríguez, Rūta Binkytė, Wafae Bakkali, Sannidhi Bookseller, Paola Tubaro, Andrius Bacevičius, Sami Zhioua, and Raja Chatila. Gender and sex bias in covid-19 epidemiological data through the lens of causality. *Information Processing & Management*, 60(3):103276, 2023.

[DT92] Dorit Dor and Michael Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*, 1992.

[DWJ13] John Duchi, Martin J Wainwright, and Michael I Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[Dwo06a] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[Dwo06b] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Hei-

delberg, 2006. Springer Berlin Heidelberg.

[EFM+19] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[EFM+20] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, and Abhradeep Thakurta. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.

[EGGL20] Vitalii Emelianov, Nicolas Gast, Krishna P Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 649–675, 2020.

[EPK14] Úlfar Erlingsson, Vasyl 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, New York, NY, USA, 2014. ACM.

[ES11] Barbara Espinoza and Geoffrey Smith. Min-entropy leakage of channels in cascade. In *International Workshop on Formal Aspects in Security and Trust*, pages 70–84. Springer, 2011.

[FFMD22] Pedro Faustini, Natasha Fernandes, Annabelle McIver, and Mark Dras. Directional privacy for deep learning. *arXiv preprint arXiv:2211.04686*, 2022.

[FNNT22] Vitaly Feldman, Jelani Nelson, Huy Nguyen, and Kunal Talwar. Private frequency estimation via projective geometry. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6418–6433.

PMLR, 17–23 Jul 2022.

[FPBD18] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *arXiv preprint arXiv:1802.09386*, 2018.

[Fuk93] Keinosuke Fukunaga. Statistical pattern recognition. In *Handbook of pattern recognition and computer vision*, pages 33–60. World Scientific, 1993.

[Fuk13] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, San Diego, CA, US, 2013.

[G+23] Ramon Gonçalves Gonze et al. A quantitative information flow model for attribute-inference attacks and utility in data releases by sampling, 2023.

[Gam21] Juan Gamella. Greedy equivalence search (ges) algorithm for causal discovery. https://github.com/juangamella/ges, 2021. Accessed: 2022-03-16.

[GBJ+22] Filippo Galli, Sayan Biswas, Kangsoo Jung, Tommaso Cucinotta, and Catuscia Palamidessi. Group privacy for personalized federated learning. *arXiv preprint arXiv:2206.03396*, 2022.

[GDBFL19] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR, 2019.

[GHP15] Yoel Gluck, Neal Harris, and Angelo Prado. Breach: reviving the crime attack (2013). *Dostupné také z: http://css. csail. mit. edu/6*, 858, 2015.

[Gir21] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[GKG+09] Kirk Glerum, Kinshuman Kinshumann, Steve Greenberg, Gabriel Aul, Vince Orgovan, Greg Nichols, David Grant, Gretchen Loihle, and Galen Hunt. Debugging in the (very) large: Ten years of implementation and experience. In *Pro-*

ceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles, SOSP '09, page 103–116, New York, NY, USA, 2009. Association for Computing Machinery.

[GLC+22] M. Emre Gursoy, Ling Liu, Ka-Ho Chow, Stacey Truex, and Wenqi Wei. An adversarial approach to protocol analysis and selection in local differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:1785–1799, 2022.

[Glo18] Global Times. Beijing to release new license plate lottery policy. *https://www.globaltimes.cn/content/1190224.shtml*, 2018.

[GLR20] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[Gra72] Ronald L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Info. Pro. Lett.*, 1:132–133, 1972.

[Grü13] Branko Grünbaum. *Convex polytopes*, volume 221. Springer Science & Business Media, New York, NY, US, 2013.

[Hau88] Dominique MA Haughton. On the choice of a model to fit data from an exponential family. *The annals of statistics*, pages 342–355, 1988.

[HB12] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

[HLGK19] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190, 2019.

[Hol19] Julian Holzel. Differential privacy and the gdpr. *Eur. Data Prot. L. Rev.*, 5:184, 2019.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[HS13] Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

[IDC20] IDC Corporate. Global storagesphere forecast, 2021-25, 2020.

[JGAP23] Mireya Jurado, Ramon G Gonze, Mário S Alvim, and Catuscia Palamidessi. Analyzing the shuffle model through the lens of quantitative information flow. *arXiv preprint arXiv:2305.13075*, 2023.

[JP19] Kangsoo Jung and Seog Park. Privacy bargaining with fairness: Privacy-price negotiation system for applying differential privacy in data market environments. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1389–1394. IEEE, 2019.

[JRUW18] Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[KBR16] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Int. Conf. on Machine Learning*, pages 2436–2444. PMLR, 2016.

[KG19] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

[KLN+08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 531–540. IEEE, October 2008.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[KMC+12] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[KOV16] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *The Journal of Machine Learning Research*, 17(1):492–542, 2016.

[KR19] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, New York, NY, US, 2019.

[KS22] Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3):1587–1595, 2022.

[KYC+19] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *arXiv preprint arXiv:1906.00389*, 2019.

[LB14] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[LCM18] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8136–

8146, 2018.

[Leo19] Sabina Leonelli. *La recherche scientifique à l'ère des Big Data: Cinq façons dont les Big Data nuisent à la science et comment la sauver*. Editions Mimésis, 2019.

[LHL+16] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

[LLMS14] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A theory of pricing private data. *ACM Transactions on Database Systems (TODS)*, 39(4):1–28, 2014.

[LLV06] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[LMKA16] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

[LTH+23] Gaoyuan Liu, Peng Tang, Chengyu Hu, Chongshi Jin, and Shanqing Guo. Multi-dimensional data publishing with local differential privacy. In *Proceedings of the 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28 - March 31, 2023*, pages 183–194. OpenProceedings.org, 2023.

[MABK+20] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842, 2020.

[MCPS12] S Alvim M'rio, Kostas Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *2012 IEEE 25th Computer Security Foundations Symposium*, pages 265–279. IEEE, 2012.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[MPBT23] Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23681–23705. PMLR, 23–29 Jul 2023.

[MZP20a] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

[MZP20b] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

[MZP21a] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, 2021.

[MZP21b] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23(1):14–23, 2021.

[NC12] Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[new15] The Guardian newspaper. Google says sorry for racist auto-tag in photo app. https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app, 2015. Accessed: 2023-06-26.

[new18] The Guardian newspaper. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, 2018. Accessed: 2023-06-26.

[Ngu14] Benjamin Nguyen. Techniques d'anonymisation. *Statistique et société*, 2(4):53–60, 2014.

[NGZ20] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[NV20] Moni Naor and Neil Vexler. Can Two Walk Together: Privacy Enhancing Methods and Preventing Tracking of Users. In Aaron Roth, editor, *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 4:1–4:20, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[NXY+16] Thông T. Nguyên, Xiaokui Xiao, Yin Yang, Siu Cheung Hui, Hyejin Shin, and Junbum Shin. Collecting and analyzing data from smart device users with local differential privacy. *ArXiv*, abs/1606.05053, 2016.

[Och19] Rodrigo Ochigame. The invention of 'ethical ai': how big tech manipulates academia to avoid regulation. *ECONOMIES OF*

VIRTUE, page 49, 2019.

[OWW22] Olga Ohrimenko, Anthony Wirth, and Hao Wu. Randomize the future: Asymptotically optimal locally private frequency estimation protocol for longitudinal data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '22, page 237–249, New York, NY, USA, 2022. Association for Computing Machinery.

[Par16] European Parliament. Regulation (EU) 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). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679, 2016. Accessed on: Insert Date Here.

[PBMB18] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials*, 21(3):2772–2793, 2018.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[PH22] Carlos Pinzón and Hwang Héber. Loloha repository, 2022. https://github.com/hharcolezi/LOLOHA.

[Pin22] Carlos Pinzón. Github repository (impossibility-fairness-non-trivial-accuracy), 2022. https://github.com/caph1993/impossibility-fairness-non-trivial-accuracy, 2022 (accessed December 12 2022).

[PJ23] Carlos Pinzón and Kangsoo Jung. Fast Python sampler for the von Mises Fisher distribution. working paper or preprint, March 2023.

[PMK+20] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of*

the 2020 Conference on Fairness, Accountability, and Transparency, pages 189–199, 2020.

[PPPV22] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7993–8000, 2022.

[PPPV23] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the incompatibility of accuracy and equal opportunity. *Machine Learning*, pages 1–30, 2023.

[PPS22] Carlos Pinzón, Cezara Petrui, and Sebastian Simon. min-leakage-padding. https://github.com/caph1993/min-leakage-padding, 2022. Accessed: February 2023.

[PQR+22] Carlos Pinzón, Santiago Quintero, Sergio Ramírez, Camilo Rueda, and Frank Valencia. Counting and computing join-endomorphisms in lattices (revisited). *arXiv preprint arXiv:2211.00781*, 2022.

[PQRV21] Carlos Pinzón, Santiago Quintero, Sergio Ramírez, and Frank Valencia. Computing distributed knowledge as the greatest lower bound of knowledge. In *Relational and Algebraic Methods in Computer Science: 19th International Conference, RAMiCS 2021, Marseille, France, November 2–5, 2021, Proceedings 19*, pages 413–432. Springer, 2021.

[PR18] Catuscia Palamidessi and Marco Romanelli. Feature selection with rényi min-entropy. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 226–239. Springer, 2018.

[PS22] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss,

V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Red21] Red Gate Blog. What's the real story behind the explosive growth of data? https://www.red-gate.com/blog/database-development/whats-the-real-story-behind-the-explosive-growth-of-data, 2021. Accessed: 2023-06-26.

[Reu18] Reuters news agency. Amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 2018. Accessed: 2023-06-26.

[Rom20] Marco Romanelli. *Machine learning methods for privacy protection: leakage measurement and mechanisms design*. PhD thesis, Institut Polytechnique de Paris; Università degli studi (Sienne, Italie), 2020.

[RR21] Andrew C. Reed and Michael K. Reiter. Optimally hiding object sizes with constrained padding, 2021.

[RSW21] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[RZG+18] Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

[SAV+18] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Po-

tential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency*, pages 5–19. PMLR, 2018.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[Sch00] Werner Schindler. A timing attack against rsa with the chinese remainder theorem. In *Cryptographic Hardware and Embedded Systems—CHES 2000: Second International Workshop Worcester, MA, USA, August 17–18, 2000 Proceedings 2*, pages 109–124. Springer, 2000.

[Scr16] Luca Scrucca. Genetic algorithms for subset selection in model-based clustering. In *Unsupervised learning algorithms*, pages 55–70. Springer, 2016.

[SCS13] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[SGCR22] Maja Schneider, Lukas Gehrke, Peter Christen, and Erhard Rahm. D-tour: Detour-based point of interest detection in privacy-sensitive trajectories, 2022.

[SGSH00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[SHH+06] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[Shi14] Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

[SIS+11] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for

learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[Smi09] Geoffrey Smith. On the foundations of quantitative information flow. In Luca de Alfaro, editor, *Proceedings of the 12th International Conference on Foundations of Software Science and Computation Structures (FOSSACS 2009)*, volume 5504 of *LNCS*, pages 288–302, York, UK, 2009. Springer.

[SMR99] Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:211–252, 1999.

[Son01] Dawn Song. Timing analysis of keystrokes and ssh timing attacks. In *Proc. of 10th USENIX Security Symposium, 2001*, 2001.

[SPPP23] Sebastian Simon, Cezara Petrui, Carlos Pinzón, and Catuscia Palamidessi. Obfuscation padding schemes that minimize rényi min-entropy for privacy. In *International Conference on Information Security Practice and Experience*, pages 74–90. Springer, 2023.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SSL+23] Maja Schneider, Jonathan Schneider, Lea Löffelmann, Peter Christen, and Erhard Rahm. Tuning the utility-privacy tradeoff in trajectory data, 2023.

[Sta21] State.gov. Diversity visa program. *http://dvprogram.state.gov/*, 2021.

[SYT20] Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International Conference on Machine Learning*, pages 8316–8325. PMLR, 2020.

[SZDK22] Jonas Seng, Matej Zečević, Devendra Singh Dhami, and

Kristian Kersting. Tearing apart notears: Controlling the graph prediction via variance manipulation. *arXiv preprint arXiv:2206.07195*, 2022.

[TFKN22] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 35:17652–17664, 2022.

[TFVH21] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939, 2021.

[TFVHY21] Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In *IJCAI*, pages 560–566, 2021.

[TK86] Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

[TKB+17] 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.

[Tsa19] Michail Tsagris. Bayesian network learning with the pc algorithm: an improved and correct variation. *Applied Artificial Intelligence*, 33(2):101–123, 2019.

[TSD20] Michael Carl Tschantz, Shayak Sen, and Anupam Datta. Sok: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 354–371. IEEE, 2020.

[Vap99] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[VCS22] Gatha Varma, Ritu Chauhan, and Dhananjay Singh. Sarve: synthetic data and local differential privacy for private fre-

quency estimation. *Cybersecurity*, 5(26), 2022.

[War65] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965.

[WBLJ17] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, Vancouver, BC, August 2017. USENIX Association.

[WCM09] Charles V Wright, Scott E Coull, and Fabian Monrose. Traffic morphing: An efficient defense against statistical traffic analysis. In *NDSS*, volume 9. Citeseer, 2009.

[WXY+19] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE, April 2019.

[XYH+22] Qiao Xue, Qingqing Ye, Haibo Hu, Youwen Zhu, and Jian Wang. DDRM: A continual frequency estimation mechanism with local differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022.

[YKCT19] Mohammad Yaghini, Bogdan Kulynych, Giovanni Cherubin, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv e-prints*, pages arXiv–1906, 2019.

[ZARX18] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[ZT21] Xingyu Zhou and Jian Tan. Local differential privacy for bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11152–11159, 2021.

ECOLE
DOCTORALE

Titre : Des études sur l'équité et la confidentialité dans l'apprentissage automatique

Mots clés : Protection de la vie privée, Équité, Aprentissage automatique, Données

Résumé : Cette thèse présente quatre articles publiés dans le domaine de l'éthique des données qui élargissent nos connaissances sur l'équité dans l'apprentissage automatique et l'état de l'art de la confidentialité dans la collecte et la transmission des données. Ce document englobe (1) le calcul de la limite de Pareto d'égalité des chances et de précision des classificateurs d'apprentissage automatique, la preuve que ces objectifs s'opposent radicalement pour certaines distributions ; (2) la preuve empirique que lors de l'utilisation d'algorithmes de découverte causale pour l'évaluation de l'équité, différents algorithmes peuvent conduire à des conclusions très différentes ; (3) la proposition d'un protocole de collecte de données longitudinales avec des garanties inspirées de la confidentialité différentielle locale ; et (4) la dérivation de deux méthodes optimales de remplissage des données transmises pour protéger la confidentialité contre les observateurs du réseau.

Title : Exploring Fairness and Privacy in Machine Learning

Keywords : Privacy, Machine learning, Fairness, Data

Abstract :
This dissertation presents four published articles in the field of data ethics that extend our knowledge of fairness in machine learning and advance the state of the art of privacy in data collection and transmission. This document encompasses : (1) a general study of the trade-off between equal opportunity and accuracy of machine learning classifiers along with the proof that these objectives may oppose each other strongly ; (2) the empirical proof that when using causal discovery algorithms for fairness assessment, different algorithms may lead to very different conclusions ; (3) the proposal of a protocol for longitudinal data collection with guarantees inspired in local differential privacy ; and (4) the derivation of two optimal methods for padding transmitted data to protect privacy against network observers.

Institut Polytechnique de Paris
91120 Palaiseau, France